Modélisation et identification de facteurs environnementaux géographiques liés à des risques morbides : Application aux séquelles développées après le traitement d'une leucémie : Cohorte LEA

Stephane Bourrelly

Ecole Doctorale Lettres, Sciences Humaines et Sociales

THESE

Présentée pour obtenir

Le grade de docteur de l'Université Nice Sophia Antipolis

Spécialité Géographie

Soutenue le 3 septembre 2014

Stéphane BOURRELLY

Modélisation et identification de facteurs environnementaux géographiques liés à des risques morbides

Application aux séquelles développées après le traitement d'une leucémie – Cohorte LEA

Composition du jury :

M.	Pascal AUQUIER	Co-directeur de thèse
M.	Arnaud BANOS	Examinateur
M.	Jean-Pierre DAURES	Rapporteur
M.	Emmanuel ELIOT	Examinateur
M.	Robin GENUER	Examinateur
M.	Gilles MAIGNANT	Rapporteur
Mme.	Christine VOIRON	Directrice de thèse

REMERCIEMENTS

Passer plus de trois ans et demi de sa vie à travailler sur la modélisation et l'identification de facteurs environnementaux liés à des risques morbides fût à la fois une aventure scientifique palpitante et incertaine, une véritable épopée maculée de joie et de tristesse, de plaisirs et de craintes, de rencontres ponctuées de longs moments de solitude. Cela amène à être redevable à un grand nombre de personnes.

Dans l'environnement scientifique, mes remerciement vont à :
Christine Voiron, ma directrice de thèse qui a su m'apporter le soutien scientifique, rédactionnel et moral nécessaire pour l'aboutissement de cette thèse en géographie.
Pascal Auquier, mon co-directeur mais aussi le mécène de cette recherche. Professeur, je vous remercie pour votre soutien scientifique, financier et moral, votre disponibilité et votre patience. Et pour m'avoir exhorté à me dépasser et m'avoir permis d'éviter de nombreuses bévues.
Jean Serra, à qui je suis redevable pour ses conseils, ses relectures systématiques de mes productions scientifiques et des parties de cette thèse sur la géostatistique.
Yannick Baraud, admirable pour son intelligence, sa patience et sa discrétion. Tout ce que je sais en statistique et en probabilité, c'est à vous que je le dois. Vous m'avez pris sous votre égide, en dépit de mon indigence, et m'avez inculqué les bases fondamentales de ces disciplines.
Robin Genuer, le père de VSURF qui me fait le privilège de participer à ce jury. Je vous sais gré d'avoir validé et relu l'intégralité de mon travail sur les Forêts Aléatoires.
Christine Malot, merci de m'avoir appris à programmer en R et aguerri à l'algorithme CART, et surtout de m'avoir orienté sur les sentiers des Forêts Aléatoires.
Julien Caudeville, merci pour avoir partagé tes connaissances en expologie et tes nombreuses astuces en SIG, ainsi que pour ton soutien psychologique. Tu avais raison, *en fin de thèse on crache du sang le matin au réveil*…
Véronique Lucas Gabrielli, merci d'avoir pris le temps de valider et de relire mon travail sur les facteurs environnementaux : sanitaires et socio-économiques

Dans l'environnement social, à :
Isabelle Claisse, je te remercie pour ton soutien et ton amour indéfectible. Je te remercie aussi d'avoir relu et corrigé l'intégralité de cette thèse. Tu es ma muse, ma déesse, ma princesse, ma sirène ; probablement l'Amour de ma vie et la raison de mon amour pour la vie.
Mes parents que je remercie : **Maman** pour tes attentions, pour ton soutien quotidien et pour avoir allégé au maximum mes obligations en fin de thèse, **Papa** pour ton soutien financier, pour m'avoir donné le goût du travail, le sens du devoir et la force d'aller jusqu'au bout.
Marion Borderon merci pour tes conseils rédactionnels et pour m'avoir initié à la géographie de la santé. Merci aussi pour ton soutien moral et ta gaîté.
Lucinda Cau merci pour ton aide en anglais, ta douceur et ton amitié. Merci aussi à Gladys et Pierrot, tes parents, chez qui la porte comme le cœur sont toujours ouverts.
Philippe Lauby, merci pour tes illustrations sont à la hauteur de tes traductions, ta bonne humeur à la hauteur de ton amitié fidèle, merci de m'avoir accompagné tout au long de cette épopée.

Je remercie aussi l'UMR 7300 (ESPACE) ainsi que l'EA 3279 (Santé publique et maladies chroniques) pour leur soutien logistique et financier. Je remercie également toute l'équipe du Professeur Auquier pour leur réactivité et leur disponibilité en particulier : Vivi, Delphine, Mohamed, Anderson et Julie. Et, tous ceux qui ont consacré de leur temps à la lecture de ce manuscrit ; tous ceux que j'ai oubliés et qui m'ont soutenu, dont les doctorants de l'Université de Nice et les membres de ma famille, font partie.

A Isabelle - la femme que j'aime - de la première lettre du premier mot jusqu'au point final

RESUME ET ABSTRACT

Résumé : Cette thèse s'inscrit dans le cadre d'une approche interdisciplinaire mêlant la Géographie à l'épidémiologie et à la Statistique. Il s'agit d'une thèse d'ordre méthodologique appliquée à une problématique de santé publique. L'idée est de concevoir une dialectique adaptée à la géographie de la santé et de proposer ou transposer des méthodes probabilistes, géostatistiques et d'apprentissage statistique afin de modéliser, puis d'identifier des Facteurs Environnementaux (FE) géographiques liés à des risques morbides. Dans cette recherche l'environnement est considéré dans ses multiples dimensions. Il est décrit par des indicateurs spatiotemporels physicochimiques, sanitaires et socio-économiques. L'environnement géographique des Caractéristiques Individuelles et Médicales (CIM) des populations ciblées est aussi pris en compte. Les propositions heuristiques visent à identifier des Déterminants Environnementaux de Santé (DES) ou des facteurs de risques contributifs. La dialectique se veut opérationnelle et reproductible à toutes les maladies. Les propositions sont appliquées à des séquelles développées après le traitement d'une leucémie chez l'enfant - Cohorte LEA. Dans les pays développés, l'accès à des traitements performants engendre une augmentation de l'incidence des séquelles - qui ont des répercussions sur la qualité de vie. Par conséquent, l'après-cancer et la santé des enfants se sont positionnés au cœur des préoccupations sociales en Europe en général et en France, en particulier. Au-delà des objectifs scientifiques visés, des contributions en santé publique sont attendues. Il s'agit de proposer des indicateurs spatiaux opérationnels permettant aux hommes politiques de mettre en place des mesures sanitaires collectives, et aux praticiens de santé de proposer des solutions préventives individuelles - visant à réduire les risques environnementaux d'exposition auxquels sont assujetties les populations du simple fait de leur situation géographique. Et ainsi, d'améliorer l'accès à une bonne santé environnementale*.

Mots clé : Géographie ; statistique, facteurs environnementaux, modélisation, santé

Abstract: This thesis is an interdisciplinary approach combining Geography, Epidemiology and Statistics. It is a methodological thesis applying to a public health issue. The concept consist in developing a dialectical adapted to the geographic health and proposing or transposing probabilistic methods, geostatistical and datamining instruments, to model and to identify geographic environmental factors related to morbid risks. In this research the environment is considered in its integrity. It is described by spatiotemporal indicators: physicochemical, health and socioeconomic. The geographical environment of the individual and medical characteristics of targeted populations is also taken into account. Heuristic proposals aim to identify environmental health determinants, or contributing risk factors. The methods have been implemented or adapted to the issue. They are applicable and reproducible in all diseases studied in Geographic Health. In order, to illustrate these proposals, they are applied to squeals observed following the treatment of childhood leukemia - Cohort LEA. In developed countries, access to effective treatments leads to an increase of squeals incidence. Those have an impact on the quality of life. Therefore, post-cancer and children's health have positioned themselves at the core of social concerns in Europe and in France. Beyond the scientific goals, contributions in public health are expected. The idea is to provide operational space indicators to politicians in order to help them to take collective health measures. And for health professionals to be able to offer individual medical solutions, devoted to reduce the risk of exposure to environmental factors of the populations due to their geographical location. In consequence, to improve access to a good environmental health.

Keywords: Geography, statistics, environmental factors, modeling, health

Les relations linéaires restent des exceptions, les faits géographiques apparaissent souvent discontinus, représentant des seuils caractéristiques, les facteurs aléatoires y jouent un grand rôle, et l'aléa géographique est spatiotemporel, ce qui devrait rapprocher le géographe du physicien, du mathématicien, … du médecin …, ou du philosophe (Peguy, 1996).

Abréviations

ACP	:	Analyse en Composante Principale
AEE	:	l'Agence Européenne pour l'Environnement (AEE)
ANSES	:	Agence française Nationale de SEcurité Sanitaire
APL	:	Accès Potentiels Localisés
ARC	:	Assistant de Recherche Clinique
ASN	:	Agence de Sureté Nucléaire
BD	:	Base de données
BMVSG	:	BoostMyVsurfGéo
BTP	:	Bâtiment Travaux Publiques
CATA	:	Cataractes
CAM	:	Caisse d'Assurance Maladie
CART	:	Classification And Regression Tree
CCE	:	Commission des Communautés Européenne
cem-ebf	:	champs électromagnétiques et magnétiques à extrême basse fréquence
CGDD	:	Commissariat Général au Développement Durable
CHU	:	Centre Hospitalier Universitaire
CIM	:	Caractéristiques Individuelles et Médicales
CIRC	:	Centre International de Recherche sur le Cancer
CKO	:	Co-Krigeage Ordinaire
CNIL	:	Commission Nationale de l'Informatique et des Libertés
CP	:	Code Postal
DES	:	Déterminants Environnementaux de Santé
DGS	:	Direction Générale de la santé
DJE	:	Doses Journalières d'Exposition
DOM	:	Départements d'Outre-Mer
DTA	:	Distances Temporelles d'Accès
e.g.	:	par exemple
ENDO.s	:	service d'Endocrinologie
EML*	:	Equipements et Matériels Lourds
ESAPCE	:	Etudes des Structures et des Processus d'Adaptations et de Changement de l'Espace
ETM	:	Eléments Métalliques Traces
FA	:	Forets Aléatoires
FE	:	Facteurs Environnementaux
FIM	:	Facteurs Individuels et Médicaux
FREC	:	Facteurs de Risques Environnementaux Contributifs
FREPA	:	Facteurs de Risques Environnementaux Potentiellement Aggravant
GENE	:	Généralistes
HAP	:	Hydrocarbures Aromatiques Polycycliques
HAS	:	Haute Autorité de Santé

HEMA.s	:	service d'Hématologie
HPST	:	Hôpital, Patient, Santé et Territoire
i.e.	:	C'est-à-dire
INERIS	:	l'Institut National de l'Environnement et des Risques Industriels
INB*	:	Installation Nucléaire de Base
InVS	:	l'Institut de Veille Sanitaire
IRM	:	Imagerie par Résonance Magnétique
irs	:	indicateurs des risques spatiaux
IRSN	:	Institut de Radioprotection et de Sûreté Nucléaire
i.st	:	indicateurs spatiotemporels
i.st.e	:	indicateurs spatiotemporels environnementaux
i.st.m	:	indicateurs spatiotemporels morbides
KO	:	Krigeage Ordinaire
LA	:	Leucémie Aigüe
LAL	:	Leucémie Aigüe Lymphoïdes
LAM	:	Leucémie Aigüe Myéloïdes
LEA	:	Leucémie chez l'Enfant et l'Adolescent
LEUC	:	Leucémie
MEP	:	Mode d'Exercice Particulier
MES	:	Matières en suspensions
MVG	:	MyVsurfGéo
NEUR.s		service de Neurologie médicale et neurochirurgie
OMS	:	Organisation Mondiale de la Santé
ONDRP	:	Observatoire National de la Délinquance et des Réponses Pénales
OOB	:	Out-Of-Bag
OPHT	:	Ophtalmologues
OPHT.s	:	service d'Ophtalmologie
ORL	:	Oto-Rhino-Laryngologues
ORL.s	:	service d'Oto-Rhino-Laryngologies
PACA	:	Provence Alpes Côte d'Azur
PEDIA	:	Pédiatrie
PCB	:	Polychlorobiphényle
PHY.CHIM	:	Physicochimique
PLAINE	:	Plateforme intégrée pour l'analyse des inégalités d'exposition environnementale
PM	:	Phénomène Morbide
PNSE	:	Plan National Santé Environnement
PRSE2	:	Plans Régionaux de Santé Environnement
PREC	:	Proxy du Risque d'Exposition à des substances Chimiques combinées
QV	:	Qualité de Vie
RADIO	:	radiologues
RCP	:	Réunion de Concertation Pluridisciplinaire
RI	:	Rayonnement Ionisant

RNI	:	Rayonnement Non Ionisant
RNM	:	Réseau National de Mesure
SAN	:	Sanitaire
SCALE	:	Science, Children, Awareness, Legal instrument, Evaluation
SCAN	:	scanner
SHS	:	Sciences Humaines et Sociales
SIG	:	Système d'Information Géographique
SOCIO.ECO	:	Socio-économique
SOeS	:	Service de l'observation et des statistiques
SROS	:	Schémas Régionaux d'Organisation des Soins
SU	:	Sans Unité
TEP	:	Tomographie par Emission de Positons
THYR	:	Tumeur Thyroïdienne
TOM	:	Territoires d'Outre-Mer
TUM2	:	Tumeur secondaire (2nd)
VSURF	:	Variable Selection Using Random Forest

GLOSSAIRE

Tous les mots du texte définis dans ce glossaire sont adjoints d'un astérisque (*)

Accès Potentiel Localisé (APL) : indicateur qui évalue la qualité d'une partie des tissus sanitaires territoriaux. Il a été conçu pour pallier la faiblesse des Distances Temporelles d'accès (DTA) à modéliser l'accès aux médecins libéraux. Il exprime le nombre de médecins disponibles pour mille habitants. APL est *a priori* l'indicateur géographique français le plus robuste puisqu'il prend en compte simultanément l'offre, la demande et le temps d'accès. Au moment où les traitements ont été réalisés, il était disponible uniquement pour les généralistes et les ophtalmologues.

Biais conditionnel : concept emprunté aux géostatistiques et transposé à la dialectique. Un biais conditionnel suppose que la variabilité spatiale d'une estimation statistique est supérieure à celle de sa valeur réelle. La minimisation du *biais conditionnel* améliore la qualité des indicateurs spatiaux. Dans cette thèse, il s'agit de coupler des méthodes, des outils et des stratégies d'intégration utilisant au mieux toute l'amplitude des sources de données disponibles. L'intérêt est de construire des indicateurs spatiotemporels morbides (i.st.m) et des indicateurs spatiotemporels environnementaux (i.st.e.) qui modélisent de façon robuste et fiable la géographie des phénomènes morbides et environnementaux pour lesquels ils sont proposés.

Boosting : sous-catégorie des méthodes d'ensemble réputée pour sa prééminence. Le boosting intervient au niveau de la phase d'échantillonnage. Les sous-échantillons sont prédéterminés par une loi de probabilité choisie *a priori.* Le principe du boosting est de focaliser la règle d'apprentissage sur la partie de l'espace des individus la plus difficile à appréhender.

BoostMyVsurfGéo (BMVG) nom de l'évolution heuristique de MVG proposée pour valider, par une approche individus-centrée, les résultats géographiques obtenus. BMVG se compose d'un algorithme qui intègre un processus de boosting, randomisé géographiquement, afin d'augmenter la puissance de MVG et de l'adapter à la complexité d'une approche à mi-chemin entre la géographie et l'épidémiologie. A l'instar de MVG, la méthode permet de disjoindre les variables de bruit de celles explicatives, et renvoie des items adaptés à l'interprétation des résultats. En revanche, elle ne permet pas d'isoler de variables prédictives.

Caractéristiques Individuelles et Médicales (CIM) : il s'agit de toutes les variables individus-centrées disponibles et pertinentes qui permettent d'expliquer un état de santé, à partir de l'historique médical ou comportemental d'un individu.

Classification And Regression Tree (CART) : modèle d'inférence statistique. Le principe de CART est de découper de façon optimale l'espace engendré par les individus et les variables d'un jeu de données. Le prédicteur CART est constant par morceaux. Sa forme dépend du contexte statistique. En régression, il s'agit d'un opérateur linéaire de type arbre de décision et en classification de l'estimateur de Bayes. Ce modèle a été imaginé par Léo Breiman et programmé avec l'assistance de l'informaticienne Cluter, en 1984.

Contexte statistique : situation statistique fixée par la nature de la variable d'intérêt, i.e. la classification pour les variables qualitatives et la régression pour les variables quantitatives.

Curieux : cf**.** Facteurs Environnementaux.

Datamining : ou Machine Learning sont des outils informatiques, i.e. des algorithmes d'inférence fondés sur des méthodes de modélisation statistiques et probabilistes. Le datamining est adapté aux jeux de données multidimensionnels caractérisant des phénomènes complexes.

Déterminants Environnementaux de Santé (DES) : ce sont des Facteurs Environnementaux (FE) ou des Facteurs Individuels et Médicaux (FIM) qui ont une influence forte sur le Phénomènes Morbide (PM) étudié.

Distance a-spatiale morbide : ce concept représente la plausibilité des effets d'une situation à risque ou d'une substance physicochimique nocive sur la santé d'un individu ou d'un groupe. Plus cette distance virtuelle est courte et plus les effets avérés ou suspectés sont bien documentés. Lorsqu'il qualifie un indicateur spatiotemporel, il prend en compte aussi la qualité des sources utilisées.

Distance Temporelle d'Accès* (DTA) : indicateur qui évalue la qualité globale des tissus sanitaires territoriaux. Il exprime la distance temporelle moyenne - en minutes - pour accéder par le réseau routier à l'ensemble des items sanitaires. Il suppose que le patient emprunte l'itinéraire le plus court. Les DTA sont cependant inconsistantes pour modéliser l'accès géographique à la majorité des praticiens libéraux. Cependant, au moment où les traitements ont été effectués les DTA constituaient les indicateurs sanitaires publics de référence.

Dyadique : relatif à une dyade, i.e. groupe de deux éléments solidaires et qui se complètent. En particulier, l'algorithme rpart permet de construire des arbres CART dans une logique dyadique, i.e. que chacun de ses nœuds permet de diviser de façon optimale – et récursive – l'espace engendré par les individus au regard de leurs coordonnées statistiques.

Effet de bord : perturbation aléatoire légère qui intervient au niveau des résultats obtenus. C'est le cas par exemple lorsqu'on lance deux fois l'algorithme *randomForest* à partir d'un jeu de données identiques. Les deux estimations statistiques vont légèrement être différentes, en raison des processus stochastiques implémentés ; on parle d'instabilité de l'algorithme.

Effet information : conséquence des incertitudes propres aux sources de données mobilisées. L'optimisation de l'effet information est une des deux phases de la minimisation du biais conditionnel. Elle consiste à combler les lacunes, uniformiser les échelles et fusionner des informations expertes ou des données géographiques auxiliaires.

Effet de support : conséquence inhérente aux caractéristiques spatiales et temporelles des données mobilisées. L*a maximisation de l'effet de support* constitue la phase d'amélioration et d'harmonisation de la granularité des sources, inhérente à la minimisation du biais conditionnel.

Equipement Matériel Lourd (EML*) : caractéristique du plateau technique des établissements de santé qui définissent des équipements médicaux techniques conséquents et nécessaires au diagnostic ou au traitement de certaines pathologies.

Fonction Aléatoire Intrinsèque (FAI) : fonction mathématique composée de deux opérateurs de statistique spatiale. Le premier est déterministe, i.e. l'espérance, dont la valeur estimable est réelle. Le second est stochastique, à valeur supposée inconnue mais finie. Lorsqu'une FAI est utilisée pour modéliser une variable régionalisée (v.r.), une partie de ses variabilités aléatoires spatiales est estimable en supposant ses accroissements stationnaires, i.e. qu'une faible variation d'espace engendre une faible variabilité de la v.r.

Espace géographique : ce concept, en géographie de la santé, permet de construire les indicateurs spatiotemporels morbides et environnementaux (i.st.m* / i.st.e) qui répondent à deux objectifs. (i) Représenter de façon fiable la géographie des Phénomènes Morbides (PM) et les Facteurs Environnementaux ou Individuels et Médicaux pertinents (FE/FIM). (ii) Analyser leurs interactions statistiques multidimensionnelles afin de comprendre les dysfonctionnements du système territorial de santé. Le point de départ pour confectionner un espace géographique consiste à choisir une échelle d'investigation adaptée à la problématique. Ces espaces sont numériques et sont élaborés via des Systèmes d'Informations Géographiques.

Facteurs Environnementaux (FE) : facteurs qui constituent tous les éléments susceptibles de caractériser l'environnement géographique des milieux de vie - dans lesquels sont subsumées les populations. Dans cette thèse il est question des environnements communaux dans lesquels sont spatialisés les individus de la Cohorte LEA. Ils peuvent être qualifiés de pertinents ou de curieux.

Facteurs Individuels et Médicaux (FIM). Cette géographie particulière modélise par le biais d'indicateurs spatiotemporels environnementaux (i.st.e) les Caractéristiques Individuelles et Médicales (CIM) d'une population ciblée, i.e. celle des individus spatialisés de la Cohorte LEA.

FE* pertinents : constituent les objets géographiques qui ont des effets documentés - néfastes ou préventifs, avérés ou suspectés - au regard d'un état de santé particulier. Ils sont discrétisés en quatre composantes environnementales : sanitaire (SAN), socio-économique (SOCIO.ECO), physicochimique (PHY.CHIM) et les Caractéristiques Individuelles et Médicales (CIM) spatiales des populations ciblées.

FE/FIM* curieux : FE/FIM* intégrés dont les interactions avec le Phénomènes Morbide (PM) d'intérêt sont incertaines, voire improbables, soit parce que les connaissances bibliographiques sont controversées, soit à cause de la granularité médiocre des sources utilisées pour les modéliser. En l'occurrence, les i.st.e* qualifiés de *curieux de test*, ont initialement été intégrés pour tester la robustesse de MVG et BMVG. La majorité a effectivement rempli son rôle. Toutefois, certains se sont « curieusement » avérés avoir un pouvoir explicatif important.

Facteurs de Risques Environnementaux Contributifs (FREC) : FE/FIM* qui ont une influence notable sur les PM, et surtout lorsqu'ils se combinent. Cette géographie particulière est statistiquement conjecturée à partir des grilles de lecture MVG - ou BMVG.

Facteurs de Risques Environnementaux Probablement Aggravants (FREPA) : FE/FIM* qui peuvent être soupçonnés, en se combinant, d'aggraver les effets délétères des expositions aux DES* et FREC. Cette géographie particulière est statistiquement conjecturée à partir des grilles de lecture MVG - ou BMVG.

Forêt Aléatoire (FA) : outil de datamining programmé qui constitue les derniers travaux d'inférence statistique menés par Léo Breiman, en 2005. Les FA se composent d'arbres CART doublement perturbés par des processus aléatoires. D'abord au niveau de leur racine, puis lors de la construction de chaque nœud. Ces processus de randomisation engendrent des prédicteurs individuels qui performent sur un sous-sous-espace des données d'apprentissage. L'ensemble des informations est ensuite agrégé par un processus ensembliste qui génère un objet statistique unique nommé Forêt Aléatoire. Les performances explicatives et prédictives des FA dépassent largement celles de CART.

Granularité : ce substantif est issu de la métrologie. Il définit l'ensemble des caractéristiques dimensionnelles constitutives d'un objet particulier. Dans cette recherche le terme est utilisé pour définir les informations des bases de données. En particulier leurs échelles - spatiales et temporelles,

leur nature – qualitative ou quantitative, leurs unités, leur précision, leurs lacunes, les quantités de valeurs disponibles.

Grille de lecture BMVG : à l'instar de MVG la méthode propose une grille de lecture qui permet d'identifier des Déterminants Environnementaux de Santé (DES) et des Facteurs de Risques Environnementaux Contributifs ou Probablement aggravants (FREC/FREPA) qui conditionnent les états de santé individuels - i.e. les séquelles développées par les patients de la Cohorte LEA.

Grille de lecture MVG : la méthode est constituée d'une grille de lecture. Elle permet d'identifier les Déterminants Environnementaux de Santé (DES) et les Facteurs de Risques Environnementaux Contributifs ou Probablement aggravants (FREC/FREPA) qui conditionnent la géographie des états de santé étudiés, i.e. des séquelles CATA, THYR, TUM2.

Indicateurs spatiotemporels environnementaux (i.st.e) : éléments fragmentaires constitutifs desdits espaces géographiques numériques. Ces indicateurs modélisent, à l'échelle des communes et sur la période recouverte par l'étude LEA, la variabilité spatiale et temporelle des Facteurs Environnementaux (FE), pertinents et curieux.

Indicateurs spatiotemporels morbides (i.st.m) : éléments fragmentaires constitutifs desdits espaces géographiques numériques. Ces indicateurs modélisent, à l'échelle des communes et sur la période recouverte par l'étude LEA, la variabilité spatiale et temporelle des Phénomènes Morbides (PM) étudiés.

Individus-centré : concept qui focalise l'intérêt sur les personnes subsumées dans un même groupe ou un même ensemble géographique. Une logique, une donnée ou une approche individus-centrée est « centrée » sur les caractéristiques intrinsèques d'un sujet. Dans l'étude menée sur LEA l'approche individus-centrée est utilisée pour valider les résultats de l'approche géographique. L'approche géographique porte sur les FIM* - i.e. sur l'agrégation des Caractéristiques Individuelles et Médicales (CIM) spatialisées dans les unités territoriales - par opposition à l'approche individus-centrée qui porte directement sur les CIM* des patients.

Imagerie par Résonance Magnétique (IRM) : les IRM représentent l'ensemble des EML* utilisant le principe de résonance magnétique nucléaire. Cette technique d'imagerie est utilisée pour observer les tissus mous, dont les cancers font partie.

Inconsistance : ce terme dénote un manque de cohésion, de robustesse ou de rigueur. Un problème d'inconsistance statistique caractérise une dialectique fondée sur des indicateurs inconsistants - i.e. des statistiques estimées sur un nombre d'individus trop petit, ou utilisant des méthodes inadaptées à la modélisation du phénomène géographique considéré.

Items sanitaires : Dans cette thèse l'expression désigne l'ensemble des objets géographiques qui caractérisant l'offre de soins à l'échelle des territoires. Les items sanitaires considérés dépendent des données publiques disponibles inhérentes à la géographie des tissus sanitaires.

Leucémie Aigues ou Chroniques (LA ou LC) : La leucémie est un cancer qui affecte la moelle osseuse. Les Leucémies Aiguës (LA) sont généralement curables et constituent le cancer de l'enfance le plus fréquent. A l'inverse, les Leucémies Chroniques (LC) engagent très souvent le pronostic vital. Avec le temps, les patients en rémission peuvent rechuter et une LA peut dégénérer en LC.

Leucémies Aigües Lymphoïdes (LAL) : il existe deux types de LA. Les LAL caractérisent une prolifération de lymphocytes défaillants et constituent environ 90% des LA infantiles. Elles sont généralement moins dangereuses que les LAM.

Leucémies Aigües Myéloïdes (LAM) : Les LAM sont des leucémies qui affectent la lignée myéloïde des leucocytes. Chez les enfants elles sont plus rares que les LAL. Cependant elles ont tendance à devenir chroniques, donc à engager le pronostic vital.

Machine Learning : cf. datamining

Mode d'Exercice Particulier (MEP) : les MEP sont des médecins généralistes ayant des compétences connexes telles que l'acuponcture, l'homéopathie, l'angiologie…

Méthodes d'ensemble : ensemble des méthodes d'inférence statistique qui consistent d'abord à ré-échantillonner les données d'apprentissage, puis à appliquer une règle, i.e. un modèle statistique. Et enfin, à agréger l'ensemble de ces prédicteurs individuels générés dans un opérateur statistique unique.

MyVsurfGéo (MVG) : nom de la méthode proposée pour caractériser les interactions spatiales statistiques entre les i.st.m* et les i.st.e, qui modélisent la géographie des PM* et des FE/FIM* considérés. MVG est la transposition de VSURF à la dialectique géographique. L'algorithme reproduit exactement les opérations de la version béta de VSURF : injection de connaissances expertes, disjonction des variables de bruit de celles explicatives, et identification des variables prédictives. L'avantage de MVG est de procéder, en amont, à la calibration du modèle FA afin de donner plus de robustesse à la procédure. MVG permet aussi de retenir un paquet de variables explicatives plus conséquent que celui par défaut, et de renvoyer des items - statistiques et graphiques - adaptés à l'interprétation géographique des résultats.

Pertinent : cf**.** Facteurs Environnementaux ;

Phénomène Morbide (PM) : maladie que l'on observe de façon objective et qui est susceptible de se produire de façon universelle. Dans cette recherche l'expression est connotée d'une dimension géographique - ou kantienne. Il est donc question d'états de santé observés expérimentalement dans l'espace et le temps.

Phénomènes Morbides (PM) d'intérêt : Dans cette recherche l'expression définit les PM* étudiés. Et plus particulièrement, la dimension géographique des séquelles cataractes (CATA), tumeurs thyroïdiennes (THYR) et tumeurs secondaires majeures (TUM2).

Protocoles de traitement : ensemble des processus médicaux qui permettent d'obtenir la guérison ou la rémission d'une maladie. Dans le cadre de l'étude menée, la Cohorte LEA, onze types de protocoles sont cités : LAL-80 ; LAL-84-85 ; LAM-80 ; EORTC ; Fralle 92-93 ; Fralle 2000 ; LAM – 89-91 ELAM-02 Atre-LAME ; Autre ; NR. Ils peuvent être subdivisés en deux sous-ensembles selon le type de leucémie traitée, i.e. LAL ou LAM. Au-delà de cette discrétisation, il n'existe pas de différence thérapeutique significative. Les méthodes médicales sont identiques. Les médicaments prescrits contiennent les mêmes principes actifs, seuls les adjuvants diffèrent.

***randomForest* :** package du logiciel R dans lequel est implémentée la version officielle de l'algorithme FA. La version utilisée dans cette thèse est *randomForest.V4.6-7.*

***rpart* :** package du logiciel R dans lequel est implémentée la version officielle de l'algorithme CART.

Santé environnementale : Le concept de santé environnementale a été défini à Frankfort lors de la conférence sur l'environnement et la santé en 1989, comme les aspects de la santé humaine et des maladies qui sont déterminés par l'environnement. Il s'agit d'un concept aux orthographes et aux

définitions protéiformes. Ici, la *santé environnementale* est employée dans le sens qui lui a été conféré lors de la conférence européenne d'Helsinki en 1994. C'est-à-dire comme les effets possibles sur la santé de l'ensemble des facteurs qui caractérisent l'environnement des milieux de vie - en se focalisant davantage sur leur dimension géographique.

Scanner (SCAN) : cette terminologie définit tous les EML* utilisant les rayons X. Ils permettent de détecter des anomalies invisibles avec des appareils standard de radiologie. Ils sont utilisés pour poser des diagnostics et aussi faciliter les traitements de nombreuses maladies.

Tissus sanitaires : ensemble des praticiens de santé libéraux et des caractéristiques du plateau technique des établissements de santé, i.e de leurs services et de leurs Equipement Matériels Lourds (EML*).

Tomographie par Emissions de Positons (TEP) : il s'agit d'un EML* d'imagerie de pointe utilisé en médecine nucléaire. Lors de la phase de diagnostic il permet notamment de préciser la gravité ou le stade de certaines pathologies, dont les cancers et les cataractes font partie.

Variable Sélection Using RandomForest (VSURF): nom d'un algorithme et éponyme d'une méthode de sélection de variables imaginée par Robin Génuer en 2010. La fonction VSURF a été programmée en R par Robin Genuer, Jean-Michel Poggi et Christine Tuleau-Malot. Une version Bêta a été mise en ligne pour la première fois en août 2013. La méthode est fondée sur les Forêts Aléatoires. Elle permet d'expliquer et de prédire les interactions entre les variables dans des jeux de données multidimensionnels complexes. Il s'agit d'une méthode de sélection dite par seuillage, i.e. que les résultats sont déterminés par des seuils statistiques. VSURF constitue une évolution de la méthode imaginée en 2010. Désormais chaque seuil peut être pondéré sur des considérations expertes. Cette évolution heuristique est une avancée notable dans la transposition de la méthode à d'autres domaines d'application.

Variables régionalisées (v.r.) : Dans la théorie géostatistique les variables régionalisées sont des objets mathématiques. Elles peuvent être modélisées de façon continue dans l'espace par le biais des Fonctions Aléatoires Intrinsèques (FAI). Toute mesure temporelle géo-localisée peut être interprétée comme la réalisation d'une v.r. en un lieu et un moment précis, i.e. les mesures météo de la température ou celles de l'activité volumique de certains éléments radioactifs.

SOMMAIRE

Les mots définis dans le glossaire facilitent l'intelligibilité du texte. Ils sont identifiés par un astérisque (*).

L'italique est utilisé pour citer littéralement l'auteur spécifié en fin de phrase ou de paragraphe. Très rarement, l'italique permet de distinguer les idées associées à la source et les constructions heuristiques personnelles.

INTRODUCTION GENERALE

Cette thèse de géographie s'inscrit dans le cadre d'une recherche interdisciplinaire sur une thématique de santé publique. Le positionnement retenu est d'ordre méthodologique.

La géographie *propose* une approche *phénoménologique et globale des problématiques et* l'espace constitue son angle d'attaque.
Phénoménologique parce qu'elle se fonde sur l'observation des rapports entre l'homme et le monde, à partir de données expérimentales, qui peuvent être quantitatives ou qualitatives, et en conférant une attention particulière à leurs significativités sociales (Merleau-Ponty, 1945).
Aussi, cette approche est globale, en ce sens que l'étude des relations entre l'homme et son milieu s'attache à toute la pluralité des dimensions géographiques : socio-culturelles, économiques, sanitaires et bio-environnementales.
Enfin, la démarche consiste à construire un *espace géographique** qui permet d'analyser, à partir d'indicateurs, les interactions spatiales entre une population et son environnement de vie – en intégrant *les enjeux sociaux* associés aux espaces.
La puissance heuristique de l'analyse s'articule autour de la façon dont la société gère – ou plutôt *contrôle* – l'espace. Par conséquent, *l'espace géographique* ne trouve sa pleine intelligence qu'à l'échelle des territoires* en leur conférant ainsi *une dimension socio-politique opérationnelle* (Salem, 1995).

Cette recherche a pour dessein la modélisation et l'identification de Facteurs Environnementaux* (FE) liés à des risques morbides. L'idée est d'apporter des éléments épistémologiques et opérationnels pour tenter de répondre à la question que se posent tous les géographes : *Pourquoi [tel phénomène, telle maladie, s'observe plutôt] ici et moins ailleurs ?* - (Durand-Dastès et Mutin, 1995).

La dialectique se veut heuristique et opérationnelle. La démarche méthodologique proposée est appliquée à la Cohorte LEA (Leucémie chez l'Enfant et Adolescent), une base de données épidémiologiques qui recense des individus traités, en France, pour une leucémie infantile, depuis 1980. Les leucémies sont des cancers qui se caractérisent par une prolifération de lymphocytes immatures dans la moelle osseuse. Il en existe deux types : Les Leucémies Aigües Lymphoïdes* (LAL) qui représentent environ 85% des leucémies infantiles et les Leucémies Aigües Myéloïdes* (LAM), plus rares, qui ont tendance à devenir chroniques, ce qui engage le pronostic vital. Actuellement 80% des Leucémies Aigües (LA) sont curables. Mais avec le temps les patients développent des séquelles qui ont un impact sur leur Qualité de Vie (QV). Le nombre ainsi que la gravité des séquelles observées sont conditionnés par l'agressivité de la leucémie et du traitement reçu. Mais pas seulement. *L'environnement a aussi, probablement, une influence* (Leplège, Ecosse et al., 1998).

Les séquelles sur lesquelles l'attention est portée sont : les cataractes (CATA), les tumeurs thyroïdiennes (THYR) et les tumeurs secondaires majeures (TUM2). Les cataractes se caractérisent par une opacification du cristallin qui rend la vision difficile. Elles sont généralement développées par des individus âgés de 65 ans ou plus. Toutefois, lorsque les sujets sont prédisposés, elles peuvent être observées dans l'enfance. (SFO, 2013). Les tumeurs secondaires sont des cancers qui se développent à distance du site initialement atteint. Ces cellules malignes sont appelées : métastases. Il s'agit d'un fléau qui s'est développé à la marge de l'élaboration de traitements anti-cancer de plus en plus efficaces. Elles apparaissent généralement entre deux et dix ans après le traitement d'un cancer. Les tumeurs thyroïdiennes font partie des tumeurs secondaires et sont grevées d'un taux de mortalité élevé, elles ont un impact important sur la QV* (HAS et INCa, 2010).

Les causes documentées pour les cataractes, les tumeurs secondaires et en particulier les tumeurs thyroïdiennes, sont liées à des anomalies génétiques, des expositions accidentelles à des rayonnements ionisants, des chimiothérapies agressives et certaines maladies infectieuses. Quant aux Facteurs

Environnementaux* qui les conditionnent – hormis les catastrophes nucléaires, quelques maladies vectorielles particulières et la présence naturelle, ou liée à des déversements illégaux, de fortes concentrations de substances toxiques dans certaines zones géographiques bien localisées –ils restent controversés (IARC, 2008).

Alors que les incertitudes qui maculent les liens entre l'environnement géographique et la santé des populations semblaient avoir clos le débat, *de récentes études épidémiologiques l'ont relancé* (Afsset, 2009a).

Ce regain d'intérêt pour l'analyse des interactions santé-environnement est lié à l'émergence de très nombreuses de bases de données protéiformes (Zeitouni, 2006), et à *de nouveaux outils mathématiques dotés d'une puissance statistique impressionnante*, *qui offrent des perspectives de recherche prometteuses*. En l'occurrence, *c'est le cas du datamining et de ses applications aux sciences humaines et sociales* (Cartier, Villani et al., 2012).

Cette recherche se positionne dans le champ de la géographie de la santé, c'est-à-dire de *l'analyse spatiotemporelle des environnements spatiaux qui influencent l'état de santé d'une population* (Picheral, 1989).

Les confusions *entre santé et médecine sont fréquentes* tant au niveau de la définition que des terminologies employées. Par définition, *la santé renvoie aux conditions d'existence sous toutes ses formes*. *Un état de santé est la résultante, à un moment donné, en un lieu donné,* d'un système d'interaction complexe entre des *facteurs endogènes ou exogènes*. Ces terminologies médicales sont équivalentes à celles de *Facteurs Individuels et Médicaux** - i.e. comportementaux, biologiques, génétiques – et de Facteurs Environnementaux* (FE) - i.e. les éléments physiques, chimiques sociaux, économiques et politiques des milieux de vie (Amat-Roze, 2011), plus utilisées en géographie.

Les disparités géographiques observées sont le produit de combinaisons différenciées de ces facteurs. Ces dernières s'étudient généralement dans une logique quantitativiste, *axée sur de l'analyse spatiale* (Voiron-Canicio, 1995).

Le rôle du géographe de la santé est d'identifier des disparités spatiales morbides et les Facteurs Environnementaux* (FE) qui les conditionnent.
*L'espace géographique** est construit à partir d'indicateurs aux composantes spatiales et temporelles.
Ceux utilisés pour modéliser la géographie des états de santé étudiés permettent de *justifier, aux yeux de la société, l'intérêt de mettre en place des mesures pour les améliorer.*
Les indicateurs représentant les faits de santé, i.e. l'environnement des milieux de vie, qui sont supposés *pertinents* au regard des pathologies étudiées.
Dans la mesure où *l'approche est globale*, elle ne se limite pas à l'étude spatiale du *système de soins*. Elle prend en compte toute la complexité du *système de santé,* i.e. toute la *pluralité de ses dimensions géographiques* : sanitaires (SAN), socio-économiques (SOCIO.ECO), physicochimiques (PHY.CHIM) et certaines Caractéristiques Individuelles et Médicales*(CIM), des populations ciblées.
L'analyse des interactions santé-environnement s'opère sur les indicateurs proposés par le biais de *méthodes statistiques adaptées*.
Les objectifs visés sont :
Participer au progrès des connaissances en épidémiologie ; Identifier les *combinaisons de facteurs* qui influencent les disparités géographiques morbides observées ; Caractériser les espaces en fonction des risques morbides associés à l'environnement spatial ; Et enfin, Proposer des *leviers* adaptés aux attentes sociales permettant de mettre en place des mesures pour favoriser l'accès à une bonne santé environnementale* (Salem, 1995).

Cependant, l'acuité du géographe de la santé, même particulièrement affutée, n'est pas assez perçante pour que ce dernier puisse à lui seul dissiper les ombres qui maculent l'espace de cette problématique

complexe de santé publique. Afin d'y parvenir il devra inéluctablement s'élever au-dessus des brumes de l'incertitude et n'aura pas d'autre choix que de se jucher sur les épaules des géants de l'épidémiologie (Kenneth et Sander, 1998) et de la statistique (Saporta, 2006).

Nous sommes des nains juchés sur des épaules de géants, outre la métaphore empruntée à Bernard de Chartres, cette expression est utilisée pour marquer l'importance de s'appuyer sur des connaissances interdisciplinaires, du présent et du passé (Jeauneau, 1967).

L'interdisciplinarité en recherche est nécessaire. Il s'agit même d'un *devoir lorsque les disciplines scientifiques se complètent et permettent* d'améliorer le processus de compréhension de *problématiques transversales complexes*. L'amélioration : de *la santé environnementale** en particulier et plus généralement de *la qualité de vie,* font partie de ce type de problématiques (Fromageot, Coppieters et al., 2005). Le concept de *santé environnementale* a été défini, lors de la conférence européenne de Helsinki en 1994, comme les effets possibles sur la santé de l'ensemble des facteurs qui caractérisent la qualité de l'environnement des milieux de vie (OMS, CC, 1994).

La Qualité de Vie après un cancer, et la santé des enfants, sont au cœur des préoccupations sociales et politiques dans les pays développés. L'espérance de vie a considérablement augmenté grâce aux progrès effectués en médecine. Mais *les cancers sont des maladies au long sillage* et comme, dans ces pays, l'accessibilité à des traitements adaptés est bonne, l'incidence des effets au long cours de type : séquelles, ne cesse d'augmenter. *Chaque patient guéri accroit le nombre de malades potentiels.* Les occurrences morbides observées sont assujetties, à des échelles locales, à de fortes disparités géographiques et - malgré la complexité des interactions spatiales entre la santé et l'environnement - il est du devoir du géographe de s'y intéresser (Peguy, 1996).

En géographie, *la complexité* caractérise *les systèmes spatiaux* constitués d'éléments qui interagissent entre eux et avec l'environnement dans lequel ils sont subsumés. Ces *interactions spatiotemporelles internes ou externes* contribuent à l'évolution des phénomènes géographiques complexes qui se dessinent par des formes spatiales, d'apparence aléatoire lorsqu'ils sont étudiés à des *échelles microscopiques*. En contrepartie, la propriété fondamentale *des systèmes complexes est l'émergence de processus immanents* dès lors qu'ils sont analysés de façon plus globale, plus imprécise, i.e. à des *échelles macroscopiques* (Pumain, 2004).

Pourtant *l'analyse de la complexité* des interactions spatiotemporelles entre la santé et l'environnement par le prisme de la géographie de la santé, dans le cadre d'une approche globale, macroscopique et vouée à la transposition de méthodes interdisciplinaires, *est encore rare malgré les perspectives théoriques promises* (Gatrell, 2005).
Or cet angle d'attaque est *a priori adapté* pour identifier des Déterminants Environnementaux de Santé* (DES) et des Facteurs Environnementaux : de Risques Contributifs (FREC) ou ceux Potentiellement Aggravants (FREPA).
Et par suite, pour proposer des solutions médicales (i.e. individuelles) et politiques (i.e. collectives) permettant de favoriser l'accès à une meilleure santé environnementale*, et générer ainsi les dynamiques économiques nécessaires au développement durable des espaces (Wallace R., Wallace D. et al., 1999).

Les fondements théoriques de la dialectique proposée dans le cadre de cette thèse ont été sommairement énoncés. En pratique, ces considérations conduisent d'une manière générale à considérer les causes des Phénomènes Morbides* (PM) étudiés – les séquelles - comme multifactorielles.
Elles sont en grande partie déterminées par des motifs génétiques, individuels et médicaux, mais ces derniers ne suffisent pas à expliquer les tendances géographiques observées. Et, il ne fait nul doute que l'environnement joue un rôle important. Ce dernier est multidimensionnel. Dans cette recherche il est considéré comme tel.

Les méthodes heuristiques proposées sont reproductibles à toutes sortes de maladies. L'idée est d'abord de décrire, par le bais d'indicateurs spatiotemporels, la géographie des PM* d'intérêt. Puis, de modéliser la géographie des Facteurs Environnementaux* (FE) susceptibles d'avoir des effets sur ces pathologies. En l'occurrence, il sera question de tous les FE* ayant des effets cliniques - déterministes ou stochastiques – dont les expositions environnementales sont avérées, ou simplement suspectées si tant est qu'elles soient documentées.
Ensuite, une méthode innovante permettra de *sélectionner* tous les indicateurs spatiotemporels environnementaux* (i.st.e) intégrés qui sont statistiquement capables d'expliquer et de prédire les indicateurs spatiotemporels morbides* (i.st.m) proposés.
L'analyse des i.st.e* explicatifs servira de base pour identifier les DES*, les FREC* et des FREPA*, à partir desquels les espaces peuvent être caractérisés par des Risques d'Expositions Géographiques* (REG) morbides.

Les résultats permettront de conjecturer des *leviers opérationnels* pour mettre en place des mesures médicales et politiques afin d'améliorer la santé environnementale* des populations – au regard du gain potentiel attendu sur les dynamiques sociales et économiques.

L'approche géographique proposée pour aborder cette thématique de santé publique est à la fois méthodologique, phénoménologique - i.e. qui repose sur des observations expérimentales considérées globalement à un niveau macroscopique (Merleau-Ponty, 1945) - et *radiale* - i.e. *qui quantifie pour étudier les inégalités d'accès à une bonne santé* (Grawtiz, 2000).

L'articulation de la dialectique proposée dans le cadre cette recherche, i.e. de la démarche retenue ainsi que des objectifs méthodologiques à atteindre, sont synthétisés dans le schéma synoptique 1, où chacun des éléments correspond aux étapes didactiques décrites dans les quatre chapitres cette thèse.

La vérité est le but, l'espace est la norme – i.e. le moyen de rationaliser, ou de se *rapprocher de la vérité* […] *par les quelques illusions qu'elle dissipe,* ou espère dissiper. Il s'agit de répondre, avec un regard spécialement spatial, à une problématique de santé environnementale* complexe dans le but de *mieux penser pour mieux vivre* (Comte-Sponville, 2000).

Conception d'une dialectique adaptée à la géographie de la santé

Choix d'une position scientifique et définition des objectifs de recherche.

Bilan des connaissances sur les interactions entre les états de santé étudiés et les causes cliniques et environnementales

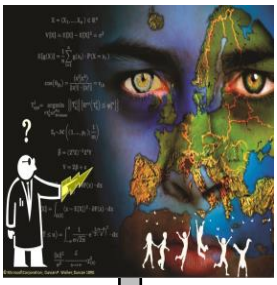
Etat de l'art sur les méthodes d'analyse statistique et de modélisation en géographie de la santé.

Hypothèses heuristiques

- Spécification de concepts pour étudier l'effet morbide de l'environnement.
- Discrétisation de l'environnement en facteurs individuels et médicaux, sanitaires, socio-économiques et physicochimiques.
- Choix d'une stratégie d'analyse statistique multidimensionnelle et d'une échelle d'investigation adéquate.

Analyse des caractéristiques de la base de données épidémiologiques LEA et des bases de données environnementales disponibles.

Propositions d'alternatives méthodologiques plus robustes utilisées dans les domaines de l'environnement et de l'apprentissage statistique.



Modélisation géographique des phénomènes morbides

Spatialisation des individus

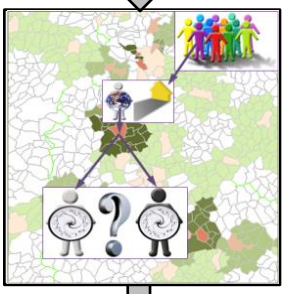
- Proposition d'une stratégie de spatialisation des patients - de la cohorte LEA - à l'échelle des communes et à partir de leur code postal.
- Modélisation de l'incertitude spatiale associée aux codes postaux.
- Estimation de la robustesse de la méthode.

Représentation des phénomènes morbides

- Proposition d'un système expert de pondérations pour atténuer les incertitudes spatiotemporelles des données épidémiologiques - LEA.
- Proposition d'un indicateur spatiotemporel modélisant la géographie des phénomènes morbides et d'un indicateur spatiotemporel renseignant sur la fiabilité du premier.

Caractérisation des espaces

- Proposition d'un indicateur spatial caractérisant les espaces par des risques d'expositions géographiques morbides à partir des données épidémiologiques observées - LEA.
- Analyse de la significativité socio-spatiale de l'indicateur et de son utilité en matière de management durable des territoires.



Modélisation géographique de facteurs environnementaux

Démarche conceptuelle

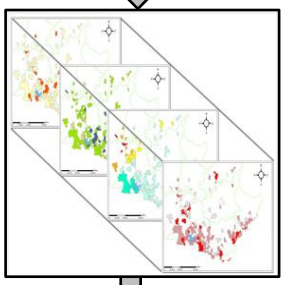
- Proposition de stratégies spatiotemporelles pour combler les lacunes et fusionner des connaissances expertes - sémantiques ou géographiques - afin d'améliorer la qualité des données sources.
- Proposition de stratégies pour uniformiser les échelles spatiales et temporelles des données inputs.

Opérationnalisation des stratégies

- Spécification des méthodes et des outils - SIG ou statistique - pour mettre en œuvre les stratégies conceptuelles.
- Spécification des particularités d'application aux données mobilisées pour modéliser les facteurs individuels et médicaux, sanitaires, socio-économiques et physicochimiques.

Analyse de la robustesse

- Etude de la capacité des indicateurs spatiotemporels à représenter les disparités de l'environnement spatial.
- Etude des limites inhérentes aux indicateurs spatiotemporels sur leur aptitude à modéliser les facteurs environnementaux pour lesquels ils ont été proposés.



Identification statistique de facteurs morbides

Processus méthodologique

- Proposition d'un algorithme de datamining adapté à la complexité des jeux de données géographiques.
- Proposition d'une méthode de sélection permettant d'analyser les interactions statistiques multidimensionnelles entre les indicateurs spatiotemporels morbides et environnementaux.
- Mise au point d'une stratégie pour transposer la procédure de sélection à la dialectique géographique - et application.

Processus heuristique

- Conception d'une méthode d'analyse géographique et de validation des résultats.
- Elaboration d'une stratégie d'identification de déterminants environnementaux de santé et de facteurs de risques contributifs ou potentiellement aggravants
- Proposition d'une méthode d'apprentissage géographique et d'analyse directement applicable sur l'état de santé des patients, i.e. sur les données épidémiologiques, afin de valider les résultats.

Processus socio-spatial

- Mise au point d'une stratégie prédictive des risques d'expositions géographiques morbides à partir des facteurs environnementaux qui expliquent la morbidité spatiale.
- Analyses des perspectives en matière de management durable des territoires.
- Analyse des incertitudes inhérentes aux indicateurs spatiotemporels proposés et de la pertinence des facteurs environnementaux morbides qui leur sont associés.

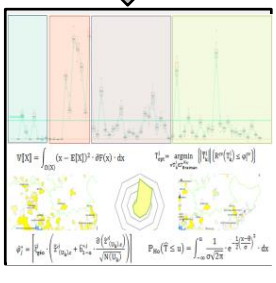


Schéma synoptique 1 : Objectifs méthodologiques

PARTIE.I : ETAT DE LA CONNAISSANCE ET MODELISATION GEOGRAPHIQUE DE PHENOMENES MORBIDES

CHAPITRE 1 : OBJECTIFS ET POSITIONNEMENT SCIENTIFIQUE

Le premier chapitre propose une approche transversale de la problématique. Il s'agit d'imaginer une dialectique permettant de mettre au point une méthode de modélisation géographique de phénomènes morbides et environnementaux, puis d'identifier ceux qui déterminent l'état de santé des populations – en particulier dans les zones où les risques d'expositions géographiques morbides sont les plus probables.
Le contexte communautaire dans lequel s'inscrit cette recherche et le positionnement scientifique retenu, permettent de spécifier les objectifs heuristiques ainsi que les contributions attendues, à l'échelle individuelle et collective.
Un état de l'art des méthodes de modélisation et d'analyse ainsi qu'un bilan des connaissances pluridisciplinaires disponibles en santé environnementale* sont dressés de façon à préciser les contours de cette problématique de santé publique. Ils constituent une approche bibliographique pour orienter la dialectique géographique, poser des hypothèses et proposer des instruments scientifiques adaptés. Ensuite une revue des données mesurées et mesurables existantes et une revue de celles mobilisées – ainsi que de leur granularité* spatiotemporelle - vont permettre de spécifier les premiers éléments pragmatiques, notamment ceux relatifs au choix de l'échelle d'investigation.

SECTION A) PROBLEMATIQUE ET PORTEE DE LA RECHERCHE

La dialectique, les méthodes et les applications proposées portent sur les expositions environnementales liés à des risques morbides, et se focalisent plus particulièrement sur l'après cancer. Ce type d'étude constitue une part importante des recherches menées en santé publique sur la thématique « cancer & environnement ». Leur cristallisation est essentiellement due aux mesures instituées par le Plan Cancer 2 (2009 – 2013) et le Plan National Santé Environnement 2 (PNSE 2, 2009 – 2013). L'analyse des interactions entre les états de santé et l'environnement est abordée à travers le prisme de la géographie, dans le cadre d'une approche spatiotemporelle.
En France comme dans le reste de l'Europe, le principe de précaution est fondamental en droit de l'environnement et où l'incidence des cancers et des maladies en général ne cesse de croitre, on observe une forte demande sociale d'informations à ce sujet. Elle est alimentée par des croyances et des incertitudes scientifiques qui déterminent un contexte particulier.

LE CONTEXTE EUROPEEN ET NATIONAL

Durant le siècle dernier l'espérance de vie a considérablement augmenté dans tous les pays de l'Union Européenne. Cette mutation démographique a engendré des modifications sociales, économiques et environnementales qui ont eu des répercussions, parfois néfastes, sur la santé des populations et qui se sont peu à peu placées au centre des préoccupations politiques (Olshansky S-J. et al., 2005) ; (Krewski, 2009). Et pour cause, puisque la fonction première de l'état est *de protéger la santé individuelle et promouvoir la santé publique*. Or la mise en place de politiques et de mesures visant à améliorer la santé environnementale* est sous le joug d'une forte pression sociale fondée sur des croyances, des principes,

des connaissances et des ambitions influencés – pour ne pas dire dictés - par un contexte communautaire (Salem, 1995).

En Europe la première grande mesure instituée par la Commission des Communautés Européennes (CEE) est la stratégie SCALE (Science, Children, Awarreness, Legal instrument, Evaluation). Son principal objectif est de mieux comprendre les interactions santé-environnement en vue de développer des politiques communautaires visant la réduction des risques sanitaires. Le premier cycle de ce programme 2004-2010 se concentrait sur les FE* ayant des effets potentiels sur l'incidence de nombreuses maladies, et portait une attention particulière aux cancers. Les objectifs scientifiques fixés consistaient à :

- Proposer des indicateurs permettant de caractériser l'état de santé des populations et celui l'environnement ;
- Développer des méthodes permettant d'établir des relations entre les indicateurs modélisant l'une et l'autre des deux composantes ;
- Elaborer des outils géographiques permettant d'évaluer les risques environnementaux auxquels sont assujetties les populations (Commission des Communautés Européennes, 2003).

Ensuite un livre blanc définissant la nouvelle stratégie communautaire de santé environnementale* a été rédigé en 2008, dans le prolongement de SCALE. Les objectifs scientifiques fixés sont à peu près identiques à ceux de la première stratégie mais une volonté de mettre en place des mesures de solidarité communautaire, avant 2013, a été rajoutée en vue de réduire les inégalités géographiques en matière de santé publique. Une des premières mesures fut de diffuser les résultats acquis par SCALE ayant permis de préconiser des mesures préventives d'une part, et de créer des bases de données contenant des indicateurs de santé d'autre part. Cependant, l'évaluation des interactions avec l'environnement est complexe et controversée. De fait, les objectifs scientifiques ont été reconduits et spécifiés. Désormais, une attention particulière est portée à l'analyse des effets sanitaires liés à l'action combinée de Facteurs Environnementaux* (FE) physicochimiques et socio-économiques particuliers. Quant à l'enjeu il s'agit toujours de détecter les populations les plus exposées à des Déterminants Environnementaux de Santé* (DES) (Commission des Communautés Européennes, 2007).

Les préconisations de la CEE en matière de santé publique inhérentes à la stratégie SCALE et celles listées dans le Livre Blanc Communautaire ont largement influencé les recommandations de l'Organisation Mondiale de la Santé (OMS) sur la thématique santé-environnement. En particulier, la conférence de Bucarest en 2004 a débouché sur un plan d'action ayant pour finalité l'amélioration de la qualité de l'environnement et celle de la santé des enfants (OMS, 2004). Quant à la conférence qui s'est tenue à Parme en 2010, elle a abouti, entre autres, à la mise en place de mesures visant : la construction d'outils voués à l'estimation des effets de l'environnement sur la santé, la protection de la santé des enfants, et la régulation des méfaits des mutations environnementales liées aux changements climatiques (OMS, 2010).

En France, les politiques destinées à mettre en place des mesures opérationnelles pour lutter contre les inégalités de santé environnementale* ont été élaborées dans une logique territoriale. Elles se sont largement inspirées des préconisations de l'OMS et se sont conformées à celles de la CCE, et plus particulièrement au Livre Blanc pour ce qui est de la réduction des inégalités géographiques de santé. Les deux mesures phares élaborées à cet effet sont le Plan National Santé Environnement (PNSE) et les Plans Cancer.

L'émergence du premier PNSE avait pour objet d'instituer 45 mesures entre 2004 et 2008. Il s'agissait essentiellement d'améliorer les connaissances sur l'impact des Facteurs Environnementaux * et l'état de santé des populations. Il était question, notamment, de mettre en place des systèmes d'information visant à surveiller les activités émettant des substances toxiques dans les compartiments

environnementaux (eau, air, sol et biologique) ; et parallèlement, de dresser un bilan des dysfonctionnements des tissus sanitaires* territoriaux. Aussi, une attention particulière avait été portée au développement d'indicateurs et de méthodes destinées à évaluer les interactions spatiotemporelles entre la santé et l'environnement en vue de caractériser les populations à risque (Ministère de la Santé et de la Protection sociale, Ministère de l'Écologie et du Développement durable, Ministère de l'Emploi, du Travail et de la Cohésion sociale, Ministère délégué à la Recherche, 2004). Cette logique est dans la lignée des engagements du Grenelle de l'environnement 2007-2012 qui se focalisait, entre autres, sur l'urgence qu'il y a à développer des outils prédictifs permettant d'évaluer l'exposition des populations à des risques morbides liés à des contextes environnementaux géographiques particuliers (Ministère de l'Ecologie, du Développement durable et de l'Energie, 2013). Par suite, le PNSE2 a été élaboré. Ses objectifs principaux reprennent ceux du Grenelle de l'environnement et spécifient ceux du PNSE1. Une attention particulière est ainsi portée à l'identification de zones géographiques où les populations sont surexposées à des risques morbides. En l'occurrence le PNSE2 met l'accent sur la réduction des inégalités des expositions environnementales en lien avec le cancer. L'enjeu est clair. Les actions induites par le PNSE2 ont été intégrées dans les politiques régionales de santé publique grâce à des Plans Régionaux de Santé Environnement (PRSE2) constituées de mesures opérationnelles adaptées aux problématiques de santé environnementale* locales (Ministère de l'Ecologie de l'Energie du Développement Durable, Ministère de la Santé et des Sports, Ministère de l'Enseignement Supérieur et de la Recherche, 2009).

Les mesures instituées par les PNSE agissent en synergie avec celles du Plan Cancer.1 (2003-2007) et celles du Plan Cancer.2 (2008-2013) qui visent à mettre en œuvre des mesures médicales et politiques opérationnelles de lutte contre les inégalités des risques associés au cancer. Ces mesures portent notamment sur le financement de travaux scientifiques visant à construire des indicateurs de santé multi-scalaires permettant de croiser l'évolution des risques liés aux cancers avec des contextes de « défaveur », démographique, socio-économique, d'exposition environnementale ou professionnelle, à des substances toxiques documentées ou encore relatifs à l'architecture des tissus sanitaires* territoriaux (Ministère de l'Enseignement Supérieur et de la Recherche, Ministère de la Santé et des Sports, 2009).

Le contexte permet de prendre conscience des enjeux sanitaires, des paradigmes intellectuels et des questions qui animent la vie politique, économique et sociale contemporaine. Il va permettre d'orienter le positionnement scientifique à adopter pour répondre à la problématique de santé publique énoncée auparavant.

POSITIONNEMENT SCIENTIFIQUE

Le positionnement scientifique est capital puisqu'il conditionne la dialectique, et donc la manière d'aborder la problématique. L'angle d'attaque est celui de la géographie de la santé qui est définie comme *l'étude spatiale de la qualité de la santé des populations, de leurs comportements et des facteurs de leur environnement qui concourent à la promotion ou à la dégradation de leur santé* (Picheral, 1996).

Cette position scientifique implique une dialectique particulière qui consiste à décrire, comprendre et expliquer *l'état de santé* d'une population – par exemple celui de personnes traitées pour une leucémie dans leur enfance - à partir de leurs caractéristiques environnementales. Pour ce faire, le géographe construit *l'espace géographique*.* Il modélise, à partir d'indicateurs spatiaux, *des faits de santé* qui sont *supposés pertinents par rapport à l'indicateur géographique de santé retenu.* Les *états de santé* sont les manifestations morbides, en l'occurrence les séquelles développées. Quant aux *faits de santé,* ils caractérisent l'environnement spatial à un moment précis, il s'agit de ce que nous appelons ici les Facteurs Environnementaux* (FE) susceptibles d'interagir avec les phénomènes morbides. Ils peuvent être *biochimiques, médicaux, sociaux, économiques, etc.* (Salem, 2003).

Ensuite le rôle du géographe est d'identifier la combinaison de Facteurs Environnementaux* (FE) qui, sur un espace donné, influencent *l'état de santé* de la population. On appelle ces facteurs explicatifs les déterminants de santé. *Le diagnostic des déterminants de santé* permet d'analyser le système de santé, i.e. l'ensemble des pratiques sociales sur l'espace qui influencent *les états de santé*. L'échelle géographique doit être *intelligemment spécifiée* de façon à conférer aux politiques une dimension opérationnelle leur permettant *une gestion orientée de l'espace* par le biais du *contrôle territorial*, i.e. par des actions directes ou indirectes sur *les faits de santé*. Par extension, il s'agit d'octroyer à l'Etat la capacité d'assurer une de ses fonctions premières : *protéger la santé individuelle et promouvoir la santé publique*. En contrepartie, le géographe a un devoir moral – aider à lutter contre les inégalités géographiques d'accès à une bonne *QV* environnementale*, par exemple en dénonçant les disparités d'accès à une bonne santé de sorte que *le contrôle politique territorial* reste sous le joug du sens commun donc sous celui *du contrôle social* (Salem, 1995).

La géographie de la santé grâce aux concepts *de distance, d'échelle et de territoire*, avec lesquels elle aborde les questions de santé, propose une approche différente et plus *globale* que celle de l'épidémiologie (Amat-Roze, 2011).

Toutefois, la géographie de la santé n'est pas sans faille et les épidémiologistes dénoncent depuis longtemps les carences d'une dialectique peu pragmatique fondée sur des méthodes discutables pour ne pas dire indigentes. Par conséquent, elle est parfois qualifiée *d'épidémiologie populaire.* Les principales attaques ne concernent pas l'approche très transversale des problématiques de santé, au contraire, mais plutôt le manque de rigueur scientifique. Le premier reproche touche à la dialectique, la réflexion de fond étant souvent dépourvue de connaissances épidémiologiques, conséquence d'un raisonnement biaisé sur les pathologies étudiées. Le second reproche concerne les méthodes *de modélisation statistique* utilisées, souvent éculées et peu efficientes (Brown, 1997).

Par conséquent, le positionnement scientifique adopté dans le cadre de cette recherche sera celui de la géographie de la santé tel qu'il est défini par Gérard Salem, mais il tiendra compte des critiques des épidémiologistes relatives à la dialectique et aux méthodes d'une approche globale qui suggèrent aussi ses limites.
Pour commencer, les indicateurs spatiaux sont *supposés pertinents par rapport à l'indicateur géographique de santé retenu*. Or pour que la spécification *des faits de santé* soit *pertinente,* il convient d'intégrer, en amont, une réflexion épidémiologique. Les études menées en géographie de la santé considèrent des Facteurs Environnementaux* (FE) très hétéroclites, parfois inadaptés à la pathologie

étudiée et généralement les Caractéristiques Individuelles et Médicales* (CIM) des populations ciblées ne sont pas intégrées alors qu'elles ont pourtant une importance majeure, *a fortiori* lorsqu'on travaille sur des séquelles (World Health Organization, 2009). D'où l'intérêt de travailler en collaboration avec des épidémiologistes et sur des données Cohorte.
En contrepartie, l'intégration de FE* hétéroclites semble plus pertinente que l'approche épidémiologique qui se limite souvent à l'analyse d'une composante environnementale particulière (World Health Organization, 2009). En géographie de la santé la vision de l'environnement est plus large, car il est considéré dans toutes ses dimensions. Il est cependant à noter qu'à ce jour il n'existe aucune définition de référence, que ce soit à l'échelle nationale, européenne ou mondiale. Il s'agit d'une prise de position scientifique. Celle retenue dans le cadre de cette thèse est celle donnée par l'OMS. L'environnement est décrit par des FE* multidimensionnels, discrétisés en quatre composantes : sanitaire, socio-économique, physicochimique et la dernière intègre les CIM* des sujets en leur conférant une dimension géographique – *autrement dit tous les facteurs possibles à l'exception de ceux génétiques* (Chasles et Fervers, 2011).
Ensuite, s'agissant des carences méthodologiques évoquées, la dialectique géographique suppose que les indicateurs spatiotemporels (i.st.) utilisés modélisent *précisément* la géographie de *l'état de santé* des populations et de leurs *faits de santé*. Sur ce sujet, les critiques des épidémiologistes ne sont pas les plus incisives - l'état de l'art spécifie ce point. Ils sont même d'accord sur le fait que la statistique est le meilleur moyen de réaliser de la *physique sociale* (Quételet, 1969). Ils ne manquent cependant pas de dénoncer *le manque de puissance* des *méthodes paramétriques classiques* utilisées pour la recherche d'interactions entre les indicateurs spatiotemporels morbides* (i.st.m) et les indicateurs spatiotemporels environnementaux* (i.st.e) (Groupe CHADULE, 1997) . Cette critique peut être étendue à la modélisation géographique des FE* physicochimiques. Mais ce reproche n'est désormais que partiellement valable car de puissantes méthodes de modélisation spatiale ont récemment été implémentées dans les Systèmes d'Informations Géographiques (SIG) et plus particulièrement dans le logiciel ArcGis.10 (ESRI, 2013). Les géographes doivent néanmoins s'aguerrir à ces méthodes qui font appel à des connaissances mathématiques poussées. En revanche, pour l'heure, les SIG ne permettent pas d'analyser de façon efficiente les interactions complexes entre un i.st.m* et de grandes quantités d'i.st.e* hétéroclites contenus dans les jeux de données géographiques *dits de grande dimension*, où la *statistique paramétrique classique* est inopérante (Tuleau-Malot, 2005). Par conséquent la phase de *diagnostic des déterminants de santé* est systématiquement biaisée. Les *méthodes d'inférence statistique* permettant de réaliser de telles prouesses restent l'apanage de logiciels mathématiques particuliers, si tant est que les procédures de sélection soient implémentées – sinon elles doivent être programmées (Institute for Statistics and Mathematics, 1997).

Par conséquent, afin de répondre de façon pertinente à cette problématique de santé publique complexe, une position scientifique interdisciplinaire est nécessaire. Et, puisqu'il s'agit d'une thèse d'ordre méthodologique, des liens ont été noués avec des épidémiologistes et des statisticiens, pour lui conférer une dimension opérationnelle.

L'angle d'attaque est celui de la géographie de la santé pour son approche transversale de la santé publique, ses méthodes d'Analyse Spatiale et sa prise en compte multidimensionnelle de l'environnement. De fait, *l'espace est intégré au cœur de la dialectique,* ce qui distingue cette recherche de celles menées en épidémiologie spatiale où il est *généralement considéré comme simple variable* (Voiron-Canicio, 2006). L'espace illumine la dialectique *sans pour autant mettre le temps à l'ombre.* (Brunet, Théry et al., 2009). Pour cette approche, l'encadrement a été assuré par Madame Christine Voiron-Canicio, Directrice de l'UMR ESPACE (Etudes des Structures et des Processus d'Adaptations et de Changement de l'Espace).

Afin d'intégrer des Facteurs Environnementaux* (FE) et des Facteurs Individuels et Médicaux* (FIM) pertinents et aussi pour donner plus de sens à la façon d'aborder cette problématique de santé publique, l'acquisition de connaissances épidémiologiques est nécessaire. La transmission des savoirs

théoriques et pratiques en épidémiologie et sur la Base LEA a été garantie par le Professeur Pascal Auquier, Directeur de l'EA-3279 et Co-Directeur du projet LEA.

Pour ce qui est de la modélisation des phénomènes morbides, des FE* et des FIM* - *jugés pertinents* - par des i.st.m* et des i.st.e* suffisamment robustes pour représenter leur réalité géographique, les techniques *classiques de statistiques spatiales* (Charre, 1995) ont été couplées à des outils probabilistes poussés (Saporta, 2006). Pour donner plus de consistance aux modélisations spatiales des FE* physicochimiques, la transposition de méthodes géostatistiques uni-variables (Matheron, 1965) et multi-variables (Wackernagel, 2003) a été nécessaire. L'état de l'art spécifie en détail, les caractéristiques, de ces méthodes permettant d'estimer, en tous points de l'espace, les valeurs inconnues de certaines variables environnementales. Les connaissances géostatistiques nécessaires à l'utilisation de ces méthodes ont quant à elles été acquises, avec la bienveillance de Monsieur Jean Serra, Directeur de Recherches à l'Ecole des Mines de Paris.

Enfin, pour pouvoir procéder au *diagnostic des déterminants environnementaux* permettant de réduire les inégalités géographiques *de santé* à partir de jeux multidimensionnels constitués d'i.st.m* et d'i.st.m* hétéroclites, le recours à la sphère des théories de l'apprentissage par des machines learning a été nécessaire (Han et Kamber, 2006). Les notions de mathématiques ainsi que la validation des programmes de sélection des déterminants environnementaux ont été effectués sous l'égide de Monsieur Robin Genuer, Maître de conférences à l'ISPED.

En se juchant sur les épaules des géants de l'épidémiologie et de la statistique, le géographe de la santé dispose de connaissances et de méthodes plus robustes qui lui permettent de dissiper une partie des ombres et des incertitudes qui nimbent l'espace d'une problématique de santé publique somme toute complexe. Il faut garder à l'esprit qu'il s'agit d'une thèse en Géographie, donc la *vérité est le but* et l'espace *est la norme*, au sens où l'entend (Comte-Sponville, 2000). Par conséquent, la dialectique est spatiotemporelle et les éléments de confusion nombreux : perte de puissance statistique liée à la phase de spatialisation, conflits inhérents aux granularités* des données géographiques inter-sources ; Incertitude sur la capacité des i.st. à modéliser les phénomènes pour lesquels ils sont créés (Brook, Lohr et al., 1984). Malgré l'intérêt évident d'un positionnement scientifique interdisciplinaire, l'approche reste géographique, l'établissement de relations causales au sens épidémiologique du terme est donc impossible dans la mesure où, en Géographie, *l'exposition aux risques* dan*s l'espace et le temps est exprimée sous forme d'une relation possible entre le milieu et l'homme* (Bailly et Beguin, 2005). D'ailleurs la condition *d'existence d'une relation effet-dose* ne sera pas abordée. En revanche celles de *reproductibilité de la méthode, de confirmation d'une relation causale chez l'animal, de prise en compte de la temporalité, et de plausibilité d'effets biologiques sur l'homme* – sont prises en compte peu ou prou (Hill, 1965).

La problématique et la considération conjointe du contexte et de ce positionnement interdisciplinaire particulier permettent de définir les objectifs scientifiques - déjà énoncés à demi-mot - et de conférer à cette recherche la dimension humaine qui lui donne tout son sens.

OBJECTIFS ET CONTRIBUTIONS ATTENDUES

En optant pour un positionnement scientifique interdisciplinaire et une dialectique géographique opérationnelle qui couple des stratégies d'intégration SIG de données morbides et environnementales par le biais d'outils de modélisation statistique, géostatistique, à des algorithmes de datamining* voués à la sélection de variables, l'identification géographique des Facteurs Environnementaux * qui déterminent *l'état de santé* des populations est possible et la géographie de la santé n'a plus à craindre d'être qualifiée *d'épidémiologie populaire* (Brown, 1997).

L'objectif ultime de cette thèse est de fournir des moyens opérationnels de sorte que la géographie de la santé occupe une place décisive en santé publique. En particulier dans l'analyse des questions qui animent la vie sociale et politique. Il ne fait nul doute qu'une approche géographique transversale de ces questions peut contribuer à l'évolution des connaissances dans les domaines de l'épidémiologie et de la médecine où elle est dépréciée parfois*, mal aimée* souvent (Salem, 1995).
Pour atteindre l'objectif fixé, il est donc nécessaire de transposer des méthodes statistiques robustes de modélisation et d'analyse, afin d'assurer au Géographe de la Santé des bases scientifiques solides pour que son point de vue sur l'effet des FE* liés à des risques morbides soit reconnu de la même façon que celui des spécialistes des autres disciplines. Cet objectif n'a cependant de valeur qu'au prorata des intérêts humains qu'il sert. Il s'agit de mieux penser pour mieux vivre.

La géographie de la santé nourrit le dessein d'avoir des répercussions sur les mentalités et les valeurs humaines dictées par une société, une époque et un contexte politique – national et européen – particuliers. Les propositions méthodologiques doivent être, elles aussi, conformes aux paradigmes scientifiques contemporains et les résultats attendus suffisamment robustes et clairs pour avoir des retentissements sur la vie politique, économique et sociale (Godin, 2007).

En conséquence de quoi il s'agira de : Raisonner à une échelle intelligible pour sensibiliser les sociétés et influencer les comportements individuels et les mentalités politiques ; Construire une dialectique opérationnelle en se fondant sur des méthodes suffisamment robustes pour que les résultats puissent valider les interactions géographiques mises en exergue entre l'environnement et la santé ; Caractériser et cartographier les zones d'exposition à des FE* géographiques qui déterminent l'état de santé des populations *in situ* ; Et aussi et surtout garantir une démarche scientifique reproductible pouvant être étendue à n'importe quelle maladie.

Pour valider la logique heuristique, estimer la robustesse des propositions méthodologiques et illustrer l'intérêt des résultats cartographiques dans le champ de la santé, la dialectique est appliquée à l'identification géographique de déterminants environnementaux liés aux séquelles - cataractes, tumeurs thyroïdiennes et tumeurs secondaires majeures - développées après le traitement d'une leucémie chez l'enfant - Cohorte LEA.

D'un point de vue scientifique il s'agit de proposer un raisonnement et des instruments géographiques validant chronologiquement les trois sous-objectifs subséquents.

Le premier objectif scientifique touche à la façon dont la géographie s'immisce dans cette thématique de santé publique en spécifiant une échelle d'investigation adaptée à la problématique, au contexte communautaire, à la modélisation géographique de phénomènes morbides et environnementaux, tout en garantissant un angle d'attaque différent de celui de l'épidémiologie spatiale. La difficulté est de choisir une échelle suffisamment fine pour modéliser la variabilité des phénomènes morbides d'intérêt mais suffisamment globale pour prendre en compte le caractère multidimensionnel de l'environnement et les incertitudes associées aux données disponibles. En outre, cette échelle doit permettre d'identifier des Déterminants Environnementaux de Santé* et de caractériser les espaces et les populations vulnérables tout en offrant une *comparaison intelligible des sociétés,* et en suggérant des *leviers*

politiques opérationnels permettant de réduire les inégalités d'accès à une bonne santé environnementale*.

Le second objectif scientifique a trait à la qualité des modélisations géographiques. Dans un premier temps il s'agit de dresser un bilan des variables mesurées et mesurables stockées dans la multiplicité des bases de données existantes. Puis dans un second temps, de proposer des stratégies d'intégration SIG par des *méthodes statistiques, probabilistes et géostatistiques* adaptées à la fois à la granularité* des données disponibles et à la construction d'i.st.m* et d'i.st.e* suffisamment *robustes et précis* pour représenter la réalité géographique des phénomènes morbides étudiés, des FIM* caractérisant la géographie des populations étudiées et pour analyser leurs interactions avec un environnement géographique multidimensionnel représenté par des FE* physicochimiques, sanitaires et socio-économiques.

Le troisième objectif scientifique est le corollaire du second. Il s'agit de proposer une méthode d'apprentissage statistique suffisamment puissante pour disjoindre, parmi tous les i.st.e* environnementaux confectionnés, ceux qui n'interagissent pas avec les i.st.m* proposés de ceux qui permettent de les expliquer et de les prédire. Seuls quelques algorithmes de datamining* permettent d'effectuer ce type d'analyse multidimensionnelle. Il conviendra de choisir le plus topique pour des jeux de données géographiques *de grandes dimensions* constitués de variables quantitatives (continues ou discrètes) et qualitatives (booléennes ou multi-classes). Et par suite, utiliser une méthode statistique capable d'identifier les Déterminants Environnementaux de Santé*, quitte à en modifier légèrement les contours, pour qu'elle puisse être appliquée à des pathologies dont l'incidence est particulièrement faible.

Toutefois, l'amélioration des connaissances scientifiques en géographie de la santé n'a de sens que si elle est mise au service de la vie. De fait, des contributions humaines sont attendues. Elles se déclinent en trois composantes.

La première contribution humaine est grevée d'une composante épidémiologique. Elle concerne le fait que les progrès thérapeutiques ont transformé le pronostic vital des individus atteints d'un cancer en posant, avec une grande acuité, le problème des effets secondaires des traitements au long cours, et plus particulièrement celui des séquelles sur l'insertion sociale et la qualité de vie. L'application de la dialectique et des méthodes proposées à des données épidémiologiques permet de valider son caractère opérationnel et d'estimer sa robustesse. Mais au-delà de *l'exercice de style* des applicatifs effectués dans le cadre de cette thèse, le but est de faire progresser l'étude Leucémies de l'Enfant et de l'Adolescent (LEA) qui s'attachait principalement à l'analyse des déterminants médicaux, socio-économiques et comportementaux du devenir à moyen et long termes des patients. Les travaux menés jusqu'à présent ont permis de répondre à certaines questions mais soulignent aussi la limite des déterminants explorés. La prise en compte des FE* dessine des perspectives de recherche très prometteuses, tant au niveau clinique qu'au niveau de l'incidence des séquelles en lien avec la gestion anthropique des espaces géographiques.

La seconde contribution est corollaire à la première et touche aux applications attendues dans le domaine médical. La méthode se veut consistante et donc transposable à n'importe quelle autre maladie. Par conséquent, lorsque l'on met en perspective les FE* géographiques identifiés comme des déterminants de santé et que l'on cartographie des zones géographiques d'exposition à ces FE, on donne aux praticiens de santé des moyens opérationnels pour leur permettre d'informer leurs patients et de leur préconiser des mesures préventives visant à réduire les risques morbides auxquels ils sont assujettis du simple fait de leur situation dans l'espace géographique*.

Quant aux contributions humaines attendues au niveau de la sphère sociopolitique, l'identification des FE* qui permettent d'expliquer et d'influencer l'état de santé des populations, couplée à des cartographies produites à des échelles intelligibles, offre aux décideurs politiques la

possibilité de mettre en place des mesures socio-économiques opérationnelles pour favoriser l'accès à une bonne santé environnementale*. En contrepartie, la lutte contre les inégalités géographiques passe par une diffusion d'informations au sein des sociétés humaines, ce qui leur confère un moyen de pression sur les politiques de santé publique lorsqu'elles s'orientent dans une direction différente du sens commun. En somme, le contrôle territorial des déterminants environnementaux qui influencent l'état de santé des populations a pour dessein l'amélioration de la qualité de vie des sociétés humaines et se voit conférer une dimension démocratique.

Afin d'établir des *diagnostics géographiques en santé environnementale*,* le positionnement scientifique choisi est celui de l'interdisciplinarité qui, dans le contexte intellectuel et communautaire actuel, s'impose comme une nécessité. La dialectique est géographique mais le cheminement intellectuel est guidé par la réflexion épidémiologique et s'appuie sur des méthodes statistiques de modélisation spatiale et d'apprentissage à la fois innovantes et robustes. Avant de modéliser la géographie des phénomènes morbides d'intérêt et d'identifier des FE* *pertinents* susceptibles de les influencer, il convient de dresser un état de l'art sur les connaissances en santé et environnement, les méthodes utilisées et proposées, et d'établir un listing des bases de données existantes en distinguant celles qui sont accessibles de celles qui ne le sont pas.

SECTION B) ETAT DE L'ART ET HYPOTHESES HEURISTIQUES

En 1900 l'espérance de vie moyenne en France était de 48 ans contre 81 ans en 2010 (INSEE, 2012c). Cette évolution est le résultat d'une amélioration générale des conditions de vie de l'homme en société et des progrès effectués dans le domaine médical. Paradoxalement a été observée une augmentation importante des maladies et plus particulièrement des cancers. Cette modification démographique a induit des changements socio-économiques et des mutations au niveau des milieux environnementaux et des modes de vie (Pison, 2005), et certaines de ces transformations s'avèrent avoir des impacts néfastes sur la santé, à tel point que les questions de santé environnementale* se sont placées au cœur des débats qui animent la scène sociale, économique et politique. (Olshansky S-J. et al., 2005) ; (Krewski, 2009). En l'état actuel des connaissances, la part de l'augmentation des maladies qui incombe à des motifs démographiques et aux progrès médicaux peut être estimée. En revanche, celle inhérente aux *expositions à des FE** est très controversée et dépend de la définition que l'on donne à l'environnement (Krewski, 2009), (IARC, 2007), (World Cancer Research Fund & American Institute for Cancer Research, 2007).

Il a été vu précédemment que le positionnement scientifique du géographe de la santé consistait à prendre en compte l'environnement dans toutes ses dimensions. Par conséquent, la définition de l'environnement retenue est celle de l'OMS. L'environnement comprend tous les agents physicochimiques potentiellement dangereux présents dans les milieux de vie, les situations à risque d'ordre socio-économique, professionnel ou encore liées au tissu sanitaire. Les Caractéristiques Individuelles et Médicales* (CIM) sont aussi prises en compte pour leurs effets avérés sur l'état de santé et leur modélisation géographique prend ici la dénomination de FIM* (Facteurs Individuels et Médicaux*). L'omission de cette composante biaiserait l'analyse statistique des interactions santé-environnement, qu'elle soit menée dans une *logique géographique* ou *individus-centrée** (World Health Organization, 2009). Par définition, tous les FE* ou FIM* sont voués à être intégrés *à l'exception des facteurs génétiques* (Chasles et Fervers, 2011). Or cela va l'encontre du bon sens et du positionnement scientifique choisi. Mais il s'agit d'un choix subi et non délibéré car les caractéristiques génétiques des individus ne sont pas disponibles dans la base épidémiologique utilisée – LEA.
D'ailleurs, ce chapitre a, entre autres, pour ambition de dresser un état de l'art des bases de données existantes, de la granularité* des informations qu'elles contiennent (nature, échelle, incertitudes, lacunes…) et de leur disponibilité, condition *sine qua non* de la modélisation géographique des FE* et des FIM* retenus par des i.st.
L'intégration des facteurs génétiques est impossible en dépit de leur impact évident *sur l'état de santé*. Cependant, le point de vue des généticiens sur le rôle FE* converge avec celui des géographes de la santé et constitue la première hypothèse fondamentale de cette recherche.

Hypothèse fondamentale 1 : *Les causes des maladies sont multifactorielles. L'environnement est multidimensionnel. Il a des répercussions sur la santé. Il doit donc être appréhendé dans toutes ses dimensions conditionnellement aux données disponibles.*

Les chercheurs en biologie moléculaire connaissent l'importance des facteurs génétiques sur les prédispositions morbides. Mais ils affirment que le génotype ne suffit pas à expliquer ces processus. D'ailleurs le génome se modifie au cours du temps. Ces mutations sont dues à des maladies d'abord, mais *indubitablement à des FE* ensuite*. L'identification des FE* qui conditionnent le génome présente un double intérêt : Prévenir des maladies et améliorer l'efficience des traitements ; Mais surtout, l'effet mutagène induit par les Déterminants Environnementaux de Santé* (DES) est *en théorie réversible*, il cesse avec la suppression du stimulus, et mieux, ces mutations ne sont pas univoques, elles peuvent être utilisées pour améliorer les propriétés du génome - *notre code génétique ne conditionne pas toute notre destinée* (Stein, 2012).

Pour les chercheurs en biologie moléculaire, les *expositions environnementales* les plus suspectées d'avoir des répercussions sur la neuro-plasticité du cerveau - chef d'orchestre des mutations génétiques (Takada, Urano et al, 1995) - sont celles liées à des substances métalliques : Cadmium, Cuivre, Plomb, Nickel, Zinc… et à la radioactivité environnementale (Reilly, 1991). Or, l'état de l'art sur les interactions santé-environnement met en évidence l'effet des expositions à des *situations à risque* en relation avec *l'accès aux soins* (Penchansky et Thomas, 1981) ou *à un contexte* socio-économique *défavorable* (Chaix, Merlo et al., 2005) qui est largement décrit dans la littérature en épidémiologie spatiale comme en géographie de la santé. En revanche les *expositions environnementale*s à des substances physicochimiques présentes dans les *milieux eau, air, sol et biologiques* sont plus controversées à cause de la difficulté à établir des relations causales significatives. *Cependant des études épidémiologiques ont récemment relancé le débat* (Afsset, 2009a). La dialectique évolue, et il convient aujourd'hui de raisonner en termes d'*expositions environnementales* combinées (Leux et Guénel, 2010). D'autant que depuis ces dernières années, l'élaboration de nombreuses bases de données environnementales et épidémiologiques protéiformes offre des perspectives de modélisation géographique intéressantes (Zeitouni, 2006). Cependant le problème de puissance statistique permettant de mettre en évidence des liens de causalité est toujours controversé. Corollairement à la première hypothèse découle la seconde hypothèse fondamentale de cette thèse qui se fonde sur le postulat suivant : *une intelligence qui*, *pour un instant donné*, *connaitrait toutes les forces dont la nature est animée*, *la situation respective de tous les êtres qui la composent* [et] *qui serait assez vaste pour soumettre ces données à l'analyse* [mathématique] – [pour une telle intelligence] - *plus rien ne serait incertain*, *le passé comme l'avenir serait présent à ses yeux* (Laplace, 1820).

Hypothèse fondamentale 2 : *Si l'on dispose de stratégies, de méthodes et de données suffisamment fiables pour modéliser, par des i.st.m* et des i.st.e, la géographie des FE* et des FIM* suspectés de conditionner les PM* étudiés, et d'outils mathématiques assez puissants pour analyser la complexité de ces interactions, alors on peut identifier des DES* et par suite, expliquer, réduire et infléchir les inégalités géographiques d'accès à une bonne santé environnementale**

Cette section a pour objet de définir ce que sont les *expositions environnementales*, d'envisager la *complexité d'évaluation des risques* qui leur sont associés, notamment pour les *substances physicochimiques*, et d'exprimer la différence qui est faite entre les *expositions environnementales potentielles* et *intrinsèques* et les hypothèses subsidiaires émises à ce sujet.
Un état de l'art est ensuite décliné à propos des méthodes de modélisation et d'analyse multidimensionnelle utilisées en géographie et des performances et limites de celles implémentées dans les SIG. A cette occasion des propositions méthodologiques extra-disciplinaires seront faites dans l'optique d'obtenir une puissance statistique suffisante et adaptée à la sélection de DES.
Enfin, un état des connaissances sur les interactions santé-environnement permettra d'identifier les FE* *supposés pertinents* par rapport aux PM* étudiés et qu'il conviendra d'intégrer dans le cadre de la dialectique proposée, selon une discrétisation de l'environnement géographique en quatre composantes.

EXPOSITIONS ENVIRONNEMENTALES ET PROPOSITIONS

L'importance des facteurs génétiques sur les prédispositions morbides et les fortes suspicions faites quant à l'action des *expositions environnementales* physicochimiques sur leur caractère mutagène ont précédemment été évoquées (Stein, 2012).

Les *expositions environnementales* dissocient les expositions à des *situations à risque* inhérentes à *l'accès aux soins* ou à *la défaveur sociale - largement documentées,* des substances physicochimiques dangereuses pour l'homme et présentes dans les *milieux environnementaux*. Ces dernières sont aussi décrites mais elles sont controversées en raison de la difficulté à établir des relations causales. Cependant le débat a été relancé (Afsset, 2009a). Pour les géographes de la santé, comme pour les généticiens, les *expositions environnementales* à des substances physicochimiques toxiques sont à la fois prégnantes et documentées dans leurs domaines respectifs, notamment pour : les métalloïdes (Reilly, 1991) ; Les pesticides, les Hydrocarbures Aromatiques Polycycliques (HAP), les Matières En Suspension (MES), et la radioactivité environnementale (IARC, 2012).

Dans cette partie, seules les expositions environnementales physicochimiques sont considérées. D'abord parce qu'elles sont controversées, ensuite parce qu'elles nécessitent des notions d'expologie qu'il convient d'énoncer et sur lesquelles des hypothèses doivent être formulées.

Les *expositions environnementales* physicochimiques sont définies comme des expositions à de faibles doses de substances toxiques, présentes dans *les milieux environnementaux.* Les substances en question ont des effets cliniques néfastes. La controverse touche au fait que les *doses environnementales moyennes* mises en jeu sont trop faibles pour avoir des effets biologiques. Or les *effets biologiques réels*, liés à des *expositions répétitives*, quand elles ne sont pas *chroniques*, sont *impossibles à évaluer par manque de recul*. De plus, les doses d'exposition subies varient tout au long des journées, avec de forts gradients géographiques, et dépendent de la sensibilité de chacun. En sus, ces substances sont présentes dans tous *les milieux environnementaux* et *les milieux d'exposition* (Afsset, 2009a).

Les milieux environnementaux sont l'eau, l'air, le sol et les matrices biologiques. Il s'agit par exemple : des eaux de pluie, de surface et de nappe, des sols, quelles que soient leur profondeur et leur nature géologique, l'air à tous les niveaux d'altitude, et tous les éléments organiques de la faune et de la flore terrestre.
Les milieux d'exposition constituent une partition des milieux environnementaux, celle susceptible d'entrer en contact avec l'homme. C'est le cas de l'air ambiant, de l'eau du robinet, des couches superficielles du sol, mais aussi des matières végétales et animales de la chaine alimentaire (Caudeville J., Boudet C. et al., 2012).

Les substances toxiques contaminent les matrices humaines par les *voies d'exposition*. On appelle *voies d'exposition* les modes de contamination organique à partir d'éléments issus d'un *milieu d'exposition*. Pour l'homme par exemple il s'agit de l'inhalation ou de l'ingestion de MES, d'eau de consommation ou d'aliments (viande, œufs, poisson, fruits, légumes, céréales) contaminés par des substances toxiques (McKone et MacLeod, 2003) (Hawley, 1985).

Les effets des expositions environnementales à des substances physicochimiques sont de deux types. Il y a d'une part les effets *déterministes*, i.e. qui engendrent des destructions cellulaires massives. Ils varient selon la dose et l'organe touché et sont caractérisés par des seuils identiques pour tous les individus. Et d'autre part les effets *aléatoires ou stochastiques* qui engendrent, à plus ou moins long terme, des mutations de cellules délétères qui peuvent ou non apparaitre, à des doses et sur des temporalités variables selon les individus. Ils sont donc particulièrement difficiles à caractériser et bien sûr très suspectés dans les expositions environnementales.

L'intensité de ces effets dépend de trois facteurs : *la source*, i.e. la nature et l'efficacité biologique de la substance, *l'exposition* - i.e. temps et le fractionnement, et *la cible*, i.e. le tissu ou l'organe touché.
De plus ces effets varient en fonction du type d'exposition qui peut être *interne*, i.e. les ingestions ou les inhalations, ou *externe*, i.e. le contact des tissus humains avec des matières souillées ou des rayonnements particuliers (IARC, 2012)
Pour mieux comprendre les processus de transfert des substances physicochimiques diffusées depuis des sources d'émission dans les milieux environnementaux, puis vers les milieux de contact, et enfin les contaminations humaines par les voies d'exposition, un schéma est proposé. Il illustre le processus de contamination humaine lié à des expositions à la radioactivité naturelle ou artificielle, i.e. liée à des activités anthropiques. La radioactivité est un exemple intelligible dans la mesure où elle est omniprésente dans l'environnement des milieux de vie et où elle engendre des contaminations *internes ou externes.*
Toutefois, ce schéma peut être étendu à la majorité des substances citées ultérieurement dans l'état de l'art : Rayonnements solaires, ondes électromagnétiques, métalloïdes, HAP, pesticides, radionucléides, MES... (IRSN, 2004)

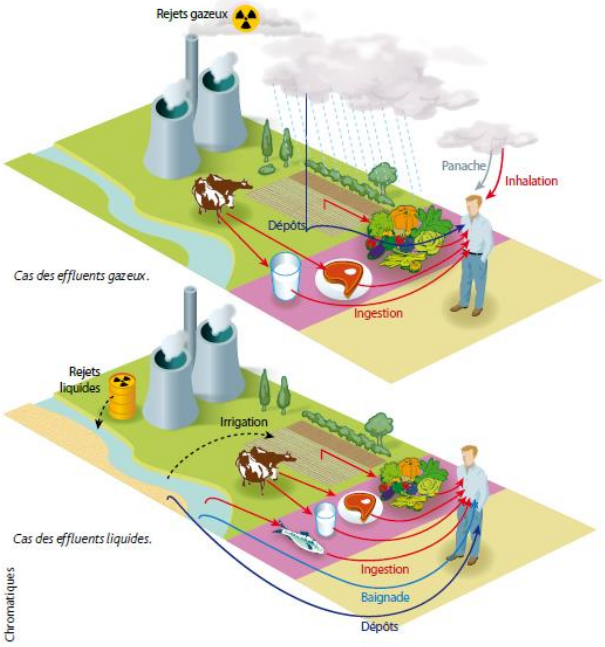


Figure 2 : Principe du processus de contamination par l'exposition à la radioactivité environnementale.
Source : (IRSN, 2009)

La présence de substances physicochimiques toxiques dans les *milieux environnementaux* s'évalue par le biais *de mesures* dans l'eau, l'air, le sol ou les milieux biologiques, par exemple pour la radioactivité liée à l'activité volumique de radionucléides, elle se mesure en Becquerel (Bq).

Pour ce qui est des expositions humaines, elles s'évaluent par le biais de doses caractéristiques et dépendent du contexte, en l'occurrence pour des expositions à la radioactivité *environnementale d'ordre professionnel ou médical,* l'unité de prédilection pour caractériser les irradiations est le Gray (Gr) ; et pour les expositions environnementales géographiques *au milieu de contact : air ambiant*, l'exposition globale *interne* à des particules radioactives et *externe* à des Rayonnements Ionisants (RI) d'origine naturelle ou artificielle s'évalue en Sievert (Sv). Le Gray, comme le Sievert, *permet d'estimer les effets déterministes et stochastiques de ces expositions* (IRSN, 2004).

Les doses géographiques des expositions à des substances toxiques s'estiment *par le biais de modèles multimédia*. Il s'agit d'outils mathématiques qui évaluent, à partir des mesures effectuées au niveau des sources d'émission polluantes, les doses transférées vers les *compartiments environnementaux*, puis *vers les milieux d'exposition*, et enfin les doses d'expositions organiques humaines en des points précis de l'espace (Pennington, Margni et al., 2009). Ces doses d'expositions géo-localisées sont ensuite couplées à des méthodes de modélisation spatiale et à des scenarii sociodémographiques interfacés dans des SIG afin de les grever d'une composante géographique (Caudeville J., Boudet C. et al., 2012).

Hypothèse subsidiaire sur la modélisation géographique des expositions environnementales. Dans cette thèse il est supposé que les expositions environnementales ont un impact sur la santé – sur les PM* étudiés – et qu'il convient de modéliser la variabilité géographique de ces FE* physicochimiques dans la limite des données disponibles. Il conviendra cependant de distinguer la notion *d'exposition géographique potentielle* de celle *d'exposition géographique intrinsèque*.

Proposition : les expositions géographiques environnementales potentielles sont modélisées à partir de mesures effectuées dans les milieux environnementaux. Elles constituent la majorité des i.st.e* physicochimiques proposés, i.e. pour caractériser la variabilité des expositions potentielles à l'activité volumique de radionucléides, la radioactivité liée à la présence d'Installations Nucléaires de Base (INB), la présence de substances toxiques multiples liée à l'occupation biophysique des sols…

Proposition : les expositions géographiques environnementales intrinsèques sont modélisées à partir de doses d'exposition à des substances physicochimiques toxiques présentes *dans les milieux de contact*. Les i.st.e* proposés permettant de caractériser la variabilité géographique de ces expositions particulières sont construits à partir d'indicateurs représentatifs du risque d'ingestion ou d'inhalation de métalloïdes divers, ou de doses efficaces globales de rayons gamma dans l'air ambiant.

Hypothèse subsidiaire sur la robustesse des i.st.e* représentatifs des expositions à des substances physicochimiques. Il convient de s'interroger sur la capacité des i.st.e* à modéliser la réalité géographique des expositions environnementales physicochimiques dangereuses. Les i.st.e* représentant des expositions potentielles sont-ils plus robustes que ceux modélisant des expositions intrinsèques ?

Les premiers sont construits à partir de mesures environnementales généralement robustes et fiables mais effectuées dans les milieux environnementaux. Ils ne présument donc pas de contaminations humaines. Toutefois, on supposera que plus les milieux sont souillés, plus les chances de contaminations humaines sont grandes.
A l'inverse, les doses d'expositions environnementales intrinsèques permettent d'estimer directement le risque de contamination. Les i.st.e* qui en découlent peuvent sembler *a priori* plus précis et plus justes. Or, l'estimation géographique de doses d'expositions est entachée de biais multiples et semble opérer uniquement à des *échelles fines (i.e. sur des petites mailles), sur des temporalités réduites, et lorsque les inputs sont d'excellente qualité*. Certains auteurs sont réfractaires à ce type de variables à cause des lacunes associées aux nombreuses hypothèses, aux paramètres, et aux variables additionnelles nécessaires pour alimenter *les modèles multimédia*. Les doses d'expositions sont même parfois qualifiées *d'inopérantes* lorsqu'elles sont agrégées à des échelles territoriales (Pistocchi, Vizcaino et al., 2010).
En effet, *la propagation des incertitudes liées aux couplages des modèles fait* que les doses d'expositions peuvent varier sur des intervalles de confiance allant de la non exposition à la surexposition. *Ces estimations ne sont pas significatives* (Mercat-Rommens, Chojnacki et al., 2008). A cela s'ajoute l'impossibilité de reconstituer les déplacements quotidiens et les mobilités résidentielles à moyen et long termes (Couet, 2006), et aussi, les incertitudes liées à la sensibilité et aux prédispositions génétiques des populations cibles (Sobol, 2004) . De surcroit, les forts gradients spatiotemporels affaiblissent la robustesse de ces indicateurs. A ce jour, il en est ainsi pour toutes les expositions environnementales à des substances chimiques associées aux cancers, aux cataractes et aux perturbateurs endocriniens (Zaidi, Bhatnargar, et al., 2000).

Dans cette recherche, la géographie de FE* physicochimiques suspectés d'avoir des interactions avec les PM* étudiés, est modélisée par des i.st.e* d'expositions *potentielles* ou *intrinsèques*. Il sera démontré que le choix des données inputs environnementales utilisées (mesures ou doses d'expositions) n'est pas uniquement conditionné par les arguments bibliographiques énoncés. En effet, il est aussi déterminé par la granularité* des données disponibles - abordée en section(c) – et par le choix de l'échelle d'investigation retenue - décliné en guise de conclusion de fin de chapitre – ainsi que par les stratégies de modélisation géographique utilisées. Il convient désormais d'aborder l'état de l'art sur les méthodes de modélisation et d'analyse spatiale multidimensionnelle utilisées en géographie de la santé ainsi que les propositions interdisciplinaires associées

METHODES UTILISEES EN GEOGRAPHIE ET PROPOSITIONS

La seconde hypothèse fondamentale assure que si l'on dispose de méthodes de modélisation robustes pour construire des i.st.m* et des i.st.e* représentatifs de la réalité géographique des PM* et des FE* retenus, et que par ailleurs, on dispose d'une méthode d'analyse suffisamment puissante pour caractériser les interactions statistiques qu'ils entretiennent ; alors on peut identifier les DES. Dans cette sous-section, il s'agit de dresser un état de l'art des méthodes utilisées en géographie de la santé. Et en tant que de besoin, en proposer des plus adaptées pour répondre à la problématique.

La géographie de la santé est généralement considérée comme une *géographie radiale*, i.e. qui étudie les inégalités d'accès à une bonne santé (Grawtiz, 2000).

L'entrée par la géographie [dans le processus de compréhension] est spatiale, la santé est la fenêtre ouverte par l'épidémiologie (Fromageot, Coppieters et al., 2005). Cette Science Humaine et Sociale (SHS) est *avant tout descriptive*, les indicateurs spatiaux caractérisent de façon simple, intelligible et fiable la géographie des *états de santé*. Simple ne veut pas dire simpliste. Les indicateurs utilisés s'avèrent assez précis pour montrer, peu ou prou, que des sociétés confrontées à des milieux environnementaux analogues peuvent, de par leur gestion particulière de l'espace, exposer leurs populations à des risques morbides différents. Le caractère intelligible des indicateurs permet d'influencer les mentalités sociales et les tendances politiques pour garantir un *contrôle territorial* visant à réduire les disparités géographiques de santé environnementale* (Salem, 1995).

Proposition n°1 :
Les méthodes de modélisation proposées doivent permettre de construire des i.st.m* et des i.st.e* adaptés à la problématique mais préserver leur caractère intelligible. Cette *intelligibilité* est associée aux méthodes et au choix de *l'échelle d'investigation* (chapitre II).

En géographie de la santé, la description spatiale des états et des faits de santé est donnée par des Atlas d'indicateurs environnementaux et morbides - constitués de tableaux, de cartes et de quelques statistiques exploratoires descriptives (Salem, Rican S et al., 2006).
Les analyses statistiques les plus poussées utilisent des *méthodes paramétriques* basées sur les *coefficients de corrélation*, et dans le meilleur des cas, sur des méthodes topologiques type *ACP*. Pour ce qui est de l'analyse spatiale, sur des méthodes du type *partitionnement de l'espace,* par exemple celles des *Polygones de Thiessen* (Rémy, Handschumacher et al., 2011) , ou des méthodes *déterministes (qui n'intègrent pas l'idée du hasard) et paramétriques calibrées a priori*, comme : *l'inverse de la distance à la puissance :d*, qui sont encore largement utilisées (Baillargeon, 2005).

En géographie, l'enseignement se limite encore parfois à des statistiques descriptives (Lahousse et Piédanna, 1998), et presque toujours, même à des niveaux avancés, à des méthodes d'analyse multidimensionnelles paramétriques classiques (Groupe CHADULE, 1997).
Il en est de même en analyse spatiale, les méthodes de modélisation enseignées sont souvent *déterministes et subjectives* (Baillargeon, 2005), et quand elles sont *stochastiques* elles sont parfois inadaptées aux données utilisées (Pumain et Saint-Julien, 1997).

Malgré les lacunes méthodologiques énoncées, les géographes de la santé parviennent à mettre en évidence, parfois avec une grande acuité, des inégalités géographiques de santé environnementale*. Il convient donc les étudier. L'objet de cette thèse est de proposer, voire de transposer des méthodes plus récentes, mieux adaptées et si possible plus robustes.

Hypothèse subsidiaire sur le choix des méthodes :
Les modèles statistiques, parce qu'ils se situent à divers degrés entre analyse et synthèse, sont a priori de bons outils d'exploration de la complexité spatiale (Charre, 1995). On suppose donc que les méthodes statistiques constituent le meilleur moyen pour répondre à la problématique. En l'occurrence le couplage de l'informatique à des processus stochastiques d'apprentissage et à des modèles probabilistes non paramétriques offre des perspectives très prometteuses (Cartier, Villani et al., 2012). L'utilisation de *méthodes statistiques puissantes* capables de traiter simultanément de grandes quantités de données hétéroclites est aujourd'hui une nécessité. En effet, avec l'essor ces dernières années de nombreuses bases de données environnementales protéiformes, les perspectives de modélisation géographique en santé environnementale* sont presque illimitées (Zeitouni, 2006).

Les barrières sont liées à l'accès aux données (abordé en section c). Les faiblesses méthodologiques mises en évidence font de la géographie de la santé la *mal-aimée* de l'épidémiologie spatiale, qui la considère à tort comme une science de l'inventaire. Pourtant, ses contributions en terme de compréhension du *fonctionnement spatial des systèmes de santé* et des enjeux *sociaux, économiques et politiques qui en découlent,* sont indubitables (Salem, 1995) Mais c'est aussi mal connaitre la géographie contemporaine et ses instruments que de croire encore cela.

Proposition n°2 : Méthodes et outils de modélisation géographique.
Les instruments d'intégration de données et de modélisation géographique utilisés actuellement par les géographes géomaticiens sont particulièrement compétitifs. En effet, des outils très performants d'acquisition, de représentation et d'intégration de données géographiques multidimensionnelles incomplètes ou incertaines, de nature qualitative ou quantitative, ont récemment été implémentés dans les SIG, et en particulier dans ArcGis.10 (ESRI, 2013).

Pour ce qui est de la robustesse des techniques d'agrégation *verticale* (Peguy, 1996), i.e. de fusion temporelle de données spatialisées, les méthodes implémentées dans les SIG suffisent. Une stratégie permettant de donner plus de consistance au processus de fusion verticale sera proposée en partie II. Sa mise en œuvre n'a pas été effectuée via un SIG, mais cela aurait très bien pu être le cas, ce qui aurait peut-être même été plus simple. Cette stratégie d'intégration se fonde sur *l'analyse probabiliste de la stationnarité temporelle apparente,* lorsque les inputs utilisés sont *des chroniques temporelles spatialisées* (Saporta, 2006).

Pour ce qui est des méthodes de modélisation géographique d'agrégation ou de désagrégation *horizontale* (Peguy, 1996), i.e. purement spatiales, *les techniques classiques utilisées en analyse spatiale* sont robustes et relativement adaptées à la représentation des phénomènes géographiques (Charre, 1995). Ces techniques sont toutes intégrées dans les SIG. Toutefois, un intérêt particulier sera porté (partie II) à celles récemment implémentées ou améliorées qui permettent d'aller encore plus loin dans le processus de modélisation. Le module *ModelBuilder* permet, entre autres, d'effectuer des calibrations itératives ou des modélisations géographiques dynamiques (ESRI, 2013). Enfin s'agissant de la reconstitution spatiale de données incomplètes ou incertaines, *les méthodes de modélisation spatiotemporelles* contemporaines qui ont récemment été implémentées dans les SIG (ESRI, 2013) seront passées en revue (partie II). L'accent sera mis sur les *géostatistiques uni-variables* (Matheron, 1965) et *multi-variables* (Wackernagel, 2003) pour leur puissance et leur robustesse, et plus particulièrement sur des techniques de *krigeage* qui permettent d'évaluer des *variables régionalisées** en tous points de l'espace en tenant compte à la fois de *l'effet de support**, i.e. de la structure spatiale des données d'échantillonnage , et de *l'effet information**, i.e. de la qualité des données disponibles (Marcotte, 2008). Ces dernières sont particulièrement prisées pour modéliser *des FE* physicochimiques* dans le domaine de *l'Environnement* : pollution des milieux (Gay et Korre, 2006), météorologie, océanographie (Gratton, 2002), géologie minière (Matheron, 1963). Elles sont aussi utilisées dans le domaine de la santé : en expologie (Goovaerts, 2006), et en épidémiologique spatiale (Gaudart, 2007). Les forces et les faiblesses des méthodes utilisées seront déclinées, ainsi que les situations et les

phénomènes où il convient d'utiliser des logiciels de modélisation statistique plus performants que les SIG (Marcotte, 2008).

Les SIG offrent donc des possibilités d'intégration et de modélisation géographique robustes et place la géographie de la santé en position de rivaliser avec les méthodes analogues, sinon moins performantes, utilisées en épidémiologie spatiale. En revanche, pour l'analyse de données multidimensionnelles et la caractérisation statistique des variables, les SIG ne sont pas compétitifs. De plus les *méthodes statistiques paramétriques* utilisées en géographie sont inadaptées, bien loin de concurrencer celles utilisées en épidémiologie (Mafijul, Alam et al., 2013).

La géographie de la santé souffre d'une sorte de retard méthodologique sur ce point. L'épidémiologie spatiale fait à peine mieux. L'identification de DES* sur des jeux multidimensionnels de données hétéroclites est une thématique à la pointe des recherches en mathématiques.

Proposition n°3 : Méthodes et outils d'analyse multidimensionnelle adaptés aux interactions géographiques de santé environnementale.
Il s'agit d'une problématique de recherche complexe qui a trait aux mathématiques : *la sélection de variables dans des jeux en grandes dimensions* (Tuleau-Malot, 2005). Les SIG sont indigents pour effectuer ce type de traitement. Pourtant, pour identifier des DES* à partir de jeux multidimensionnels de données géographiques protéiformes, il existe des solutions. En l'occurrence, le champ *des théories de l'apprentissage* par des *machines learning* offre des *perspectives explicatives et prédictives* très prometteuses et jusqu'à ce jour inégalées (Han et Kamber, 2006). Mais le choix de *l'algorithme de datamining** ne se fait pas au hasard. D'ailleurs les algorithmes dévolus à la *sélection de variables* sont peu nombreux. Il convient d'en utiliser un qui soit adapté au jeu de données utilisé, i.e. aux i.st.m* et aux i.st.e* proposés. Ce type d'outils de datamining* reste l'apanage de logiciels mathématiques spécifiques. Et qui plus est, l'algorithme ne permet pas à lui seul de résoudre le problème. Il convient aussi de choisir une méthode de sélection adaptée. Bien sûr, ces méthodes ne sont pas, ou partiellement seulement, programmées dans les *packages* (Institute for Statistics and Mathematics, 1997).

Le choix de la méthode est intimement lié à l'algorithme, aux caractéristiques du jeu de données d'apprentissage et à la finalité recherchée. A l'heure actuelle, seules certaines méthodes statistiques, basées sur du *scoring,* parviennent, à leur façon et en partie au moins, à démêler l'écheveau de cette problématique complexe. Les caractéristiques de certaines d'entre elles ont été analysées (Ghattas et Ben Ishak, 2008). Mais le problème est que les scores d'importance des variables ne présument pas de la manière dont elles interagissent avec le phénomène étudié. De plus, toutes ces méthodes ne sont pas prouvées sur le plan théorique. Les seules preuves scientifiques disponibles sont empiriques et se fondent sur des *jeux de données jouées* ou *issues d'expériences contrôlées*. Récemment, une stratégie de sélection par seuillage permettant de disjoindre les variables de bruit des variables explicatives a été proposée. (Genuer, 2010).
Elle sera décrite dans le chapitre 3 et les développements démontreront qu'elle est adaptée à l'identification des DES* à partir des i.st.m.* et i.st.e.* proposés. Il sera également expliqué qu'elle a dû être adaptée pour les pathologies dont l'incidence est particulièrement faible.

Il s'agit donc, dans un premier temps, de modéliser la géographie des PM* et des FE-FIM*, *jugés pertinents* par des i.st.m* et des i.st.e* robustes. Puis dans un second temps, d'identifier, au moyen d'une méthode de sélection de variables adaptée, les DES.
Avant de clore cette section consacrée à l'état des connaissances, il convient de définir ce que sont les FE* et les FIM* *pertinents* à intégrer

SANTE ENVIRONNEMENTALE ET FACTEURS ENVIRONNEMENTAUX RETENUS

La définition de l'environnement retenue est celle utilisée par l'OMS, i.e. *tous les FE* à l'exception des facteurs génétiques* (World Health Organization, 2009). Or, la position du géographe de la santé consiste à s'intéresser à tous les FE* *jugés pertinents* et donc à ne négliger aucune piste. (Salem, 1995). Les facteurs génétiques conditionnent pourtant les prédispositions morbides (Stein, 2012). Leur mise à l'écart résulte d'une contrainte pratique et non d'un positionnement particulier. Malgré l'émergence ces dernières années de Bases de Données (BD) environnementales et épidémiologiques qui offrent des perspectives de modélisation géographique presque illimitées, l'accessibilité à ces données reste la principale contrainte (Zeitouni, 2006).

La phase préalable à la modélisation géographique de PM* et de FE* par des i.st.m* et des i.st.e, à partir de méthodes robustes et de données fiables, consiste à identifier les *FE* pertinents* qu'il convient d'intégrer en vue de déterminer les DES*, grâce à des *méthodes datamining* adaptées*. Par définition, la modélisation géographique des FE* caractérise la variabilité spatiotemporelle des expositions environnementales *potentielles* ou *intrinsèques* à des *substances toxiques présentes dans les milieux : eau, air, sol, et biologiques*…, ou à des expositions *potentielles* à *des situations à risque* caractéristiques dans les milieux de vie familiaux, sociaux, économiques, sanitaires… (Hill, 1965). Il en découle l'hypothèse suivante.

Hypothèse subsidiaire sur les *FE* jugés pertinents :*
Il s'agit de tous les facteurs susceptibles de caractériser *les expositions environnementales* géographiques, *internes* ou *externes*, *avérées* ou simplement *suspectées*, à des *substances toxiques* présentes dans les milieux : eau, air, sol et biologiques, et qui ont des effets documentés, *déterministes ou stochastiques,* sur la santé, ou encore, les expositions *potentielles à des situations à risque*, i.e. à des CIM* ou des *contextes* territoriaux particuliers, si tant est qu'elles soient décrites dans la littérature.

Proposition 1 :
L'état de l'art s'attache à toutes les connaissances disponibles sur les interactions santé-environnement en épidémiologie et en Géographie. Comme l'environnement géographique est protéiforme, les FE* caractérisant les milieux sont discrétisés en quatre composantes : Les FIM* qui modélisent la géographie des CIM, et les FE* à connotation : Sanitaire (SAN) ; Socio-économique (SOCIO.ECO) ; et Physicochimique (PHY.CHIM).

Proposition 2 :
Dans le cadre d'une approche géographique multidimensionnelle globale, l'état de l'art balaye l'intégralité des expositions environnementales *potentielles ou intrinsèques* qui ont des effets avérés ou suspectés sur les PM* d'intérêt - uniquement. Or il n'existe pas de littérature sur les séquelles étudiées et le projet LEA a justement pour objet de faire le point à ce sujet. Par conséquent, il est supposé que les expositions à risque documentées pour les individus *a priori* sains, le sont *a fortiori* chez les sujets prédisposés, i.e. les patients.

L'état de l'art se focalise donc sur tous les FE* ayant des interactions plausibles avec les PM* d'intérêt – séquelles. Il convient de les décrire d'abord dans les grandes lignes, avant de les spécifier par composante environnementale. Les leucémies (LEUC) sont des cancers qui se caractérisent par une prolifération de lymphocytes immatures dans la moelle. Il en existe deux types : Les Leucémies Aigües Lymphoïdes* (LAL), soit 85% des leucémies infantiles et les Leucémies Aigües Myéloïdes* (LAM), plus rares mais plus dangereuses. Actuellement 80% des leucémies sont curables. Toutefois, avec le temps les individus traités développent des séquelles qui ont un impact sur leur Qualité de Vie (QV). Le nombre et la gravité des séquelles dépendent principalement de l'agressivité de la leucémie et des traitements ainsi que de prédispositions génétiques. Ces facteurs sont englobés dans la dénomination CIM* (Leplège, Ecosse et al., 1998).

Les cataractes (CATA) surviennent généralement à partir de 65 ans. Chez les sujets prédisposés, elles peuvent se manifester beaucoup plus tôt. Les causes médicales qui favorisent les cataractes sont les irradiations thérapeutiques, certains traitements agressifs et des maladies infectieuses (SFO, 2013). Quant aux FE* : le tabagisme, les expositions environnementales à la radioactivité ou à de rayonnements solaires, l'accès aux soins, l'effet *de la défaveur sociale,* ainsi qu'une alimentation pauvre en fruits et légumes favorisent les cataractes (Blondeau, 2009).

Les tumeurs secondaires sont des organismes cellulaires pathogènes ou tumoraux qui se développent à distance du site initial. Elles sont engendrées par la diffusion de métastases. Il s'agit d'un véritable fléau qui se développe à la marge de traitements anti-cancer de plus en plus efficaces. Elles apparaissent généralement entre deux et dix ans après la prise du premier traitement. Ici, la dénomination de tumeurs secondaires (TUM2) définit toutes les tumeurs secondaires majeures, i.e. malignes. Les tumeurs thyroïdiennes (THYR) font partie des tumeurs secondaires, la particularité de leur étude étant que l'intérêt est porté à toutes les THYR, bénignes ou malignes. La thyroïde est une glande du système endocrinien dont la principale fonction est de secréter les hormones qui garantissent le bon fonctionnement de l'organisme. De fait, la séquelle THYR a un impact considérable sur la QV* (HAS et INCa, 2010).

D'une manière générale les causes médicales de ces séquelles de type cancer sont liées à des prédispositions génétiques, à la prise de traitements agressifs, et à des expositions à de fortes irradiations thérapeutiques (Leplège, Ecosse et al., 1998). Les FE* avérés ayant un impact sur ces PM* sont : les expositions accidentelles à de fortes doses de rayonnements ionisants par suite de catastrophes nucléaires ou de retombées atmosphériques d'essais nucléaires, et des séquelles induites par des maladies vectorielles spécifiques (Henry-Amar. M, 1999). Quant aux FE* les plus suspectés d'avoir un impact sur le développement des cancers en général, et des THYR en particulier, il s'agit de toutes les expositions environnementales aux substances toxiques des deux premiers groupes de la nomenclature du Centre International de Recherche sur le Cancer (CIRC), comme le benzène, certains pesticides, les Hydrocarbures Aromatiques Polycycliques (HAP), le formaldéhyde, les Polychlorobiphényles (PCB) et les nitrates (IARC, 2008). Enfin, s'agissant des *expositions potentielles* à des situations à risque, les *tissus sanitaires* déficitaires* ou les *conjonctures défavorables* sont citées dans la littérature épidémiologique pour avoir des effets évidents sur l'incidence et la gravité des PM* étudiés (Penchansky et Thomas, 1981). Ce point de vue est partagé par les géographes de la santé qui qualifient ces expositions de *prédispositions géographiques à des phénomènes morbides* (Haddad, 1992).

Cependant, les expositions à des FE-PHY.CHIM* restent controversées. Les études géographiques irréfutables à ce sujet sont rares. Pour les PM* d'intérêt on peut citer l'exemple d'une augmentation accrue des tumeurs de la thyroïde chez les femmes en Slovénie, laquelle a été expliquée par la présence d'usines de fabrication de PCB à proximité des lieux de vie (Pavuk, Certhan et al., 2004) ; ou celui d'une augmentation anormale de l'incidence des cancers en Colombie qui avait été induite par la contamination des réseaux d'adduction d'eau potable par des sédiments naturellement chargés en plomb, en phtalate et en sulfure (Gaitan, 1983). Quelques cas similaires avaient été rapportés un peu partout dans le monde (Brucker-Davis, 1998), si bien que les *expositions géographiques chroniques à de faibles doses* de substances toxiques sont particulièrement étudiées depuis une décennie en géographie de la santé. Cependant, en France, on assiste à un regain d'intérêt de la part des épidémiologistes pour ce sujet pourtant longtemps mis à l'écart à causes des problèmes inhérents à la significativité des résultats (Afsset, 2009a). Ce changement de paradigme s'étend à l'Europe. Très récemment les données épidémiologiques d'une trentaine de cohortes ont permis de montrer de façon *significative* les effets *des expositions environnementales géographiques à des substances physicochimiques* sur la santé des populations et *a fortiori* sur celle des enfants (Gehring, Casas et al., 2013).

A l'heure actuelle, l'exposition à des FE* PHY.CHIM ne peut plus être négligée dans l'analyse des interrelations santé-environnement. Il convient d'ailleurs de raisonner – dans la mesure du possible -

en *terme d'expositions combinées* (Leux et Guénel, 2010) à toutes les substances présentes dans les milieux environnementaux et ayant des effets sur la santé, sans pour autant négliger l'effet conjoint des FE-SAN* et FE-SOCIO.ECO*, ni omettre d'intégrer l'influence des CIM*.

Cet état de l'art sur les interactions santé-environnement a pour objet de dresser une liste aussi exhaustif que possible des *FE* jugés pertinents* vis-à-vis des PM* d'intérêt – séquelles : les CATA, les THYR et les TUM2. La perspective reste la modélisation géographique de ces FE* qui sont déclinés selon la discrétisation proposée : SAN, SOCIO.ECO et PHY.CHIM, sans omettre les FIM*, la première qui est investiguée.

FACTEURS INDIVIDUELS ET MEDICAUX

On entend par Facteurs Individuels et Médicaux* (FIM) la géographie des Caractéristiques Individuelles et Médicales* (CIM) des personnes dont les variabilités spatiotemporelles sont susceptibles de présenter une forme *d'exposition intrinsèque* morbide. Les CIM* sont accessibles par le biais d'informations épidémiologiques à connotation personnelle et/ou comportementale, et relatives à l'historique médical des individus – en l'occurrence il s'agit des données patients extraites de la BD-LEA, et *suspectées* d'avoir un lien avec les Phénomènes Morbides* (PM) d'intérêt - séquelles CATA, THYR et TUM2.

En géographie de la santé, le paradigme actuel consiste à raisonner en termes d'expositions *environnementales combinées*. Il s'agit de prendre en compte simultanément toutes les expositions géographiques possibles – sporadiques ou chroniques – à des doses élevées ou faibles de substances physicochimiques, présentes dans les milieux environnementaux, qui ont des effets documentés sur la santé (Leux et Guénel, 2010). Parallèlement il s'agit aussi d'estimer les *caractéristiques territoriales conjoncturelles socio-économiques* ou celles du *tissu sanitaire* qui induisent des *prédispositions aux phénomènes morbides* (Haddad, 1992).

Toutefois, les études menées en géographie de la santé négligent souvent l'importance des CIM* pour évaluer les expositions *à des situations à risque* ou à des *substances toxiques*. Elles se limitent à la modélisation géographique globale des *milieux et* omettent celle *des microcosmes géographiques inhérents* aux styles de vie, aux conduites individuelles à risque, et à l'historique médical des individus. Par conséquent, la spécification de DES* est biaisée. (World Health Organization, 2009).

Pourtant, les CIM* conditionnent l'état de santé et la façon propre à chaque individu d'appréhender et donc de subir l'environnement géographique des milieux de vie. L'omission des CIM, lorsque l'information est disponible, est une entrave à l'analyse des interactions santé-environnement (Abramson J.H., Abramson Z.H., 1988).

Il convient donc d'intégrer tous les FIM* permettant de modéliser la géographie *microcosmique* des *expositions environnementales intrinsèques*, estimable à partir des CIM* (Voltaire, 1490).

En géographie de la santé, les FIM* sont souvent négligés ou intégrés de façon partielle dans les autres composantes environnementales. A ce jour les FE* qualifiés de DES*, au *sens géographique du terme,* intégrant partiellement l'idée des CIM* sont : Les expositions professionnelles au benzène, formaldéhyde, PCB, pesticides, radionucléides ; Les expositions ménagères ou socio-urbaines à certains solvants et pesticides. Quant aux expositions des populations concernées à ces substances, ce sont celles qui résident à proximité d'installations : nucléaires, de traitement et de stockage des déchets, de fabrication de papier, de cuir, de bois, de bitumes.., d'exploitations agricoles, ou de réseaux de transports (Leux et Guénel, 2010).

Les études épidémiologiques conduites par l'EA-3279, dans une logique *individus-centré*e - i.e non géographique*, portent directement sur les CIM* à partir de la BD-LEA et ont permis d'identifier des DES* sociaux et médicaux (Leplège, Ecosse et al., 1998). Mais ces études soulignent aussi les limites des CIM* qui ne suffisent pas à expliquer l'incidence des séquelles étudiées. Et il fait nul doute que l'environnement géographique joue un rôle important (Michel, Auquier et al., 2007).

Le Professeur Michel et le Professeur Auquier de l'EA-3279 ont préconisé d'intégrer dans le cadre de cette approche géographique certaines CIM* ayant des effets très plausibles sur les séquelles étudiées, telles que le sexe, l'âge au moment du diagnostic de la leucémie et la durée du suivi, le type de leucémie, le protocole de traitement, les rechutes, et le recours à des traitements agressifs nécessitant des greffes ou des irradiations corporelles totales.

Le Professeur Auquier a soulevé *a posteriori* un problème d'interprétation au niveau de la durée du suivi. En effet, avec le temps, le risque de séquelle augmente. Cependant plus le suivi des patients est long et mieux leurs séquelles sont diagnostiquées. Or les séquelles des patients peu assidus sont diagnostiquées tardivement. L'ambiguïté de la durée du suivi est d'être identique au temps d'exposition à l'environnement des patients les mieux suivis. Ainsi cette temporalité peut être connotée positivement comme *de l'accessibilité aux soins* ou négativement comme un *effet environnemental délétère*.

Les connaissances épidémiologiques documentées sur les PM* étudiés et dont la géographie est modélisable à partir des CIM* sont :

- Les *facteurs génétiques* et l'influence conjointe de l'environnement sur les mutations du génome – par exemple les THYR ont une prévalence *deux fois plus forte chez la femme* (IARC, 2007). Mais un grand nombre d'autres gênes *ont un impact* sur les PM* étudiés et ne sont pas encore identifiés – il convient de s'y intéresser (Takada, Urano et al, 1995) ;

- Les *conduites individuelles à risque liées* à la consommation d'*alcool* et de *tabac* (IARC, 2013), *au stress oxydatif* lié à une *alimentation déséquilibrée* ou à *l'inactivité physique* et la s*édentarité* (Ghanbari-Niaki, Saghebjoo, et al., 2009), et aussi à l'altération de la santé mentale (Ahola, Honkonen et al., 2006), qui ont des effets *directs ou indirects* sur certains gènes.

- Les expositions professionnelles ou thérapeutiques à la radioactivité à des doses accidentellement fortes ou faibles mais chroniques qui sont avérées pour les THYR, TUM2 (IARC, 2013) et aussi pour les CATA (Jacob, Bertrand et al., 2010).

Quant aux effets de la radioactivité présente dans l'environnement, ils sont fortement suspectés et ont récemment été mis en évidence, dans certaines circonstances, sur les données d'une trentaine de cohortes européennes. L'étude *Environnement-Exposure* en question a aussi permis de démontrer les effets significatifs d'un grand nombre *d'expositions environnementales à des substances physicochimiques* et, par la même occasion, la vulnérabilité accrue des enfants (Gehring, Casas et al., 2013).

Ces expositions, autrefois fortement suspectées, tendent à se vérifier - en l'occurrence pour les rayonnements non ionisants, *type microondes,* liés à la téléphonie mobile (Kyung, Yul et al., 2010), aux champs magnétiques induits par les lignes haute tension (Inserm - expertise collective, 2005), à certains solvants, peintures, produits de beauté, au tabagisme passif dans les milieux familiaux ou sociaux (IARC, 2008), à l'usage de pesticides, de HAP, de caoutchouc, d'éléments radioactifs… dans certains secteurs d'activité (agricole, vétérinaire, militaire, pressing) ou des industries (mécanique, traitement des déchets,…) situées à proximité du lieu de résidence (Inserm - expertise collective, 2005).

Un grand nombre de ces expositions environnementales peut être modélisé dans l'espace géographique* par des FE, à partir de données communautaires. Mais elles peuvent l'être aussi de façon conjointe et par le biais des FIM. La modélisation des FIM* est assujettie à la disponibilité des informations décrivant les CIM* – contenues dans la BD-LEA. L'omission des FIM* crée un biais dans l'analyse géographique. En revanche les i.st.e* modélisant la géographie des FIM* ont une distance *a-spatiale virtuelle* avec les individus beaucoup plus petite que celle des autres i.st.e.

Le concept de *distance a-spatiale morbide** représente la plausibilité des effets sur l'état de santé d'une situation à risque ou d'une substance physicochimique. Plus cette distance est petite et plus les effets avérés ou suspectés des expositions environnementales en question sont bien documentés. En effet, les FE* sont modélisés à partir de *données environnementales* plus éloignées, i.e. plus incertaines,

imprécises, hétéroclites et empreintes de bruits de fond environnementaux (Mandin, 2004). Le concept de *distance a-spatiale virtuelle* sera pris en compte dans le processus de modélisation des FE-SAN.

FACTEURS ENVIRONNEMENTAUX SANITAIRES

Les Facteurs Environnementaux* (FE) sanitaires (FE-SAN) sont largement documentés pour leurs effets *directs* ou *indirects* sur l'état de santé des populations. Ils ont pour objectif de décrire *l'exposition géographique potentielle* à des *situations à risque à connotation sanitaire*.

La géographie de la santé s'intéresse tout particulièrement à *l'accès aux soins* qui en constitue la dimension socio-spatiale et qui permet d'évaluer la *qualité des tissus sanitaires** territoriaux (Coldefy, Lucas-Gabrielli et al., 2011). L'effet de *l'accès aux soins* sur l'incidence et la gravité des maladies est sans équivoque aussi pour les épidémiologistes (Penchansky et Thomas, 1981). *L'effet environnemental* du tissu sanitaire sur les séquelles - CATA, THYR, TUM2 – est particulièrement suspecté (Auquier, 2010).

Pour cette thématique de santé publique le point de vue des Géographes de la santé et celui des Epidémiologistes convergent (Carretier et Luporsi, 2011). Les FE-SAN* *pertinents* sont ceux qui permettent d'estimer la *qualité géographique des tissus sanitaires**.

En géographie de la santé le *recours aux soins* est défini comme l'usage que font les populations de *l'offre de soins*. Tous les individus n'ont pas accès aux mêmes services et aux mêmes équipements sanitaires car l'offre de soins est inégalement répartie sur le territoire. *L'accessibilité aux soins* est un concept pluridisciplinaire dans lequel s'enchevêtrent des composantes sociales, économiques et géographiques (Litva et Eyles, 1995).

On définit *l'accessibilité aux soins* comme *la capacité matérielle d'accéder à des ressources sanitaires et aux services de santé.* [Elle est] *considérée comme un déterminant de santé ou un éventuel facteur de risque* (Picheral, 2001). *L'accessibilité physique* constitue la dimension spatiale d'accès à l'offre de soins (Penchansky et Thomas, 1981). Et c'est justement à cette dimension spatiale de la problématique qu'il convient de s'intéresser.

Les Géographes de la santé sont des orfèvres pour quantifier *l'accessibilité aux soins*. *Les densités médicales* et *les distances d'accès aux soins* sont les principaux indicateurs spatiaux utilisés. Ils sont conçus dans une logique territoriale et fournissent aux politiques des informations sur les *niveaux de médicalisation* territoriaux. La cartographie constitue un outil d'aide à la décision qui permet d'améliorer les systèmes sanitaires (Picheral, 2001).
En épidémiologie spatiale, les indicateurs spatiaux sont les mêmes *et l'échelle des communes* est définie comme la plus topique pour *capter les variabilités spatiales* de recours aux soins en France (Chaix, Merlo et al., 2005).

La modélisation de *l'accès aux soins* sur le territoire confère une dimension opérationnelle aux états pour garantir *la santé individuelle et promouvoir la santé publique* (Salem, 1995). La notion d'accessibilité géographique aux soins est un élément clé de la constitution des Schémas Régionaux d'Organisation des Soins (SROS).
Ces derniers ont pour but de réduire les inégalités *géographiques d'accès aux soins*, et plus particulièrement de supprimer la présence de *déserts médicaux* en France (Coldefy, Lucas-Gabrielli et al., 2011).
Les SROS constituent la principale mesure de la loi Hôpital, Patients, Santé et Territoire (HPST) du 21 juillet 2009 (DILA, 2002).

L'accessibilité aux soins s'estimait historiquement par la densité médicale, i.e. par le rapport entre *le nombre de professionnels de santé* et *le nombre de malades potentiels*, dans un lieu et à un moment précis (Peguy, 1996). Or, cet indicateur spatial ne représente pas la réalité géographique. Il *colporte un biais*

d'échelle lié au *poids de la demande* qui ne tient pas compte des déplacements de population par-delà les frontières administratives (Lahousse et Piédanna, 1998). Il induit aussi implicitement l'hypothèse selon laquelle tous les professionnels, à l'intérieur d'une même zone géographique, présentent une accessibilité équivalente (Talen et Anselin, 1998) en supposant que les limites administratives constituent une *barrière spatiale* pour accéder à l'offre de soins des territoires voisins (Harrouin, Aligon et al., 2012).

Il ne prend pas en compte non plus la distance temporelle qui sépare les individus des ressources sanitaires, qui est pourtant déterminante en géographie puisqu'il s'agit du vecteur privilégié pour estimer l'accès aux soins - *qui change dans le temps selon les modes* [et les infrastructures] *de transport* (Brunet, Théry et al., 2009)

En géographie de la santé, les FE-SAN* sont modélisés dans l'espace et le temps par des i.st.e* représentatifs *de l'accès aux soins* et d'évidence, les notions d'espace et de distance-temps doivent être intégrées au cœur du processus de modélisation. De plus, ce processus doit tenir compte du fait que les frontières administratives n'entravent pas les déplacements dans l'espace géographique*. L'idéal serait de parvenir à embrasser simultanément, dans un même i.st.e* : La logique territoriale de la demande de consommation de soins, et sur des *secteurs flottants,* i.e. en s'affranchissant des limites administratives, celle de *l'offre de soins* et des distances-temps permettant, en moyenne, d'accéder *aux items sanitaires**. *Les items sanitaires* constituent* l'offre de soins, i.e., les spécialités médicales exercées par des praticiens de santé libéraux, ainsi que le plateau technique des établissements de santé, i.e. les services médicaux et les Equipements et Matériels Lourds (EML**).
D'une manière générale *l'accès aux soins* a une *influence environnementale potentielle* fortement suspectée pour les PM*. La modélisation des disparités géographiques des FE-SAN, par des i.st.e* représentatifs, s'attache uniquement *aux indicateurs spatiaux d'accès aux soins disponibles* les plus robustes afin de pallier la source le problème de redondance statistique. Il conviendra aussi d'identifier *les items sanitaires** les plus topiques aux séquelles d'intérêt.

En géographie la *notion de distance* est *le principal attribut des systèmes spatiaux* (Grataloup, 1996). Les distances spatiales peuvent être euclidiennes, i.e. *à vol d'oiseau*, *routière géodésique..*. ou transformées en *distances spatiales temporelles*.

Les FE-SAN* abordent *l'accessibilité géographique aux soins* par le prisme de la notion de *distance spatiale* aux items sanitaires*. Quant aux FE-SOCIO.ECO, ils constituent la dimension *a-spatiale de l'accessibilité géographique aux soins* - composante géographique des expositions *environnementales potentielles* à des *situations* à risque à laquelle il convient de s'intéresser.

FACTEURS ENVIRONNEMENTAUX SOCIO-ECONOMIQUES

L'accessibilité aux soins est un concept interdisciplinaire où s'enchevêtrent des notions sociales, économiques et géographiques.
Les FE-SOCIO.ECO* constituent une autre notion de *l'accès aux soins* qui s'articulent autour des concepts de distance : *perçue*, *économique, sociale, culturelle*... et dont l'impact sur l'état de santé des populations n'est pas négligeable (Litva et Eyles, 1995).

Le comportement des individus vis-à-vis *du recours aux soins* n'est pas conditionné uniquement par les CIM* des individus, i.e. par leur situation socio-économique ou leur historique médical. L'analyse géographique des FE-SOCIO.ECO* s'attache conjointement *à des dimensions individuelles* et *collectives*, sur la santé. En épidémiologie spatiale, les FE* à connotation socio-économique sont caractérisés par *l'effet de contexte* i.e. l'impact du milieu de vie communautaire sur les croyances et les comportements individuels (Chaix, Merlo et al., 2005).

Les FE-SOCIO.ECO* constituent la dimension *a-spatiale de l'accès aux soins*. Mais pas seulement. Ils confèrent une attention particulière aux *conjonctures territoriales défavorables,* qui sont aussi largement décrites en géographie de la santé, et qui constituent ce qu'Anderson et Newman définissent comme des *éléments de prédisposition aux phénomènes morbides* (Haddad, 1992)

L'intégration des FE-SOCIO.ECO* a pour objet de décrire la géographie *des expositions potentielles à des conjonctures territoriales* favorisant *les risques morbides*. A l'instar des FE-SAN, les FE-SOCIO.ECO* sont décrits, dans la littérature des interactions santé-environnement, comme ayant des répercussions sur tous les PM, dont les séquelles d'intérêt font partie (Carretier et Luporsi, 2011).

En géographie de la santé la modélisation du concept de distance *a-spatiale d'accès aux soins* s'évalue par le biais d'indicateurs communautaires qui caractérisent la variabilité des conjonctures socio-économiques territoriales (Haddad, 1992). En épidémiologie spatiale le principe est le même, les indicateurs spatiaux sont construits à partir de données géographiques socio-économiques. L'échelle la plus topique pour capter *les variabilités géographiques* de l'*effet de contexte* sur les états de santé est celle des communes (Chaix, Merlo et al., 2005).

Une pléthore d'indicateurs spatiaux est proposée dans ces littératures. Les plus classiques sont simples mais particulièrement robustes pour mesurer des distances *a-spatiales d'accès aux soins* à connotation culturelle, sociale et économique (Powell, 1995). Certains permettent de quantifier des variabilités géographiques spécifiques, comme les niveaux de vie, le pouvoir d'achat ou la répartition des richesses (Carstairs et Morris, 1989). D'autres sont plus atypiques, moins fiables, mais évaluent *des faits de santé particuliers* comme des distances *psychologiques d'accès aux soins* (Benach et Yasui, 1999).

Enfin, il existe des modélisations géographiques plus complexes qui font appel à des combinaisons de variables et des transformations topologiques comme les niveaux *de défaveur sociale*. Chacun performe selon des caractéristiques géographiques particulières comme : L'occupation du sol : rurale/urbaine, l'échelle géographique d'investigation, les populations ciblées - sexe, âge, etc. - et surtout les pathologies étudiées (Townsend, 1987).

Les FE-SOCIO.ECO* les plus *pertinents*, en l'état des connaissances actuelles en géographie de la santé et en épidémiologie spatiale, qu'il convient d'intégrer au regard des PM* d'intérêt sont ci-après déclinés.
La pauvreté territoriale qui induit un climat d'aversion à la consommation de soins et aux recours aux soins, avec des individus qui attendent d'être dans un état alarmant avant de consulter. La pauvreté territoriale contribue aussi à la création de déserts médicaux, donc à une diminution de la qualité des

soins (Penchansky et Thomas, 1981), à une augmentation des *délais d'obtention d'un rendez-vous* et à une difficulté accrue d'accès aux *traitements de référence* (Klein, 1989).

Les niveaux d'espérance de vie ou les taux d'accroissements démographiques – naturels ou globaux - permettent d'estimer l'efficience des politiques menées en matière de durabilité. Par exemple, l'augmentation des *migrations internes* ou *externes* constitue un marqueur d'attractivité géographique, généralement suscité par les aménités *socio-économiques* et la qualité *socio-sanitaire* des territoires (Dumond, 2004).

Les niveaux culturels géographiques renseignent sur les croyances et les valeurs populaires *vis-à-vis du recours aux soins* et de la *consommation de soins.* Le niveau individuel de culture est fortement corrélé avec celui des personnes que l'on côtoie, i.e. celui de l'entourage familial et socioprofessionnel (Zorman. Michel, 2001).

La géographie des catégories professionnelles est un moyen indirect d'évaluer les proportions de personnes exerçant des activités où il existe des *risques potentiels d'exposition* à des substances physicochimiques néfastes. Les personnes occupant ces emplois sont formées à la prévention des risques, ce qui influence positivement *leur consommation de soins* et celle de leur ménage. En contrepartie, des processus *d'accommodation au risque* se mettent en place, et de mauvaises habitudes induites par des conditions de stress s'installent et créent des contextes de surexposition aux risques (Anses, 2012).

La spécialisation économique des espaces à caractère agricole ou industriel permet d'estimer indirectement *les expositions géographiques potentielles* aux *pesticides* ou à des *substances toxiques* (métalloïdes, HAP, benzène, MES,…) en conférant à la notion de *risque* une *dimension collective* (Bailly et Beguin, 2005). Ces agents sont vectorisés par des processus de transfert dans les compartiments environnementaux (eau, air, sol, et biologique) des milieux de vie, exposant ainsi les populations locales à des risques morbides multiples (Anses, 2012).

Les comportements alimentaires géographiques : *L'évolution des modes* de vie a conduit à une *surconsommation d'aliments gras, salés, sucrés ou alcoolisés*, ce qui a induit une augmentation de l'obésité mais aussi des cancers (Ministère des affaires sociales et de la santé, 2012). Aussi, l'alimentation est une source d'exposition majeure aux *radionucléides* et aux *pesticides*. Ces expositions à risque sont liées à la nature et aux quantités d'aliments consommés – qui varient en fonction *d'habitudes alimentaires régionalisées* (Unité Cancer et Environnement, 2012).

Les niveaux géographiques de *défaveur sociale* caractérisent des accumulations spatiotemporelles de *désavantages socio-économiques* qui ont un impact fort sur les comportements individuels et des effets néfastes documentés sur la santé psychique des populations (Rey, Jougla et al., 2009). La défaveur sociale peut être interprétée comme une carence politique en matière de lutte contre : *la pauvreté, la délinquance, l'exclusion, l'équité d'accès à l'éducation et la culture* (Godin, 2007).
Les climats géographiques engendrés par la *défaveur sociale* ou *l'insécurité territoriale,* peuvent fragiliser la *sensibilité psychologique des populations* et induire des états de *souffrance psychique d'origine sociale* dont les effets cliniques sur la santé sont largement documentés (Furtos, 2007). En effet, *les contextes géographiques* propices à des *états individuels de stress chronique* engendrent des altérations de la santé mentale (Ahola, Honkonen et al., 2006), *des burnout* (Twellar, Winants et al., 2008), des troubles de l'humeur (Godin, Kittel, et al., 2005), de l'anxiété (Melchior, Caspi et al., 2007), et de la qualité du sommeil (Akerstedt, 2006), des conduites addictives comme la consommation de psychotropes (Head, Standsfeld et al., 2004) et la libération de neurotransmetteurs et d'hormones (Inserm - Expertise collective, 2011). Et par extension ils favorisent le développement de nombreuses maladies dont les cancers font partie, et aussi *des comportements à risque* vis-à-vis du recours aux soins.

Les FE-SOCIO.ECO* caractérisent la dimension a-*spatiale* de *l'accès aux soins* et des *conjonctures socio-économiques territori*ales. Ils ont des *répercussions indirectes sur* : le recours aux soins, les croyances, les conduites à risque, le stress, l'exposition à des substances toxiques. La dimension *collective de l'effet de contexte* se répercute sur les états de santé individuels. La modélisation des FE-SOCIO.ECO, par des i.st.e* robustes, est toutefois conditionnée par la granularité* des données communautaires disponibles, c'est-à-dire qu'elle est limitée par la qualité des BD socio-économiques étatiques (section.C)

FACTEURS ENVIRONNEMENTAUX PHYSICOCHIMIQUES

L'exposition géographique à des Facteurs Environnementaux* (FE) physicochimiques (FE-PHY.CHIM) est controversée. D'un côté, les géographes de la santé s'y intéressent depuis des décennies en raison du lien prégnant qu'il y a entre la présence de substances dangereuses dans les milieux de vie et la santé des populations.
Et de l'autre, les épidémiologistes sont plus modérés, parfois même réfractaires à cause de la difficulté d'établir des relations causales significatives. *Cependant de récentes études épidémiologiques ont relancé le débat* (Afsset, 2009a).

Les FE-PHY.CHIM* ont pour objet de modéliser la variabilité géographique des expositions environnementales, *internes ou externes*, à des substances toxiques – présentes dans les *milieux environnementaux* ou les *milieux de contact* - qui ont des effets documentés *avérés* ou simplement *suspectés*, *déterministes* ou *stochastiques*, sur les PM* d'intérêt - séquelles.
La controverse liée à l'analyse géographique des FE-PHY.CHIM* repose sur le fait qu'il existe des niveaux de seuil en deçà desquels les agents physicochimiques, même fortement toxiques, n'ont *a priori* aucun effet nocif. Or les quantités de substances toxiques ou les doses de contamination induites depuis les différents compartiments environnementaux sont généralement très faibles. A cela s'ajoutent des problèmes de puissance statistique, de qualité des données environnementales et de taille des effectifs spatialisés qui conduisent les épidémiologistes à émettre tant de réserves à ce sujet (Chasles et Fervers, 2011).

Nonobstant le développement récent de nombreuses bases de données environnementales et épidémiologiques (Zeitouni, 2006) et la puissance des outils mathématiques *de sélection* qui croît de façon exponentielle (Cartier, Villani et al., 2012), actuellement les perspectives offertes sont telles que les deux dernières réserves ne constituent plus réellement des barrières infranchissables. En revanche, le premier argument est de taille. En effet, les analyses des FE-PHY.CHIM* menées dans une logique purement géographique ne permettaient pas de conjecturer l'état de santé des populations avec une *évidence suffisante*. Pourtant, quelques études sans équivoque ont bien été rapportées un peu partout dans le monde (Brucker-Davis, 1998).

Toute l'ambivalence de l'attrait des géographes de la santé pour les FE-PHY.CHIM* réside dans le fait qu'en dépit de la faiblesse des doses mises en jeu, les populations sont exposées de façon incessante à des substances toxiques. En outre, les études cliniques manquent de recul pour évaluer avec certitude les risques réels des expositions chroniques à de faibles, voire de très faibles doses de substances toxiques. Surtout qu'il existe des substances *cancérogènes génotoxiques*, particulièrement difficiles à identifier, qui agissent sans seuil d'effet (Afsset, 2009a).
Les expositions environnementales à des substances physicochimiques sont multiples et diffuses. Les sources de contamination sont : la nourriture, l'air que l'on respire, l'eau que l'on consomme... De plus, ces substances sont omniprésentes dans les milieux de vie et l'intégralité des populations y est exposée. Et bien que les doses d'exposition soient effectivement parfois très faibles, les gradients géographiques sont entachés d'une forte variabilité spatio-temporelle. Les doses d'exposition varient en effet, avec les déplacements des individus dans l'espace géographique* (Inserm - expertise collective, 2005)

De surcroît, tous les milieux de vie sont contaminés simultanément par de nombreux agents physicochimiques néfastes liés à des effluents industriels et urbains, au lessivage des terres agricoles, à l'incinération des déchets, aux ondes électromagnétiques, aux radionucléides artificielles ou naturelles... Le rôle contributif de la qualité de l'environnement géographique sur les états de santé doit être abordé en termes d'expositions *combinées* (Leux et Guénel, 2010)
D'ailleurs, très récemment les données épidémiologiques d'une trentaine de cohortes européennes ont été croisées afin d'évaluer, entre autres, les effets des expositions environnementales géographiques à des substances toxiques combinées présentes dans les milieux de vie. Cette approche collaborative intégrée s'est révélée être un succès. Elle a permis d'établir des relations environnementales

significatives de certaines *expositions environnementales* à des substances toxiques combinées. Et l'efficience de cette significativité est d'autant plus forte chez les personnes les plus vulnérables i.e. les femmes enceintes, les enfants et les sujets prédisposés (Gehring, Casas et al., 2013) – dont les patients de la Cohorte LEA font partie.

Le Centre International de Recherche sur le Cancer (CIRC) évalue scientifiquement depuis 1971 le caractère cancérogène des différentes expositions à des substances physicochimiques toxiques et les classifie selon une nomenclature constituée de cinq groupes. Parmi près d'un millier d'agents étudiés, presque la moitié a été classée dans les deux premiers groupes, i.e. *comme cancérogènes avérés* (groupe1) ou *cancérogènes probables* (groupe2) pour l'homme (IARC, 2013). Le renvoi à cette nomenclature sera souvent utilisé par la suite.7

L'analyse des effets sanitaires des FE-PHY.CHIM, i.e. des expositions environnementales géographiques à des substances toxiques combinées dans les milieux de vie, est aujourd'hui nécessaire tant sur le plan scientifique que pour conférer aux états des supports leur permettant de justifier socialement leurs actions territoriales préventives en vue de *protéger la santé individuelle et promouvoir la santé publique* (Salem, 1995)

Les FE-PHY.CHIM* *pertinents* sont ceux qui ont des effets cliniques sur la santé et qui sont particulièrement suspectés dans la littérature en géographie de la santé, épidémiologie spatiale, médecin, pour leur toxicité et en particulier ceux discutés pour leurs effets dans le cadre d'expositions géographiques chroniques - pour les PM* d'intérêt, i.e. CATA, THYR, TUM2.

SUBSTANCES CHIMIQUES TOXIQUES

Il existe une pléthore de substances chimiques toxiques présentes dans les milieux environnementaux et toute la population y est exposée. S'agissant des principales substances chimiques, liées aux PM* d'intérêt, et présentes dans les milieux de vie, les plus *pertinentes* sont ci-après explicitées.

Le benzène :
Ce composé organique volatile est présent dans l'air. Les principales sources d'émission sont : la sylviculture, la pétrochimie, les transports routiers et les usines de fabrication de caoutchouc et de solvants. Les doses d'expositions environnementales au benzène sont soumises à de fortes variabilités géographiques. Sa présence dans les logements est accentuée par l'utilisation de solvants et de chauffages électriques, le tabagisme et la combustion de bois, de bougies et d'encens (Afsset, 2009a). Le benzène est un perturbateur endocrinien (Wong, 1995). Il est classé par le CIRC comme *cancérigène avéré* pour les THYR et les LEUC (Groupe 1) ; Et l'exposition environnementale au Benzène est classée *cancérigène probable* (Groupe 2) pour tous les cancers (IARC, 2013).

L'oxyde d'éthylène :
Il s'agit d'un gaz inodore présent dans la plupart des solvants commercialisés en France. Il est utilisé dans les secteurs de l'industrie, l'agroalimentaire, la pharmaceutique et le milieu médical. L'oxyde d'éthylène est une substance *classée* groupe 2B pour les LEUC. Les expositions environnementales à cette substance dans le processus de cancérogénèse sont particulièrement débattues (Unité Cancer et Environnement, 2012).

Les Hydrocarbures Aromatiques Polycycliques (HAP) :
Les HAP sont présents dans tous les compartiments environnementaux et notamment dans l'air, à des doses particulièrement élevées dans les grandes villes ou à proximité des axes routiers, dans les sols de certaines friches industrielles (carbochimie, pétrochimie) ou commerciales (anciennes stations d'essence).

Les HAP sont classés dans le Groupe 1 pour tous les cancers (IARC, 2013) et les effets des expositions environnementales sont particulièrement suspectés pour les THYR (Braverman, He et al., 2005).

Le nitrate (NO_3) :
Cette substance chimique est présente dans tous les compartiments environnementaux. Les principales sources de diffusion des nitrates sont d'origine : agricole car le nitrate est présent dans de nombreux pesticides et produits de fertilisation des sols, industrielle puisque tous les secteurs utilisent des produits contenant du nitrate, en particulier celui de l'automobile. Ils sont aussi présents dans les produits domestiques tels la lessive, les détergents, les solvants. Le nitrate est classé comme substance du Groupe 2 pour tous les types de cancers (IARC, 2013) . C'est un perturbateur endocrinien et les expositions environnementales au nitrate sont fortement suspectées d'avoir une incidence sur les THYR (Braverman, He et al., 2005)

Les Eléments Traces Métalliques (ETM)
Les ETM sont présents dans tous les milieux environnementaux et milieux de contact, et en particulier dans la chaine alimentaire. Les ETM caractérisent toutes les substances métalliques, qu'elles soient à l'état liquide ou solide. Les principales sources d'exposition aux ETM sont l'ingestion de denrées alimentaires et l'inhalation d'air ambiant. D'une manière générale les métaux lourds comme le plomb et le mercure sont considérés comme des agents cancérigènes avérés ou probables selon les cancers (IARC, 2008).
Quant aux expositions à des doses environnementales d'ETM combinées – dont les plus présents dans les milieux eau, air, sol et biologiques sont le Cadmium, le Cuivre, le Plomb, le Nickel et le Zinc – elles sont sans équivoque cancérigènes pour certains auteurs (Reilly, 1991).
Des études géographiques ont montré l'effet d'expositions combinées à des ETM sur l'incidence des cancers. Par exemple, en Colombie avec la présence naturelle de sédiments fortement chargés infiltrés dans les systèmes d'adduction d'eau potable (Gaitan, 1983).

Les dioxines
Ces composés organiques englobent de nombreuses substances. Les plus citées dans la littérature sont le Polychlorobiphényle (PCB) et le Formaldéhyde.

Le PCB est un agent organique qui était particulièrement utilisé dans la fabrication de pesticides. Sa production et son utilisation sont interdites en France depuis 1987. Mais son caractère particulièrement persistant fait qu'il s'est accumulé dans les sols et qu'il est toujours présent dans les milieux environnementaux. L'alimentation constitue 90% de l'exposition totale au PCB, lequel se trouve principalement dans le poisson, les fruits de mer, la viande, les œufs et le lait. (Unité Cancer et Environnement, 2012)

Le formaldéhyde est un gaz inodore et incolore présent de façon *inquiétante* dans l'air extérieur et dans les logements. Sa dénomination commerciale est le *formol* et il est actuellement présent dans de nombreux pesticides, solvants, désinfectants, peintures, résines et vernis à ongle (Unité Cancer et Environnement, 2012).

Le PCB et le Formaldéhyde ont un pouvoir cancérogène avéré sur l'homme et notamment pour les THYR (Brauer VF., Below H., Kramer A., Furthrer D., Paschke R., 2006). Ces dioxines sont classées dans le Groupe 1 de la nomenclature du CIRC pour les THYR et les LEUC (IARC, 2013). Les expositions environnementales au PCB sont largement documentées pour tous les types de cancers (Afsset, 2009a). Quant aux expositions à des doses environnementales, des études ont permis de montrer que l'incidence des THYR chez les femmes s'expliquait en grande partie par la présence d'usines de fabrication du PCB à proximité de leurs lieux de vie (Pavuk, Certhan et al., 2004).

Les pesticides

Les pesticides sont composés de près d'un millier de molécules chimiques. Elles sont toutes, sans exception, classées dans la nomenclature du CIRC. La dénomination de pesticides concerne tous les insecticides, fongicides, herbicides et antiparasitaires. Les pesticides sont en vente libre dans les supermarchés et les coopératives agricoles. Ils sont utilisés par les ménages et les professionnels. Les quantités mises en jeu dans l'agriculture sont astronomiques comparativement à celles destinées à des usages privés. Mais curieusement, la toxicité des pesticides agricoles est moindre. Les processus de diffusion successifs dans les milieux environnementaux, puis les milieux de contact, impliquent que des résidus, i.e. les molécules toxiques qui les composent, contaminent tous les maillons de la chaîne alimentaire. Les principales sources d'exposition sont les eaux de boisson et l'alimentation. Ces substances toxiques sont bio-accumulées dans le sang et les graisses sous-cutanées. Les risques sur la santé liés à des expositions environnementales aux pesticides sont difficiles voire impossibles à évaluer tant le nombre de substances toxiques est grand et les expositions assujetties à de fortes disparités géographiques, à la présence d'activités agricoles, à des doses ingérées déterminées par la catégorie socioprofessionnelle, au style de vie et aux habitudes alimentaires de chacun. Mais d'une manière générale l'exposition à des doses élevées de pesticides induit de nombreuses maladies, dont les CATA et les cancers (THYR, TUM) font partie (Afsset, 2009a).

Les expositions chroniques à de faibles doses sont particulièrement discutées pour les LEUC et les THYR (Garry, Danzl et al., 1992). Pour cause, les pesticides contiennent des substances chimiques qui classées dans les Groupes 1 et 2 de la nomenclature du CIRC pour les TUM et les THYR, e.g : le Formaldéhyde, le Trichloréthylène, le Nitrate, le Benzène, le Tetrachlorodibenzopara ou dioxine SEVESO, le Chlorophenoxy (IARC, 2013).

SUBSTANCES PHYSIQUES DELETERES

Les principales substances chimiques liées aux PM* d'intérêt ont été énoncées. Il est maintenant question de lister les facteurs physiques les plus suspectés pour leur toxicité et qui sont pareillement omniprésents dans les milieux de vie.

Les Rayonnements Non Ionisants (RNI)

Les RNI définissent tous les champs électromagnétiques et magnétiques à extrêmes et basses fréquences. Les RNI *hautes fréquences* se propagent dans l'environnement géographique depuis des sources disparates, et majoritairement depuis des lignes à haute tension et des caténaires (Afsset, 2009a). De nombreuses études épidémiologiques ont montré qu'il existait une corrélation évidente entre les expositions environnementales aux RNI et l'incidence des LEUC chez l'enfant (Inserm - expertise collective, 2005).

D'une manière générale les expositions à de fortes doses de RNI sont classées 2B pour tous les cancers (IARC, 2013). Les expositions à des RNI *basses fréquences type microondes* sont dues à la présence d'antennes relais et à l'utilisation de téléphones portables. Les études conduites jusqu'à présent ont permis de montrer leurs effets déterministes sur certains cancers. Quant aux expositions environnementales, elles restent fortement suspectées (Kyung, Yul et al., 2010). En milieu urbain les émetteurs de RNI sont omniprésents et les doses d'exposition sont soumises à de forts gradients géographiques. Par conséquent tous les citadins y sont assujettis (Draper et al., 2005).

La radioactivité environnementale

Cette dénomination qualifie la quantité incommensurable de substances radioactives présentes dans l'environnement. Leur origine peut être naturelle ou artificielle. La majeure partie de la radioactivité présente dans l'environnement est d'origine naturelle. L'air est le compartiment environnemental le plus contaminé et le plus contaminant.

Les Rayonnements Ionisants (RI) dans l'air sont essentiellement dus au rayonnement cosmique et aux processus telluriques de la désintégration de l'Uranium en gaz Radon. Les apports artificiels sont dérisoires, sauf parfois à proximité de certaines Installations Nucléaires de Base (INB) et à l'exception d'accidents nucléaires type Tchernobyl ou Fukushima (ASN, 2010).

Tous les atomes instables se dégradent en émettant de la radioactivité jusqu'à atteindre une forme stable. On appelle ces éléments des radionucléides. Ce processus s'effectue par libération de particules radioactives : α ou β et s'accompagne de RI* de type : X ou γ (ASN, 2013).

Les expositions à la radioactivité sont d'ordre environnemental, professionnel ou médical. Elles peuvent être *chroniques* ou *accidentelles* et les doses mises en jeu sont variables. Les effets *déterministes* sur l'état de santé sont connus. En revanche *les effets aléatoires* ne le sont pas et sont particulièrement suspectés pour les expositions environnementales. *La contamination interne* s'opère par ingestion ou inhalation de particules α *ou* β ; La contamination *externe* par contact des RI* X *ou* γ avec les tissus humains. Le *Gray* (Gr) et le *Sievert* (Sv) sont des unités qui permettent de comparer les effets des irradiations en fonction de la nature, l'organe cible, la dose, Et la durée d'exposition. Le Gray *est utilisé en médecine* pour contrôler la dangerosité des traitements. Le Sievert permet d'estimer les doses d'exposition à la radioactivité environnementale. Enfin, le Becquerel (Bq) sert à mettre en évidence et quantifier l'activité volumique des radionucléides dans les milieux environnementaux (ASN, 2010).

La radioactivité présente dans l'air est vectorisée à la surface terrestre par la pluie. Puis les radionucléides se propagent par les eaux de surface et les eaux souterraines dans tous les milieux environnementaux, jusque dans la chaine alimentaire. La principale source d'exposition s'opère par contact avec l'air ambiant. Les eaux de boisson et l'alimentation constituent la seconde source d'exposition environnementale.
Ces expositions sont chroniques. Les doses mises en jeu sont faibles. Cependant, elles peuvent être accidentellement élevées à cause d'apports artificiels liés à la présence d'INB. Les quantités propagées sont petites mais les radionucléides émises sont très dangereuses, c'est le cas du Strontium90, du Césium137 ou du Plutonium 238 (ASN, 2013).

L'exposition à la radioactivité environnementale de la population française est estimée en moyenne à 3,7 mSv/an. mais cette valeur - avec des gradients inter-régionaux variant de 5 à 15 - n'est pas représentative de la réalité géographique. Cette volatilité s'explique par des motifs météorologiques, géologiques comme la formation et l'épaisseur des couches de sols, topographiques avec notamment l'altitude, atmosphériques eu égard à l'épaisseur et la qualité de la couche d'ozone (IRSN, 2009).

<u>Les effets cliniques de la radioactivité sur la santé :</u>
Au-delà d'une certaine dose, l'exposition *interne* à des particules α ou β est classée groupe1 par le CIRC pour THYR et les LEUC. Et d'une manière plus générale : *all types of ionizing radiation are carcinogenic to humains*, i.e. qu'ils sont dans le Groupe.2b pour tous les cancers TUM (IARC, 2012).
S'agissant des CATA, la cohorte Hiroshima-Nagasaki a permis de montrer que cette pathologie était anormalement élevée chez les personnes exposées. L'effet des expositions chroniques professionnelles à de faibles doses de radioactivité a aussi été démontré chez *les radiologues, les cardiologues, les pilotes d'avion et les astronautes*. Quant aux expositions à la radioactivité environnementale, le lien est presque évident mais reste à démontrer (Jacob, Bertrand et al., 2010).
<u>Les expositions médicales :</u> Les expositions thérapeutiques représentent 35% de l'irradiation moyenne annuelle. Les patients irradiés font partie de la population *à risque* pour tous les PM* d'intérêt (IRSN, 2009).

<u>Les expositions géographiques à la radioactivité environnementale</u>
La définition de cette exposition a été énoncée précédemment. La nature des RI* et des radionucléides ainsi que les doses d'expositions induites par les milieux de contact sont assujetties à de fortes variabilités spatiotemporelles. Les expositions à la radioactivité environnementale sont controversées,

cependant dans la multitude de radionucléides présents dans l'environnement un grand nombre a des effets documentés ou suspectés sur les PM* d'intérêt – séquelles. En l'occurrence on peut citer : L'iode 131 - ^{131}I ; Le Carbone -^{14}C ; le Strontium90 - ^{90}Sr ; Le Césium137 - ^{137}Cs ; L'Antimoine 125 - ^{125}Sb ; Les isotopes radioactifs du Tritium ^{3}H $\stackrel{\text{def}}{=}$ T ; Ceux du Plutonium - ^{238}Pu, ^{239}Pu, ^{239}Pu. ; Ainsi que, d'une manière plus générale, les RI* de type γ (IARC, 2012).
Parmi les expositions environnementales les plus décrites dans la littérature, celles en lien avec les PM* d'intérêt – séquelles sont l'exposition au Radon dans les habitations et le fait de résider à proximité d'INB*. En dépit des controverses méthodologiques, ces expositions environnementales sont particulièrement étudiées pour leurs effets cancérigènes, notamment chez les jeunes gens (IARC, 2008).

Les expositions géographiques au Radon :
Le radon est un gaz naturel issu de la désintégration de l'uranium et du radium. Il est présent dans la croûte terrestre et a tendance à s'accumuler dans les maisons. L'exposition moyenne annuelle des individus en France à la radioactivité environnementale liée au Radon représente environ 38% de l'exposition totale (IRSN, 2009). Mais cette valeur reste peu significative car elle est sous le joug de fortes disparités géographiques. En effet, *l'exposition des populations au radon dans les habitations peut atteindre des niveaux proches de ceux qui ont été observés dans les mines d'uranium en France* (IRSN, 2001).
L'exposition environnementale au Radon a été prouvée pour le cancer du poumon. Mais il ne s'agit pas du seul cancer incriminé puisque cet élément radioactif est classé dans le Groupe.1 pour les LEUC et les THYR (IARC, 2013).

Les expositions géographiques liées à la présence d'INB :
Toutes les INB* en fonctionnement normal diffusent continuellement des radionucléides dans l'environnement. Les quantités mises en jeu sont faibles. Cependant, il est fréquent que ces rejets dépassent les limites autorisées. Ce type d'exposition environnementale est particulièrement suspecté en géographie de la santé et de nombreuses études mettent en lumière une augmentation curieusement élevée de l'incidence des cancers chez les populations riveraines. C'est le cas par exemple dans la ville de Pickering, au Canada, où une recrudescence évidente de l'incidence des LEUC infantiles a été constatée après la mise en service d'un réacteur nucléaire (AECB, 1991).

Quant aux études épidémiologiques, elles s'intéressent exclusivement aux expositions professionnelles du personnel. Elles ont permis de déterminer des *liens de causalité* (Zablobska, Ashmore et al., 2004).
Cependant, de récentes études épidémiologiques ont également mis en évidence une incidence accrue des LEUC et des THYR chez les enfants vivant à proximité d'INB* - *le débat a été relancé en France* (Afsset, 2009a).

Les Matières En Suspension (MES)
Les MES sont des particules solides présentes dans l'air ambiant *extérieur* et dans l'air ambiant *intérieur* des habitations. Les principales sources d'émissions environnementales des MES sont naturelles (feux de forêts, éruptions volcaniques, érosions éoliennes) ou anthropiques (rejets gazeux des activités industrielles, des transports, des chauffages résidentiels et exploitation de carrières). Les doses de MES dans l'air *intérieur* des habitations résultent de transferts d'air *extérieur* et elles sont amplifiées par la cuisson d'aliments, l'utilisation de certains solvants, le tabagisme et les ventilations mal entretenues. L'exposition environnementale aux MES s'opère par *inhalation* ou *ingestion*. Elle varie selon la zone géographique, les conditions météorologiques, l'occupation biophysique des sols et l'état des logements. Les MES sont discrétisées selon leur diamètre et se constituent par agrégation de particules chimiques - souvent toxiques – dont les plus courantes sont : Les HAP, les métalloïdes, les pesticides, les phtalates des particules composées d'un noyau benzénique, et les dioxines comme le formaldéhyde et le polybromodiphényléther… Les expositions environnementales aux MES favorisent les troubles respiratoires et les cancers du poumon. Mais les MES sont aussi documentées pour avoir des impacts sanitaires contributifs à spectre large (Afsset, 2009b).

Au regard des éléments qui constituent les MES et des disparités géographiques des expositions, ces dernières sont fortement suspectées pour tous les PM* d'intérêt.
D'ailleurs, des études ont permis d'établir des liens évidents entre l'exposition géographique chronique à de faibles concentrations de MES et l'incidence des cancers sur des populations vivant en milieu urbain (Pope et Dockery, 2006).

Le lien a également été établi concernant des expositions occasionnelles à de fortes concentrations de MES dans l'air après des feux de forêt de grande ampleur, où une augmentation irréfutable de l'incidence des maladies, et des cancers en particulier, a été constatée. (Johnston Fay, Henderson et al., 2012).

Les caractéristiques géophysiques
Ces facteurs sont pour certains assujettis à de fortes variabilités spatiales et ont des effets potentiels sur les PM* d'intérêt. Parmi eux, ont notamment été retenus des paramètres climatiques et le relief terrestre.

Les paramètres climatiques
Les paramètres météo, ont déjà été énoncés à demi-mot et deux ont été retenus. Il s'agit en premier lieu du *rayonnement solaire global* pour lequel les effets des expositions environnementales sont documentés pour les CATA (SFO, 2013). Il en est de même pour ses effets sur certains cancers de la peau (ifss, 2012). Le rayonnement global se compose de RI* de types X et γ et mais aussi de RNI comme les Ultraviolets (UV). Les RI* sont classés dans le Groupe.1 pour les THYR et dans le Groupe.2 pour tous les cancers (IARC, 2012).
Les expositions au rayonnement solaire global sont assujetties à de fortes disparités géographiques qui dépendent des conditions météorologiques et des niveaux altimétriques. Cependant, les régions ensoleillées ont des effets psychologiques bénéfiques – donc peuvent aussi avoir des répercussions positives sur l'état de santé des populations (Blondeau, 2009), (à ce sujet, voir aussi la partie sur les FE-SOCIO.ECO). Mais les personnes vivant en altitude, là où la couche d'ozone est plus fine, sont surexposées aux UV et l'exposition artificielle aux UV en institut se cumule à l'exposition naturelle (Blondeau, 2009).
Le paramètre climatique retenu en second lieu est la *pluviométrie*. D'évidence il est corrélé avec le premier. Cependant, ce paramètre météorologique peut être interprété comme un marqueur géographique de qualité des milieux de vie. En effet, les répercussions de la pluviométrie sur l'état psychologique des individus (Afsset, 2009a), mais aussi sur le niveau de contamination des milieux environnementaux par la radioactivité environnementale (IARC, 2012), les ETM (IARC, 2012), les dioxines (Unité Cancer et Environnement, 2012) et les MES (Afsset, 2009b) dont les risques sont associés aux PM* étudiés, ont déjà été évoquées.

La topographie
L'effet du relief sur les PM* d'intérêt est une hypothèse corollaire et évidente aux notions précédemment énoncées. En effet, les expositions aux rayonnements solaires globaux et aux UV - pour les CATA - (SFO, 2013) et à la radioactivité environnementale naturelle liée à des apports cosmiques et telluriques (IRSN, 2009) – pour les CATA, THYR, TUM2 - (Jacob, Bertrand et al., 2010) ; (IARC, 2012) ont déjà été évoquées et sont *indirectement* liées à l'altimétrie des lieux de vie.

L'analyse des processus d'expositions environnementales à des substances physicochimiques et des hypothèses énoncées a grevé ces dernières d'un caractère *potentiel* ou *intrinsèque*.
L'état des connaissances sur les interactions santé-environnement a permis d'identifier les FE* *pertinents* qu'il convient d'intégrer au regard des PM* étudiés et de la dialectique proposée, vouée à la modélisation géographique et à l'identification statistique de DES*.

L'état de l'art a servi de point de départ pour lister et proposer des méthodes de modélisation géographique et d'analyse multidimensionnelle adaptées à la problématique.

Par conséquent, il est désormais possible de modéliser, par des i.st.m* et des i.st.e* robustes, la variabilité géographique des PM* d'intérêt et des *expositions environnementales* : *Potentielles i.e.* à des FE-SAN*, FE-SOCIO.ECO* et des FE-PHY.CHIM* estimés à partir de mesures environnementales, et *intrinsèques* - i.e. aux FIM* et aux FE-PHY.CHIM* évalués sur des doses d'exposition.

Cependant la modélisation géographique des PM* d'intérêt et des FE/FIM* pertinents ne se limite pas à des considérations bibliographiques. En effet l'estimation des i.st. est aussi fortement conditionnée par la qualité des données disponibles stockées dans les BD existantes. La qualité des variables utilisées est passée en revue dans la section suivante.

SECTION C) BASES DE DONNEES EPIDEMIOLOGIQUES ET ENVIRONNEMENTALES

Les méthodes de modélisation énoncées ou proposées dans l'état de l'art sont vouées à la modélisation géographique des PM* et des FE/FIM* *pertinents* par le biais d'i.st.m* et d'i.st.e* robustes.

Avec l'essor, ces dernières années, de multiples Bases de Données (BD) de toutes sortes, les perspectives de modélisation géographique en santé environnementale* sont quasi illimitées. Cependant, les deux principaux écueils à la construction d'i.st. robustes sont : L'accès à ces données, i.e. leur disponibilité ; Et leur qualité, i.e. la fiabilité des données mesurées et mesurables utilisées (Zeitouni, 2006).

Objectif : Passer en revue les principales grandes BD retenues et les variables utilisées en vue de modéliser la géographie des PM* d'intérêt et des FE/FIM* jugés *pertinents*. La granularité* des variables impliquées dans les processus d'estimation des i.st.m* et des i.st.e* est déclinée.
Définition : la granularité* des variables représente leurs caractéristiques intrinsèques : Nom ; Echelle, Temporalité, Précision, Unité, Lacunes, Informations auxiliaires ; Incertitudes ; Techniques ; Disponibilité ; Fournisseur des données…

Remarque 1: Le choix des données est aussi lié à leur capacité à caractériser les interactions santé-environnement. Et parfois la modélisation géographique des FE* nécessite l'utilisation de plusieurs variables. L'usage privilégié et les stratégies de fusion de certaines d'entre elles seront spécifiés en Partie.2.

Remarque 2: Certains FE/FIM* ne peuvent pas être modélisés car les données font défaut. Pour chaque composante environnementale, quelques propositions de BD environnementales et de variables existantes mais indisponibles seront faites.

Remarque heuristique liminaire : L'état de l'art a permis de mettre en évidence l'importance de l'analyse mathématique des interactions entre les i.st.m* et les i.st.e. Les progrès effectués dans le couplage de l'informatique avec des processus stochastiques d'apprentissage et des statistiques non paramétriques offrent la possibilité de caractériser de façon consistante ces interactions (Cartier, Villani et al., 2012). La méthode de sélection retenue : VSURF, permet de disjoindre les variables de bruit, i.e. celles qui n'influencent pas la réponse, de celles qui permettent de l'expliquer. Si les i.st.e* sont suffisamment représentatifs de la réalité géographique, ils permettront d'identifier les DES*, i.e. les FE* qui déterminent les PM* d'intérêt. Or la *méthode d'ensemble* sur laquelle VSURF s'appuie n'est pas prouvée scientifiquement même si sa robustesse et sa puissance, dans la pratique, sont manifestes (Genuer, 2010)

Proposition heuristique :
Introduire dans chaque composante environnementale un *i.st. Curieux* de test* dans le but de vérifier la robustesse de VSURF. Autrement dit, il s'agit pour chaque FE/FIM* de proposer des i.st.e* qui doivent être qualifiés de variables de bruit, soit parce qu'ils sont très éloignés du PM* d'intérêt, soit parce qu'ils utilisent des données de moindre qualité qui *bruitent la modélisation géographique des FE/FIM* qu'ils sont censés représenter* (Mandin, 2004).
Remarque heuristique : Comme la procédure de datamining* se fonde sur une méthode d'ensemble statistique non paramétrique et que la procédure de sélection est éliminatrice, l'introduction de *i.st. Curieux* de test* n'induit pas de perte de puissance.

Les FE.SAN, FE.SOCIO-ECO et les FE.PHY-CHIM seront modélisés à partir de données géographiques contenues dans les BD environnementales. Quant aux PM* et aux FIM* *pertinents,* ils utilisent des variables de la BD épidémiologiques LEA

BASE DE DONNEES EPIDEMIOLOGIQUES

La BD LEA est la base de données utilisée pour modéliser la géographie des PM* d'intérêt – séquelles – et celle de la géographie des CIM, i.e. les FIM.

La géographie des PM* d'intérêt est modélisée à partir de variables séquelles, ou réponses, notées : y_i^j. Elles caractérisent le fait que le patient « i » a développé ou pas la séquelle « j ».
Les FIM* sont décrits par des variables représentatives des CIM, dont la notation vernaculaire statistique est généralement donnée par : $x_i^{l:CIM}$. Autrement dit la Caractéristique Individuelle ou Médicale « l » associée au patient « i ».

D'une manière générale l'avantage d'utiliser des données Cohorte touche à leur fiabilité et permet d'estimer certaines prédispositions individuelles morbides telles que le sexe, l'âge, l'activité physique ou l'historique médical : type de LEUC, agressivité du traitement reçu…

PROJET LEA ET CARACTERISTIQUES DE LA BASE

Généralités : la BD LEA fait pendant à l'étude Leucémies de l'Enfant et Adolescent (LEA) mise en place en 2004. Actuellement cette cohorte prospective multicentrique comptabilise plus de 2000 patients traités en France pour une Leucémie Aiguë (LA), pendant leur enfance et après janvier 1980. Chaque année des patients sont inclus, soit sporadiquement par *les centres investigateurs de référence*, soit massivement avec le ralliement d'autres *centres de référence.* Les *centres de référence* sont des établissements de santé habilités à traiter les leucémies chez l'enfant par la Haute Autorité de Santé (HAS) (HAS et INCa 2010)

Les progrès thérapeutiques ont transformé le pronostic vital des individus atteints de LA en posant, avec une grande acuité, le problème des effets secondaires comme celui des séquelles.
Le projet LEA s'attache principalement à la Qualité de Vie (QV) des individus par l'analyse des déterminants de santé du devenir à moyen et long termes.

Tous les deux ans des *auto-questionnaires de QV* permettent de* recueillir auprès des patients et de leurs familles des informations sociodémographiques, économiques, cliniques (incluant les séquelles) et thérapeutiques. Ces informations mesurent la perception subjective du patient selon plusieurs dimensions : *relations avec la famille, estime de soi, vitalité, état de santé perçu, relation avec les amis…*

Les auto-questionnaires sont remplis par les patients lorsqu'ils sont majeurs et leur QV* est évaluée par le biais de scores et d'une *échelle* internationale *: SF-38* (Leplège, Ecosse et al., 1998). En revanche, lorsque les patients sont mineurs, les auto-questionnaires sont remplis conjointement avec les parents et la QV* est évaluée par le biais d'une échelle différente : VSP-A (Vécu et Santé Perçue de l'Adolescent et l'enfant) (Michel, Auquier et al., 2007).

Les données acquises jusqu'à présent ont permis de répondre à certaines questions, mais les recherches menées soulignent aussi la limite des déterminants explorés. *La prise en compte des FE* et bientôt des facteurs génétiques dessine des perspectives explicatives ou contributives intéressantes* (Auquier, 2012).

Caractéristiques de la BD LEA

La BD LEA utilisée contient les données recueillies en 2009 qui ont été consolidées, *a posteriori*, par les données de 2010. De fait, sont concernés les 943 patients inclus dans la BD-LEA de 2009 des centres de

référence de : Nice, Marseille, Grenoble, Clermont-Ferrand et Nancy. Cet échantillon constitue *a peu près* l'essentiel des enfants traités pour une LA, depuis 1980 et domiciliés, au moment du diagnostic, dans les régions : PACA, Corse, Alsace-Lorraine, et une partie de ceux situés en Auvergne et en Rhône-Alpes.

Le processus de consolidation a permis de combler de nombreuses lacunes sur des variables relatives aux CIM* : $x_i^{l:CIM}$ et aux PM* : y_i^j. Les variables réponses ou séquelles sont booléennes et de nature qualitative, tel que :

$$y_i^j = \begin{cases} 1 \stackrel{\text{def}}{=} \text{OUI} & \text{lorsque: le patient. i a développé la séquelle. j} \\ 0 \stackrel{\text{def}}{=} \text{NON} & \text{lorsque: le patient. i n'a pas développé la séquelle. j} \end{cases}$$

La modélisation géographique des CIM* et des séquelles représente des *clusters* - ou *agrégats spatiaux de faits ou d'états de santé* recouvrant *une période temporelle* (Tillaut, 2005). En l'occurrence cette période s'étale de 1980 à 2010.

La seule information spatiotemporelle disponible dans la BD-LEA est le Code Postal (CP) du lieu de résidence du patient tel que ce dernier l'a décliné au moment de sa dernière évaluation de QV* ou de sa dernière consultation médicale. La variable LEA associée est notée : x_i^{CP}. Lorsqu'un patient déclinait un nouveau CP l'ancien était écrasé. Ce n'est plus le cas aujourd'hui. Mais la reconstitution diachronique n'a pas été possible pour des raisons économiques (Auquier, 2010).
La variable x_i^{CP} de la BD-LEA 2009 consolidée 2010 compte 196 lacunes. Les patients concernés ont été recontactés et ces lacunes ont été comblées en 2012. Or à ce moment la quantité des traitements effectués ne permettait plus de les intégrer.
Parmi les 943 patients de la BD LEA de 2009 consolidée, la cohorte compte 521 garçons pour 422 filles. La moyenne d'âge au moment du diagnostic de la LA est de 6,5 ans. En moyenne ces individus étaient âgés de 20,5 ans en 2010. A cette date seulement 4 d'entre eux étaient déclarés décédés.

S'agissant des séquelles étudiées, les données utilisées ont permis d'estimer les incidences suivantes : 12,5% de cataractes (CATA) ; 9,44% de tumeurs thyroïdiennes (THY) et 4,03% de tumeurs secondaires majeures (TUM2).

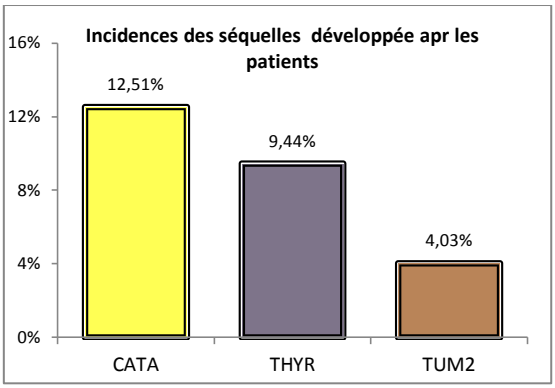


Figure 3 : Incidence des séquelles sur les patients de la BD LEA 2009 consolidée avec les données 2010

Dans cet échantillon 85,7% des patients ont été traités pour une Leucémie Aigüe Lymphoïde (LAL) contre 14,3% pour une Leucémie Aigüe Myéloïde (LAM).
Parmi les principaux facteurs médicaux ayant une influence sur le développement ou la latence des séquelles d'intérêt, on dénombre : 27,3% des patients ayant subi une greffe ; 19,8% une irradiation corporelle totale et 16,1% ayant fait une rechute.

Enfin, concernant les comportements individuels à risque, une variable LEA permet d'estimer le niveau d'activité physique pratiqué. Parmi les patients participants activement à l'étude de QV, 25% déclarent

ne pas pratiquer de sport – cette estimation concerne uniquement les patients pour lesquels la variable ne contenait pas de lacune.

Les principales caractéristiques du projet LEA et de la BD-LEA ont été décrites, il s'agit maintenant de lister les variables utilisées et de spécifier leur granularité* - pour la modélisation géographique PM* et des FIM* *pertinents et Curieux**.

VARIABLES UTILISEES POUR LA MODELISATION DES PM* ET DES FIM*

Les variables LEA : y_i^j x_i^l déclinées ci-après sont celles de la BD-LEA 2009 consolidée avec les données 2010. Par conséquent elles sont supposées fiables entre 1980 et 2010. Ces variables caractérisent des patients – donc la granularité* d'échelle est du type *individus-centrée**.

PM* des modélisables & granularités* des variables disponibles

PM	Variable	Description	Nature	Unité / modalité	Lacune	Commentaire
Géographie	*CIM*	*CIM*	*CIM*		*proportion*	
CATA	y_i^{CATA}	séquelle	Qualitative Booléenne	OUI/NON	24,1%	
THYR	y_i^{THYR}	séquelle	Qualitative Booléenne	OUI/NON	7,40%	
TUM2	y_i^{TUM2}	séquelle	Qualitative Booléenne	OUI/NON	2,05%	
Code Postal	x_i^{CP}	géographique	Quantitative Discrète	SU	21,05%	Clé de spatialisation des données LEA

Tableau 1 : Variables morbides épidémiologiques mobilisées

FIM* *pertinents* modélisables et granularités* des variables utilisées

FIM	Variable	Description	Nature	Unité / modalité	Lacune	Commentaire
Géographie	*CIM*	*CIM*	*CIM*		*proportion*	
GENRE	x_i^{SEXE}	Sexe du patient	Qualitative Booléenne	Fille/Garçon	0%	
AGE AU DIAGNOSTIC	$x_i^{AGE\_DIAG}$	Age au moment du diagnostic de la LA	Quantitative Continue	Année	0%	
DUREE DU SUIVI	x_i^{DSUIVI}	Durée entre le diagnostic de la LA et la dernière consultation	Quantitative Continue	Année	0%	Ambivalence sur l'interprétation épidémiologique de cette variable - voir section b état de l'art sur les FIM
TYPE DE LEUCEMIE	$x_i^{TYPLEUC}$	Type de LA traitée	Qualitative Booléenne	LAL/LAM	0%	
PROTOCOLE DE TRAITEMENT	x_i^{PROTOC}	Protocole de traitement reçu	Qualitative Multi-classes	11 modalités – types de traitement*	0,2%	(*) LAL-80 ; LAL-84-85 ; LAM-80 ; EORTC ; Fralle 92-93 ; Fralle 2000 ; LAM – 89-91 ELAM-02 Atre-LAME ; Autre ; NR
RECHUTE	x_i^{RECHUT}	Rechute après traitement	Qualitative Booléenne	OUI/NON	0%	
GREFFE	x_i^{GREF}	Greffe de moelle	Qualitative Booléenne	OUI/NON	0%	
IRADIATION THERAPEUTIQUE	x_i^{IRCAT}	Irradiation corporelle totale	Qualitative Booléenne	OUI/NON	0%	

Tableau 2 : Variables individuelles et médicales épidémiologiques mobilisées

FIM* *Curieux* de test* et granularités* des variables utilisées

FIM	Variable	Description	Nature	Unité / modalité	Lacune	Commentaire
géographie		*CIM*			*proportion*	
ACTIVITE PHYSIQUE SPORTIVE	x_i^{ACPHY}	Nature de l'intensité des activités sportives	Qualitative Multi-classes	3 modalités(*)	28,63%	(*): Pas de sport ; Sport scolaire ; Sport scolaire et extra-scolaire

Tableau 3 : Variable individuelle comportementale mobilisée.

Remarque :
Selon le Professeur Auquier, la variable LEA x_i^{ACPHY} *est inconséquente. A côté des autres variables intégrées elle est moins fiable, plus éloignée des patients, et n'a a priori aucune incidence sur les séquelles.*
Par conséquent x_i^{ACPHY} présente bien les caractéristiques nécessaires à l'estimation des i.st.e* Curieux* de test. Cependant, l'état de l'art met en évidence le fait que l'activité physique : *régule certains gènes qui contribuent à la fabrication des globules blancs* (Ghanbari-Niaki, Saghebjoo, et al., 2009), diminue le stress oxydatif donc le risque de croissance tumorale, prévient des maladies en général et influence positivement l'humeur. L'OMS recommande même de pratiquer une activité physique modérée 50 min 5 fois / semaine ou une activité sportive 20 min 3 fois / semaine (Lesur, 2011).
En contrepartie x_i^{ACPHY} *est effectivement plus éloignée des patients et le nombre de lacunes est important à l'aune des autres CIM.* Elle devrait donc bien être identifiée comme une variable de bruit.

***FIM* non modélisables* - variables LEA indisponibles**
Remarque liminaire :
Les variables LEA de QV* pourraient *a priori* permettre de modéliser certains FIM* comme l'état psychologique des patients. Or elles utilisent des échelles de QV* différentes qui les rendent incomparables. Leur utilisation induirait une ségrégation des patients, donc des effectifs spatialisés encore plus petits. Ces variables ne sont donc pas utilisables.

FIM	Référence à l'état de l'art	Motifs	Commentaire
Géographie	Décrire la géographie		
ETAT PSYCHOLOGIQUE DES* PATIENTS	De l'état psychologique qui a des répercussions sur l'état de santé	Variable de QV	Ségrégation des patients spatialisés
FACTEURS GENETIQUES	Des patients porteurs génétiquement prédisposés	N'existe pas dans la BD	Facteurs pas encore intégrés dans LEA
TABAGISME PASSIF	Des expositions aux fumées de cigarette dans les milieux familiaux	N'existe pas dans la BD	Facteurs pas encore intégrés dans LEA
EXPOSITION A DES* SUBTANCES TOXIQUES	Des expositions à risques liées à l'activité professionnelle des parents	Pourrait être reconstituée	La profession des parents est toutefois renseignée dans LEA
EXPOSITION GEOGRAPHIQUE	Vivre à proximité d'une : ligne haute tension ; INB* ; Usine de fabrication de peintures, de solvants, d'incinération des déchets…	N'existe pas dans la BD	Facteurs pas encore intégrés dans LEA

Tableau 4 : Variables épidémiologiques pertinentes indisponibles.

Remarque :
Les FIM* proposés sont tous documentés pour avoir des effets *directs ou indirects* sur les CATA, les THYR et les TUM2 (Inserm - expertise collective, 2005) ; (Kyung, Yul et al., 2010) ; (IARC, 2008).

BASES DE DONNEES ENVIRONNEMENTALES MOBILISEES

La position du Géographe de la Santé consiste à s'intéresser à tous les FE* *jugés pertinents* et donc à ne négliger aucune piste (Salem, 1995).
L'environnement géographique est multidimensionnel. La modélisation géographique de PM* et de FE* par des i.st.m* et des i.st.e* consistants nécessite l'utilisation de méthodes de modélisation robustes mais aussi de données environnementales fiables – si tant est qu'elles soient accessibles.
Les BD environnementales, leurs caractéristiques utilisées ainsi que la granularité* des variables qu'elles contiennent seront déclinées par composantes environnementale : SAN, SOCIO.ECO et PHY.CHIM

Proposition :
L'intérêt sera porté uniquement sur les bases de données recouvrant l'intégralité du territoire français métropolitain et susceptibles de modéliser les FE* pertinents et Curieux* qu'il convient d'intégrer.

FACTEURS ENVIRONNEMENTAUX SANITAIRES ET BASES DE DONNEES

Les FE-SAN* représentent la dimension spatiale de l'accès aux soins et sont largement documentés en géographie de la santé et en épidémiologie spatiale pour leur effets directs et indirects sur l'incidence ou la gravité des PM* étudiés (Penchansky et Thomas, 1981). Ces deux disciplines sont passées maîtres en matière de quantification de *l'accessibilité aux soins territoriale* (Chaix, Merlo et al., 2005). *Les densités médicales* et *les distances d'accès aux soins* sont les principaux indicateurs spatiaux utilisés (Picheral, 2001).
Mais les densités médicales classiques ne sont pas représentatives de la réalité géographique (Peguy, 1996). Elles n'intègrent pas la *dimension spatiale* et les notions d'offre et de demande – de soins de santé – sont biaisées par les frontières administratives en supposant que tous les *items sanitaires*, i.e.* : praticiens libéraux, services hospitaliers et EML** ont une accessibilité équivalente à l'intérieur d'une même zone (Talen et Anselin, 1998).

La conclusion de l'état de l'art décliné précédemment suppose que, l'idéal serait un indicateur spatial qui embrasserait simultanément l'offre de soins, la demande, et le temps d'accès aux items sanitaires* - dans logique territoriale et en éludant le biais que constituent les limites administratives des unités spatiales.

A ce jour, il existe un indicateur spatial capable d'une telle prouesse. Il est cependant controversé sur certains points inhérents à son estimation. Par conséquent, son utilisation est restreinte. Peu de données sanitaires en France sont publiques - les seules données disponibles sont celles mises à disposition par la Direction de la Recherche et des Etudes, de l'Evaluation et des Statistiques (DREES).

BASE DE DONNEE DREES : CARACTERISTIQUES & GRANULARITE

En France, les indicateurs spatiaux *d'accès aux soins* sont estimés par les Agences Régionales de Santé (ARS). En dehors des densités médicales classiques ces données ne sont pas publiques. Or ces indicateurs ne sont pas robustes sur le plan spatio-temporel. Récemment des KIT_AS (Accès aux Soins) ont été développés pour estimer les *distances temporelles par la route d'accès aux items sanitaires** et, plus récemment encore*, des Accès Potentiels Localisés (APL).*

Ces indicateurs sont beaucoup plus consistants d'un point de vue géographique. Ils sont estimables à l'échelle des cantons et des communes. Cependant, ils ne sont pas forcément mobilisables et quand ils le sont, ils doivent être collectés dans une logique régionale, et les délais d'obtention se comptent en années… Cette barrière d'accès aux données sanitaires s'explique par le fait que les KIT_AP sont

développés dans une logique mercantile et les indicateurs spatiaux en question sont destinés à être vendus aux collectivités territoriales, aux établissements de santé et aux professionnels de santé libéraux (ARS, 2012).

Les ARS dépendent de la Direction de la Recherche et des Etudes, de l'Evaluation et des Statistiques (DREES). La mission de la DREES est de concevoir, collecter, et diffuser des statistiques. Elle est aussi chargée de veiller à la cohérence de ces statistiques contenues dans les systèmes d'informations étatiques, pour les différentes thématiques de santé publique. La DREES est chargée aussi d'élaborer de concert avec les ARS des indicateurs spatiaux caractérisant les états de santé et les tissus sanitaires* territoriaux.

L'Institut de recherche et de documentation en économie de la santé (Irdes) est l'organisme en charge de la mise au point des méthodes d'estimation des indicateurs d'accès géographiques aux items sanitaires* (Irdes, 2012). L'Institut National de la Statistique et des Etudes Economiques (INSEE) contribue à ce travail en tant que fournisseur de données, puis assure la reproductibilité des méthodes – une fois validées par la DREES et les ARS - par la création des KIT_AP qui sont ensuite vendus aux ARS (INSEE, 2012c).

Dans le cadre d'une démarche scientifique *d'un désir économique* à peine caché, la DREES met à disposition du public, de façon très restreinte, ces indicateurs spatiaux ainsi que les documents scientifiques, les fiches techniques et les analyses géographiques qui les décrivent et les spécifient.

Actuellement les seuls indicateurs spatiaux disponibles - sur l'intégralité du territoire français métropolitain et suffisamment robustes d'un point de vue géographique – pour estimer *l'accès aux soins* en fonction de la qualité des tissus sanitaires* géographiques sont les Distances Temps pour Accéder (DTA) par la route aux items sanitaires* en 2007 (DREES, 2012) - depuis le 5 juin 2013, les indicateurs APL sont publiques pour deux professions de santé libérales en 2010 (DREES, 2013).

Les i.st.e* proposés sont voués à la modélisation géographique des FE-SAN* (chapitre.3), utilisent des données sanitaires – les seules à la fois disponibles et robustes – dont les spécificités sont décrites sommairement ci-après

VARIABLES DRESS EXISTANTES ET DISPONIBLES

Les deux variables géographiques mobilisables afin d'estimer l'accès aux soins en France sont les DTA d'accès aux items sanitaires* et les indicateurs APL. Dans cette sous-section il est question de décrire sommairement leurs caractéristiques.

Les Distances Temps d'Accès (DTA) :
Les *Distances Temps d'Accès (DTA)* ont été proposées en 2007 à cause de l'incapacité des *densités médicales classiques* à représenter l'accès géographique aux soins, et dans la lignée des *Distances* d'Accès Euclidiennes (DAE) proposées en 1990. Les DAE estiment les distances euclidiennes moyennes qui séparent les individus des items sanitaires* territoriaux. La mise en évidence de l'inconsistance* spatiale des DAE, i.e. *à vol d'oiseau*, et du caractère intemporel de cet indicateur spatial a conduit les Géographes de la Santé à proposer les DTA – qui s'expriment en minutes.

Caractéristiques des DTA :
La DTA postule que l'accès par la route est plus pertinent que *la distance à vol d'oiseau* pour estimer l'accessibilité aux items sanitaires* – puisque les déplacements géographiques sont plus sensibles au différentiel : distance temps qu'au différentiel : de distance métrique (McGuirk et Porell, 1984)
La méthode de calcul des DTA estime une distance kilométrique routière moyenne qui sépare les populations des items sanitaires* les plus proches. Cette opération est effectuée à partir du logiciel

Chronomap, des réseaux routiers de la BD TeleAtlas, et des statistiques routières Navted. La distance temporelle routière est celle qui sépare la mairie de chaque commune de celle où se trouve l'item sanitaire d'intérêt le plus proche. Par hypothèse le patient se déplace en voiture et emprunte l'itinéraire le plus court. Par définition, cette distance est nulle lorsque la commune du patient contient l'item en question.
Afin de tenir compte indirectement du poids de la demande, les temps d'accès aux items sanitaires* sont pondérés par la population située dans chaque bassin de vie - à partir de données INSEE.
La présence des professionnels de santé libéraux dans les communes est évaluée par le biais des données du Système national d'information inter-régimes (Sniiram) et de la Caisse nationale d'assurance maladie des travailleurs salariés (Cnamts) ; Spécialités hospitalières par croisement des bases Statistiques Annuelles des Etablissements de santé (SAE) et des données de consommation de soins du Programme de Médicalisation de Systèmes d'Information (PMSI). Ensuite, la localisation spatiale de l'établissement de santé est effectuée grâce au répertoire Fichier national des établissements de santé et sociaux (Finess). La stratégie utilisée pour les EML* : i.e. les Scanner*s (SCAN), les Images par Résonnance Magnétique (IRM), les caméras à scintillation (CAM) et les Tomographes à Expulsion de Postions (TPE) est analogue à celle décrite précédemment, i.e. en croisant les données géographiques de : Finess, SAE et celles de la Cnamts – uniquement (Coldefy, Com-Ruelle et al., 2011).

Remarques :
Les DTA mettent en exergue des disparités géographiques d'accès aux soins pour tous les services hospitaliers et les EML*. A l'échelle régionale elles sont d'autant plus fortes et des *déserts médicaux* se dessinent.

En revanche, ce n'est pas le cas pour les praticiens libéraux. En effet, la DTA est portée à zéro lorsqu'une spécialité libérale est repérée dans une unité géographique. Mais, en 2007 presque toutes les communes comptaient au moins un praticien de santé, pour chaque spatialité considérée. De fait, la DTA est inconsistante pour cet item sanitaire.

De plus, les DTA ne tiennent pas compte non plus de la quantité de personnes âgées qui influence le délai d'obtention d'un rendez-vous, des horaires d'ouverture des cabinets, ni des déplacements infra-communaux. (Coldefy, Lucas-Gabrielli et al., 2011)

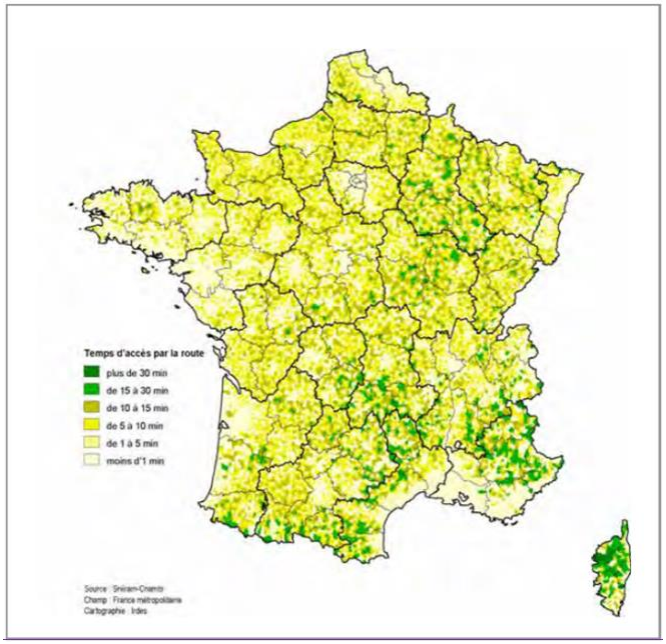


Figure 4 : Distances Temporelles d'Accès* à un médecin généraliste au 1er janvier 2007.
Source : DREES

Toutes ces lacunes font que les DTA sont inconsistantes sur le plan géographique pour caractériser l'accès à des praticiens de santé libéraux. La principale faille des DTA est de supposer que tous les items sanitaires*, à l'intérieur d'une même zone géographique, présentent une accessibilité équivalente (Talen et Anselin, 1998).
Ce qui a conduit la DREES, l'INSEE et l'Irdes à proposer l'indicateur APL

L'Accès Potentiel Limité (APL) :
Il s'agit d'un indicateur, de type densité, qui s'exprime en nombre de médecins pour mille habitants. L'APL est grevé d'une composante : spatio-temporelle puisqu'il prend en compte les distances temps d'accès par la route aux items sanitaires* ; socio-spatiale puisqu'il tient compte de la demande par le biais des densités de population communales ; médico-spatiale car l'offre est évaluée à partir des densités médicales localisées et de l'attractivité potentielle des médecins estimée sur des secteurs flottants et pondérée par la surface *géographique des bassins de vie*. L'APL se calcule à partir de l'Equivalent Temps Plein (ETP) de travail des praticiens de santé –qui permet d'évaluer indirectement le temps d'obtention d'un rendez-vous – donc de l'offre territoriale réelle. L'ETP est estimé à partir des données de la Cnamts. La demande est supputée en explorant la pyramide des âges des populations communales INSEE, Enfin les distances temporelles sont les DTA telles qu'elles ont été décrites précédemment.
En somme, l'APL est un indicateur permettant de considérer simultanément l'offre, la demande et la Distance Temporelle d'Accès* aux praticiens de santé libéraux. Il s'appuie sur la méthode : *Two-step floating catchment area* (Barlet, Lucas-Gabrielli et al., 2012).

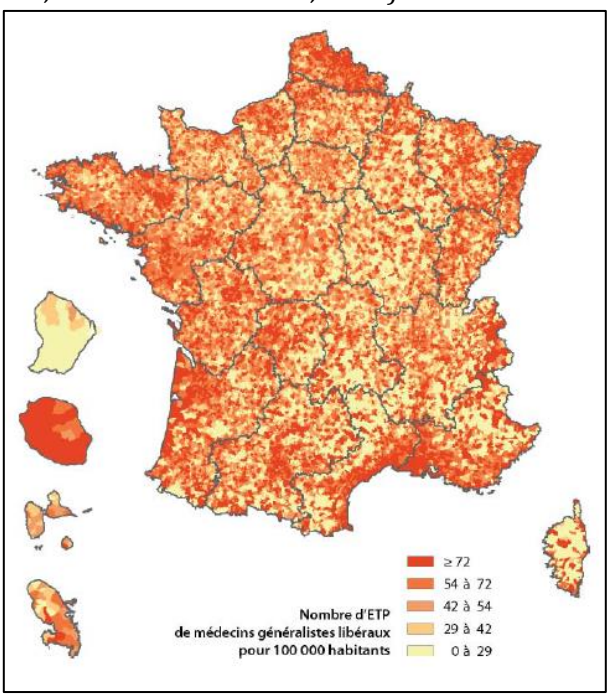


Figure 5 : Indicateurs APL 2010 pour les médecins généralistes libéraux en France métropolitaine
Source : DREES

Remarque :
A l'heure actuelle APL est l'indicateur spatial le plus consistant d'un point de vue spatiotemporel. Cependant, son système de pondération permettant de modéliser l'attractivité potentielle de l'offre est controversé. En effet, ce dernier est estimé sur dires d'expert, dans une logique géographique, et il ne peut pas être calibré. Comme le système de pondération influence fortement les valeurs prises par l'APL, cela pose problème (Harrouin, Aligon et al., 2012).

L'indicateur APL 2010 est disponible à l'échelle des communes françaises métropolitaines et uniquement pour les médecins généralistes (GENE) et les ophtalmologues (OPHT) (DREES, 2013).

VARIABLES UTILISEES POUR LA MODELISATION

FE-SAN* *pertinents* & granularité des variables disponibles

Les variables utilisées pour construire les i.st. $x_{U_k}^{l:SAN}$ qui permettront de modéliser la géographie des FE-SAN* sont, parmi tous les indicateurs spatiaux DTA 2007 et APL 2010 disponibles, celles dont l'item sanitaire a un rapport avec le PM* d'intérêt – séquelle.

Echelle géographique :
Les DTA : $x_{(U_u)}^{DTA}$ et les $x_{(U_u)}^{APL}$ sont disponibles sur l'intégralité de la France métropolitaine à l'échelle des communes, à l'exception des villes de Paris, Marseille et Lyon pour lesquelles elles sont déclinées par arrondissement.

Echelle temporelle :
les $x_{(U_u),t}^{DTA}$ et les $x_{(U_u),t}^{APL}$ mises à disposition sur le site de la DREES sont déclinées respectivement pour l'année : {t = 2007} pour les DTA et {t = 2010} pour les APL.

FIM	Variable	Description	Nature	Unité	Commentaire
géographie	*CIM*	*CIM*			
DTA AUX ITEMS SANITAIRES	$x_{(U_u),t}^{DTA\_MED}$	Temps d'accès par la route à un praticien libéral (*)	Quantitative Discrète	minutes	(*):Médecins : Généralistes ; Spécialistes : ORL, ophtalmologistes (OPHT), radiologistes (RADIO), pédiatres (PEDIA).
	$x_{(U_u),t}^{DTA\_SH.}$	Temps d'accès par la route à un Service Hospitalier Spécialisé	Quantitative Discrète	minutes	Hématologie (HEMA) ; ORL ; Ophtalmologie (OPHT); Endocrinologie (ENDO), Neurologie médicale et neurochirurgie (NEUR)
	$x_{(U_u),t}^{DTA\_EML*}$	Temps d'accès par la route à un EML*	Quantitative Discrète	minutes	IRM, Scanner* (SCAN), caméra à scintillation (CAME), TEP
APL AUX MEDECINS LIBERAUX	$x_{(U_u),t}^{APL\_MED}$	Accès Potentiel Localisé* à un praticien de santé libéral	Quantitative Continue	Nombre de médecins pour 1000.habs	(*):Médecins : Généralistes ; Spécialistes : Ophtalmologistes (OPHT),

Tableau 5 : Variables sanitaires mobilisées.

Spécification :
Une demande a été faite auprès du service SIG inter-ARS en vue d'obtenir les DTA, dans les communes de 1ère espèce, pour les temporalités : 1997 et 2000.

FE-SAN* *curieux* de test* & granularité* des variables disponibles

Compte tenu du fait qu'en dehors des $x_{(U_u)}^{DTA}$ et des $x_{(U_u)}^{APL}$ il n'existe pas d'autres indicateurs spatiaux, à l'exception des densités médicales classiques qui sont inconsistantes et redondantes avec ces derniers, et pour ne pas perdre de puissance statistique, aucune autre variable supplétive ne sera intégrée.
Les FE-SAN* *Curieux* de test* seront simplement les i.st.e* sanitaires : $x_{(U_k)}^{DTA}$ - construits à partir des données mobilisées $x_{(U_k)}^{DTA}$ mais dont l'interaction avec les PM* d'intérêt est inconséquente. Par exemple les DTA à un OPHT seront intégrées pour la séquelle THYR, ou la DTA à un ORL pour la séquelle CATA.

FE-SAN* *non modélisables* - variables indisponibles

Une pléthore d'indicateurs géographiques d'accès aux soins est proposée dans la littérature. Or les données sanitaires et socio-économiques permettent rarement de les calculer à l'échelle des communes et pour l'intégralité de la France (Powell, 1995). Et dans le cas contraire leur robustesse n'est pas validée par la DREES.

Par conséquent à ce jour, les seuls indicateurs supplétifs pertinents qu'il conviendrait d'intégrer sont les APL pour les autres spécialités médicales. Aussi la méthode de calcul des APL peut être facilement étendue aux plateaux techniques des établissements de santé, i.e. aux services hospitaliers et aux EML*. Par conséquent, lorsque ces derniers seront mis à disposition du public il conviendra de les substituer aux DTA.

FACTEURS ENVIRONNEMENTAUX SOCIO-ECONOMIQUES ET BASES DE DONNEES

Les FE-SOCIO.ECO* caractérisent la dimension a-*spatiale* de *l'accès aux soins* et des *conjonctures socio-économiques territoriales. Le recours aux soins* est un concept interdisciplinaire où s'enchevêtrent aussi des composantes sociales, économiques et géographiques. (Litva et Eyles, 1995).
L'*effet de contexte* a des répercussions *collectives et individuelles* (Chaix, Merlo et al., 2005). Lorsque les FE-SOCIO.ECO* sont *défavorables* ils induisent des *prédispositions géographiques aux phénomènes morbides.* La variabilité géographique des FE-SOCIO.ECO* sera modélisée par des i.st.e* qui ont pour objet de caractériser les conjonctures territoriales (Haddad, 1992).

Les FE-SOCIO.ECO* retenus dans *l'état de l'art* pour leurs *effets collectifs indirects* sur les états de santé *individuels* étudiés sont : *Le recours aux soins*, *les croyances collectives, la culture, l'économie, la défaveur sociale, les habitudes alimentaires, l'insécurité*, *les conduites à risques*, *le stress*, *l'exposition spatiale à des substances toxiques* liées à la *spécialisation des espaces ou aux activités professionnelles dominantes.*

La modélisation des FE-SOCIO.ECO* par des i.st.e* robustes s'effectue à partir d'indicateurs communautaires et elle est conditionnée par leur granularité*. Les BD retenues sont celles qui contiennent des variables *fiables* et disponibles sur l'intégralité du territoire français métropolitain.

Les BD utilisées sont celles de l'Institut National de la Statistique et des Etudes Economiques (INSEE) ; L'Observatoire National de la Délinquance et des Réponses Pénales (ONDRP).

BASES DE DONNEES UTILISEES : CARACTERISTIQUES & GRANULARITE

Les BD subséquentes contiennent des données communautaires qui caractérisent les conjonctures sociales et économiques des territoires français.

INSEE

Caractéristiques générales :
L'INSEE est sous l'autorité du Ministère de l'Economie, des Finances et de l'Industrie. Sa mission consiste à construire des bases de données statistiques représentatives du panorama démographique, social et économique en France. Cet organisme étatique collecte, organise, traite et diffuse les données du recensement national, et mène des enquêtes ciblées sur certaines catégories d'entreprises ou de ménages pour construire des indices conjoncturels.
L'INSEE est aussi en charge de gérer les Codes INSEE qui confèrent une dimension spatiale à ses indicateurs et aussi de les intégrer dans des SIG (IGN, 2004). Les données INSEE sont publiques et téléchargeables sur le site internet de l'institut. Leurs caractéristiques granulométriques dépendent de la qualité des sources et des restrictions juridiques associées (INSEE, 2012c).

Granularité des variables :
Il existe une pléthore de données socio-économiques disponibles sur le site de l'INSEE. Les principaux fournisseurs sont sous la tutelle : de la Direction Générale des Finances Publiques (DGFiP) ; du ministère de l'Agriculture, de l'agroalimentaire et de la forêt (Agreste) ; du ministère des Affaires sociales et de la Santé ; du Ministère de l'écologie, du développement durable et de l'énergie ; et du Ministère du Travail, de l'Emploi, de la Formation Professionnelle et du Dialogue Social.
Les variables INSEE sont diffusées à différentes échelles géographiques : nationale, régionale, départementale, communale et à celle des Ilots Regroupés pour l'Information Statistique (IRIS). Les IRIS sont définis comme des ensembles *de petits quartiers* contigus et constituent le niveau d'information légal le plus fin. Le système unitaire dépend de la nature des variables INSEE considérées : $x_{(U_u),t}^{l:INSEE}$. Les temporalités des variables portent sur certaines années spécifiques et sont espacées par des *sauts* (INSEE, 2012c).

La loi du 6 janvier 1978 *relative à l'informatique, aux fichiers et aux libertés* confère à la Commission Nationale de l'Informatique et des Libertés (CNIL) le pouvoir de restreindre la diffusion de certaines variables dans les unités géographiques où le nombre d'individus recensés est *trop petit,* ce qui pourrait porter atteinte à leur vie privée. Dans ces *zones* les données sont lacunaires. Les IRIS sont particulièrement impactés. Le *secret statistique* conditionne aussi les temporalités disponibles. Par exemple, *les revenus fiscaux médians* sont disponibles uniquement pour les cinq dernières années : $t = \{2008; 2009; 2010; 2011; 2012\}$; Par contre, *les niveaux de diplôme* sont disponibles pour les six derniers recensements nationaux $t = \{1968; 1975; 1982; 1990; 1999; 2008\}$ (INSEE, 2013) - *les autres sont systématiquement effacés* (Leduc, 2011).

Les indicateurs INSEE sont disponibles sur le site Internet de l'institut.

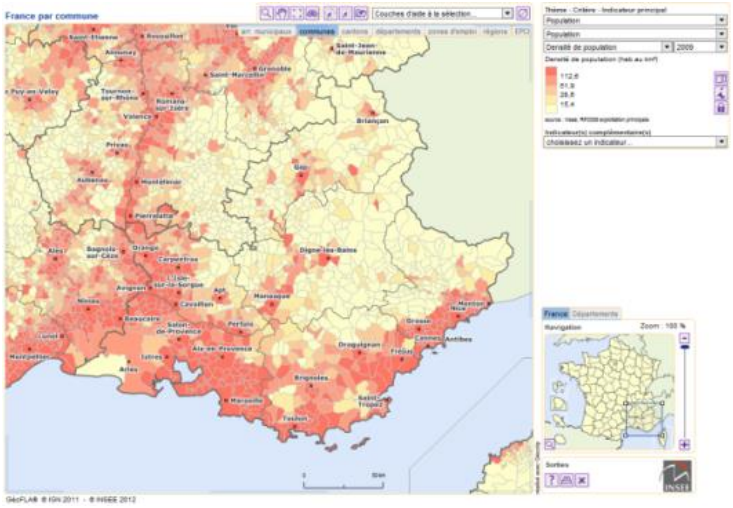


Figure 6 : Densité des populations communales dans les communes de PACA et aux alentours
Source : INSEE

ONRDP

Caractéristiques générales :

L'ONDRP est un département de l'Institut National des Hautes Etudes de la Sécurité et de la Justice. Sa mission est de recueillir et de diffuser des données statistiques relatives à la délinquance et la criminalité territoriale. Cet observatoire gère et assure la maintenance d'une une base de données géographiques publiques : *cartocrime.net*. (INHETSJ, 2012).

Granularité* des variables :

La BD ONDRP contient quatre *indicateurs spatiaux composites* construits par agrégation *d'index thématiques d'infractions*, i.e. *crimes et délits*.

Ces données géographiques sont publiques et disponibles à l'échelle des départements : D_u et des régions. Les index thématiques d'infractions sont mensuels ou annuels et recouvrent une période qui s'étale sur les temporalités : $t = \{1996, \ldots, 2011\}$. Les indicateurs spatiaux composites sont disponibles sous forme de cumuls ou de densités, i.e. de cumuls rapportés à la population *in-situ*.

Les variables ONDRP : $x^{l:ONDRP}_{(D_u),t}$ caractérisent quatre types d'infractions : *atteintes aux biens* (index : vols à main armée, cambriolages, incendies volontaires, attentats…), *atteintes volontaires à l'intégrité physique* (index : viols, harcèlements, homicides, violences, …), *escroqueries et infractions économiques et financières* (Faux en écritures publiques et authentiques, fausse monnaie, contrefaçons…), *infractions relevées par l'action des services* (recels, proxénétisme, trafics et revente de stupéfiants…) (ONDRP, 2012)

Les indicateurs sont disponibles depuis le portail cartocrime.net.

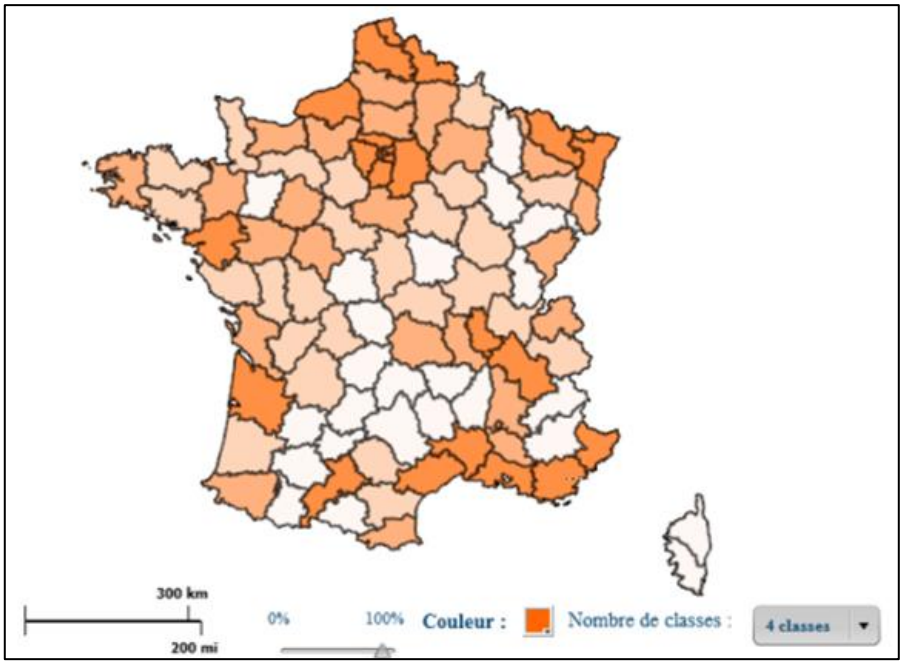


Figure 7 : Cumul des index d'atteintes volontaires à l'intégrité physique commises en 2010. Source : ONDRP

Les trois principales limites des indicateurs spatiaux sont les suivantes :
(i) Les index des infractions relevées par l'action des services sont redondants et aussi moins exhaustifs que les autres ;
(ii) Les cumuls sont biaisés car ils agrègent des index qui dénombrent toutes les infractions enregistrées sur le territoire et non pas celles perpétrées par la population résidente ;
(iii) Les ratios sont doublement biaisés car il existe un delta entre le nombre de personnes résidentes recensées et celles réellement présentes sur le territoire (ONDRP, 2011).

VARIABLES UTILISEES POUR LA MODELISATION GEOGRAPHIQUE

Remarques liminaires :
Les données LEA utilisées sont déclinées une période allant de 1980 à 2010. Or la stratégie d'agrégation *verticale* des données communautaires se fonde sur le concept de *variabilité temporelle apparente* (Peguy, 1996).
De fait, une temporalité supplémentaire est ajoutée dans la mesure du possible.

FE-SOCIO.ECO* *pertinents*

Caractéristiques granulométriques communes :
L'échelle géographique : les communes U_u ou les arrondissements Ar_u

FE-SOCIO.ECO	*Estimateur*	Variables	Description	Nature	Temporalité	Base INSEE	Spécifications
géographie	*Etat de l'art*	*INSEE*	*Variables*	*& Unité*	*Disponibles*	*Nom*	*& variables auxiliaires*
COMPORTEMENTS VIS-A-VIS DU RECOURS ET DE L'OFFRE DE SOIN TERRITORIALE	Niveaux de vie	$x_{(\cdot),t}^{FFI}$	Foyers Fiscaux Imposables	Quantitative discrète (U)	1999, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2009, 2010	Indicateurs de distribution par Ménage	VA : $x_{(\cdot),t}^{FoyFisTot}$ foyers fiscaux totaux
	Niveaux de revenus	$x_{(\cdot),t}^{RevMed.m}$	Revenu fiscal médian par ménages	Quantitative discrète (peut-être négatif) (€)	2001, 2002, 2003, 2004, 2005, 2006, 2007, 2009, 2010	Indicateurs de distribution par Ménage	Lacune : les unités géographiques comptant moins de 50 ménages (recensement de 1999)
		$x_{(\cdot),t}^{RevMed.p}$	Revenu fiscal médian par personnes				
	Répartition des richesses	$x_{(\cdot),t}^{Gini.m}$	Indices de Gini estimé pour les ménages	Quantitative continue SU à valeur dans $[\![0 ; 1]\!]$	1999, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010	Revenus fiscaux des ménages	Lacune : les unités géographiques comptant moins de 2000 habitants (recensement de 1999)
		$x_{(\cdot),t}^{Gini.p}$	Indices de Gini estimé pour les personnes				
	Niveaux de précarité	$x_{(\cdot),t}^{CHOM}$	chômeurs de 15 à 64 ans	Quantitative discrète (U)	1999, 2006, 2009, 2010	Chiffres clés Emploi; Population	VA : $x_{(\cdot),t}^{POP.ACTIVE}$ population active âgée de 15 à 64 ans
	Catégories socioprofessionnelles	$x_{(\cdot),t}^{OUV}$	Ouvriers de 15 à 64 ans	Quantitative discrète (U)	1999, 2006, 2009, 2010	Chiffres clés Emploi; Population	VA $x_{(\cdot),t}^{POP>15}$; population totale âgée de 15 à 64 ans
EFFORTS POLITIQUES EN MATIERE DE DURABILITE ET DE LEURS REPERCUSSIONS SUR LES ATTRAITS : SOCIAUX, ECONOMIQUES ET SANITAIRES DES* TERRITOIRES	Niveaux de mortalité	$x_{(\cdot),t}^{DECES}$	Personnes décédées	Quantitative discrète (U)	1975, 1982, 1990, 1999, 2009, 2010	Evolution et structures de population	VA: $x_{(\cdot),t}^{POP}$ Population sans double compte
	Accroissements démographiques Globaux & Naturels	$x_{(\cdot),t}^{NAISS}$	Nombre de naissances	Quantitative discrète Nouveau-nés (U)	1975, 1982, 1990, 1999, 2009, 2010	Evolution et structures de population	VA : $x_{(\cdot),t}^{DECES}$; $x_{(\cdot),t}^{POP}$
	Niveaux de diplôme	$x_{(\cdot),t}^{Niv.Dip}$	Individus en fonction du dernier diplôme obtenu	Quantitative discrète Individus de plus de 15 ans	1968, 1975, 1982, 1990, 1999, 2009, 2010	Tableau rétrospectif communal population par sexe, âge et diplôme au lieu de résidence	VA: $x_{(\cdot),t}^{POP>15}$
EXPOSITIONS POTENTIELLES A DES* SUBSTANCES TOXIQUES	Proportion d'Emplois dans des secteurs d'activité à risque	$x_{(\cdot),t}^{Emp.A.Ris}$	Cumul des emplois dans un secteur d'activité à risque	Quantitative discrète d'emploi selon les secteur d'activité	1968, 1975, 1982, 1990, 1999, 2009, 2010	Analyse fonctionnelle des emplois	VA : $x_{(U_{n}),t}^{Emp.Tot}$ Nombre total d'emplois tous secteurs d'activité confondus
EXPOSITIONS POTENTIELLES AUX PESTICIDES* LIEES A LA SPECIALISATION ECONOMIQUE DES* TERRITOIRES	Taux de SAU (Surface Agricole Utilisée)	$x_{(\cdot),t}^{SAU}$	Quantité de Surface Agricole Utilisée	Quantitative discrète Hectares (ha)	1968, 2000	Exploitations agricoles	La SAU est un estimateur statistique européen ; VA : $x_{(\cdot),t}^{SG}$
	Niveaux d'intensité de l'activité agricole	$x_{(\cdot),t}^{UTA}$	Unité de Temps Annuel	Quantitative continue (U)	1968, 2000	Exploitations agricoles	Une UTA équivaut à un emploi à temps complet, dans une exploitation agricole, pendant un an
		$x_{(\cdot),t}^{Exp.AGRI}$	Nombre d'exploitations agricoles	Quantitative discrète (U)	1968, 2000	Exploitations agricoles	La définition utilisée d'une exploitation agricole est celle d'Agrest
PREDISPOSITION COMPOSITE AUX PHENOMENES MORBIDES	Niveaux de défaveur sociale	Transformation topologique	Indice de défaveur sociale FDep.XX	Quantitative continue (SU)	2001 ; 1999 2009 ; 2010	Diverses	Variables utilisées : $x_{(\cdot),t}^{DECES}$; $x_{(\cdot),t}^{POP}$; $x_{(\cdot),t}^{OUV}$; $x_{(\cdot),t}^{RevMed.m}$; $x_{(\cdot),t}^{Niv.Dip}$; $x_{(\cdot),t}^{POP>15}$

Tableau 6 : Variables socio-économiques mobilisées.

Remarque :
Les stratégies de fusion, la capacité des variables à limiter la redondance (Saporta, 2006) et maximiser *l'effet information** (Marcotte, 2008) ainsi des compléments bibliographiques sur l'adéquation à des FE* *pertinents* seront spécifiés par la suite (chapitre.3).

FE-SOCIO.ECO* curieux* de test :
Les caractéristiques granulométriques communes :
Echelle géographique : les départements De_u ; *Période temporelle* : $t = \{1996, \dots, 2011\}$; Nature : Quantitative discrète ; Unité: cumul des index thématiques d'infraction – puisque les ratios sont biaisés.

FE-SOCIO.ECO	*Estimateur*	Variables	Description	Spécifications
géographie	*Etat de l'art*	*ONDRP*	*Variables*	*Variables utilisées*
NIVEAUX DE STRESS POTENTIELLEMENT PERÇUS INDUITS PAR L'INSECURITE TERRITORIALE CONTEXTUELLE	Niveaux d'atteintes aux biens	$x^{\text{att.BIENS}}_{(\text{D}_u),t}$	*Cumul de l'ensemble des index d'infractions thématiques*	*Cumuls annuels*
	Niveaux d'atteintes volontaires à l'intégrité physique	$x^{\text{att.PHY}}_{(\text{D}_u),t}$	*Cumul de l'ensemble des index d'infractions thématiques*	*Cumuls annuels*

Tableau 7 : Variables socio-économiques mobilisées.

Remarques :
(i) D'un point de vue philosophique l'insécurité territoriale caractérise aussi une carence politique en matière de lutte contre : La pauvreté, la délinquance, l'exclusion, l'équité d'accès à l'éducation et la culture (Godin, 2007).
(ii) De plus, la distance a-spatiale induite par l'imprécision d'échelle et l'éloignement du phénomène avec les PM* d'intérêt –séquelles font que les variables ONDRP sont topiques pour supputer des i.st.e* Curieux* de test.
(iii) Comme les *infractions relevées par l'action des services* sont inconsistantes et que les index : Escroqueries et Infractions économiques ne sont pas représentatifs du *stress lié à l'insécurité* territoriale – ces variables ont été écartées.

FE-SOCIO.ECO* non modélisables :

Les trois principaux motifs limitant la modélisation des FE-SOCIO.ECO* pertinents sont liés à des contraintes d'accès aux données :

FE-SOCIO.ECO	*Estimateur proposé*	BASE	Motifs	Spécification
géographie	*Etat de l'art*	*proposée*	*Limitant la modélisation*	*Variables proposées*
EXPOSITIONS POTENTIELLES A DES* CONDUITES A RISQUE	*Comportements alimentaires liés aux habitudes et aux spécialités culinaires locales*	*Nielsen*	Variables mesurées mais non disponibles	Les ratios Nielsen agrégeant des denrées alimentaires spécifiques (*)
EXPOSITIONS POTENTIELLES A DES* SUBSTANCES TOXIQUES	*Quantités achetées de produits commerciaux contenant des substances toxiques*	-	Variables non mesurées	Les ratios agrégeant des produits dangereux utilisés par les ménages
POLITIQUES DURABLES ET LEURS EFFETS SUR LA QUALITE SANITAIRE ET SOCIO-ECONOMIQUE DES* TERRITOIRES	*Niveaux géographiques d'espérance de vie* ou des *proportions d'individus visant en-dessous du seuil de pauvreté*	*INSEE*	Variables mesurées, disponibles mais non fiables	*L'Espérance de vie ; Les taux de personnes vivant en dessous du seuil de pauvreté - à des échelles plus fines (**)*

Tableau 8 : Variables socio-économiques indisponibles.

Compléments informationnels :

(*) Nielsen est une société internationale dans le secteur du marketing de la consommation. En 2011 elle a effectué une collecte de données au niveau des caisses dans les grandes surfaces. L'étude statistique établie a permis de mettre en évidence que les quantités et la nature de denrées alimentaires achetées varient avec des gradients géographiques importants. Et qu'elles sont liées aux traditions et aux habitudes alimentaires locales (Nielsen, 2012).

Les statistiques Nielsen sont des ratios. Les quantités de produits achetés sont rapportées à la moyenne nationale des ventes. L'échelle géographique est celle des départements. Dans celui de Paris, les ventes de lait frais sont 3 fois supérieures à la moyenne. Dans les Alpes de Haute Provence les ventes de spécialités locales : nougat, miel, fromage de chèvre… engendrent des écarts relatifs de 467%. (Deluzarche, 2012).

Cependant, les quantités achetées ne correspondent pas exactement aux quantités consommées *in-situ*. Et bien qu'une bonne partie des ventes soit attribuable aux populations locales, une autre est imputable à l'activité touristique. Les statistiques Nielsen auraient pu permettre d'estimer les disparités géographiques des ventes d'aliments : alcoolisés, gras et salés, gras sucrés ou protéinés.

(*) Dans le même esprit que les statistiques Nielsen, il serait intéressant de mesurer les niveaux géographiques des ventes, ou des consommations, de produits commerciaux comme : le tabac ; les cosmétiques, les peintures, les hydrocarbures, certains solvants et les pesticides. Ces derniers sont tous cités dans l'état de l'art pour avoir des effets potentiels sur les PM* d'intérêt – séquelles (IARC, 2008).

(**) Les variables citées sont disponibles uniquement à l'échelle régionale (INSEE, 2012c). Or, la granularité* d'échelle est inopérante pour capter les variabilités spatiales des effets du contexte communautaire sur les expositions individuelles à l'insalubrité territoriale ou vis-à-vis du recours aux soins (Chaix, Merlo et al., 2005).

FACTEURS ENVIRONNEMENTAUX PHYSICOCHIMIQUES ET BASES DE DONNEES

Les expositions géographiques *environnementales*, i.e. à des doses chroniques mais faibles de substances physicochimiques toxiques sont encore controversées (Chasles et Fervers, 2011).

Or ces substances sont présentes dans tous les *milieux environnementaux* et aussi de *contact* (Caudeville J., Boudet C. et al., 2012) . Elles sont diffuses, les sources de contamination sont protéiformes, et les doses d'expositions soumises à de forts gradients géographiques – pouvant être accidentellement très élevées- et conditionnées par les déplacements spatiotemporels des individus (Inserm - expertise collective, 2005).

De plus, les études cliniques manquent de recul et l'identification des substances *pathogènes génotoxiques* est complexe (Afsset, 2009a). Le rôle déterminant ou au moins contributif de l'environnement géographique sur les états de santé doit être abordé en termes *d'expositions combinées* (Leux et Guénel, 2010). Actuellement on assiste à un regain d'intérêt pour ce paradigme (Afsset, 2009a). Les multiples BD environnementales offrent des perspectives de modélisation géographique prometteuses (Zeitouni, 2006).

Les FE-PHY.CHIM* retenus dans *l'état de l'art suspectés d'avoir des effets environnementaux sur les PM* d'intérêt* sont *:* Le benzène ; L'oxyde d'éthylène ; Les HAP ; Les nitrates ; Les ETM ; Les dioxines (PCB, formaldéide) et les pesticides, pour ce qui est des substances chimiques. Les RNI notamment : les *hautes fréquences* induites par les lignes à haute tension ; *basse fréquence* liées à la téléphonie mobile mais aussi les rayonnements solaires dont les variabilités spatiales sont indirectement liées à la topographie ou la pluviométrie. Les expositions géographiques à la radioactivité environnementale naturelle ou artificielle sont particulièrement citées pour les PM* d'intérêt. Les expositions à des radionucléides purement artificiels liés à la présence d'INB* fait polémique. Les expositions naturelles au Radon sont plus connues. Les conditions météorologiques ont un impact important sur la qualité environnementale des milieux de vie et les variabilités des expositions géographiques. La modélisation des FE-PHY.CHIM* retenus est conditionnée par la robustesse des i.st.e* proposés.

La modélisation des FE-PHY.CHIM* par des i.st.e* robustes s'effectue à partir de mesures environnementales – pour ce qui est *des expositions géographiques potentielles* – et à partir de doses d'expositions à des substances toxiques ou encore d'indicateurs spatiaux de risque pour ce qui est des *expositions géographiques qualifiées d'intrinsèques*. La qualité de la modélisation des expositions environnementales potentielles ou intrinsèques à des substances physicochimiques est conditionnée par la granularité* des données environnementales utilisées. Les BD utilisées retenues à cet effet sont :

Celle de Météo-France ; Le Réseau National de Mesure de la radioactivité environnementale (RNM) ; L'atlas Radon ; Le répertoire des Installations Nucléaires de Base (INB) ; La Plateforme intégrée pour l'analyse des inégalités d'expositions environnementales (PLAINE) ; Et CORINE Land Cover (CLC).

BASES DE DONNEES UTILISEES : CARACTERISTIQUES & GRANULARITE

Les BD subséquentes contiennent les données environnementales à partir desquelles les i.st.e* proposés modélisent la variabilité géographique des expositions environnementales *potentielles ou intrinsèque*s à des substances physicochimiques toxiques (chapitre.3).

Compte tenu du nombre de BD existantes, seules seront décrites en détail celles qui ont été utilisées. Et de façon plus sommaire celles qui auraient pu permettre de modéliser des FE-PHY.CHIM* *pertinents* mais qui n'étaient pas accessibles. La première BD environnementale décrite est celle de Météo-France.

METO-FRANCE

Caractéristiques générales :

Météo-France est le service national de météorologie. Il s'agit d'un établissement scientifique et technique qui a pour mission d'informer les populations des variabilités climatiques géographiques et d'alerter les autorités civiles lorsque des aléas peuvent être anticipés. Météo-France est une vigie du climat qui dispose d'un réseau national de stations mesurant en continu des paramètres atmosphériques, pluviométriques et océanographiques (Météo-France, 2011)

Granularité* des variables :

Il existe une pléthore d'indicateurs climatiques disponibles. L'intérêt est porté sur les mesures météo $m_{(s_g,t)}^{l:\{Météo-France\}}$ effectuées par le service interrégional. La variable $m_{(s_g,t)}^{l:\{Météo-France\}}$ est une mesure du paramètre météo : l au temps : t et au niveau du site s_g. Elles sont disponibles sur l'intégralité du territoire français et en accès libre pour les laboratoires de recherche. Il s'agit de mesures géo-localisées, i.e. définies pour des sites géo-référencés : $s_g = (x.geo_g;\ y.geo_g)\ \forall\ g = \{1, \dots,\ n_s^l\}$. Pour chaque station des chroniques temporelles sont disponibles à différents *pas de temps* et sur une période allant de 1981 à 2011 (Météo-France, 2010). Le nombre de points de mesures: n_s^l disponibles dépend des caractéristiques techniques des stations et du paramètre météo mesuré. La pluviométrie par exemple est systématiquement mesurable. En revanche, le rayonnement global nécessite des instruments plus élaborés (Jacq, 2012).

Figure 8 : Portail interactif des données météorologiques disponibles.
Source : Météo-France

RESEAU NATIONAL DE MESURE DE LA RADIOACTIVITE ENVIRONNEMENTALE

Caractéristiques générales :

Le Réseau National de Mesure (RNM) est un portail consacré à la radioactivité environnementale. Il est sous l'égide de l'Agence de Sécurité Nucléaire (ASN), une autorité administrative de radioprotection, i.e. qui s'affaire à la prévention et à la surveillance des rayonnements ionisants pouvant porter atteinte aux personnes et à l'environnement. L'ASN est aussi en charge de la réglementation des Installations Nucléaires de Base (INB) (ASN, 2013).
L'ASN a délégué la gestion du portail interactif RNM* à l'Institut de Radioprotection et de Sûreté Nucléaire (IRSN) qui s'occupe de l'actualiser et de l'alimenter. L'IRSN est un établissement public et commercial expert en matière de surveillance radiologique de l'environnement et de prévention des risques nucléaires (IRSN, 2012b).

Les autres acteurs du RNM* sont nombreux. Le réseau est guidé par cinq directions ministérielles opérant dans le contrôle sanitaire ainsi que par l'Institut de Veille Sanitaire (InVS) et L'Agence française Nationale de Sécurité Sanitaire (ANSES) qui contribuent à la fois au pilotage et à l'alimentation en mesures de la radioactivité environnementale. La majorité des données provient du portail de la mesure de l'IRSN. Mais d'autres organismes possèdent des réseaux de mesure et y contribuent aussi, c'est le cas de : Electricité de France (EDF) ; AREVA ; Commissariat à l'Energie Atomique (CEA) ; La Marine nationale ; L'Agence Nationale pour la gestion des Déchets Radioactifs (ANDRA) ; Certaines collectivités territoriales ; Le Groupement des Scientifiques pour l'Information sur l'Energie Nucléaire (GSIEN) ainsi que l'Association pour le Contrôle de la Radioactivité dans l'Ouest (ACRO) (RNM, 2010).

Granularité* des variables RNM* :

Le RNM* est une plateforme interactive qui rassemble des chroniques temporelles localisées $m_{(s_g,t)}^{l:\{RNM\}}$ mesurant des niveaux de radioactivité dans les milieux environnementaux : eau, air, sol et biologique. Ces mesures sont effectuées dans des prélèvements gazeux, solides ou liquides et sont généralement exprimées en Bq/u. matière. Elles permettent d'évaluer l'activité volumique de radionucléides spécifiques.

Particularité, dans l'eau des niveaux globaux de l'activité volumique des particules α ou β sont disponibles. Dans l'air *des doses efficaces* liées à la dégradation de la globalité des éléments radioactifs par rayonnements X et γ et par émission de particules α et β, sont évaluées. Ces dernières sont exprimées en Nano.Sievert/heure et représentent les doses globales moyennes absorbées par les tissus organiques humains. Le réseau de stations recouvre l'intégralité du territoire français. Par contre la densité spatiale des stations s_g n'est pas uniforme. Elles sont regroupées dans les *zones géographiques sensibles*, i.e. à proximité des INB* ou dans celles où la radioactivité naturelle est plus forte. Le nombre de points de mesure n_s^l, leur précision, ainsi que le pas de temps des chroniques varient en fonction de la nature des radionucléides mesurés, du milieu et des caractéristiques techniques des stations. La période temporelle la plus longue disponible s'étend de 2008 à 2013 (ANS et IRSN, 2013).

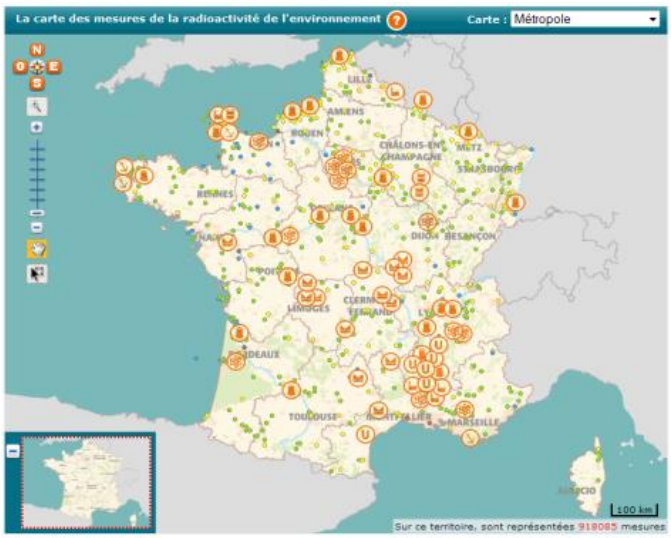


Figure 9 : Portail interactif des mesures publiques de la radioactivité environnementale. Source : RNM

Remarque :

(i) Aucune extraction massive n'est possible à partir du portail RNM* et toutes les chroniques temporelles utilisées ont été extraites manuellement – point par point (ANS, IRSN, 2013).

(ii) Le *portail de la mesure de la radioactivité environnementale de l'IRSN* propose des mesures temporelles, parfois plus précises, avec des *pas de temps plus courts* et surtout, sur des temporalités plus anciennes – pouvant dater de 1960 (IRSN, 2012a). Ces données sont plus consistantes sur le plan temporel au regard de la période couverte par LEA. Or pour l'heure, il n'est pas possible de les extraire, même pas manuellement. Cependant une plateforme de transfert est en cours de construction. Mais actuellement les variables IRSN ne sont pas accessibles pour des raisons techniques et juridiques (Pierrard, 2013).

Le portail interactif RNM, comme celui de L'IRSN, permet de connaitre aussi la localisation spatiale des INB. Mais les données attributaires associées ne sont pas aussi exhaustives que celles du répertoire des déclarations et des autorisations d'exploitation des INB.

REPERTOIRE DES DECLARATIONS D'EXPLOITATION DES INB*

Caractéristiques générales :

L'enregistrement des déclarations et des autorisations d'exploitation des Installations Nucléaires de Base (INB) est assuré par l'ASN. Les INB* concernent toutes les installations nucléaires civiles, i.e. liées à des activités industrielles, médicales ou de recherche, dont le fonctionnement est soumis à autorisation dans le cadre du décret n°2007-1557.

Granularité* des variables du répertoire de l'ASN :

Le répertoire des déclarations d'exploitation des INB* dénombre 126 INB* sur le territoire français métropolitain. Il contient des informations spatiotemporelles. Certaines d'entre elles sont publiques. En l'occurrence, le répertoire renseigne sur : la commune d'implantation, le nom de l'exploitant, sa date de mise en service, i.e. l'année où l'exploitant a obtenu l'autorisation de l'ASN, la date de fin d'exploitation, le type d'installation nucléaire… La variable ASN type : permet de se faire une idée plus précise de ce que sont les INB* : des centrales de production d'énergie, des stations de traitement des déchets, des zones de stockage de produits de fission, des établissements médicaux ou centres de recherche... (ASN, 2011)

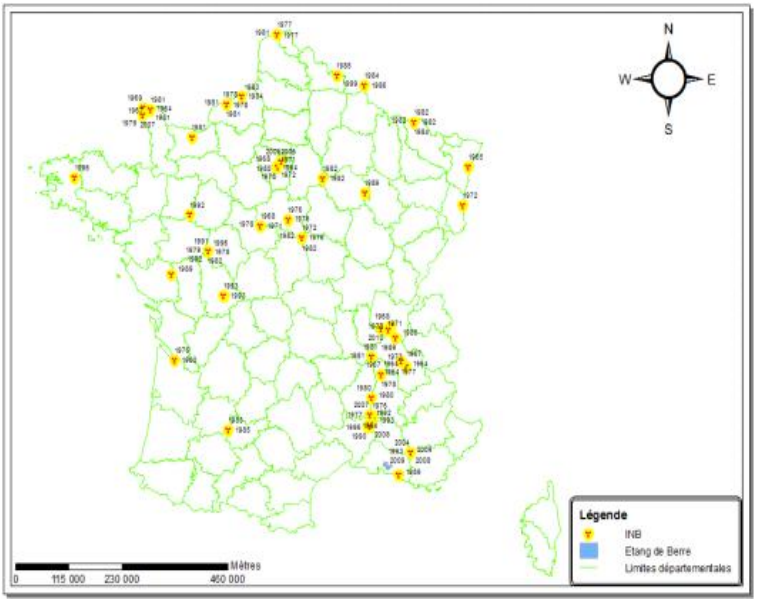


Figure 10 : Cartographie des 126 INB* présentes sur le territoire français métropolitain – documentées par les années de mise en service.
Source : Répertoire ASN

Remarque :
(i) le nombre d'INB* visuellement présentes sur la carte est inférieur au nombre d'INB* réellement présentes. Il s'agit simplement d'un biais d'échelle. Les caractéristiques d'affichage ne permettent de visualiser qu'une seule INB* par commue. Or, les INB* sont souvent regroupées, et une même commune peut en contenir jusqu'à vingt.

Toujours dans l'optique d'évaluer les expositions à la radioactivité environnementale, l'Atlas Radon permet d'évaluer l'exposition géographique des populations à ce qui constitue l'essentiel de la radioactivité naturelle tellurique.

ATLAS RADON

Caractéristiques générales :
En France, les premières campagnes de mesure du Radon menées par l'IRSN remontent à 1980. A cette époque, seules quelques régions étaient concernées. Mais les mesures n'étaient pas comparables. Aujourd'hui elles sont normalisées AFNOR. Entre 1982 et 1990, 38 départements font l'objet d'investigations complémentaires. En 1992, la Direction Générale de la Santé (DGS) porte un intérêt particulier aux expositions domestiques au gaz Radon. En 1997, l'IRSN est missionné pour compléter sa campagne de mesures en France métropolitaine. L'Atlas Radon a été constitué en 2001. Il s'agit de la seule source de données nationales publiques disponibles. Il permet d'évaluer l'activité volumique du radon dans les habitations (IRSN, 2001).

Granularité* des variables Atlas Radon :
L'activité volumique du Radon dans les habitations telle qu'elle est évaluée par l'Atlas Radon est disponible à l'échelle des communes sises en France métropolitaine, sauf pour Paris où la variable IRSN est déclinée à l'échelle des arrondissements. Dans chaque commune des mesures ponctuelles – parfois uniques – ont été effectuées dans certaines habitations. Puis, elles ont été agrégées par l'estimateur de la moyenne. Ces valeurs ne sont pas publiques. Et la seule variable mobilisable, à l'échelle des communes, est l'Activité Volumique du Radon, discrétisée en classe modales continues : $AVR_{(U_u)}$ – associée à un code couleur.

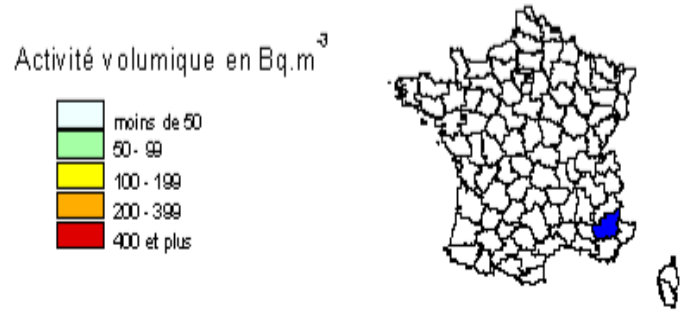


Figure 11 : Valeurs associes à chaque classe pour la variable : $AVR_{(U_u)}$
Source : IRSN

Les gammes de valeurs sont exprimées en : $Bq \cdot m^{-3}$. Ces mesures ont été effectuées dans une logique temporelle départementale et la période d'investigation s'est étalée de 1982 à 2000. Les seules valeurs quantitatives disponibles le sont à l'échelle des départements (IRSN, 2001).

Remarques :
(i) Les investigations menées sur l'activité volumique du Radon dans les habitations montrent de fortes disparités géographiques à l'échelle des communes.
(ii) Depuis 2008, les études des expositions environnementales au Radon sont menées conjointement par L'IRSN et par l'Institut National de l'Environnement et des Risques Industriels (IRSN, INERIS, 2008).

L'Institut National de l'Environnement et des Risques Industriels élabore actuellement une gigantesque plateforme SIG permettant d'évaluer l'exposition des populations à des substances physicochimiques toxiques combinées connue sous le nom de PLAINE.

PLAINE

Caractéristiques générales :
La Plateforme intégrée pour l'analyse des inégalités d'expositions environnementale (PLAINE) est un support informatique permettant : d'intégrer des données environnementales hétéroclites, de développer des outils permettant de modéliser les expositions m*ultimédia* à des substances physicochimiques toxiques, et d'identifier des zones à risques et des populations vulnérables.
PLAINE est un projet en cours d'élaboration porté par l'Institut National de l'Environnement et des Risques Industriels (INERIS). Cet établissement scientifique a pour mission d'évaluer et de prévenir les risques accidentels ou chroniques liés à des pollutions ubiquitaires comme : les déchets, les sols pollués, les sites industriels, dans un objectif de réduction de leurs impacts sur la santé et l'environnement, et dans une logique de durabilité (INERIS, 2012).

Granularité* des variables :
PLAINE intègre des : outils confectionnés par l'INERIS visant à l'amélioration des paramètres et des équations de transfert de modèles d'expositions multimédia ; stratégies d'incorporation de données environnementales, sanitaires et socio-économiques ; méthodes de détection de zones soumises à des risques d'expositions chroniques à des substances chimiques. A l'heure actuelle PLAINE se focalise essentiellement sur les expositions spatiotemporelles environnementales aux Eléments Métalliques Trace (ETM) et permet de caractériser les zones de surexposition des populations, ou *hotspots*.
Les variables INERIS disponibles dans PLAINE sont des Doses Journalières d'Exposition (DJE) ou de indicateurs de risques spatiaux (irs) d'exposition multi-milieux à des ETM particuliers. Et plus récemment, des Proxy du Risque d'Exposition à des substances Chimiques combinées (PREC) vectorisées par un *milieu de contact* particulier. Ces variables sont disponibles sur l'ensemble de la France : soit dans une logique adaptée à l'expologie, i.e. par le biais de raster dont la maille la plus fine

est de 1km, soit dans une logique territoriale avec des échelles géographiques allant de l'IRIS à celle des régions. Ces variables sont estimables à différentes dates dans une période comprise entre1990 et 2010. (Caudeville, 2011) ; (Caudeville J., Boudet C. et al., 2012) ; (Caudeville, Bonnard et al., 2012).

Les variables INERIS permettent d'estimer des expositions environnementales *intrinsèques*. Quant aux expositions environnementales *potentielles* elles peuvent par exemple être évaluées à partir de la biophysique des sols – par le biais de CORINE Land Cover.

CORINE LAND COVER

Caractéristiques générales :

CORINE Land Cover (CLC) est une base de données géographiques décrivant l'occupation biophysique des sols. Elle a été constituée dans le cadre d'un programme européen regroupant 38 nations. Le pilotage du projet est assuré par l'Agence Européenne pour l'Environnement (AEE). La diffusion ainsi que la maintenance des données sont déléguées aux états. En France c'est le Service de l'observation et des statistiques (SOeS) qui dépend du Commissariat Général au Développement Durable (CGDD), lui-même rattaché au Ministère de l'écologie, de l'énergie, du Développement Durable et de l'Aménagement du Territoire, qui est chargé de cette mission (CGDD, SOeS, 2009).

Granularité* des variables CLC :

CLC est une BD SIG qui représente l'occupation biophysique des sols à l'échelle 1/100 000 - qui s'est imposée par compromis entre une qualité de modélisation suffisamment précise et un coût d'actualisation supportable. Ce niveau de précision correspond à la base *CLC complète* qui recouvre l'intégralité la France. Deux autres échelles sont disponibles : le 1/50 000 et le 1/25 000. Elles sont l'apanage des *bases CLC de changement* qui ne concernent que quelques zones géographiques précises. L'intérêt est donc porté à la *base CLC complète qui* est constituée d'unités géographiques vectorielles obtenues par vectorisation d'images satellites de précision : 20 mètres. La taille minimale des éléments vectoriels pour une occupation biophysique homogène représente une surface de 25 ha. La base CLC est calée celles de l'IGN. L'occupation biophysique des sols est décrite pour trois temporalités spécifiques {1990, 2000, 2006}.

CLC-1990 a été réalisée à partir d'ortho-photographies acquises entre 1987 et 1994. CLC-2000, à partir d'images satellites prises dans le courant de l'année en question. Et finalement, CLC-2006 a été effectuée à l'aide d'images satellites de 2006 et de raster IRS et SPOT. Toutefois, entre 1990 et 2006 les techniques d'acquisition de données numériques ont évolué.

Conséquence, des conflits inter-bases font que les CLC-1990 et CLC-2000 peuvent être mises en perspective. Or ce n'est pas le cas avec CLC-2006. De fait, il existe deux bases pour l'année 2000 : CLC-2000 et une seconde base révisée : CLC 2000.r. Aussi, il convient de noter que l'analyse de l'occupation biophysique des sols avec CLC est pertinente jusqu'à l'échelle des communes mais pas en dessous. Le code couleur des entités géographiques CLC est normé par une nomenclature qui se hiérarchise en trois niveaux de précision. Le niveau 1 comprend 5 postes, le niveau 2 en compte 15 et le troisième niveau décrit 44 postes. Chaque *poste* est numéroté et le nombre de chiffres dépend du niveau d'investigation choisi. Par exemple : 221 avec 2 pour : territoires agricoles, 2 pour : cultures permanentes, et 1 pour : vignobles. En dépit du désir de *cohérence spatiale* de CLC, des zones hétéroclites ont dû être parfois groupées pour satisfaire la contrainte des unités spatiales minimales de 25 hectares (CGDD, SOeS, 2009).

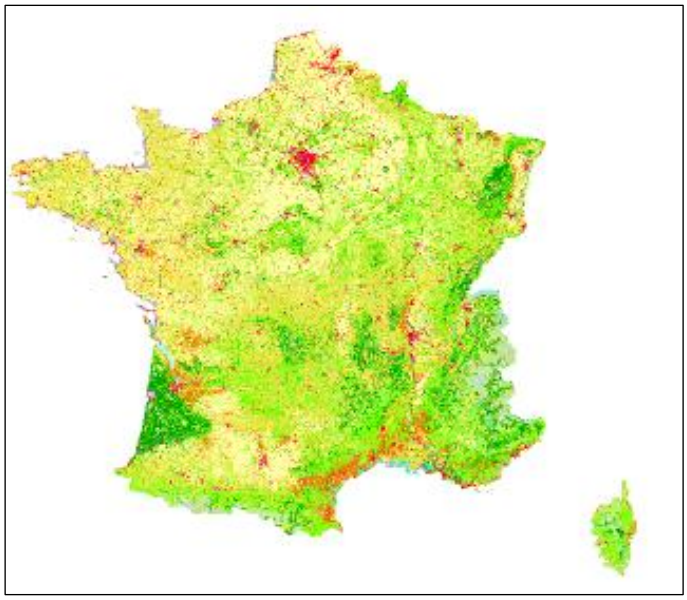


Figure 12 : Cartographie de l'occupation biophysique des sols en France métropolitaine
CLC 2000 – niveau 3
Source : CLC

Remarques

(i) Dans le cadre cette recherche ce sont les postes de niveau 3 qui sont utilisés. Par la suite les entités surfaciques vectorielles du poste l seront notées : $e.clc_t^l$. Et l'ensemble des entités vectorielles agrégeant des postes divers d'une même thématique sera noté $\{l \in \varrho(\mathrm{CLC}) = \mathrm{THEME}\}$, soit encore $e.clc_t^{\mathrm{THEME}}$.

(ii) Un descriptif précis des zones incluses dans chaque *poste* de la nomenclature est disponible en annexe et de façon encore plus exhaustive dans la documentation technique de CLC.

Toutes les BD environnementales utilisées ont été déclinées. Ils convient maintenant de donner la Granularité* des variables utilisées dans les processus d'estimation des i.st.e*. voués à la modélisation géographique des FE-PHY.CHIM

VARIABLES UTILISEES POUR LA MODELISATION GEOGRAPHIQUE

Les données LEA utilisées déclinent une période allant de 1980 à 2010. La stratégie d'agrégation *verticale* des données environnementales se fonde sur le concept de *variabilité temporelle apparente* (Peguy, 1996). De fait, une temporalité supplémentaire est ajoutée quand les données disponibles le permettent.

FE-PHY.CHIM* pertinents :

Granularité* des variables Météo-France :
L'échelle géographique : mesures environnementales géo-localisées s_g ; *Pas de temps* : mensuel ; *Lacunes :* aucune ; *Champ* : territoire français ; *Répartition spatiale des données d'échantillonnages* : uniforme

FE-PHY.CHIM	EXPOSITION	PARAMETRE	VARIABLE	DESCRIPTION	SITES	TEMPORALITES	AUTRE
géographie	*type / nature*	géophysique	unité	*variables*	n_s^l	disponibles	*particularités*
EXPOSITIONS ENVIRONNEMENTALES GEOPHYSIQUES	Potentielle : Climatique	Rayonnement solaire global	$m^{RAYG}_{(s_g,t)}$ (joule/cm²)	Moyenne annuelle des cumuls mensuels	27	1981, …, 2011	Densité spatiale d'échantillonnage faible liée aux caractéristiques techniques des stations météo
		Températures	$m^{TEMP}_{(s_g,t)}$ (Celsius)	Moyenne annuelle	54	1981, …, 2011	-
		Précipitations	$m^{NJAP}_{(s_g,t)}$ (U)	Nombre de Jours Annuels Pluvieux	54	1981, …, 2011	Les pluies dont la pluviométrie est supérieure à 1mm
	Potentielle : Topographique	Altitude	$m^{TOP}_{(s_g,t)}$ (m)	Niveau altimétrique moyen communal	421	2003	Localisée au centroïde

Tableau 9 : Variables géophysiques mobilisées.

Remarques :
(i) (*) Une description plus détaillée des données utilisées est disponible dans les fiches techniques (Météo-France, 2010) – annexe.1 ; annexe.1 ; annexe.1

(ii) (*) Les caractéristiques techniques de la BD SIG Géofla utilisée sont déclinées par la suite (chapitre.2) – un complément informationnel est aussi disponible (IGN, 2004) – (annexe.2).

Granularité* des variables RNM* :
Echelle géographique : variables localisées à la commune ; *Périodes annuelles couvertes* : de 2008 à 2010 ; Lacune : *aucune.*

FE-PHY.CHIM	EXPOSITION	MESURE	VARIABLE	SPECIFICATION	SITES	NATURE
géographie	*Type / nature*	*Type / unité*	notation	mesure	n_s^l	*prélèvement*
EXPOSITIONS A LA RADIOACTIVITE ENVIRONNEMENTALE	Intrinsèque Dans l'air	Doses efficaces (NanoSievert/heure)	$m_{(s_g,t)}^{GAMMA}$	Globale : émise par l'ensemble des radionucléides	88	Mesurée par densitométrie active ; Echantillon air ambiant : gaz et poussières
	Potentielle dans l'eau	Activité radioactive volumique (Becquerel/litre)	$m_{(s_g,t)}^{ALPHA}$	Globale : tous les émetteurs α	42	Echantillons : eau douce, i.e. des eaux de nappe, de surface ou de pluie
			$m_{(s_g,t)}^{BÊTA}$	Globale : tous les émetteurs β	54	
			$m_{(s_g,t)}^{3H}$	Totale : des isotopes radioactifs du Tritium	41	
	Potentielle dans les sols	Activité radioactive volumique (Becquerel/kilogramme de matériau sec)	$m_{(s_g,t)}^{238Pu}$	Dégradation du Plutonium 238	25	Echantillons de Sols sédiments : graviers, sables limons
			$m_{(s_g,t)}^{137Cs}$	Dégradation du Césium 137	50	
			$m_{(s_g,t)}^{125Sb}$	Dégradation du l'Antimoine 125	26	
	Potentielle dans les milieux biologiques	Activité radioactive volumique (Becquerel/litre)	$m_{(s_g,t)}^{131.i}$	Dégradation du l'Iode 131	86	Echantillons organiques : de lait de : vache, brebis ou chèvre
			$m_{(s_g,t)}^{90Sr}$	Dégradation du Strontium 90	61	
			$m_{(s_g,t)}^{137Cs}$	Dégradation du Césium 137	84	

Tableau 10 : Mesures de radioactivité environnementale mobilisées.

Remarques sur les variables RNM* :
(i) La précision et le pas de temps varient en fonction de la nature des substances radioactives, du type de prélèvement, du milieu, de l'exploitant, et des caractéristiques techniques des stations.
(ii) La densité spatiale des mesures disponibles est parfois faible. Cette lacune est prise en compte dans les stratégies d'estimation des i.st.e* proposées.
(iii) La période temporelle est plus courte que de celle de l'étude LEA mais pas d'autres données disponibles.
(iv) Les mesures ont été géo-localisées au niveau des centroïdes communaux puisque les coordonnées géographiques des stations ne sont pas communiquées sur le portail RNM.

Compléments informationnels sur les mesures Météo-France et RNM* :
La pertinence des variables mobilisées repose sur un compromis tridimensionnel :
a) Sur le plan temporel, les longueurs des chroniques utilisées doivent être à peu près semblables et recouvrir au mieux la période d'étude LEA.
b) Sur le plan spatial elles doivent pouvoir être interprétées comme des *variables régionalisées* (v.r.)* afin d'être interpolées sur l'intégralité du territoire français métropolitain par le biais de méthodes géostatistiques fiables. Cela dépend de la répartition et de la densité spatiale du réseau de stations mesurant les paramètres environnementaux jugés pertinents (Matheron, 1965).
c) Sur le plan médical, la distance a-spatiale morbide* entre les PM* d'intérêt - séquelles et les paramètres mobilisés doit être jugée *pertinente*.

(i) La qualité du réseau de mesures Météo-France a permis de mobiliser des chroniques temporelles dont les caractéristiques granulométriques répondent aux conditions (a) et (b) énoncées précédemment ; quant à la condition (c), i.e. leur *pertinence* au regard des PM* d'intérêt, elle est spécifiée explicitement dans l'état de l'art.

(ii) Le réseau RNM* est très dense mais la nature des mesures disponibles est disparate et leur granularité* est maculée par des conflits spatiotemporels. Les chroniques mobilisées répondent aux conditions (a) et (b), cependant, leur *pertinence* au regard des PM* d'intérêt, i.e. le respect de la dimension (c), n'est pas spécifiée – ce qui est justement l'objet de ce complément. D'une manière générale *all types of ionizing radiation are carcinogenic to humains*, i.e. comme cancérigène probable pour tous les cancers – THYR, TUM2 (IARC, 2012), et les expositions chroniques à de faibles doses de radioactivité ont un effet contributif fortement suspecté sur l'incidence des CATA (Jacob, Bertrand et al., 2010).

- Mesures RNM* effectuées dans l'air ambiant : $\left(m_{(s_g),t}^{GAMMA}\right)$

Cette variable mesure la dose efficace de rayons γ liée à la dégradation des 70 radionucléides présents dans l'air ambiant. Elle est exprimée en NanoSv. Globalement, l'exposition à de fortes doses de rayons γ est classée comme cancérigène - groupe 1 - pour les cancers et celle à des doses environnementales comme cancérigène probable - groupe 2B (IARC, 2012).

- Mesures RNM* effectuées dans des prélèvements d'eau douce: $\left(m_{(s_g),t}^{ALPHA}, m_{(s_g),t}^{BETA};\ m_{(s_g),t}^{3H}\right)$

Ces variables mesurent l'activité globale de tous les émetteurs de particules α et β ainsi que tous les isotopes radioactifs du Tritium. Elles sont exprimées en Bq/litre (ASN, 2013).
Les précipitations sont à la fois un excellent moyen d'épuration de l'air et le vecteur de contamination principal de tous les autres milieux environnementaux. Les eaux douces ont des activités volumiques radioactives de 10 à 100 fois supérieures à l'eau de mer. Les mesures retenues varient avec des gradients géographiques importants liés à la proximité des INB* – le Tritium total étant particulièrement représentatif de cette variabilité spatiale (ANS et IRSN, 2013).
D'une manière générale les expositions *internes* à des particules α et β sont classées dans le groupe 1 et celles liées à des expositions environnementales appartiennent au groupe 2b (IARC, 2012)
Les risques liés à l'exposition au Tritium : ^{3}H sont discutés. Les études effectuées sur des animaux contaminés au Tritium montrent sans équivoque que cet élément à des effets cancérigènes déterministes (Johnson, Myers et al., 1995). Il en est de même pour les cultures effectuées in vitro sur des cellules humaines (Kamiguchi, Tateno et al., 1990). S'agissant des expositions environnementales, quelques études rapportent une augmentation de l'incidence des cancers en général et de celle des leucémies chez des enfants vivant à proximité d'INB* et dégageant d'importantes quantités de Tritium (AECB, 1991), mais elles restent très controversées.

- Mesures RNM* effectuées dans des prélèvements de sols : $\left(m_{(s_g),t}^{238Pu}, m_{(s_g),t}^{137Cs};\ m_{(s_g),t}^{125Sb}\right)$

Ces variables mesurent l'activité volumique, exprimée en Bq / kg de matière sèche, du
- **Plutonium 238**, un émetteur de rayons *γ* et α presque purs dont la période radioactive est de 90 ans (ASN, 2013); Les expositions environnementales au ^{238}Pu sont classées comme cancérigènes probables (groupe 2b) pour tous les types de cancers (IARC, 2012).
- **Césium 137**, un émetteur de rayons *γ* et β presque purs dont la période radioactive est de 30 ans (ASN, 2013). Les expositions environnementales au ^{137}Cs sont classées groupe 2b pour tous les types de cancers (IARC, 2012).
- **Antimoine 125**, un émetteur de rayon *γ* et de particule β dont la période radioactive est de 3 ans. Le ^{125}Sb est particulièrement présent sur le territoire français et avec des forts gradients géographiques. La contamination des sols est liée à des effluents liquides diffusés par le fonctionnement normal des INB* ou accidentel depuis celles de traitements de déchets radioactifs (GRNC, 1999). Les expositions à des doses environnementales à ^{125}Sb ne sont pas *a priori* dangereuses. En revanche, il s'agit d'un radionucléide mesurable facilement et parfaitement caractéristique de la présence d'autres agents radioactifs particulièrement toxiques et liés à la contamination des sols par des produits de fission dont l'exposition environnementale est classée 2b (IARC, 2012).

➢ Mesures RNM* dans les milieux biologiques – i.e. le lait : $\left(m^{137Cs}_{(s_g),t}, m^{131i}_{(s_g),t}; m^{90Sr}_{(s_g),t}\right)$

Ces variables mesurent l'activité volumique, exprimée en Bq / litre, de radionucléides présents dans les milieux biologiques pour des motifs militaires, énergétiques ou économico-industriels (Colle, Adam et al., 2005). Les mesures de la radioactivité dans les milieux biologiques sont effectuées dans des prélèvements de diverses denrées, des poissons jusqu'aux fruits et légumes. Celles choisies ont été effectuées dans le lait. Un aliment de base que l'on trouve partout, dans les pâtisseries, les biscuits, les yaourts, les sauces, les fromages et consommé à tout âge et depuis l'enfance (ANS et IRSN, 2013).
Les radionucléides présents dans les sols et les eaux douces sont absorbés par les végétaux – par contaminations foliaires ou transferts trophiques – dans lesquels ils s'accumulent par le bais des phénomènes de bioaccumulation et de bioamplification pour atteindre ensuite toute la chaine alimentaire. Les productions agricoles végétales sont contaminées de cette façon. Les ovins et les bovins le sont par suite de la consommation de végétaux et d'eaux elles-mêmes contaminées. Et finalement l'homme est atteint (ASN, 2010).
Le Césium 137 a déjà été décrit pour les sols.
L'iode 131 est un émetteur de rayon γ et β quasi pur et dont la période radioactive est de 8 jours mais il est bio-absorbé et bio-accumulé très rapidement et se retrouve dans le lait (ASN, 2013). Les expositions environnementales à ^{131}i sont classées groupe 1 pour les THYR et groupe 2 pour tous les types de cancers (IARC, 2012).
Le Strontium 90 est un radionucléide exclusivement artificiel et un émetteur de rayon γ et β quasi pur et dont la période radioactive est de 60 ans (ASN, 2013). En cas d'ingestion il est assimilé à 99% par les organismes humains et s'accumule au niveau des tissus osseux et des dents. ^{90}Sr est classé groupe 1 pour les cancers et en particulier les THYR et les LEUC. Quant aux expositions environnementales, elles ont récemment été classées groupe 2 pour tous les types de cancers (IARC, 2012)

Granularité* des variables du répertoire des autorisations d'exploitation des INB* de l'ASN

FE-PHY.CHIM	EXPOSITION	VARIABLE	NOTATION	SPECIFICATION	COMMENTAIRES
géographie	*nature*	nombre	notation	données	*données*
EXPOSITIONS GEOGRAPHIQUES A DES* RADIONUCLEIDES* ARTIFICIELLS (EGRA)	Potentielle : par des diffusions légales ou accidentelles de radionucléides dans les milieux environnementaux liées à la proximité géographique des INB	$INB_i^{info.ASN}$ Informations associées aux 126 INB	CI_i	Commune d'Implantation	France métropolitaine
			$EXPL_i$	Nom de l'exploitant	ANDRA, AREVA, CEA…
			DMS_i	Date de Mise en Service	1964, …, 2010
			DFS_i	Date de Fin de Service	1965, …, 2010
			$TYPE_i$	Type d'Installation nucléaire	Centre de stockage, Usine, Centrale nucléaire, station de traitement…

Tableau 11 : Variables sémantiques ASN mobilisées.

Remarques sur les variables ASN :
(i) Ces informations spatiotemporelles sémantiques seront notées par la suite : $INB_i^{info} = (CI_i; EXPL_i; DMS_i; TYPE_i)$
(ii) Les INB* ont été géo-localisées, avec un léger décalage stochastique, au niveau des centroïdes des communes d'implantation CI_i.

Granularité* des variables Atlas Radon :

Champ : *territoire français métropolitain* ; Période temporelle : entre 1982 et 2000 ; unité : $Bq \cdot m^{-3}$; *Lacunes : aucune* .

FE-PHY.CHIM	EXPOSITION	VARIABLE	SPECIFICATION		Echelle
géographie	*Type et Nature*	notation	Valeur ou description		*disponible*
EXPOSITIONS GEOGRAPHIQUES A LA RADIOACTIVITE TELLURIQUE, GAZ : RADON	Potentielle : Liée à l'activité volumique moyenne du Radon dans les habitations	$AVR_{(U_u)}$	Blanc	Moins de 50	Commune U_u, ou arrondissement AR_{j_u}
			vert	50-99	Commune U_u, ou arrondissement AR_{j_u}
			Jaune	100-199	Commune U_u, ou arrondissement AR_{j_u}
			Orange	200-399	Commune U_u, ou arrondissement AR_{j_u}
			Rouge	400 et plus	Commune U_u, ou arrondissement AR_{j_u}
		$M^{RADON}_{(\{U_u \subset DEP_d\})}$	Minimum, Maximum, Moyenne et Ecart-type des valeurs communales		Département : DEP_d

Tableau 12 : Variables de l'Atlas Radon mobilisées.

Remarque sur les variables IRSN :

Compte tenu du fait que dans de nombreuses communes un seul point de mesure a été installé, l'IRSN émet des réserves quant à la qualité des variables à l'échelle des communes. C'est certainement la raison pour laquelle les mesures sont rendues sous forme de classes de valeurs.

Granularité* des variables PLAINE: *Echelle* : les communes : *Champ* : territoire français métropolitain ; *Période temporelle couverte* : 1990, …, 2009 *; Lacunes : aucune*

FE-PHY .CHIM	EXPOSITION	VARIABLE	SPECIFICATION	Echelle
géographie	*Type et Nature*	Notation / unité	estimateur	Voie d'exposition
RISQUE D'EXPOSITION A DES* ELEMENTS METALLIQUES TRACES	Intrinsèque Contamination des organismes humains par des métalloïdes spécifiques ou combinés	$x^{DEJ(Cr)}_{U_k}$ (Mg/kg)	Dose Journalière des Expositions au Chrome	Par ingestion ou inhalation rapportée à la masse moyenne d'enfants et d'adolescents
		$x^{irs(Pb)}_{U_k}$ (Sans Unité)	Indicateur des risques spatiaux d'exposition au Plomb	Par ingestion de particules de sol dans l'eau potable et les denrées alimentaires ainsi que par l'inhalation de particules présentes dans l'air et rapportées aux doses journalières maximales tolérées par des organismes humains
		$x^{irs(Ni)}_{U_k}$ (Sans Unité)	Indicateur des risques spatiaux d'exposition Nickel	
		$x^{irs(Cd)}_{U_k}$ (Sans Unité)	Indicateur des risques spatiaux d'exposition Cadmium	
		$x^{PREC}_{U_k}$ (Sans Unité)	Proxy du Risque d'Exposition à des substances Chimiques	Par inhalation de substances chimiques présentes dans l'air ambiant et émises depuis les installations industrielles recensées par l'IrEP (Inventaire des émissions polluantes)

Tableau 13 : Variables INERIS mobilisées.

Remarque sur les variables INERIS :

(i) Les indicateurs spatiotemporels INERIS ont été fournis sous une forme ingérable directement. Pour des renseignements plus spécifiques sur ces indicateurs, il convient de se référer à : (Caudeville, 2011) ; (Caudeville J., Boudet C. et al., 2012) ; (Caudeville, Bonnard et al., 2012).

Granularité* des variables CLC : *Echelle géographique* : 1/100 000 ; *Taille de l'unité géographique vectorielle minimale* : 25ha ; *Nature variable* : Quantitative discrète ; *Postes utilisés* : le niveau 3 de la nomenclature CLC ; *Temporalités* : $t = \{1990; 2000\ ; 2000.\mathrm{r}\ ;\ 2006\}$

FE-PHY.CHIM	EXPOSITION	VARIABLE	POSTE
géographie	*Type et Nature*	Notation / unité	représentant
EXPOSITIONS A DES* SUBSTANCES PHYSICOCHIMIQUES DELETERES COMBINEES	Potentielle : Aux pesticides	e. $clc_t^{l:PEST}$ (SU : postes CLC correspondant*)	Zones géographiques spécialisées dans l'agriculture*
	Potentielle A des agents physicochimiques toxiques liés des activités humaines polluantes	e. $clc_t^{l:URIN}$ (SU : postes CLC correspondant*)	Zones fortement urbanisées ou industrialisées*

Tableau 14 : Postes CLC mobilisées.

Remarques :

(i) (*) Un descriptif des postes de niveau 3 est donné en annexe et dans la documentation technique associée (CGDD, SOeS, 2009).

(ii) (*) Les risques d'expositions environnementales à toutes les substances physicochimiques citées dans l'état de l'art, dont les pesticides font partie. Les espaces vulnérables sont liés à l'occupation biophysique des sols par des activités agricoles ou des zones fortement urbanisées et industrialisées (Unité Cancer et Environnement, 2012) (Afsset, 2009a).

FE-PHY.CHIM* curieux de test :

Granularité* des variables CLC :

Identiques à ceux énoncés-ci-avant.

FE-PHY.CHIM	EXPOSITION	VARIABLE	SPECIFICATION
géographie	*Type et Nature*	Notation / unité	estimateur
EXPOSITIONS A DES* SUBSTANCES PHYSICOCHIMIQUES DELETERES COMBINEES	Potentielle Aux : formaldéhyde, HAP et MES	e. $clc_t^{l:FEFO}$ (SU : postes CLC correspondant*)	Zones où ont eu lieu des feux de forêt de grande ampleur*
EXPOSITIONS A DES* ENVIRONNEMENTS A PRIORI* PREVENTIFS	Potentielle A des espaces naturels où les pollutions anthropiques sont *a priori* limitées	e. $clc_t^{l:PREV}$ (SU : postes CLC correspondant*)	Zones recouvertes par des végétations : forestières, abusives ou sclérophylles**

Tableau 15 : Postes CLC mobilisées.

Remarque :

(iii) les feux de forêt de grande ampleur induisent des émissions de MES toxiques et engendrent, selon certains auteurs, une augmentation de toutes les maladies et des cancers (Johnston Fay, Henderson et al., 2012).

(iv) les zones naturelles sont *a priori* des zones où les pollutions anthropiques sont limitées et constituent des espaces géographiques *préventifs* (Unité Cancer et Environnement, 2012).

FE-PHY.CHIM* non modélisables :

Les trois principaux FE-PHY.CHIM* pertinents qu'il aurait été intéressant de modéliser sont :

FE-PHY.CHIM	*Estimateur proposé*	BASE	VARIABLES	NON INTEGRABLES
géographie	*Etat de l'art*	*gestionnaire*	*existantes*	*motif*
EXPOSITIONS AUX RNI	Liées à la présence des cem-ebf	BD : *Cartoradio* ANFR (*)	Informations spatiotemporelles géo-localisées	Impossible d'établir de conventions
EXPOSITIONS ALIMENTAIRES AUX PESTICIDES*	Potentielles, liées à la consommation d'eau potable en France	BD : *interne du BQE* DGS (*)	Indices géographiques qualitatifs de qualité des eaux de boisson	Données non publiques
EXPOSITIONS A DES* SUBSTANCES TOXIQUES PRESENTES DANS L'AIR	Potentielles, liées à l'inhalation d'air ambiant	BD : *Buldair* ADEME (**)	Mesures de concentration en : Ozone, Dioxyde d'azote, MES ; Et indice : ATMO	Données publiques mais aucun partenariat possible pour une extraction massive dans des délais raisonnables

Tableau 16 : Variables physicochimiques indisponibles.

Remarques sur les BD environnementales non accessibles et qui auraient pu permettre de modéliser des FE.PHY.CHIM pertinents :

(*) Les effets sur l'état de santé des RNI* émis par les champs électromagnétiques et magnétiques à extrême et basse fréquence (cem-ebf) ont été décrits dans l'état de l'art. Afin d'estimer ce type d'exposition, en 2001 l'Agence Nationale des Fréquences (ANFR) crée un portail interactif *Cartoradio*. Deux types d'entités spatiales sont disponibles : des stations de mesure qui déclinent des chroniques temporelles géo-localisées permettant d'évaluer l'intensité combinée des cem-ebf, des émetteurs géo-localisés de radiotéléphonie, radiodiffusion et autres éléments radioélectriques ou radars ainsi que leurs fréquences d'émissions théoriques (ANFR, 2012).

(*) Les effets sur l'état de santé des expositions *internes* aux pesticides ont été déclinés dans l'état de l'art. L'indice géographique de qualité des eaux de boisson est un avis annuel de conformité, ou de non-conformité, des eaux liées à la contamination des réseaux d'eaux potable par des pesticides. Il a été élaboré par la DGS et le Bureau de la Qualité des Eaux (BQE) dans une logique géographique territoriale (DGS / Bureau de la qualité des eaux, 2008).

(**) Les effets sur les PM* d'intérêt des expositions aux dioxines, aux MES ainsi qu'à un grand nombre de substances chimiques présentes dans l'air ont été déclinés dans l'état de l'art. La base de données Buldair a été créée par l'ADEME qui se charge aussi de son actualisation. Les variables existantes dans cette base se déclinent sous la forme de chroniques temporelles géo-localisées de concentrations moyennes journalières dans l'air extérieur en : Ozone, Dioxyde d'azote, MES ; Ainsi que de la variable : ATMO. Cette dernière est un indice composite global de qualité de l'air qui évalue l'ensemble des pollutions atmosphériques liées à l'industrie, au traitement des déchets, aux transports, à l'urbanisation et aux activités agricoles, à la combustion de biomasses et à l'érosion éolienne (ADEME, 2012).

Les BD environnementales accessibles ainsi que la granularité* des variables utilisées pour la modélisation géographique des PM* d'intérêt – séquelles - par des i.st.m, et des FE/FIM* *pertinents* par des i.st.e* ont été déclinées. Il s'agit désormais de proposer des stratégies adaptées à la modélisation géographique et à l'identification des DES. Pour clore ce chapitre, l'échelle d'investigation retenue – nécessaire et préalable à toute modélisation géographique – est spécifiée au regard de considérations théoriques bibliographiques et de la granularité* des variables qui ont été mobilisées.

SYNTHESE DU CHAPITRE 1

Ce chapitre proposer une dialectique adaptée à l'identification géographique des FE* qui déterminent *l'état de santé* des populations. L'angle d'attaque est méthodologique et il a trait à la géographie de la santé. Toutes les propositions heuristiques sont ensuite appliquées aux données de la Cohorte LEA.

Pour répondre à cette problématique de santé publique complexe, le positionnement scientifique choisi est interdisciplinaire et la géographie se mêle à l'épidémiologique et à la statistique. La démarche se veut reproductible et extensible à d'autres pathologies. L'objectif *in fine* est de donner aux médecins et politiques des moyens opérationnels leur permettant de proposer des mesures préventives, individuelles ou collectives, afin de limiter l'exposition à des Déterminants Environnementaux de Santé* (DES), du simple fait de la position des individus dans l'espace géographique*.

Dans cette recherche, l'environnement géographique est considéré dans toute sa plénitude et a été discrétisé en quatre composantes : sanitaire (SAN), socio-économique (SOCIO.ECO) et physicochimique (PHY.CHIM) – sans omettre la géographie des Caractéristiques Individuelles et Médicales* (CIM) des individus qui d'évidence déterminent en partie leur état de santé.

L'analyse des interactions santé-environnement a conduit à distinguer les expositions environnementales géographiques *potentielles, i.e. à des situations à risque* ou à *la présence de substances toxiques dans les milieux environnementaux.* Et les expositions environnementales géographiques *intrinsèques,* i.e. à partir de doses de contamination par des agents physicochimiques depuis *les milieux de contact.*

L'état de l'art sur les méthodes utilisées en géographie de la santé a permis, d'une part de dissocier les méthodes classiques de modélisation géographique et suffisamment robustes pour être utilisées, et dans le cas contraire d'en proposer des plus appropriées. Et d'autre part, de soulever l'indigence des méthodes d'analyse multidimensionnelle utilisées en géographie et de suggérer une procédure *de sélection de variables*, basée sur du datamining*, plus adaptée aux jeux de données spatiotemporels contemporains. Il est donc *a priori* possible de modéliser, par des i.st.m* et des i.st.e* robustes, la variabilité géographique des PM* d'intérêt - séquelles (cataractes : CATA, tumeurs thyroïdiennes : THYR, et tumeurs secondaires majeures : TUM2) - et des expositions environnementales *potentielles* ou *intrinsèques* à des FIM* et des FE* à connotation : SAN, SOCIO.ECO et PHY.CHIM, et par suite, d'identifier rigoureusement des DES.

L'état des connaissances sur les PM* d'intérêt a permis d'identifier les FE/FIM* *pertinents* qu'il convient d'intégrer. Comme il n'existe pas de littérature sur les séquelles développées après le traitement d'une leucémie (LEUC), les *FE/FIM* pertinents* sont ceux décrits dans la littérature pour avoir des effets *avérés ou suspectés* sur les maladies citées. Ensuite, il a été supposé que ce est qui vrai chez les *sujets sains* l'est *a fortiori* chez les *personnes prédisposées*.

Par ailleurs, la qualité des modélisations géographiques ainsi que la capacité d'intégrer les FE/FIM* *pertinents* sont fortement conditionnées par la granularité* des variables stockées dans les BD épidémiologiques et environnementales. L'accès à des sources de données de bonne qualité est le principal écueil des modélisations géographiques.

Au regard des données accessibles, il est possible de modéliser la géographie des : FIM* *pertinents* par le genre ou sexe, l'âge au moment du diagnostic, la durée du suivi des patients, le protocole de traitement reçu ainsi que son agressivité ; FE-SAN* *pertinents* par les temps d'accès par la route à des praticiens libéraux, à des services hospitaliers et à des Equipements Matériels Lourds (EML*), ainsi que l'Accès Potentiel Localisé* (APL) à des généralistes et des ophtalmologues ; FE-SOCIO.ECO* *pertinents* tels des comportements vis-à-vis du recours aux soins, des politiques territoriales menées en matière de durabilité, des expositions à des substances toxiques liées aux secteurs d'activité professionnelle

dominants, les expositions potentielles aux pesticides induites par la spécialisation des espaces, et enfin *la défaveur sociale* ; FE-PHY.CHIM* *pertinents* par des paramètres géophysiques, l'exposition à la radioactivité environnementale, à des radionucléides artificiels liés à la proximité spatiale des INB, au Radon dans les habitations, aux ETM dans les milieux de contact, à des substances physicochimiques multiples potentiellement engendrées par l'usage biophysique des sols.

Certains *FE/FIM* pertinents* sont modélisables mais à partir de données de moindre qualité ou par des variables environnementales *a-spatialement* éloignées FE/FIM* qu'elles sont censées représenter. Le concept de *distance a-spatiale morbide** est défini comme la plausibilité théorique, i.e. au regard de l'état des connaissances actuelles, des effets avérés ou suspectés sur l'état de santé des expositions environnementales à des situations contextuelles à risque ou à des substances physicochimiques.

A cela s'ajoute l'absence de preuve théorique *de la méthode de sélection* des DES* proposée. La considération conjointe de ces deux allégations a conduit à proposer des FE/FIM* *Curieux* de test* – qui doivent normalement être détectés du bruit environnemental ou, dans le cas le plus favorable, avoir un pouvoir explicatif faible.

La revue des données utilisées offre la possibilité de modéliser la géographie des : FIM* *Curieux** avec l'intensité de l'activité physique pratiquée par les patients ; FE-SAN* *Curieux** grâce à des distances d'accès à des *items sanitaires* incohérents,* i.e. qui n'ont pas de lien avec le PM* ; FE-SOCIO.ECO* *Curieux** par le stress potentiellement induit par l'insécurité territoriale ; FE-PHY.CHIM* *Curieux** avec l'exposition à des substances toxiques engendrées par des feux de forêts de grande ampleur ou, à l'inverse, l'effet préventif des zones naturelles *a priori* moins polluées.

Avant proposer une méthode de spatialisation adaptée aux données épidémiologiques et des stratégies de modélisation géographique des PM* d'intérêt, il convient de spécifier une échelle géographique d'investigation adaptée à la problématique.

CONCLUSION DU CHAPITRE 1 ET CHOIX DE L'ECHELLE D'INVESTIGATION

Le choix de l'échelle géographique est primordial, préalable et nécessaire à la spatialisation des données épidémiologiques – de la Cohorte LEA – et à la modélisation géographique des PM* ainsi que des FE/FIM* par des i.st.m* et des i.st.e* robustes. Il résulte d'un consensus entre les considérations théoriques bibliographiques énoncées, le positionnement scientifique choisi, les hypothèses posées, et les considérations pragmatiques inhérentes à la granularité* des données disponibles utilisées, i.e. aux échelles, temporalités, précisions, unités de mesure, lacunes, incertitudes, restrictions juridiques et conflits spatiotemporels inter-sources.

Le choix de l'échelle doit d'abord prendre en considération les théories du domaine dans lequel s'inscrit la recherche. En l'occurrence celui des Sciences Humaines et Sociales (SHS), dont l'objet est d'établir des relations entre l'homme et son milieu. Il est donc question de phénoménologie et de fait, *les symétries (i.e. les relations) n'apparaissent qu'à des observateurs prudents qui s'en tiennent à des données macroscopiques* (Merleau-Ponty, 1945).

Les échelles macroscopiques sont privilégiées en SHS et chaque discipline prône une approche particulière. En géographie, et plus particulièrement en analyse quantitative, il s'agit d'intégrer *l'espace transversalement et non comme simple variable* (Voiron-Canicio, 2006).
La lecture géographique d'un phénomène s'évalue *donc par son expression spatiale*, dont la projection d'indicateurs *sur une carte en est l'expression la plus triviale* (Amat-Roze, 2011).

La Géographie appliquée à une problématique de santé publique est la géographie de la santé. Elle a pour objet de caractériser, par des i.st.*, l'état de santé des sociétés dans le but d'avoir des répercussions sur les comportements sociaux. En conséquence, l'analyse des Facteurs Environnementaux* (FE) qui influencent la promotion ou la dégradation de la santé doit permettre d'orienter les mesures politiques visant à réduire les expositions environnementales à des substances délétères, et donc intégrer une *logique territoriale* (Salem, 1995).

En France, les géographes de la santé ont conscience de cette double spécificité des modélisations géographiques qui doivent permettre de décrire et d'analyser les disparités spatiales morbides mais aussi d'influencer les courants de pensée. L'échelle régionale permet de dresser un bref diagnostic des défaillances des systèmes sanitaires spatiaux. Puis, celle des communes permet de spécifier plus précisément les disparités territoriales (Salem, Rican S et al., 2006). Enfin, quand les données le permettent, la démarche classique consiste à analyser les inégalités de santé sur des espaces plus petits – i.e. à des échelles fines – afin d'étudier plus précisément l'effet de certains Facteurs Environnementaux* (FE) sur l'état de santé des populations *in situ* (Rican, Salem G et al., 2009).

En géographie de la santé l'échelle des communes est donc privilégiée. Cependant, bien qu'elle soit plus grossière, l'échelle des cantons est parfois utilisée. C'est le cas par exemple lorsqu'elle est jugée suffisamment pertinente pour analyser les disparités spatiales du recours aux soins (Rémy, Handschumacher et al., 2011). En effet, les variables décrivant le tissu sanitaire sont diffusées à l'échelle des communes et à celle des cantons, mais pas en dessous (ARS, 2012).

Cependant en épidémiologie spatiale, l'échelle des cantons est controversée car elle ne permet pas de capter *la variabilité géographique des inégalités sanitaires* aussi bien que celle des communes (Chaix, Merlo et al., 2005). A cela s'ajoute le fait qu'en France le recours aux soins est fortement conditionné par des facteurs socio-économiques dont l'analyse des interactions spatiales avec l'état de santé des populations performe à l'échelle des communes (Rey, Jougla et al., 2009). Quant aux indicateurs socio-

économiques, ils sont disponibles à l'échelle des communes, et de façon plus restreinte à celle des *IRIS* (INSEE, 2012c).

Les coûts d'évolution et d'actualisation de ces bases géographiques socio-économiques ou sanitaires sont importants et sont supportés par l'État. Et plus l'échelle d'une variable est fine, plus la base est onéreuse (Elliott, Briggs et al., 2001). De plus, les restrictions liées à la protection de la vie privée - *le secret statistique* – engendrent de nombreuses lacunes dans les unités géographiques faiblement peuplées (INSEE, 2013). Ces deux raisons expliquent la logique territoriale des indicateurs et l'absence totale de données d'accès aux soins à l'échelle des IRIS, la faible disponibilité des indicateurs socio-économiques à cette même échelle, et le fait que certaines bases géographiques étatiques soient disponibles uniquement à des échelles encore plus grossières. C'est le cas par exemple des variables d'insécurité territoriale qui sont spécifiées à l'échelle des départements (ONDRP, 2012).

En science de l'environnement en revanche, les échelles macroscopiques et *a fortiori* l'approche territoriale sont controversées. La modélisation des substances physicochimiques s'effectue à partir de mesures géo-localisées de façon : soit à couvrir uniformément le territoire, c'est le cas des données météorologiques (Météo-France, 2010), soit à évaluer des flux de substances polluantes depuis leurs sources d'émission, c'est le cas des mesures de radioactivité environnementale à proximité des Installations Nucléaires de Base ou dans les zones géographiques où la radioactivité environnementale naturelle est plus importante (ANS et IRSN, 2013).

Les méthodes d'interpolation spatiale permettent de reconstituer les valeurs inconnues par des couches de surface dont les mailles sont généralement assez fines. Les géostatistiques sont particulièrement appréciées, en l'occurrence les techniques de krigeage uni-variables (Matheron, 1965) et multi-variables (Wackernagel, 2003). Dans l'état de l'art, il a été spécifié que leurs champs d'application sont nombreux dans le domaine de l'Environnement (Gay et Korre, 2006) ; (Gratton, 2002) comme dans celui de la Santé (Gaudart, 2007) (Goovaerts, 2006). L'agrégation de couches de surface à l'échelle des communes ou à celle des IRIS engendre une perte de précision considérable. Certains auteurs qualifient cette pratique de *grossière* et d'autres sont encore plus sévères. En expologie environnementale, lorsque des modèles multimédia interfacés par un SIG permettent d'estimer des doses d'expositions, la restitution des résultats est faite sur des mailles. Et pour cause, les frontières administratives n'ont aucun effet sur la diffusion spatiale de substances toxiques – *la logique territoriale est insensée* (Pistocchi A. Sarigiannis DA., Vizcanio P., 2010).

Cependant, le gain de précision apparent des modèles de *type dose-réponse* utilisés en expologie environnementale est spécieux. Lorsqu'on prend en compte la propagation des incertitudes dans le couplage de ces modèles, les doses d'expositions estimées varient sur des intervalles de confiance allant de la non exposition à la surexposition – *ces modélisations ne sont pas significatives* (Mercat-Rommens, Chojnacki et al., 2008)

De plus, l'analyse des interactions nécessite de fait une approche individus-centrée* et suppose donc d'isoler des zones géographiques *petites*. A cette échelle, une seconde incertitude surgit. Elle est due aux barrières techniques, économiques et juridiques liées à la reconstitution des *déplacements* et des *mobilités résidentielles* à court, moyen et long termes (Couet, 2006).

L'évaluation des risques d'exposition à des doses environnementales de substances nocives est très difficile à valider, voire même impossible. En contrepartie, l'agrégation à des échelles territoriales – plus grossières - des couches de surfaces modélisant les quantités de substances toxiques présentes dans les milieux environnementaux offre parfois un pouvoir explicatif curieusement élevé (World Health Organization, 2009).

Les échelles territoriales sont chères aux géographes car la géographie est une SHS et son but est précisément d'avoir des répercussions sur la vie sociale. En géographie de la santé il est question d'influer sur les conduites individuelles, les comportements collectifs socio-économiques et les pratiques médicales par des mesures politiques *justifiées* (Bloch et Ricordeau, 1996). En France, la lutte contre les inégalités en matière de santé environnementale* s'opère dans une logique territoriale. Les indicateurs spatiaux *des états de santé* permettent d'élaborer *des mesures politiques opérationnelles* légitimes puisqu'elles visent à réduire les disparités géographiques morbides *en agissant autant que faire se peut* sur les déterminants de santé (Salem, 1995).

Il en est ainsi pour l'application des mesures du PNSE et des Plans Cancer, dont les dimensions sociales et politiques de la logique territoriale sont aussi évidentes que nécessaires. Cependant, la logique des sciences de l'environnement relative au caractère irrationnel des frontières territoriales a atteint les sphères de la géographie de la santé. Les nouveaux indicateurs spatiaux d'accès aux soins sont désormais estimés sur des secteurs flottants avant d'être agrégés dans une logique territoriale (Muriel, Lucas-Gabrielli et al., 2012).
Et inversement, l'approche territoriale et son influence sur la vie politique et sociale ont atteint les sphères des sciences de l'environnement. En dépit des controverses, des indicateurs territoriaux sont construits à partir de couches de surface. Ces indicateurs spatiaux n'en sont pas moins robustes et intègrent même – par le biais de scenarii – des composantes socio-économiques et comportementales jusqu'alors inexplorées. Peuvent être cités à titre d'exemple les indicateurs de risques spatiaux (irs) liés à l'inhalation ou l'ingestion de substances chimiques qui sont agrégés à l'échelle des IRIS (Caudeville, Bonnard et al., 2012), et ceux d'exposition à l'activité volumique du radon dans l'air intérieur des habitations, de l'IRSN, agrégés à l'échelle des communes (Baysson,Tirmarche et al., 2004)

La dernière considération, et non des moindres, concerne la modélisation géographique des phénomènes morbides d'intérêt et aussi celle des FIM. Ces modélisations sont conditionnées par le repère géographique disponible dans la base LEA, en l'occurrence le Code Postal (CP) du lieu de résidence. En effet, cette information est plus grossière que le code commune mais permet néanmoins de spatialiser les patients à cette échelle modulo une incertitude résiduelle (chapitre.2).

Le choix de l'échelle est sous le joug d'une autre réalité de support : la Cohorte LEA constitue une population de petite taille dont les sujets sont éclatés sur une grande partie du territoire français. Par ailleurs, les causes des séquelles sont multifactorielles et les Facteurs Environnementaux* (FE) ont sensément un effet. Les épidémiologistes soupçonnent particulièrement les effets de la composante sanitaire. Le choix de l'échelle doit ainsi être adapté à cette prénotion, donc à la granularité* des indicateurs modélisant le recours aux soins (Michel, Auquier et al., 2007). Il convient de ne pas perdre de vue non plus le fait qu'il s'agit d'une recherche en géographie de la santé et qu'aujourd'hui il est indubitablement nécessaire de raisonner en terme *d'expositions environnementales combinées* (Leux et Guénel, 2010).

Ce chapitre a permis de montrer l'intérêt de la dialectique géographique dans les thématiques de santé environnementale* - *tant pour ses concepts que pour ses outils* et pour *son approche globale des phénomènes observés* (Amat-Roze, 2011). Et en mettant en perspective l'importance phénoménologique de raisonner à une échelle macroscopique, les prénotions épidémiologiques sur l'influence soupçonnée de *l'accès aux soins*, le fait que le CP de résidence est le seul repère géographique immédiatement disponible dans la Base LEA, l'importance de modéliser par des i.st.e* robustes la diversité de toutes les composantes environnementales à partir de données géographiques - sanitaires, socio-économiques, physicochimiques et épidémiologiques - dont la granularité* des sources est conflictuelle, l'importance de travailler avec une échelle uniforme et unique afin de pouvoir procéder à des analyses statistiques multidimensionnelles robustes, et enfin, le fait qu'il s'agit d'une recherche en géographie de la santé, et donc la nécessité de mener une analyse scientifique orientée vers des intérêts humains avec des

répercussions médicales et sociales susceptibles d'influencer les tendances politiques visant à réduire les inégalités de santé environnementale*. La prise en compte conjointe de toutes ces considérations conduit à proposer :

L'échelle communale, comme étant celle qui offre le meilleur compromis entre considérations théoriques et bibliographiques et réalité de support, i.e. inhérente à la granularité* des données disponibles mobilisées.

L'ensemble des considérations théoriques et pratiques permettant de fonder une dialectique méthodologique robuste a été spécifié. Celle-ci est vouée à la modélisation géographique des PM* et des FE/FIM* pertinents et Curieux*, puis à l'identification de DES. Les méthodes proposées seront appliquées aux données de la cohorte LEA pour les PM* (chapitre.2) et les FIM* (chapitre.3), et aux données environnementales mobilisées pour les FE* (chapitre.3).

CHAPITRE 2 : MODELISATIONS GEOGRAPHIQUES DE PHENOMENES MORBIDES

Précédemment une échelle d'investigation a été spécifiée au vu des hypothèses heuristiques émises, du positionnement scientifique adopté, d'arguments bibliographiques, et de considérations pratiques sur la granularité* des données disponibles.

Il convient désormais de proposer une méthode permettant de spatialiser des individus – en l'occurrence des patients – et qui soit adaptée à des données épidémiologiques de type Cohorte. Puis il s'agira d'élaborer des stratégies de modélisation géographique adaptées aux Phénomènes Morbides (PM), par le biais d'i.st.m* robustes.

Et enfin, d'imaginer un moyen de caractériser les espaces en fonction des Risques d'Expositions Géographiques (REG) morbides auxquels sont assujetties les populations locales. Pour illustrer ces propositions et garantir leur caractère reproductible, elles seront appliquées aux données de la Cohorte LEA.

SECTION A) METHODE DE SPATIALISATION ADAPTEE A DES DONNEES EPIDEMIOLOGIQUES

La méthode proposée s'appelle SpaLea. Elle est l'éponyme d'un algorithme dont l'objectif est de spatialiser, à l'échelle des communes, les individus de la Cohorte LEA. La stratégie proposée se veut robuste, parcimonieuse et reproductible. Elle permet aussi d'évaluer l'incertitude de spatialisation associée aux hypothèses et à la qualité des données disponibles. La méthode *SpaLea* est ensuite appliquée aux données de Cohorte LEA.

CARACTERISTIQUES, CONFLITS ET TRAITEMENTS DES DONNEES

La spatialisation des individus de la Cohorte LEA est conditionnée par les informations géographiques rattachées à chaque patient. En l'occurrence la seule information disponible est le Code Postal : x_i^{CP} du patient : I_i décliné lors de sa dernière évaluation de Qualité de Vie (QV) ou lors de sa dernière consultation médicale. De fait, certains Codes Postaux (CP) sont anciens. Ils changent au gré des trajectoires de vie des patients – mais cette information temporelle n'est pas disponible pour des raisons à la fois techniques et économiques (Afsset, 2009a).

La qualité de la spatialisation des patients dans l'espace géographique* français métropolitain est aussi conditionnée par la base SIG utilisée. En l'occurrence il s'agit de la BD-géofla2003. Il s'agit d'une base SIG spatiotemporelle statique. Autrement dit les unités spatiales U_u correspondent au découpage administratif des communes tel qu'il était en 2003. A chaque U_u ou commune sont appariés son code INSEE : $\mathrm{V}_{(\mathrm{U}_u)}$, sa population résidente : $\mathrm{P}_{(\mathrm{U}_k)}$ et sa surface géographique $\mathrm{S}_{(\mathrm{U}_k)}$ (IGN, 2004).

Il s'agit donc de spatialiser des patients disposant d'un x_i^{CP} à une date comprise entre 1980 et 2009 dans une base SIG temporellement statique. Deux biais fondamentaux de spatialisation ont été pris en compte : la différence de granularité* entre le x_i^{CP} et le $\mathrm{V}_{(\mathrm{U}_u)}$, et la variabilité spatiotemporelle de l'une et l'autre de ces deux informations géographiques.

CONFLITS SPATIOTEMPORELS CONSIDERES PAR SPALEA

La méthode SpaLea permet de traiter numériquement et de façon autonome deux types de conflits.

Le conflit de granularité* spatiale

Les codes INSEE de la BD-géofla2003 - notés $V_{(U_k)}$ - ne correspondent pas forcément aux codes INSEE associés aux Codes Postaux de résidence des patients. Les CP recouvrent parfois, même souvent, des zones géographiques plus vastes que les limites territoriales communales. Ils sont plus grossiers que les $V_{(U_k)}$ qui sont quant à eux régis par un *principe d'unicité temporelle* avec les communes. Autrement dit, il n'existe qu'un seul code INSEE par commune à une date précise (INSEE, 2012b).

Les CP sont rarement adaptés aux limites communales et encore moins aux *zones de vie*. Ils peuvent recouvrir plusieurs communes et même être « à cheval » sur d'autres. Le code postal est un code à cinq chiffres initialement inventé par la Poste pour convenir à la tournée d'un facteur à bicyclette depuis son bureau de référence. Les CP dénotent une imprécision spatiale qu'il convient de prendre en compte. Ainsi qu'il sera vu dans des développements ultérieurs, ils présentent toutefois l'avantage de recouvrir des zones parfaitement contigües (La Poste, 2012).

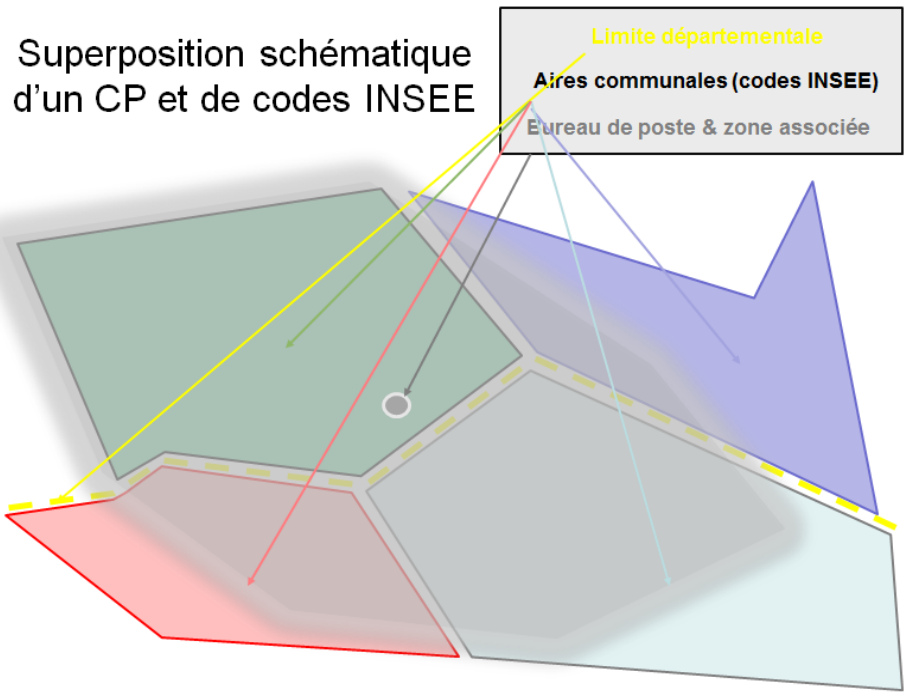


Figure 13 : Principe de superposition des CP et des Communes en France métropolitaine

Le conflit de granularité* temporel

Les codes postaux évoluent dans le temps, notamment en fonction de l'implantation des bureaux de poste et des moyens de transport. Cette volatilité s'observe aussi au niveau des codes INSEE dont la variabilité est encore plus importante. Pour des raisons socio-économiques et politiques, chaque année les découpages administratifs de certaines communes sont modifiés par le biais de processus de création/suppression ou de fusion/division (Bellin, Morin et al., 2011).

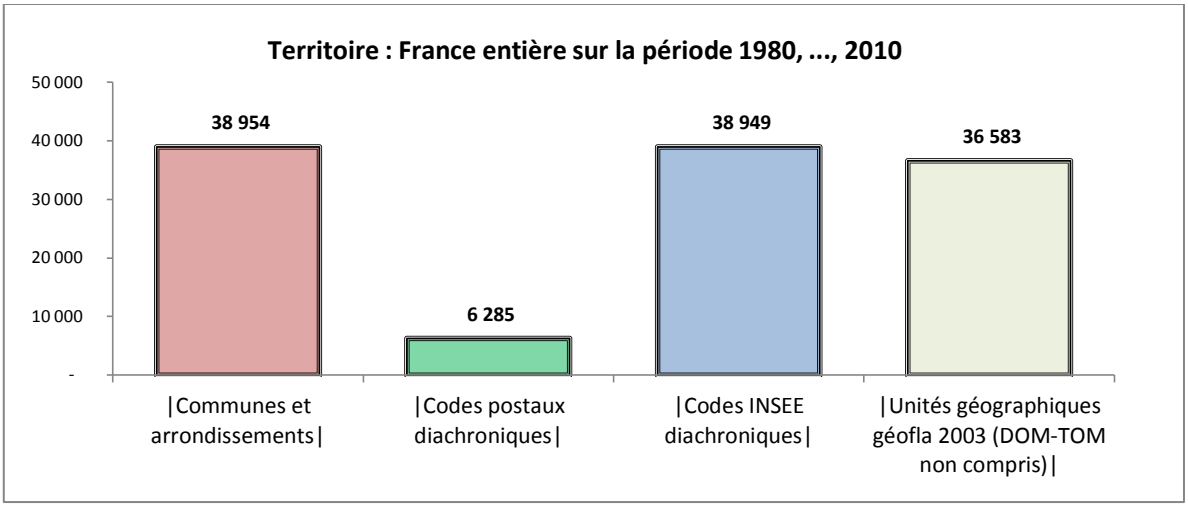


Figure 14 : Dénombrement diachronique ou temporel des communes, des codes INSEE, des CP et des Unités géographiques de la BD-géofla.2003 - en France métropolitaine.

Il existe un code INSEE unique par commune mais les grandes villes sont découpées en arrondissements, c'est le cas de Paris, Lyon, et Marseille. Les codes INSEE géofla 2003 sont ceux correspondant aux communes situées en France métropolitaine en 2003 uniquement, d'où le delta significatif avec les codes INSEE qui sont dénombrés sur l'intégralité du territoire français, i.e. Départements d'Outre-Mer (DOM) et Territoires d'Outre-Mer (TOM) compris.

TRAITEMENTS LIMINAIRES ET PATIENTS INCLUS

L'objectif est d'identifier, dans la mesure du possible, la commune de résidence U_k de chacun des patients à partir de son x_i^{CP} tout en tenant compte des conflits spatiotemporels inter-bases. Pour ce faire, de nombreux traitements ont été effectués, à commencer au niveau de la base LEA.

Traitements liminaires de la base LEA : certains x_i^{CP} sont « douteux » car ils n'ont pas de contrepartie dans la BD de La POSTE. Ceux ainsi concernés sont liés à des erreurs de remplissage des auto-questionnaires de QV*, des erreurs de saisie dans la base et, plus rarement, des Codes Postaux étrangers ou situés dans les DOM-TOM. La vérification de l'intégrité des CP *douteux* est rapide car peu de patients sont concernés. Les Assistants de Recherche Clinique (ARC) des centres de référence des patients ont accès aux fiches individualisées des patients qui permettent de savoir s'il s'agit d'un x_i^{CP} situé en dehors de la France métropolitaine ou d'un CP erroné, et dans ce dernier cas, de recontacter le patient. En revanche pour les lacunes les choses sont différentes car près de 200 individus de la Base LEA de 2009 étaient concernés. Il s'agit des patients qui n'ont pas été revus depuis 2003 et auprès desquels aucune information géographique n'a été collectée. De fait, un travail de reconstitution a été entrepris par l'EA-3279 et désormais tous les individus de la Base LEA 2010 disposent d'un x_i^{CP}. Mais la durée de reconstitution des lacunes a été longue et ces patients n'ont pas pu être intégrés dans l'analyse géographique.

Individus inclus dans l'analyse géographique :
Il s'agit de tous les patients dont le x_i^{CP} correspond à une zone géographique sise en France métropolitaine. Il y a deux possibilités pour qu'un individu soit exclu de l'analyse géographique. D'une part, il s'agit des ressortissants français résidant à l'étranger, en Principautés ou dans les DOM-TOM. Ces individus sont exclus car les variables ou les mesures utilisées pour modéliser la géographie des FE* de leurs communes sont, soit indisponibles, soit estimées de façon différente (au niveau des

techniques, des précisions, des temporalités, des densités d'échantillonnage...) et donc pas en mesure de caractériser l'environnement géographique d'une manière comparable à celles utilisées pour les patients situés en France métropolitaine. C'est le cas par exemple pour la radioactivité environnementale (ANS et IRSN, 2013), pour l'accessibilité aux soins (ARS, 2012), des i.st.e* d'exposition à des métalloïdes (INERIS, 2012) et pour certaines variables socio-économiques (INSEE, 2012c).

D'autre part, ont été exclus de l'analyse géographique les patients dont le CP était une lacune : $\{x_i^{\mathrm{CP}} = \phi\}$. En dépit du fait que cette variable a été reconstituée par les ARC, la quantité incommensurable des traitements numériques effectués à ce moment était telle qu'ils n'ont pas pu être réintroduits par la suite. De plus, il convient de remarquer que les x_i^{CP} reconstitués sont des CP de contact actualisés en 2010, et non ceux de la commune de résidence déclinés par les patients lors de leur dernière consultation médicale ou leur dernier entretien de QV*. L'information spatiotemporelle associée à chacune de ces variables est donc radicalement différente. Une reconstitution plus adéquate aurait induit des coûts exorbitants.

CONCEPTS, INDICATEURS ET SYNOPTIQUE DE LA METHODE SPALEA

Remarques liminaires

Les codes postaux peuvent sembler *a priori* inadaptés à une dialectique communale. Mais les incertitudes spatiales *affectent principalement les communes rurales faiblement peuplées* (La Poste, 2012). Or actuellement *plus de 80% des Français vivent dans des zones urbaines* (Blanpain et Chardon, 2008).

Ainsi qu'il a été vu, pour des raisons financières et temporelles, il est impossible de mobiliser les ARC pour procéder à une reconstitution diachronique des x_i^{CP} afin de reconstituer les trajectoires de vie des patients (Auquier, 2010).

En somme, la méthode SpaLea doit s'adapter à la parcimonie économique et aux lacunes de granularité* énoncées et se fonder sur des hypothèses et des concepts suffisamment robustes pour que la spatialisation des individus de la cohorte corresponde *à peu près* à la réalité géographique.

Le parti retenu

Il s'agit de spatialiser les individus de la cohorte dans des communes de 1ère espèce : $\mathrm{U_k}$, i.e. la commune qui a le plus de chances de correspondre au x_i^{CP}. On posera l'hypothèse selon laquelle il s'agit de la plus peuplée. Ensuite, il convient de modéliser l'incertitude géographique associée à la spatialisation des patients – dans des communes de 2nde espèce $\mathrm{U_{ko}}$ - i.e. toutes celles recouvertes par le x_i^{CP} mais qui ne sont pas éligibles au concept de 1ère espèce. Ce dernier se fonde sur *les probabilités qui sont, a priori, le meilleur moyen pour raisonner en situation d'ignorance partielle* (Borel, 1928).

Proposition 1 :

Le concept de spatialisation de 1ère espèce suppose que les patients résident dans la commune la plus peuplée de la zone géographique représentée par le code postal décliné. On appellera code INSEE de 1er ordre: $\boldsymbol{V_{i,1}}$, celui correspondant à la commune de 1ère espèce. Ce concept suggère aussi que même si le patient ne réside pas dans la commune de 1ère espèce en question, il y a une forte probabilité que ce soit dans celle-ci qu'il exerce son activité professionnelle ou des pratiques culturelles, sportives ou encore sociales.

En somme, on fait l'hypothèse que les individus s'y rendent régulièrement et par conséquent qu'ils sont assujettis, peu ou prou, aux FE* qui caractérisent leur état de santé. Par la suite on notera l'ensemble des unités géographiques de 1ère espèce comme suit :

$$\mathcal{U}_{.}^{1^{er}} = \left\{ \bigcup_{u=1}^{n_{(U_u)}} \left(U_u \middle| V_{(U_u)} \in \mathcal{V}_i^1\right) \middle| \mathcal{V}_i^1 = \bigcup_{i=1}^{n} \left(V_{i,1}\right) \right\}$$

Avec: $n_{(U_u)}$ l'ensemble des communes sises en France métropolitaine U_u ; Et $V_{(U_u)}$ le code INSEE de la Base-géofla2003 associé à chacune d'elle ; $\mathcal{V}_i^1$ l'ensemble des codes INSEE de 1er ordre associés aux patients spatialisés, au regard de leur x_i^{CP}.

- **Indicateur géographique associé à la spatialisation de 1ère espèce** : $SpaLea_{(U_k)}^1$ représente, dans chacune des unités géographiques communales U_k, la proportion de patients spatialisés et inclus dans l'analyse géographique, tel que :

$$SpaLea_{(U_k)}^1 = \mathbb{P}_{F_n}\left(V_{i,1} = V_{(U_k)} \middle| \{n \leq n_{LEA}\}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{V_{i,1}=V_{(U_k)}\}}$$

  Avec : n le nombre de patients inclus dans l'analyse géographique ; n_{LEA} le nombre total de patients inclus dans LEA. Il est défini : $\forall\ k \in \{1, \ldots, q_1\}$ avec q_1 le nombre total de communes de 1ère espèce.
- **Intérêts et limites :** le concept de spatialisation de 1ère espèce concentre les patients dans les unités géographiques les plus peuplées. De fait, le nombre d'individus spatialisés dans les U_k est surestimé. En contrepartie, cette stratégie augmente la consistance statistique, i.e. les effectifs de patients spatialisés, et permet d'effectuer une spatialisation à la fois rapide et parcimonieuse.

Proposition 2 : le concept de spatialisation de 2nde espèce permet de modéliser et de quantifier l'incertitude spatiale associée à la stratégie de spatialisation de 1ère espèce. Il s'agit d'identifier les communes de 2nde espèce, i.e. celles qui correspondent à un code INSEE de 2nd ordre, notées $V_{i,(k_o+1)}\ \forall\ k_o \in \{1, \ldots, q_{o,i}\}$. Autrement dit, les communes françaises U_u dont le code INSEE n'est pas éligible au concept de 1ère espèce mais qui sont néanmoins intersectées par les zones représentatives des x_i^{CP}. Ensuite l'idée du concept est d'évaluer, par le biais d'une probabilité d'occurrence, l'incertitude associée à la spatialisation des patients dans les communes de 1ère espèce : U_k qui se trouvent en contiguïté des communes de 2nde espèce. Par la suite on notera l'ensemble des unités géographiques de 2nde espèce comme suit :

$$\mathcal{U}_{.}^{2^{nd}} = \left\{ \bigcup_{u=1}^{n_{(U_u)}} \left(U_u \middle| V_{(U_u)} \in \mathcal{V}_i^2\right) \middle| \mathcal{V}_i^2 = \bigcup_{i=1}^{n} \bigcup_{k_o=1}^{q_{o,i}-1} \left(V_{i,(k_o+1)} \neq V_{i,1}\right) \right\}$$

Avec: $\mathcal{V}_i^2$ l'ensemble des codes INSEE de 2nd ordre associés à l'ensemble des patients inclus dans l'analyse géographique ; $q_{o,i}$ le nombre de codes INSEE de 2nd ordre attribués aux patients: I_i.

- **Indicateur géographique associé à la spatialisation de 2nde espèce** : $SpaLea_{U_{k_o}}^2$ permet de modéliser dans les communes de 2nde espèce : U_{k_o}, par le biais d'une probabilité conditionnelle empirique, l'incertitude associée au concept de spatialisation de 1ère espèce.

$$SpaLea_{U_{k_o}}^2 = \mathbb{P}_{F_n}\left(V_{(U_k)} \in \mathcal{U}_{.}^{2^{nd}} \middle| V_{(U_k)} \notin \mathcal{U}_{.}^{1^{er}}\right) = \frac{\sum_{i=1}^{n}\sum_{k_o=1}^{q_{o,i}} \mathbb{1}_{\{V_{i,(k_o+1)}=V_{(U_k)}\}\cap\{V_{i,(k_o+1)}=V_{i,1}\}}}{\sum_{k=1}^{q_1}\sum_{i=1}^{n}\sum_{k_o=1}^{q_{o,i}} \mathbb{1}_{\{V_{i,(k_o+1)}=V_{(U_k)}\}\cap\{V_{i,(k_o+1)}\neq V_{i,1}\}}}$$

- **Intérêts et limites :** le concept de spatialisation de 2nde espèce présente l'avantage de supputer un indicateur de contigüité spatiale permettant de discuter la robustesse de la spatialisation des patients dans les communes de 1ère espèce : U_k. Il permet d'évaluer indirectement l'erreur spatiale associée à la granularité* des x_i^{CP} mais ne présuppose rien sur l'incertitude spatiale de la variabilité temporelle des CP et des codes INSEE.

SYNOPTIQUE DES TRAITEMENTS ET PRESENTATION DES RESULTATS

L'algorithme SpaLea, éponyme de la méthode, permet de spatialiser les patients de la Cohorte LEA dans les « communes de 1ère espèce ». Il peut être appliqué à n'importe quelle autre cohorte, ou groupe, si tant est que l'information associée à chaque individu soit un CP. L'algorithme est entièrement autonome dès lors que les données inputs sont uniformisées. SpaLea a été programmé en Visual Basic (Premium Consultants, 2008)

Les données inputs :
Les inputs utilisés par l'algorithme sont : une table diachronique des Codes Postaux entre 2003 et 2010 (La Poste, 2012) ; une table des correspondances Codes Postaux - codes INSEE existants entre 1980 et 2010 ; une table actualisée des modifications diachroniques des codes INSEE entre 2003 et 2010 (INSEE, 2012b) ; et l'extraction SIG de la table attributaire de la couche commune de la BD.géofla2003 (IGN, 2004).

Synoptique de l'algorithme SpaLea

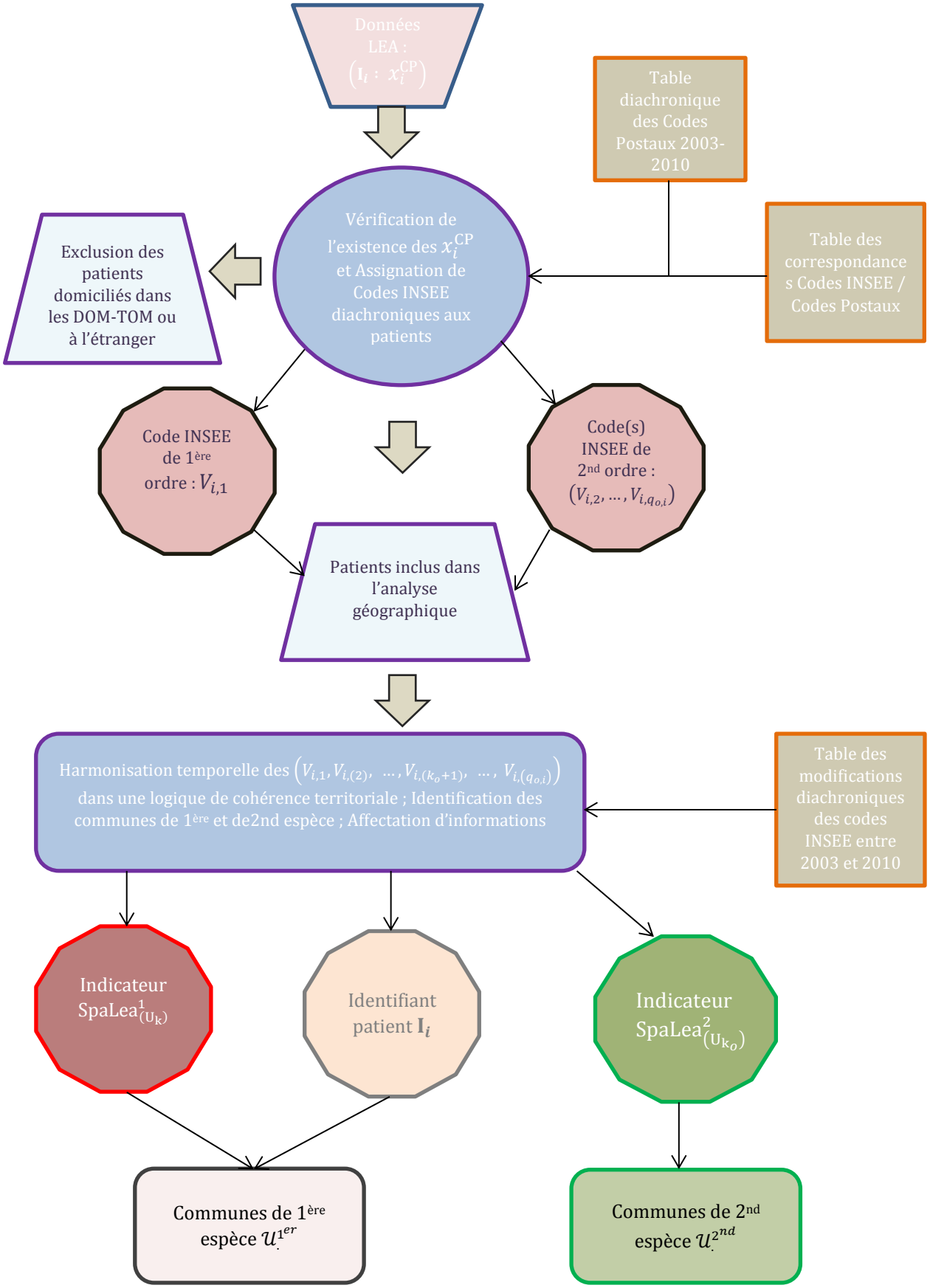


Figure 15 : Chaine synoptique des traitements numériques effectués par l'algorithme SpaLea

L'algorithme SpaLea permet de spatialiser les patients de la Cohorte LEA dans les unités géographiques dites de 1ère espèce $(U_k)_{k=\{1, ..., q\}}$ d'un SIG avec une logique de cohérence territoriale. L'indicateur $SpaLea^{1}_{(U_k)}$ permet d'identifier les U_k sur les cartes.

Il n'y a pas lieu de l'interpréter car le nombre de patients spatialisés est lié à la *file active,* donc aux dates d'inclusion des centres de référence dans le projet LEA. Pour cette raison $SpaLea^{1}_{(U_k)}$ ne permet pas de comparer les U_k ni de poser d'hypothèse sur les zones où il y a plus de leucémies en rapportant cette valeur à la population *in situ*. En revanche, il permettra par la suite d'estimer une incertitude statistique qui interviendra lors de la modélisation géographique des phénomènes morbides d'intérêt.

$SpaLea^{2}_{U_{k_o}}$ permet d'apprécier et de modéliser l'incertitude spatiale associée aux hypothèses de la méthode SpaLea et à la granularité* des données épidémiologiques. L'analyse de ce dernier permet d'apprécier la robustesse de cette méthode parcimonieuse.

APPLICATION AUX DONNEES DE LA COHORTE LEA, INCERTITUDES ET ROBUSTESSE DE LA METHODE

La méthode SpaLea a été appliquée aux données de la Cohorte LEA. Il s'agit de présenter les résultats de la spatialisation des patients et d'analyser les incertitudes associées aux hypothèses et à la granularité* des *inputs* ainsi que d'étudier la robustesse de SpaLea.
La spatialisation des individus dans l'espace géographique* est une phase préalable nécessaire lorsqu'on veut analyser les interactions spatiotemporelles entre la santé et l'environnement.

RESULTAT DE L'APPLICATION DE SPALEA AUX DONNEES LEA

SpaLea a été appliquée aux données épidémiologiques de la Base LEA 2009 consolidée. Sur 943 patients, seulement 747 ont été spatialisés dans 421 communes de 1ère espèce. Les 196 patients exclus se subdivisent en deux groupes.
Le premier est constitué de 5 individus qui ont déclaré vivre soit à l'étranger, soit dans les DOM-TOM. Et le second comprend les 191 patients qui n'ont pas été revus depuis 2003 et ne disposant pas de x_i^{CP}, puisqu'à cette date aucune information géographique n'était demandée. Bien que les lacunes de la variable x_i^{CP} aient été reconstituées par les ARC, en fin d'année 2012 le nombre de traitements effectués – à cette date - sur les données épidémiologiques et environnementales ne permettait plus de les intégrer.

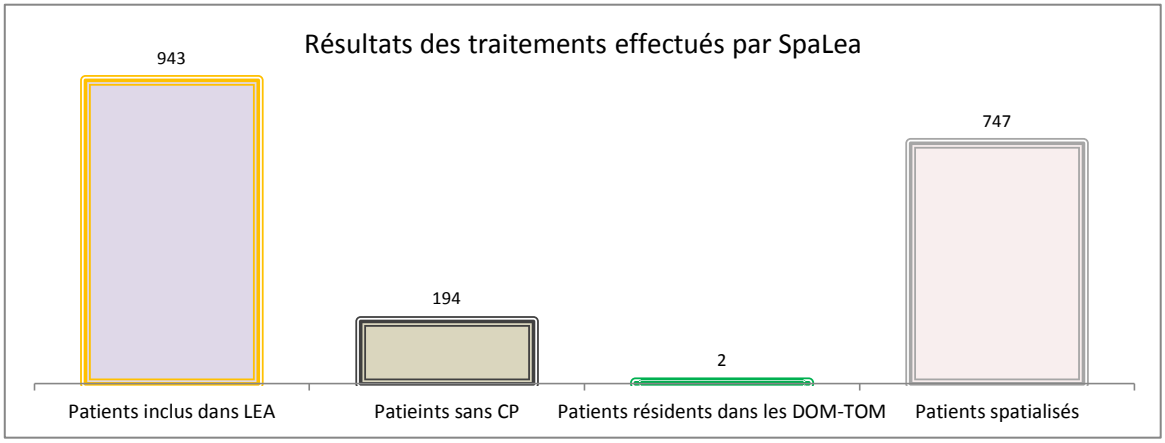


Figure 16 : Résultats statistiques obtenus par application de SpaLea sur la Base LEA 2009.

Les Caractéristiques Individuelles et Médicales*(CIM) de l'échantillon de patients spatialisés sont très proches de celles de l'intégralité des patients inclus dans la Base LEA 2009 – qui ont été rappelées ci-après entre parenthèses.

Parmi les 747 patients inclus dans l'analyse géographique qui ont été spatialisés dans 421 de 1ère espèce, on dénombre 417 garçons pour 330 filles. La moyenne d'âge au moment du diagnostic de la leucémie est de 6,2 ans (6,5 ans). En 2010 les individus concernés étaient, en moyenne, âgés de 19,1 ans (20,5 ans) et tous déclarés en vie à cette époque. Or sur l'intégralité de la cohorte 4 patients sont décédés.

S'agissant des séquelles d'intérêt estimées sur les patients spatialisés, 13,0% d'entre eux ont développé des cataractes (CATA : 12,5%) ; 10,2% des patients ont été traités pour une tumeur thyroïdienne (THYR, 9,4%) et enfin 4,02% pour une tumeur secondaire majeure (TUM2, 4,03%).

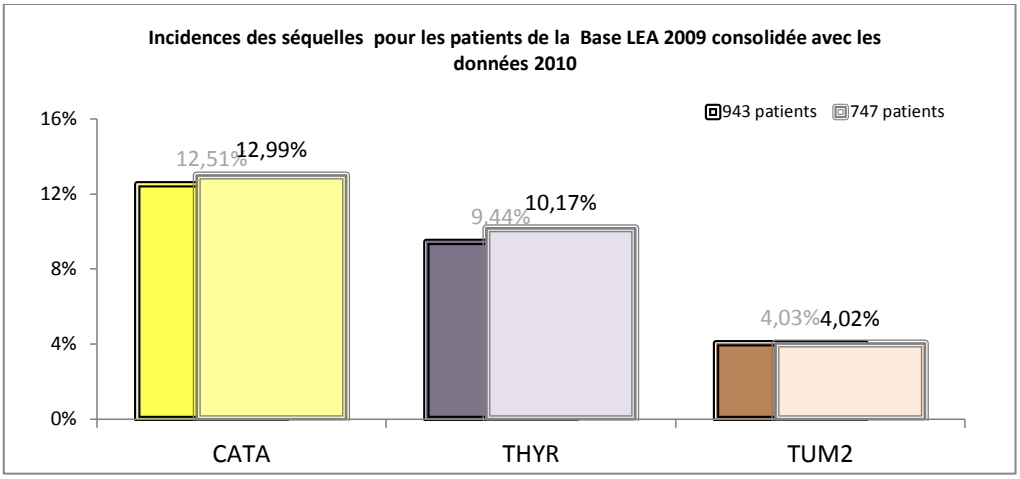


Figure 17 : Histogramme des incidences morbides calculées sur les patients spatialisés comparées à celles calculées sur l'intégralité de la cohorte.

Les incidences estimées sur les patients spatialisés sont légèrement plus fortes que celles estimées sur l'intégralité de la cohorte. La raison vient du fait que les patients spatialisés ont une participation plus active au sein de la cohorte. Leurs séquelles sont donc mieux diagnostiquées ou diagnostiquées plus tôt.

Parmi la population spatialisée, 86,3% (85,7%) des patients ont été traités pour une LAL* contre 13,7% (14,3%) pour une LAM. Concernant les CIM* qui augmentent potentiellement les risques de séquelle, la population spatialisée est constituée de 27,0% (27,3%) de sujets qui ont subi une greffe, et 19,4% (19,8%) qui ont reçu une irradiation corporelle totale. Quant à la proportion de rechute, elle est de 16,3% (16,0%). Enfin, parmi les patients spatialisés participant activement à l'étude de qualité de vie, 48% disent ne pas pratiquer de sport du tout – alors que cette proportion s'élève à seulement 25% lorsqu'elle est estimée sur l'intégralité de la cohorte.

Sur le plan géographique les répartitions régionales des patients spatialisés dépendent des dates d'inclusion des centres de référence dans le projet LEA. Pour rappel, en 2009 les centres concernés se situent dans les villes de Nice, Marseille, Grenoble, Clermont-Ferrand et Nancy. Les proportions de patients spatialisés dans chaque région sont données dans le tableau 17 :

Nom des régions	∑(SpaLea1er_Uk)
AUVERGNE	10,17%
ALSACE-LORRAINE	19,54%
PACA	49,93%
CORSE	2,14%
RHONE-ALPES	8,70%
AUTRES REGIONS	11,65%

Tableau 17 : Proportion de patients dans les régions où sont situées des communes de 1ère espèce

Parmi les 747 patients spatialisés, l'essentiel la cohorte est localisé dans des communes sises en Alsace-Lorraine (20%) et en PACA (50%). Il s'agit des patients inclus par le biais des trois centres de référence historiques pour le traitement des leucémies infantiles : Marseille, Nice et Nancy. Le projet LEA est né en PACA, les centres qui s'y trouvent disposent donc d'une file active beaucoup plus longue que les autres, ce qui explique le nombre élevé de patients. Il est donc impossible de raisonner en terme d'incidence géographique régionale et encore moins communale. De plus, parler d'incidence des

leucémies nécessiterait de ramener ces ratios à l'échelle des populations communales du même âge. En outre, la région PACA est aussi la plus peuplée de toutes celles citées.

Dans les régions Rhône-Alpes et Auvergne les patients spatialisés sont ceux des centres de Grenoble et de Clermont-Ferrand. Quant au pourcentage des patients spatialisés dans les autres régions (11%), il résulte essentiellement de déménagements ou de patients traités et suivis dans les centres de Marseille ou de Nice qui sont très réputés.

Par conséquent les cartographies résultant des modélisations géographiques associées aux phénomènes morbides et par suite, aux FE* ayant des interactions potentielles, seront présentées uniquement pour le Sud-Est et le Nord-Est de la France. Elles se focaliseront plus précisément sur les $\mathcal{U}_{.}^{1^{er}}$ sises en PACA et en Alsace-Lorraine.

L'indicateur $\text{SpaLea}^1_{(\text{U}_\text{k})}$ permet donc uniquement d'identifier l'ensemble des $\mathcal{U}_{.}^{1^{\text{er}}}$ dans lesquelles des patients sont spatialisés, rien de plus. Il constitue toutefois la première étape inhérente à la modélisation de la géographie des séquelles développées après le traitement d'une leucémie chez l'enfant et à l'analyse de leurs interactions avec l'environnement. La section suivante explique en quoi le couplage d'une transformation en patients-années et d'une métrique floue géographique permettent de donner plus de consistance à la modélisation géographique des séquelles, d'augmenter virtuellement la taille de la cohorte et par la même occasion de pallier, en partie au moins, le problème d'inconsistance* statistique liée à la phase de spatialisation.

Toutefois il s'agit là d'une proposition méthodologique. La stratégie d'estimation des i.st.m* proposés est donc discutable. Mais la dialectique est géographique et elle doit avant tout être pensée dans une logique spatiotemporelle. De plus, la méthode proposée est indubitablement valide pour des cohortes de taille plus importante. Par conséquent, l'identification des Déterminants Environnementaux de Santé* sera d'abord appliquée dans une logique géographique, i.e. sur les i.st.m* et les i.st.e* proposés (chapitre.4). Puis dans une logique individus-centrée*, donc directement sur les séquelles des patients y_i^j et après leur avoir attribué les FE/FIM* caractéristiques de l'environnement communal de leur lieu de résidence : $x^{\text{j:FE/FIM}}_{(i|\text{U}_\text{k})}$.

En contrepartie, l'analyse de l'indicateur $\text{SpaLea}^2_{\text{U}_{\text{k}_o}}$ présente un intérêt immédiat certain puisqu'il permet d'estimer la robustesse associée à l'incertitude spatiale de la méthode SpaLea et d'identifier les patients spatialisés de façon incertaine et auxquels il convient de porter une attention particulière.

ESTIMATION DE LA ROBUSTESSE DE LA METHODE SPALEA

L'objectif est d'évaluer la robustesse de la méthode SpaLea et par la même occasion d'identifier les patients spatialisés de façon incertaine et auxquels il convient de porter une attention particulière. *Dans les sciences de l'environnement, l'analyse d'incertitude est devenue une étape incontournable dans la restitution des résultats, tant en terme de rigueur scientifique que de crédibilité sociale* (Mercat-Rommens, Chojnacki et al., 2008)

L'indicateur spatial $\text{SpaLea}^2_{\text{U}_{\text{k}_o}}$ permet d'évaluer l'incertitude de spatialisation des patients par une probabilité d'occurrence matérialisant, dans les communes de 2nde espèce U_{k_o}, la possibilité qu'un ou plusieurs patients soient spatialisés à tort dans une commune de 1ère espèce : U_k limitrophe.

Il a été vu précédemment que les $\text{SpaLea}^2_{\text{U}_{\text{k}_o}}$ sont estimés à partir des codes INSEE de 2nd ordre : $\mathcal{V}_\text{i}^2$ appariés à chacun des patients. L'analyse de la distribution des $\mathcal{V}_i^2$ permet de se faire une première idée

de l'incertitude de spatialisation commise. Le graphique ci-dessous représente les principaux paramètres statistiques de cette distribution.

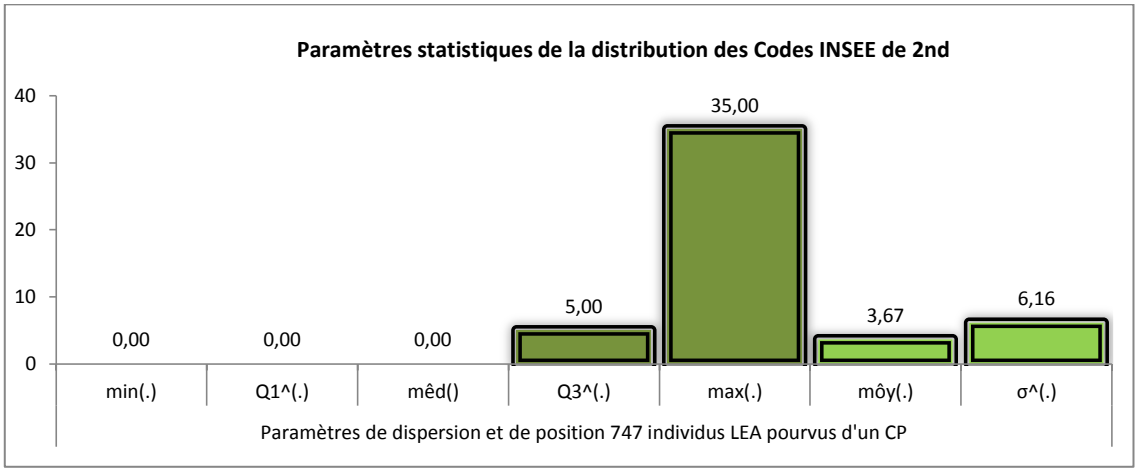


Figure 18 : Valeurs des principaux paramètres statistiques représentatifs de l'appariement des codes INSEE de 2nd ordre aux patients spatialisés

Ce graphique montre que plus de 50% des patients inclus dans l'analyse géographique sont spatialisés de façon *certaine*. Et que plus de 75% des patients sont spatialisés de façon quasi-certaine puisqu'ils ont au maximum cinq codes INSEE de 2nd ordre qui leur ont été appariés.
Par contre, environ 25% d'entre eux sont spatialisés dans des U_k mais auraient pu être spatialisés dans des U_{k_o} dont le nombre varie entre 5 et 35.

Une analyse plus fine des $\mathcal{V}_i^2$ a permis de montrer que près de 7.5% des patients ont un nombre important de codes INSEE de 2nd ordre appariés $q_{o,i}$. En l'occurrence, la méthode SpaLea peut être soupçonnée d'être :

- *Apparemment douteuse* lorsque le nombre de codes INSEE de 2nd ordre est tel que : $q_{o,i} \in [\![15, 25 [\![$ ce qui représente 6.1% des individus spatialisés, soit 45patients.
- *Apparemment incertaine* lorsque le nombre de codes INSEE de 2nd ordre est tel que : $q_{o,i} \in [\![25, 36 [\![$, i.e. 1.5% des individus spatialisés, soit 11patients.

A ce stade de l'analyse on remarque que la proportion de patients *soupçonnés* d'avoir été spatialisés de façon *apparemment incertaine* est dérisoire. Et que celle *soupçonnée d'être apparemment douteuse* est faible.

Toutefois, il convient de remarquer qu'il s'agit là d'une analyse effectuée dans une logique *individus-centrée** or la dialectique est spatiale, donc cette incertitude doit être analysée dans une logique *géographique* et c'est justement ce que permet de faire $SpaLea^2_{U_{k_o}}$, d'autant que la phase de spatialisation implique que les patients sont regroupés dans des U_k et que par conséquent, l'analyse effectuée sur les codes INSEE de 2nd ordre surestime l'erreur de spatialisation.

Afin d'étudier plus précisément la robustesse des hypothèses associées à la méthode SpaLea, il convient d'étudier les propriétés spatiales de $SpaLea^2_{U_{k_o}}$, i.e. ses valeurs intrinsèques mais aussi et surtout la localisation des U_{ko}, leur nombre et les surfaces associées. Pour ce faire deux analyses SIG ont été effectuées :

- La première est une analyse spatiale itérative de contiguïté visant à évaluer la robustesse spatiale de la méthode SpaLea.
- La seconde est une analyse spatiale statistique d'identification par seuillage des Uk pour lesquelles il convient de vérifier la spatialisation des patients

ANALYSE SPATIALE ITERATIVE DE CONTIGUÏTE

L'objectif de l'analyse spatiale itérative de contiguïté

Il s'agit de vérifier que les communes de 2nde espèce : U_{k_o} forment bien des zones géographiques contigües – ou au moins partiellement contigües autour des communes de 1ère espèce : U_{k}, lorsqu'on injecte séquentiellement les U_{k_o} sachant leur rang et que cette hypothèse reste vraie lorsque les surfaces territoriales autour des U_{k} sont déjà saturées par les injections itératives des U_{k_o} de rang plus faible.

Cette procédure d'analyse spatiale itérative a été mise œuvre grâce au logiciel SIG ArcGis.10 via l'outil ModelBuilder qui sera à nouveau énoncé dans la partie dévolue à la modélisation des FE* (ArcGIS: ModelBuilder, 2013).

Le principe de l'algorithme spatial pour estimer la *contiguïté spatiale approximative de rang* se fonde sur la procédure suivante :

1. Les indicateurs $\mathrm{SpaLea}^2_{\mathrm{U}_{k_o}}$ sont injectés de façon itérative sachant le rang du code INSEE $\mathcal{V}_{\mathrm{i}}^2$ associé à chacun des patients. Autrement dit, pour toutes les U_{k_o} associées au code INSEE de 2nd ordre du patient, i.e. au regard de l'ensemble des $V_{i,(k_o)}$ appariés aux patients spatialisés, $q_{i,o} \in \{1, \dots, 35\}$. Pour rappel, les $V_{i,(k_o)}$ ont été appariés en fonction du rang de la commune, i.e. de sa quantité de population $\mathrm{P}_{(\mathrm{U}_{k_o})}$.
2. A chaque itération $\{\mathrm{step}\} \in \{0, \dots, (q_{i,o} - 1)\} = \{0, \dots, 34\}$ l'algorithme calcule et stocke la proportion de U_{k_o} ne se trouvant pas en contiguïté spatiale des blocs formés par l'ensemble des U_{k} et des U_{k_o} au *pas de calcul* {step} et au *pas de calcul* précédent{step − 1}. Cette proportion représente un facteur de discontigüité spatiale, il est noté : δ_{step} et s'estime tel que :

$$\delta_{\mathrm{step}} = \frac{\left|\mathrm{U}_{\{0,\dots,\mathrm{step}\}} \ominus \mathrm{U}_{\{0,\dots,\mathrm{step}-1\}}\right|}{\left|\mathrm{U}_{\{0,\dots,\mathrm{step}\}}\right|}; \quad \mathrm{step} \in \{1,..,35\}$$

   Avec {|·| = "la fonction cardinale"} ; Et $\mathrm{U}_{\{0,\dots,\mathrm{step}\}}$ l'union des U_{k}et les U_{k_o} agrégées en blocs lorsqu'elles forment des zones géographiques contigües, *au pas de calcul* step.

Cet algorithme spatial permet d'estimer la robustesse de l'indicateur $\mathrm{SpaLea}^2_{\mathrm{U}_{k_o}}$ et sa capacité à modéliser l'incertitude associée à la spatialisation des patients dans les U_{k}.

Proposition : Si la valeur maximale des facteurs de discontigüité spatiale : δ_* reste *petite* alors l'hypothèse de *contiguïté spatiale approximative* de $\mathrm{SpaLea}^2_{\mathrm{U}_{k_o}}$ est validée, i.e. :

$$\delta_* = \left\{ \max_{\forall \mathrm{step} \in \{1,\dots,34\}} \left(\delta_{\mathrm{step}=} \left| \mathcal{V}_{\cdot}^2 = V_{(i=1|k_o,q_{i,o})}, .., V_{(i=\mathrm{n}|k_o,q_{i,o})} \right. \right) = \mathrm{petit} \right\}$$

Schéma de principe cartographique du processus : estimation du facteur de discontigüité spatiale à l'itération numéro 3, i.e. lorsque {step = 2}.

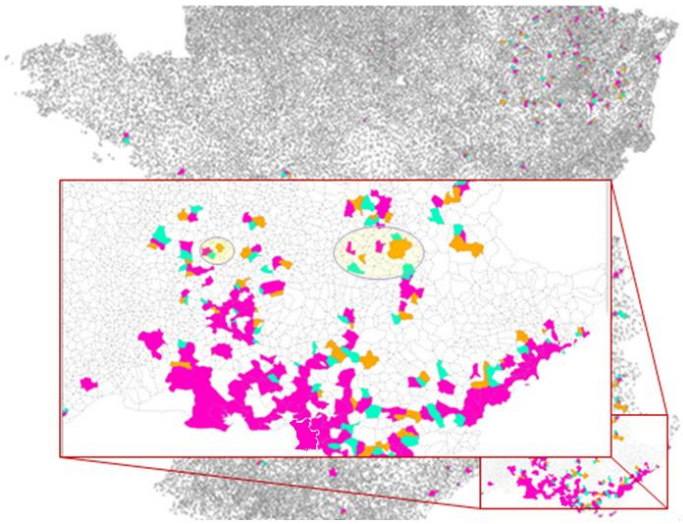


Figure 19 : Schéma de principe de l'injection séquentielle des U_{k_o} dans le SIG et mise en exergue des zones discontigües.

Les ronds jaunes, dans le *zoom-in,* mettent en exergue les zones géographiques où U_{k_o} injectées au pas de calcul : 1, puis : 2 ne forment pas des blocs ; Les U_k sont injectées à l'initialisation. La procédure se déroule de la façon suivante :

- A l'initialisation $\{step = 0\}$, seules les communes de 1ère espèce sont injectées $U_{\{step=0\}} \stackrel{def}{=} U_k$ et constituent l'ensemble U_0 [en magenta] ; δ_0 n'est pas estimable.
- Au premier pas de calcul $\{step = 1\}$ $U_{\{step=1\}} \stackrel{def}{=} \left(U_k, U_{\{k_o|q_{i,o}=1\}}\right)$ lesU_k et les U_{k_o} de rang : 2 sont injectées, ces dernières sont matérialisées [en cyan]. L'algorithme calcule le facteur de discontinuité spatiale, qui à ce moment vaut $\delta_1 \approx 0{,}5\%$
- Au second pas de calcul $\{step = 2\}$ $U_{\{step=E\}} \stackrel{def}{=} \left(U_k, U_{\{k_o|q_{i,o}=2\}}, U_{\{k_o|q_{i,o}=3\}}\right)$ lesU_k et les U_{k_o} de rang:3 sont injectées, ces dernières sont matérialisées [en orange]. L'algorithme calcule le facteur de discontinuité spatiale, qui à ce moment vaut $\delta_2 \approx 0{,}4\%$
- …
- La procédure est iétéré ainsi jusqu'au pas de calcul $\{step = 34\}$

Remarque : le δ_{step} n'est pas une fonction croissante du pas d'injection des $U_{\{k_o|q_{i,o}\}}$. A chaque itération de nouvelles contigüités et discontigüités spatiales se forment, raison pour laquelle $\delta_2 < \delta_1$.

Résultats obtenus :
L'analyse visuelle des blocs formés par les $U_{\{step\}}$ et des valeurs de δ_{step} a permis de montrer que l'injection itérative des U_{k_o} , en fonction de leur rang, forme des zones géographiques approximativement contiguës avec U_k.
A la dernière itération les blocs formés par les U_k et l'ensemble des U_{k_o} constituent des blocs de communes parfaitement contiguës, de sorte que $\{\delta_{34} = 0\%\}$.
De plus, Les valeurs prises par δ_{step} ont toujours été très faibles, à tel point que le facteur de discontiguïté maximale, est tel que : $\delta_* \ll 1\%$.

Analyse des résultats :
Au vu des résultats obtenus et si l'on concède à l'hypothèse selon laquelle moins une commune est peuplée et moins les patients ont de chances d'y résider, alors l'indicateur $SpaLea^2_{U_{k_o}}$ permet d'estimer de façon robuste l'incertitude spatiale à la phase de spatialisation des patients dans les U_k.

Conclusion :
La validation de l'hypothèse de contiguïté spatiale approximative de la méthode SpaLea, à partir de l'indicateur $SpaLea^2_{U_{k_o}}$ permet d'affirmer que même si la spatialisation d'un patient est erronée, il y a de très grandes chances pour qu'il soit spatialisé dans une commune limitrophe. Et que cette probabilité peut être approchée par $SpaLea^2_{U_{k_o}}$ qui fait de lui un indicateur d'incertitude robuste d'un point de vue spatial.
Par conséquent, la méthode SpaLea est adaptée à la spatialisation d'individus à l'échelle des communes et elle est reproductible à toutes les populations lorsque la seule information disponible est un x_i^{CP} à un moment donné.
Par la suite ces conclusions seront reprises pour valider la robustesse de SpaLea. Mais il convient d'abord d'effectuer une analyse statistique des valeurs prises par $SpaLea^2_{U_{k_o}}$ et de démontrer sa capacité à détecter les patients dont la localisation géographique doit être vérifiée.

IDENTIFICATION DES PATIENTS MAL LOCALISES

L'analyse spatiale statistique des valeurs prises par $SpaLea^2_{U_{k_o}}$ a pour dessein d'identifier, par seuillage, les communes de 1ère espèce : U_k se trouvant en contiguïté spatiale de blocs communes de 2nde espèce U_{k_o} dont les valeurs de $SpaLea^2_{U_{k_o}}$ sont *fortement élevées*. Par conséquent, on peut identifier les patients dont le risque statistique d'une spatialisation erronée est fort.

Il convient de remarquer que le risque d'une spatialisation erronée ne dépend pas uniquement des valeurs prises par $SpaLea^2_{U_{k_o}}$ mais du nombre de U_{k_o} associées et des surfaces communales concernées.
En proposant une stratégie de sélection par seuillage statistique sur $SpaLea^2_{U_{k_o}}$ on peut détecter les communes U_k où l'erreur de spatialisation des patients est la plus grande. Mais il est nécessaire de demander aux ARC de vérifier l'adéquation du code INSEE de 1er ordre avec la commune où résidait réellement le patient à cette époque – ce qui nécessite de le contacter et de l'interroger à ce sujet - notamment au regard du nombre et des surfaces des U_{k_o} associées aux U_k les plus incertaines.

En effet comme on s'intéresse à l'effet de l'environnement géographique communal sur les états de santé à cette même échelle, si certains des patients de U_k sont *très probablement mal spatialisés* il convient de s'assurer que ce risque ne concerne pas une seule et malheureuse U_{k_o} limitrophe de petite taille.

Les cartographies subséquentes présentent simultanément les indicateurs $SpaLea^1_{U_k}$ et $SpaLea^2_{U_{k_o}}$ et montrent que le risque d'une spatialisation erronée est marqué par une forte composante régionale.

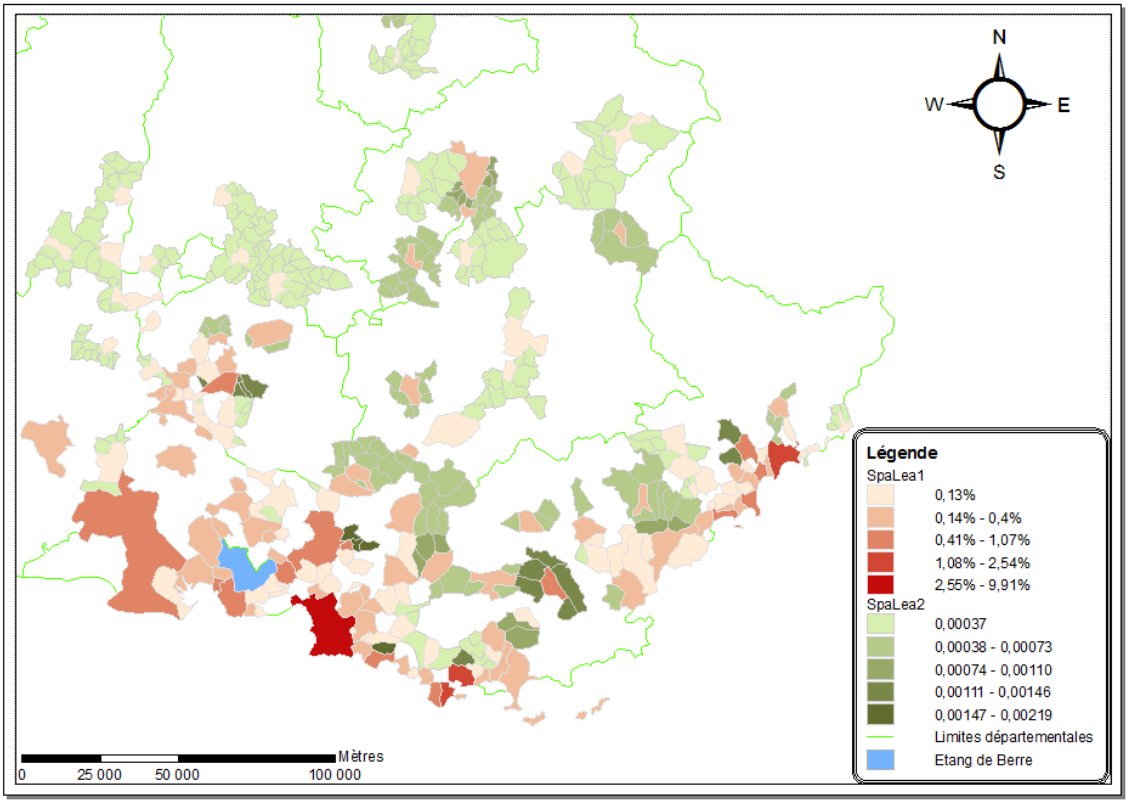


Figure 20 : Cartographie des indicateurs $SpaLea^{1}_{U_k}$ et $SpaLea^{2}_{U_{k_o}}$ dans les communes de la région PACA

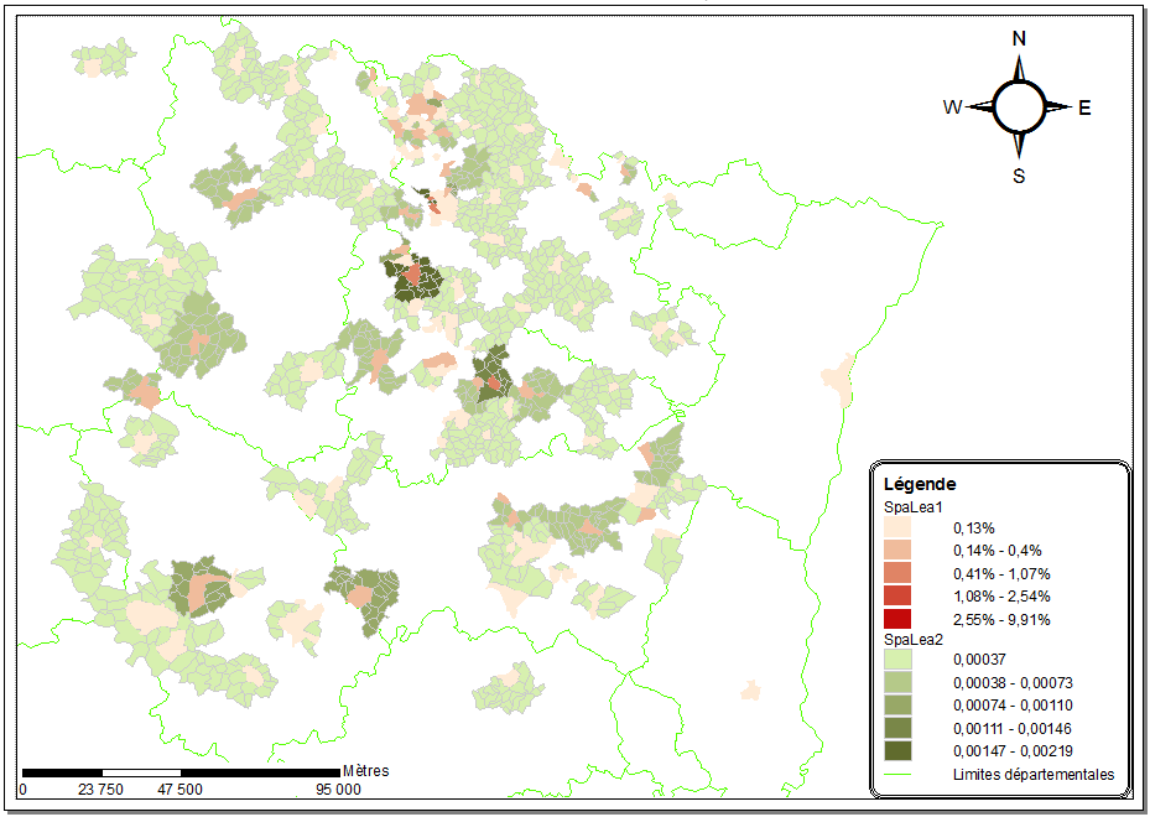


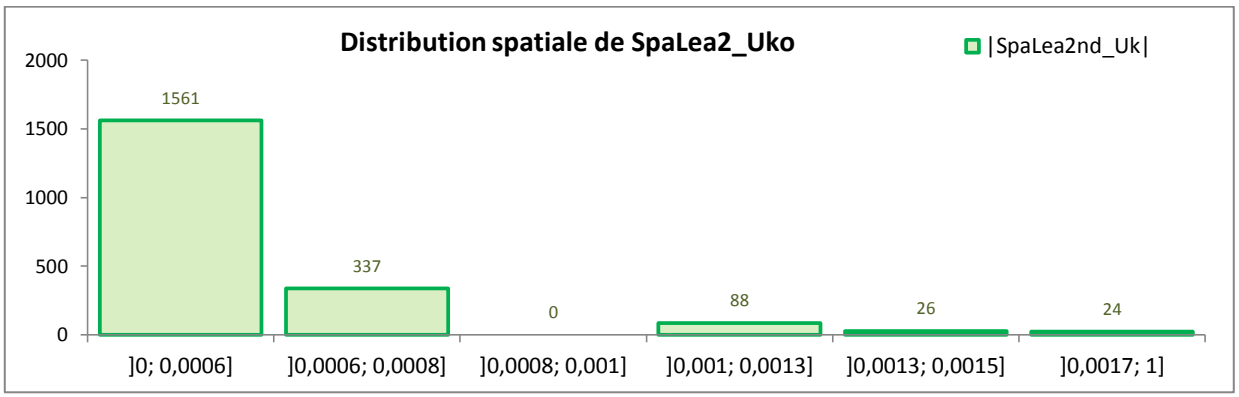
Figure 21 : Cartographie des indicateurs $SpaLea^{1}_{U_k}$ et $SpaLea^{2}_{U_{k_o}}$ dans les communes des régions Alsace et Lorraine

Ces cartographies montrent que les valeurs prises par $SpaLea^{2}_{U_{k_o}}$ et surtout le nombre de U_{k_o} ainsi que leur taille, forment des zones géographiques d'incertitude singulièrement différentes selon que les patients ont été spatialisés en région Alsace-Lorraine ou en PACA. Ce constat renforce l'idée que la

sélection des patients à réinterroger ne doit pas uniquement porter sur les valeurs prises par $SpaLea^2_{U_{k_o}}$.

Toutefois $SpaLea^2_{U_{k_o}}$ reste un indicateur spatial robuste pour caractériser les incertitudes de spatialisation associées à la méthode SpaLea. L'objectif est donc de l'utiliser afin détecter les U_{k_o}, donc les patients dont la localisation du lieu de résidence *est très probablement erronée*.

Afin de fixer un seuil statistique de sélection des U_{k_o} les plus propices à accueillir des patients mal spatialisés, il convient de s'intéresser aux valeurs prises par $SpaLea^2_{U_{k_o}}$. La figure 21 représente l'histogramme de sa distribution spatiale et un tableau des principaux indicateurs statistiques estimés sur l'intégralité de la France métropolitaine.



Unités géofla	Paramètres de dispersion et de position de SpaLea2er_Uk						
Paramètres statistiques	min(.)	Q1^(.)	mêd()	Q3^(.)	max(.)	môy(.)	σ~(.)
Estimation sur SpaLea2	0,00036	0,00036	0,00036	0,00036	0,00219	0,00062	0,00044

Figure 22 : Distribution spatiale et synthèse statistique des indicateurs $SpaLea^2_{U_{k_o}}$ sur l'intégralité de la France.

Afin faciliter la lecture de l'histogramme la dernière classe a été agrégée, elle comprend toutes les valeurs de $\left\{SpaLea^2_{U_{k_o}} > 0{,}0017\right\}$.

Résultats obtenus :

Le nombre d'U_{k_o} s'élève à 2036 communes, ce qui est largement supérieur à celui des U_k qui est de 421 communes.

Il est rappelé que plus la valeur de $SpaLea^2_{U_{k_o}}$ est élevée et plus les chances qu'un patient mal spatialisé réside dans une U_{k_o} sont grandes.

Le nombre d'U_{k_o} pour lesquelles l'incertitude de spatialisation est dérisoire est très grand. Plus de 1500 des U_{k_o} ont un $\left\{SpaLea^2_{U_{k_o}} \leq 0{,}0006\right\}$.

Proposition du seuil de sélection statistique des communes les plus propices à accueillir des individus dont la spatialisation est erronée:

Afin d'identifier les U_k qui ont de fortes chances de contenir des patients mal spatialisés, il convient de détecter les U_{k_o} associées à une *forte probabilité* d'accueillir ces derniers. Une stratégie statistique classique de sélection par seuillage, basée sur l'écart-type, est proposée afin d'identifier les ensembles d'U_{k_o} concernées.

$$U_{risque.SpaLea2} = \bigcup_{k_o=1}^{2036} \left(U_{k_o} \middle| SpaLea^2_{U_{k_o}} \geq \varphi_{SpaLea2}\right)$$

Le seuil de sélection des U_{k_o} à *forte probabilité* d'accueillir ces sujets est donné par :

$$\varphi_{\text{SpaLea2}} = \text{môy}\left(\text{SpaLea}^2_{U_{k_o}}\right) + 2 \cdot \hat{\sigma}\left(\text{SpaLea}^2_{U_{k_o}}\right) = 0{,}00105$$

Il s'agit d'un seuil statistique classique mais il n'en demeure pas moins subjectif (Saporta, 2006). De toute façon, en analyse spatiale, *il y a un réel problème d'identification de la valeur mathématique d'un paramètre avec sa signification géographique* (Sanders, 1992)

En appliquant la stratégie proposée pour la sélection des U_{k_o} à *probabilité forte* d'accueillir ces patients dont la spatialisation est erronée, on obtient les résultats cartographiques suivants :

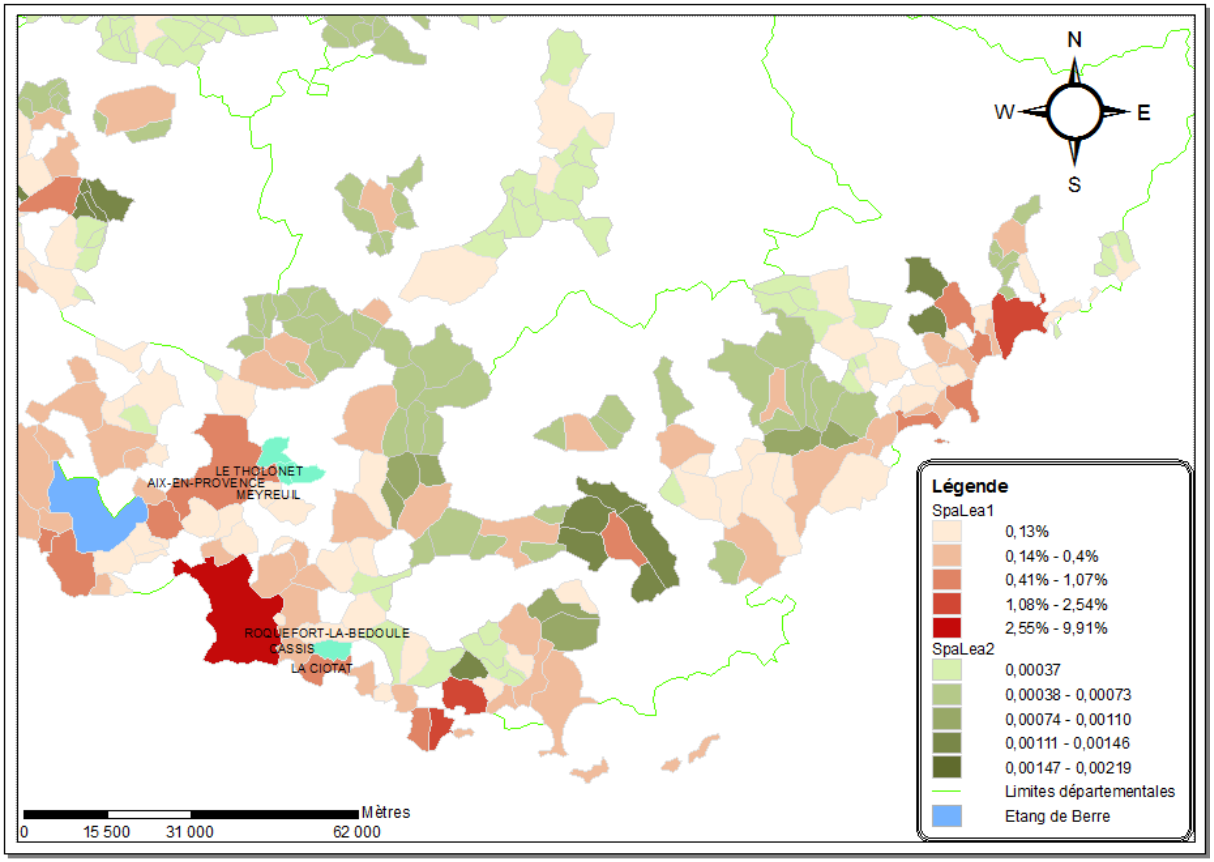


Figure 23 : Cartographie - zoom-in - de l'ensemble $U_{\text{risque.SpaLea2}}$ situées dans en région PACA

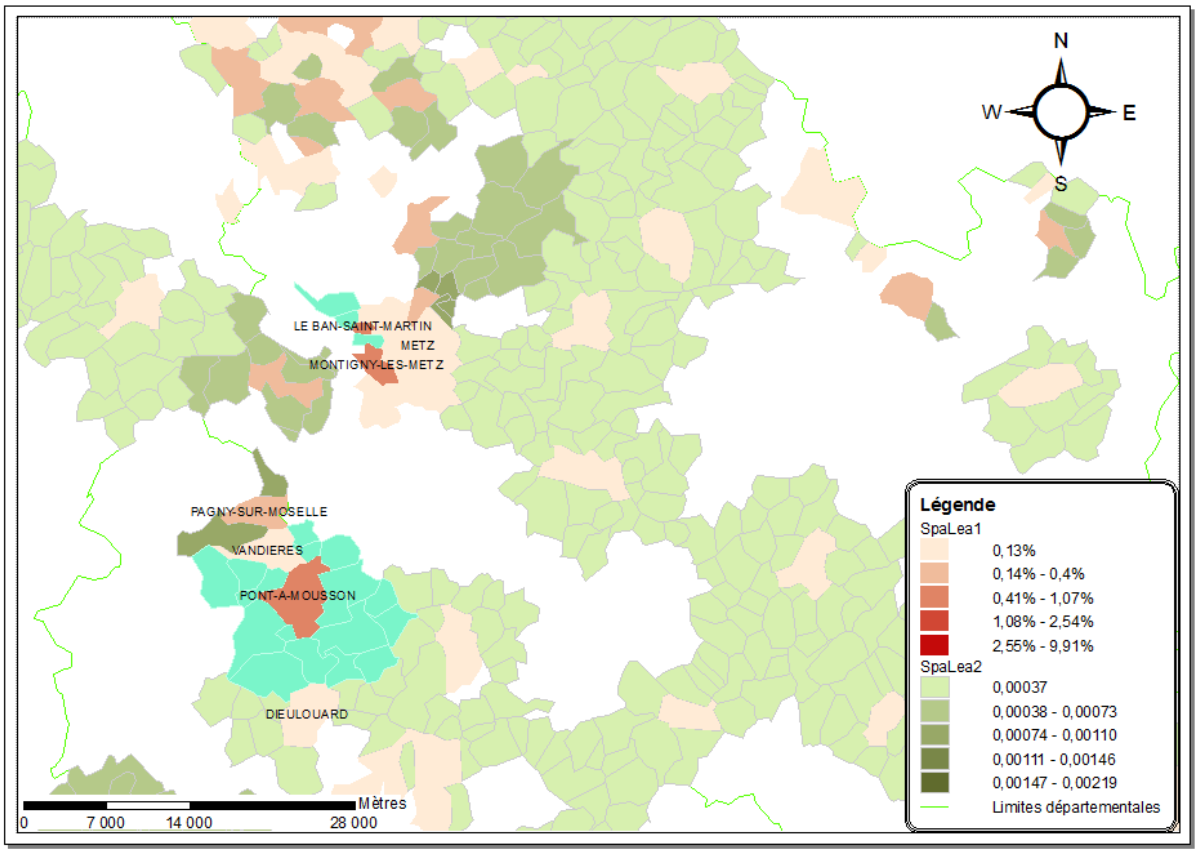


Figure 24 : Cartographie - zoom-in - de l'ensemble $U_{risque.SpaLea2}$ situées dans la région Alsace-Lorraine

Les communes de 2nde espèce concernées par une très *forte probabilité d'accueillir des patients spatialisés à tort*, i. e. l'ensemble $\mathbf{U_{risque.SpaLea2}}$ a été mis en emphase par une couleur cyan.

Analyse des résultats :
La stratégie de sélection par seuillage des U_{k_o} à *forte probabilité d'accueillir des patients mal spatialisés* permet de détecter les $\mathbf{U_k}$ où des patients ont *probablement été spatialisés à tort*. Ce sont celles qui se trouvent en contiguïté spatiale directe avec les U_{k_o} appartenant à $\mathbf{U_{risque.SpaLea2}}$.

Sur l'intégralité des 421 U_k de 1ère espèce, seulement 14 sont concernées. Ce qui correspond à un total de 41 individus situés en France métropolitaine, soit 5,4% de l'effectif spatialisé :
8/14 U_k sont situées en Alsace-Lorraine, soit 22 patients

- Département: Meurthe et Moselle, les communes {Pont-à-Mousson (6), Vandières (1), Pagny-sur-Moselle (3), Dieulouard (1)}.
- Département: Moselle, les communes {Montigny-lès-Metz (5), Metz (1) et Le-Ban-Saint-Martin (5)}.
-

6 U_k en région PACA, soit 19 individus.

- Département: Bouches du Rhône, les communes {Aix-en-Provence (4), Le Thoronet (5). Et celles de : La Ciotat (5), Cassis (2), Roquefort-la-Bédoule (1) et Saint-Cyr-sur-Mer (2)}

Les chiffres écrits entre parenthèses représentent le nombre de patients spatialisés.

Identification des patients à réinterroger :
Il a été vu que dans la mesure où l'on s'intéresse à l'effet de l'environnement géographique sur l'état de santé des patients, i.e. sur leurs séquelles, et où le fait de recontacter des patients engendre des démarches coûteuses sur le plan financier et temporel, il convient de solliciter les ARC uniquement pour

les patients spatialisés dans des U_k dont les U_{k_o} limitrophes sont incluses dans $\mathbf{U_{risque.SpaLea2}}$ et dont le nombre ou les surfaces géographiques mises en jeu sont importantes.

Par conséquent, il convient d'analyser les cartographies des $\mathbf{U_{risque.SpaLea2}}$ pour se rendre compte finalement que le risque d'erreur en matière d'exposition environnementale ne concerne que les communes de Pont-à-Mousson (6), Vandières (1) et Le-Ban-Saint-Martin (5). Ce qui représente un total de 12 individus, soit seulement 1,6% des patients spatialisés.

La spatialisation de ces individus est qualifiée *d'incertaine et avérée*. Ils ont été recontactés, et étonnamment, ils étaient correctement spatialisés à l'exception de trois d'entre eux.

CONCLUSIONS ET REMARQUES

Manifestement les hypothèses inhérentes à la méthode SpaLea, et par extension au concept de spatialisation de 1ère espèce, sont particulièrement robustes.

De plus, SpaLea est parcimonieuse et elle est reproductible à toutes sortes de populations lorsqu'on désire spatialiser des individus à l'échelle des communes et que la seule information géographique disponible est le CP à un moment donné.

Aussi et surtout, elle permet d'estimer l'erreur de spatialisation commise et d'identifier les individus associés à une *forte probabilité d'être spatialisés à tort* dans une commune de 1ère espèce et auxquels il convient d'accorder une attention particulière.

Il a été montré par ailleurs que *presque sûrement* la commune de 1ère espèce est bien la commune de résidence. Et lorsque ce n'est pas le cas, la commune de résidence se situe toujours *approximativement en contiguïté spatiale* de celle-ci.
En sus, l'idée que *les patients mal spatialisés* pratiquent au moins une activité sociale, professionnelle ou sportive dans la commune disposant de plus grand nombre de fonctionnalités – donc par hypothèse la plus peuplée, i.e. celle de 1ère espèce - n'est pas insensée. Donc même un patient spatialisé à tort dans celle-ci est confronté peu ou prou aux FE* qui la caractérisent.

Par la suite, lorsque les Déterminants Environnementaux de Santé* auront été identifiés, un affichage cartographique conjoint des i.st.e* qui les modélisent et de l'indicateur $SpaLea^2_{U_{k_o}}$ permettra d'estimer de façon plus précise le risque d'exposition réel à ces déterminants, à partir de l'appréciation visuelle de l'incertitude associée à chacune de ces communes. Mais on peut déjà conjecturer que l'imprécision qui pèse sur l'environnement réellement subi, induit par la phase de spatialisation, est relativement faible.

Enfin pour finir, les petites communes de 1ère espèce sont associées aux CP les plus incertains, donc à un nombre de communes de 2nde espèce important ou recouvrant des zones géographiques contigües mais très étendues. En faisant l'hypothèse que dans ces communes les individus se déplacent plus dans les communes limitrophes, et en vertu de la première loi de la Géographie selon laquelle *ce qui est proche se ressemble et ce qui est éloigné dissemble* (Tobler, 1970), alors les caractéristiques environnementales de ces ensembles sont relativement stationnaires. On peut supposer que ces individus subissent des expositions environnementales dont la variabilité est aussi forte que celles des patients spatialisés de façon certaine dans des communes de très grande taille ou très peuplées et dans lesquelles les conjonctures environnementales sont à la fois volatiles et fragmentées.

La méthode SpaLea est donc à la fois simple, rapide, consistante et parcimonieuse. Et surtout, elle est adaptée à l'analyse des interactions géographiques entre la santé et l'environnement.

SECTION B) MODELISATIONS GEOGRAPHIQUES DE PHENOMENES MORBIDES

La méthode SpaLea permet de spatialiser les individus à partir de données Cohorte dans l'espace géographique*. Il convient donc désormais de proposer une méthode de modélisation géographique des PM* auxquels sont assujettis ces individus. Pour ce faire deux indicateurs spatiaux temporels morbides (i.st.m) seront proposés. Afin de parvenir à une représentation plus robuste et plus juste de la réalité géographique, un système de pondération spatiotemporel est intégré dans l'estimation des i.st.m.

D'abord, par le biais d'une *métrique floue géographique* qui prend en compte les incertitudes spatiales associées : aux hypothèses de la méthode de spatialisation (SpaLea), au choix de l'échelle d'investigation retenue, à la granularité* des variables épidémiologiques utilisées (nature, temporalités, précisions, lacunes), et à la finalité des i.st.m* (e.g. voués à être croisés avec des données environnementales). De cette façon l'espace est intégré *transversalement* dans la dialectique.

Ensuite, le temps d'exposition à l'environnement des individus est inscrit dans le processus d'estimation des i.st.m* afin d'intégrer, *verticalement* cette fois, la temporalité dans la dialectique, toujours dans le but de construire des i.st.m* robustes et en adéquation avec la finalité recherchée.

Un premier i.st.m* est proposé. Il est classique, de nature quantitative, et modélise la variabilité géographique des phénomènes morbides par le biais de prévalences spatiotemporelles pondérées et rapportées au temps d'exposition à l'environnement.
L'intégration de *connaissances expertes* par le biais d'une *métrique floue géographique* permet d'harmoniser la phase de fusion ensembliste des données morbides, dans les unités géographiques, en fonction de leur qualité spatiotemporelle. Mais le système de pondération affecte la variable épidémiologique et un bruit de fond peut être induit et avoir l'effet inverse de celui espéré (Abramson J.H., Abramson Z.H., 1988).

De fait, un second i.st.m* est proposé. Il est atypique, de nature qualitative, et son rôle est de spécifier la robustesse du premier. Il modélise des propensions spatiotemporelles à développer une pathologie particulière dans les unités géographiques, tout en accordant une attention particulière à la qualité des données utilisées en se basant aussi sur un ratio entre la connaissance temporelle médicale acquise sur les individus et celle de leur temps d'exposition à l'environnement.

Les propositions méthodologiques sont appliquées aux données de la Cohorte LEA. Elles ont pour dessein de modéliser la géographie des séquelles d'intérêt : cataractes (CATA), tumeurs thyroïdiennes (THYR) et tumeurs secondaires majeures (TUM2). La méthode proposée est intimement liée aux données utilisées. Elle peut cependant être facilement étendue à toutes les séquelles développées après le traitement d'une leucémie (LEUC). Elle peut être également transposée à n'importe quelle autre maladie, conditionnellement à la *granularité* des données épidémiologiques* utilisées et aux *connaissances expertes* disponibles et intégrables (Kenneth et Sander, 1998).

PROPOSITION D'INDICATEURS SPATIOTEMPORELS MORBIDES

L'objectif :
Proposer des stratégies d'estimation d'i.st.m* suffisamment robustes pour modéliser de façon précise la réalité géographique des phénomènes morbides. Par conséquent, il convient de prendre en compte les incertitudes spatiotemporelles associées aux méthodes et aux données utilisées de sorte que ces i.st.m* soient en adéquation avec la finalité poursuivie, i.e. adaptés à l'analyse spatiale des interactions santé-environnement.

L'estimation d'i.st.m* robustes doit prendre en compte conjointement les incertitudes associées aux hypothèses inhérentes à la méthode de spatialisation, à la qualité spatiotemporelle des variables épidémiologiques ainsi qu'à la consistance statistique de l'information spatialisée.

Hypothèse :
Une possibilité consiste à utiliser la théorie *des ensembles flous* dont le but est d'améliorer la qualité du processus de fusion de données de natures, de degrés précision, de niveaux d'incertitude, d'échelles spatiales et temporelles - très différents. L'idée consiste à injecter *a priori* de la *connaissance experte* par le biais de *fonctions mathématiques*. La théorie *des ensembles flous* est applicable à toutes les problématiques de fusion de données hétéroclites. Cependant elle doit être systématiquement adaptée *au domaine scientifique*, *à la finalité des outputs*, *aux connaissances expertes et aux données auxiliaires intégrables*, susceptibles d'améliorer la qualité du processus d'agrégation ensembliste (Dubois et Prade, 2004)

Proposition :
Imaginer une stratégie de pondération afin de construire *une métrique floue géographique*, permettant d'injecter, par le biais de fonctions mathématiques, des *connaissances expertes* afin d'harmoniser le processus d'agrégation ensembliste des variables épidémiologiques dans les unités géographiques, i.e. Uk, au regard de la qualité spatiotemporelle épidémiologique, géographique, et statistique (EpiGéoStat) associée à chacune d'elles.

La métrique floue géographique a pour dessein d'augmenter la robustesse spatiotemporelle de la phase d'agrégation des données morbides dans les communes. Par conséquent l'espace est intégré au cœur du processus d'estimation des i.st.m* et donnera plus de consistance à l'analyse spatiale (Voiron-Canicio, 1995). La dialectique est géographique, l'espace est primordial certes, mais il s'agit de ne pas mettre, pour autant, *le temps à l'ombre* (Brunet, 1968). Pour ce faire, en épidémiologie spatiale, une stratégie de conversion en patients-années est couramment utilisée. Elle permet d'augmenter virtuellement les effectifs spatialisés et parallèlement de prendre en compte la latence des maladies en fonction du temps d'exposition à l'environnement (Bernard et Lapointe, 2003).

En somme il s'agit de parvenir à des modélisations plus justes et plus proches de la réalité géographique des Phénomènes Morbides (PM) d'intérêt en intégrant d'abord, de façon robuste, l'espace *horizontalement*, puis en spécifiant dûment le rôle de la temporalité, en intégrant *verticalement* le temps au cœur du processus d'estimation des i.st.m* (Peguy, 1996).

La prise en compte du temps d'exposition à l'environnement donne nécessairement plus de consistance aux i.st.m. En revanche, l'intégration de *connaissances expertes* peut introduire un bruit de fond, donc des biais. Pour pallier cette éventuelle défaillance, deux i.st.m* sont proposés.
Le premier est classique et de nature quantitative. Il modélise la géographie du PM* d'intérêt par des prévalences pondérées *EpiGéoStat* converties en patients-années.
Le second est atypique et de nature qualitative. Il est spécialement conçu pour spécifier le niveau de qualité associé au premier et en même temps pour informer sur la propension qu'ont les individus à

développer cette même pathologie, mais en attachant une attention particulière à la qualité spatiotemporelle des données épidémiologiques sans affecter la nature qualitative de la variable morbide input - à l'aune de son prédécesseur.

La synoptique d'estimation de la métrique floue géographique est décrite subséquemment dans ses grandes lignes de sorte à garantir une représentation intelligible et un caractère reproductible. Ensuite le principe théorique de l'intégration du temps d'exposition à l'environnement est explicité sommairement pour les mêmes raisons. Puis de la même façon, l'idée du principe d'estimation des deux i.st.m* est rapidement déclinée.

L'estimation des facteurs d'incertitude spatiotemporelle : π_i^j qui constituent la métrique floue géographique, et l'estimation du temps d'exposition à l'environnement : tee_i^j et des i.st.m* de nature quantitative $z'^j_{(\text{U}_k),c}$ et qualitative $z'^j_{(\text{U}_k),q}$sont intimement liées à la granularité* des données utilisées.

Par conséquent les stratégies d'estimation seront spécifiées explicitement dans la section.C car parfaitement interdépendantes de l'échelle d'investigation retenue, de la granularité* des données épidémiologiques, des connaissances expertes intégrées ainsi que de la consistance statistique de l'information spatialisée.

Désormais il convient d'énoncer les principes de construction des *facteurs d'incertitude spatiotemporelle* π_i^j de *la métrique floue géographique* à partir de *connaissances expertes* sur la qualité spatiotemporelle Epidémiologique, Géographique et Statistique (EpiGéoStat) des méthodes, des hypothèses et des inputs utilisés.

PRINCIPE D'INTEGRATION DES INCERTITUDES SPATIOTEMPORELLES

Remarque liminaire :
Précédemment il a été montré que la spatialisation d'individus à partir leur CP était parfois grossière et pouvait induire une incertitude. Cette incertitude spatiale à connotation géographique intervient au niveau de la phase d'agrégation ensembliste des données morbides : y_i^j dans les communes de 1ère espèce. Chaque indicateur morbide y_i^j est associé à un individu dont l'incertitude liée à sa spatialisation peut être évaluée à partir de connaissances expertes, en l'occurrence du nombre de codes INSEE de second ordre : $q_{o,i}$ associés à son CP : x_i^{CP}. Cette incertitude est inhérente à la qualité de la méthode de spatialisation et doit être prise en compte dans le processus de modélisation géographique des phénomènes morbides. Mais ce n'est pas la seule incertitude. Il en existe d'autres à connotation épidémiologique, géographique ou statistique qu'il convient d'intégrer aussi dans le processus d'estimation des i.st.m* (Abramson J.H., Abramson Z.H., 1988).

Objectif :
Proposer une stratégie d'estimation d'une *métrique floue géographique* en injectant, par le biais de fonctions mathématiques, des connaissances expertes *a priori* en vue de parvenir à une représentation plus robuste et plus juste de la réalité géographique des phénomènes morbides que l'on veut modéliser (Dubois et Prade, 1994).

Hypothèse :
La stratégie de pondération suppose qu'une valeur y_i^j tend vers son dual proportionnellement aux incertitudes EpiGéoStat qui lui sont associées. En d'autres termes une absence de séquelle $\{y_i^j = \text{NON}\}$ lors du processus d'agrégation ensembliste n'est pas tout à fait un « non » selon la qualité spatiotemporelle EpiGéoStat qui lui est associée. Plus le facteur d'incertitude EpiGéoStat - constitutif de la métrique floue géographique – est grand et plus un : « NON incertain tend vers un OUI ». Et inversement pour ce qui est des $\{y_i^j = \text{OUI}\}$.

Spécification de l'hypothèse :
Les variables morbides épidémiologiques permettant de savoir si le patient I_i a développé ou pas la séquelle.j sont booléennes : $y_i^j = \{\text{OUI} \cup \text{NON}\}$. L'idée consiste à se dire que lors de la phase d'agrégation ensembliste des variables y_i^j dans les U_k, ces dernières ne sont plus tout à fait des OUI ni des NON parfaitement sûrs.

Et qu'il s'agit d'altérer la variable en fonction de la qualité spatiotemporelle EpiGéoStat qui lui est associée. Par exemple si un patient n'a pas développé une séquelle $y_i^j = \text{NON}$ mais qu'il n'a pas été revu depuis dix ans, alors peut-être est-t-il en train ou l'a-t-il déjà développée ? De plus, ce même patient est spatialisé dans une U_k à partir d'un CP qui peut être associé à une dizaine d'autres communes – il y a donc des chances pour que sa commune de résidence ne soit pas celle dans laquelle il est localisé.

Il a donc peut-être développé sa séquelle dans un environnement radicalement différent de celui qu'il est supposé côtoyer. Dans cette même commune, un autre patient est spatialisé. Celui-ci n'a pas non plus développé la séquelle $y_i^j = \text{NON}$. Or il vient d'effectuer un entretien de QV* le mois dernier et il est spatialisé de façon sûre puisque son CP est associé à une seule U_k, celle dans laquelle il est spatialisé. D'un point de vue géographique, ce patient a très probablement subi l'environnement de la commune dans laquelle il est spatialisé et la qualité spatiotemporelle épidémiologique associée à $y_i^j = \text{NON}$ est ô combien plus sûre que celle de l'autre patient évoqué. Voici illustrée l'idée des facteurs d'incertitude constitutifs de la métrique floue géographique qui va permettre de pondérer des variables morbides y_i^j,

à partir de connaissances EpiGeoStat expertes *a priori* afin de calculer des prévalences spatiales pondérées consistantes plus proches de la réalité géographique des phénomènes morbides que l'on veut modéliser.

Proposition n°1 :
La métrique floue est composée de facteurs d'incertitude associés à chaque patient et notés π_i^j qui ne peuvent pas être calibrés dans la mesure où l'on ne connait pas *a priori* la réalité géographique des phénomènes morbides que l'on cherche justement à modéliser. Ils doivent donc être estimés à partir de connaissances expertes.

Proposition n°2 :
En théorie, la valeur absolue des poids d'incertitude est bornée, π_i^j ne peuvent pas excéder 0,5. Cela correspond à l'incertitude maximale, valeur à partir de laquelle on ne peut plus distinguer les patients qui ont ou pas développé une séquelle. Au-delà de cette valeur l'interprétation devient insensée.

Proposition n°3 :
Les facteurs EpiGéoStat constitutifs de la métrique floue géographique dépendent de trois types d'incertitude spatiotemporelle. Le premier poids est à connotation géographique : $\pi_i^{j,geo}$, le second est épidémiologique $\pi_i^{j,epid}$ et le dernier est statistique $\pi_i^{j,stat}$. Les valeurs de ces poids sont bornées. La valeur maximale de chacun dépend du type de séquelle mais aussi de la connotation EpiGeoStat, e.g. les incertitudes géographiques inhérentes aux hypothèses de la méthode SpaLea pèsent plus lourd dans la balance que celles associées à la qualité des données de la cohorte LEA.

Proposition n°4 :
Les facteurs d'incertitude EpiGéoStat ont pour but de faire tendre proportionnellement les variables épidémiologiques morbides – séquelles – dont la forme codée est : $y_i^j = \{1 \stackrel{\text{def}}{=} \text{OUI}\} \cup \{0 \stackrel{\text{def}}{=} \text{NON}\}$ - vers la modalité duale, lors de la phase d'agrégation dans les U_k. Par conséquent les π_i^j sont pourvus d'un signe noté : ξ_i^j. Ce dernier dépend de la réponse associée à la variable morbide

$$\xi_i^j = \begin{cases} -1 & \text{lorsque: } \quad y_i^j = 1 \\ 1 & \text{lorsque: } \quad y_i^j = 0 \end{cases}$$

Les facteurs d'incertitude spatiotemporelle EpiGeoStat embrassent les trois facteurs d'incertitude et la variable binaire ξ_i^j dans une fonction d'agrégation simple.

$$\pi_i^j = f\left(\xi_i^j;\ \pi_i^{j,geo}, \pi_i^{j,epid}, \pi_i^{j,stat}\right) \in [\![-0{,}5; 0{,}5]\!]$$

Principe de la stratégie d'estimation des facteurs d'incertitude EpiGéoStat
Le principe d'estimation des π_i^j se fonde sur des connaissances expertes épidémiologiques, géographiques et statistiques (EpiGeoStat) qui permettent de quantifier les incertitudes liées à la qualité spatiotemporelle des hypothèses, des méthodes et des données utilisées.

Synoptique d'estimation des facteurs d'incertitude à connotation Géographique
Les facteurs d'incertitude géographique s'estiment par la somme de trois fonctions fragmentaires $\pi_i^{j,geo}$ fondées sur des indicateurs géographiques et construites à partir des incertitudes spatiotemporelles liées à :

$$\pi_i^{j,geo} = \left(g_1(q_{o,i}) + g_2\left(S_{i(U_k)};\ \hat{\alpha}, \hat{\beta}\right) + g_3\left(T_i^j, \hat{\Theta}_m\right)\right)$$

i. La méthode de spatialisation des patients dans les U_k qui peut être évaluée à partir du nombre de codes INSEE de 2nd ordre attribués aux patients : $i.geo_i^j(l=1) = g_1(q_{o,i})$. Cette incertitude est indirectement appréciable visuellement sur les cartographies via l'indicateur $SpaLea^2_{U_{ko}}$ dans les communes de 2nde espèce.

ii. La variabilité des FE* caractéristiques des milieux de vie qui dépend de l'échelle d'investigation retenue. Cette incertitude est une fonction croissante de la superficie des U_k, tel que : $i.geo_i^j(2) = g_2(S_{(i|U_k)};\ \hat{\alpha}, \hat{\beta})$.

iii. La probabilité liée aux trajectoires de vie des individus. Dans la mesure où les variables de la BD-LEA ne permettent pas de reconstituer les mobilités résidentielles des patients, elles sont intégrées dans la composante géographique et s'estiment aussi à partir de données communautaires. $i.geo_i^j(l=3) = g_3(T_i, \hat{\Theta}_m)$.

Schéma de principe de la synoptique du processus d'estimation des $\pi_i^{j,geo}$

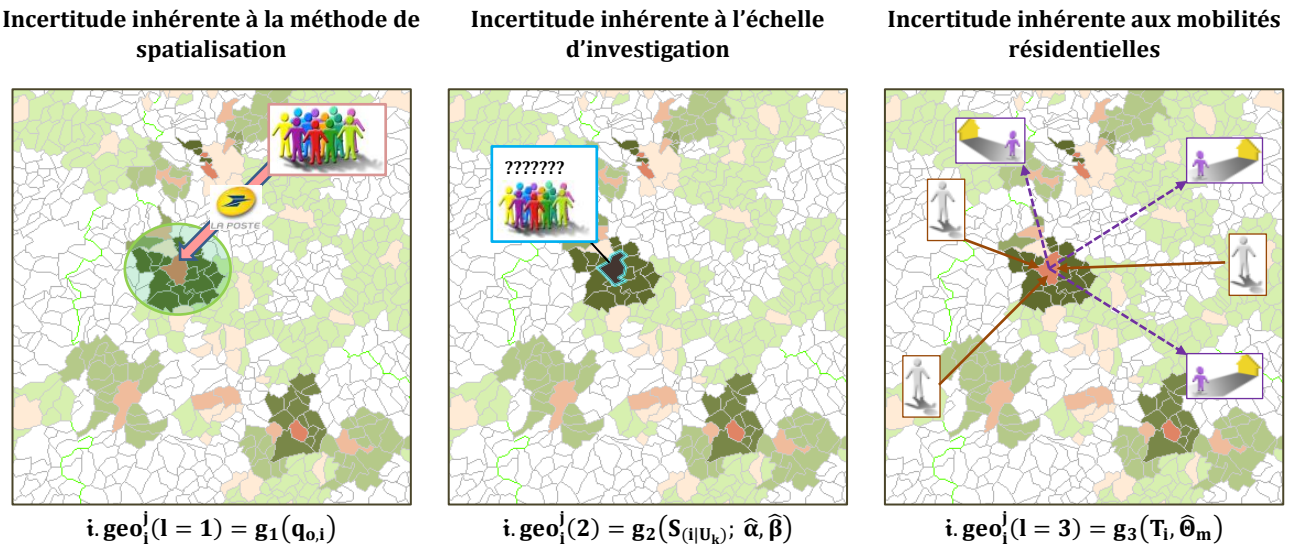


Figure 25 : Synoptique d'estimation du facteur d'incertitude spatiotemporelle à connotation géographique

Synoptique d'estimation des facteurs d'incertitude à connotation Epidémiologique :
La stratégie d'estimation des facteurs d'incertitude épidémiologique a été construite d'après des *recommandations expertes* (Auquier et Michel, 2012).

Le parti retenu
Il est supposé que les variables morbides y_i^j ont des qualités spatiotemporelles différentes qui doivent être prises en compte lors de leur agrégation dans les U_k. Les facteurs d'incertitude épidémiologique sont évalués grâce à une fonction mathématique dont les paramètres sont spécifiés à partir de variables épidémiologiques permettant d'estimer le niveau de variabilité de ces incertitudes, tel que :

$$\pi_i^{j,\mathrm{Epi}} = f_{épi}\left(l_i^j; s_i^j; \mathrm{m}_{i\,;} Q(\mathrm{ti}_i) \middle| y_i^j\right).$$

Deux cas de figure sont possibles :

(i) - Le patient a développé la séquelle $y_i^j = 1$, cette information est sûre d'un point de vue épidémiologique			
(ii) - Le patient n'a pas développé la séquelle $y_i^j = 0$, cette information est grevée d'incertitudes liées à :			
$\mathbf{l_i^j}$ - La qualité informationnelle de la BD épidémiologique	$\boldsymbol{s_i^j}$ - La qualité informationnelle du suivi	$\mathbf{m_i}$ - La capacité du patient à développer des séquelles	$\mathbf{ti_i}$ - La qualité temporelle de l'information morbide
Les variables séquelles *lacunaires* y_i^j =? sont systématiquement comblées, dans la BD-LEA, par une absence de séquelle : $y_i^j = 0$ mais cette information est plus incertaine qu'une absence de séquelle avérée	Les patients qui ont reçu un traitement agressif sont systématiquement suivis pour certaines séquelles, l'incertitude sur le dépistage et sur la remontée de l'information dans la BD épidémiologique diffère selon que le patient est suivi ou pas.	Certains patients sont décédés. De fait, ils ne peuvent plus développer de séquelles ni être exposés à l'environnement géographique. Une absence de séquelle chez un défunt n'a pas la même signification que chez un sujet déclaré en vie	L'incertitude temporelle associée à une absence de séquelle : $Q(ti_i)$ est proportionnellement croissante avec la temporalité : ti_i qui sépare la dernière consultation du patient et celle à laquelle l'information est supposée exacte.

Tableau 18 : Hypothèse d'estimation du facteur d'incertitude spatiotemporelle à connotation épidémiologique

Schéma de principe de la synoptique du processus d'estimation de $\pi_i^{j,geo}$

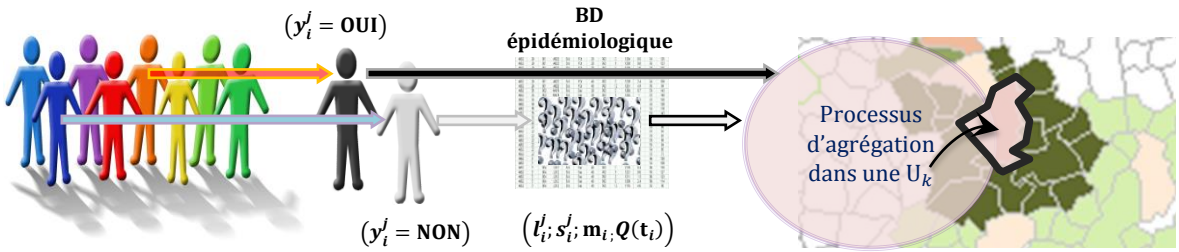


Figure 26 : Synoptique d'estimation du facteur d'incertitude spatiotemporelle à connotation épidémiologique

Synoptique d'estimation des facteurs d'incertitude à connotation Statistique

Schéma synoptique du processus d'estimation des $\pi_i^{j,stat}$

La stratégie d'estimation des facteurs d'incertitude statistique repose sur l'idée selon laquelle les i.st.m* ne peuvent pas contribuer de la même façon à l'analyse statistique multidimensionnelle vouée à l'identification des DES* géographiques.

La valeur de $\pi_i^{j,stat}$ est proportionnellement croissante avec *l'inconsistance* statistique* associée au processus d'estimation l'i.st.m, i.e. au nombre de patients $n_{(.)}^{I_i}$ spatialisés dans chaque U_k permettant de les estimer (Saporta, 2006) – de sorte que :

$$\pi_i^{j,stat} = f_{Stat}\left(\mathbf{n}_{(U_k)}^{I_i}\right)$$

Incertitude d'inconsistance statistique

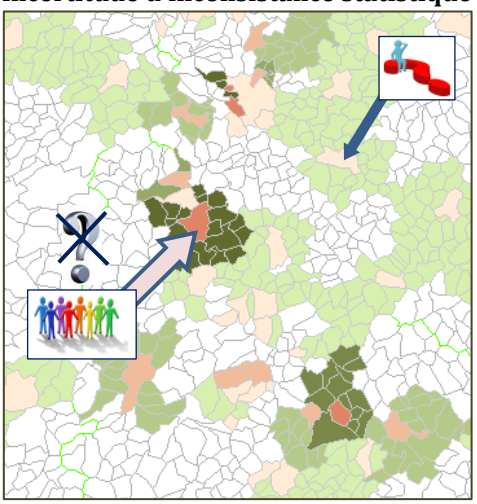


Figure 27 : Synoptique d'estimation du facteur d'incertitude spatiotemporelle à connotation statistique

Conclusion :

Pour l'indicateur quantitatif :

Les π_i^j permettent de pondérer la valeur des y_i^j en les faisant tendre *raisonnablement* vers leur contrepartie duale, au prorata du degré d'incertitude *EpiGéoStat* associé aux hypothèses de la méthode de spatialisation et à la qualité des données disponibles.

Pour l'indicateur qualitatif :

Les π_i^j n'ont pas d'influence directe sur les y_i^j mais ils permettent de sélectionner, dans chaque U_k, un sous ensemble de patients pour lesquels la qualité spatiotemporelle EpiGeoStat associée y_i^j est suffisamment fiable pour estimer la propension des patients à développer, ou pas, un PM* – séquelle.

La métrique floue géographique :

Elle est constituée des π_i^j et est une façon d'intégrer *transversalement l'espace* dans le processus d'estimation des i.st.m. L'injection de *connaissances expertes a priori* permet de grever aux i.st.m., à l'échelle des U_k, i.e. d'un point de vue purement spatial, une notion de qualité spatiotemporelle et d'en améliorer la robustesse. L'espace est donc intégré *transversalement au cœur de la dialectique géographique* et il est maintenant question d'y intégrer *verticalement* le temps.

PRINCIPE D'INTEGRATION D'EXPOSITION TEMPORELLE

Remarques liminaires :

La métrique floue constitutive des π_i^j permet d'introduire *transversalement* la notion de qualité spatiale dans le processus d'estimation des i.st.m. Cependant la dialectique est géographique, donc spatiotemporelle.

Les i.st.m* classiques prennent la forme de prévalences spatiales. Celle-ci, est définie en épidémiologie comme le nombre de personnes atteintes par une pathologie, rapporté à la taille de la population *in situ* considérée.

Or si les facteurs d'incertitudes EpiGéoStat permettent effectivement d'augmenter la robustesse spatiale, ils n'offrent cependant pas la possibilité *d'intégrer verticalement* la notion de temporalité au cœur du processus d'estimation. De plus, ce type d'estimateur est peu adapté lorsque les effectifs spatialisés sont *petits* (Peguy, 1996).

Objectifs :

Intégrer *verticalement* la notion de temporalité au cœur du processus d'estimation des i.st.m, et augmenter v*irtuellement*, par la même occasion, grâce à une stratégie de transformation temporelle, la taille des effectifs spatialisés

Hypothèses :

L'intégration de la temporalité dans le processus d'estimation consiste à supposer qu'il faille raisonner non plus sur des patients, mais sur le nombre d'années durant lesquelles ces derniers ont subi des *expositions environnementales potentielles ou intrinsèques.*

Tout en distinguant par ailleurs, dans l'espace et le temps, les individus atteints des sujets sains.

Intégration verticale de la *temporalité* par transformation topologique : Patients-Années

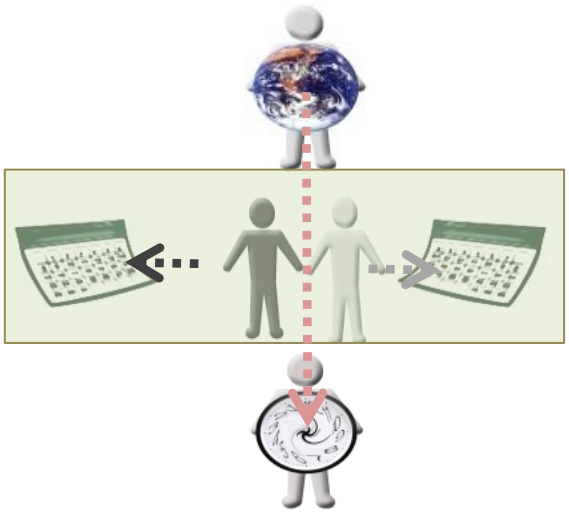


Figure 28 : Synoptique du processus de la transformation Patient-Années

Proposition 1 :
Prendre en compte le *temps d'exposition à l'environnement géographique,* noté : tee_i^j. En épidémiologie spatiale cet artifice permet de procéder à une conversion en *Patients-Années*.
Cette stratégie permet de valider le double objectif énoncé. D'une part elle *permet d'augmenter virtuellement* la taille des effectifs spatialisés. Et d'autre part, elle permet de dissocier, dans le temps, *la population à risque*, i.e. celle qui n'a pas ou pas encore développé la pathologie - de celle chez qui elle a été décelée. En sus, les transformations en Patients-Années permettent de prendre en compte la latence de la maladie, i.e. le caractère précoce du PM* développé. (Bernard et Lapointe, 2003).

Proposition 2 :
Il s'agit d'une proposition heuristique qui consiste à calculer un ratio Exposition-Participation, noté : rpe_i^j. Celui-ci a trait au processus d'estimation du second i.st. dans le cadre d'une approche qualitative. Il a pour objet d'intégrer conjointement le caractère temporel des expositions environnementales et celui de la fiabilité des connaissances acquises sur l'historique médical des individus.

Principe d'estimation :
Pragmatiquement la conversion en Patients-Années est assez simple à effectuer. Il s'agit de diviser le nombre de personnes atteintes par le temps d'exposition des patients à l'environnement (Bernard et Lapointe, 2003).

i. Le temps d'exposition à l'environnement $\mathbf{tee_i^j}$ correspond à la durée de participation pour les individus sains, et à la durée à laquelle la maladie est diagnostiquée chez les personnes atteintes.
ii. Le temps de participation $\mathbf{tp_i}$ est la durée qui sépare la date à laquelle l'étude commence – à laquelle tous les individus de la population cible sont supposés sains – et la date à laquelle l'étude est arrêtée ou est interrompue.
iii. Le ratio Exposition-Participation : $\mathbf{rpe_i^j}$ est le rapport entre le temps de participation tp_i et le temps d'exposition à l'environnement tee_i^j.

Conclusion :
L'intégration du temps d'exposition à l'environnement $\mathbf{tee_i^j}$ dans le processus d'estimation des i.st.m* quantitatifs permet : d'augmenter virtuellement la taille des effectifs spatialisés, de d'intégrer la notion de temporalité aux *expositions environnementales potentielles ou intrinsèques* subies, et de distinguer par ailleurs, dans le temps, la population à risque des individus atteints par le PM* – séquelle. S'agissant de l'i.st.m. qualitatif, le principe est le même sauf que l'idée de qualité inhérente aux informations médicales et individuelles disponibles est ajoutée.

La stratégie d'intégration du temps d'exposition à l'environnement $\mathbf{tee_i^j}$ – et de la qualité temporelle des informations disponibles - dans le processus d'estimation des i.st.m* permet d'intégrer *verticalement le temps* au cœur de la dialectique géographique. Il est désormais question d'énoncer le principe d'estimation des i.st.m. quantitatifs et qualitatifs proposés en vue de modéliser, dans l'espace et le temps, la géographie de PM.

PRINCIPE D'ESTIMATION DES INDICATEURS SPATIOTEMPORELS MORBIDES

Remarques liminaires :
Les variables épidémiologiques binaires disponibles sont $y_i^j = \{\text{OUI} ; \text{NON}\}$. La dialectique est géographique et ces données doivent être fusionnées dans les unités géographiques d'un SIG. Or elles sont maculées d'incertitudes spatiales et opèrent sur des temporalités différentes.

Objectif :
Il s'agit de représenter, par le biais d'i.st.m* robustes et fiables, la géographie de phénomènes morbides à partir de variables épidémiologiques y_i^j aux consistances spatiotemporelles très différentes. Dans un premier temps il convient d'intégrer *horizontalement* l'espace par le biais de la notion de qualité spatiotemporelle, au moment de la phase d'agrégation dans les U_k. Puis dans un second temps, d'incorporer *verticalement* la notion d'exposition temporelle dans le but de parvenir à une représentation plus robuste et plus juste de la réalité géographique des PM* étudiés (Peguy, 1996).

Proposition°1 :
L'intégration *verticale* de la temporalité s'effectue par une conversion en Patients-Années qui tient compte à la fois du temps d'exposition à l'environnement et de la vitesse à laquelle les individus développent leurs pathologies - séquelles.

Proposition°2 :
L'intégration *horizontale* de notion de qualité spatiotemporelle s'effectue par le biais d'une *métrique floue géographique* constituée de facteurs d'incertitudes EpiGéoStat π_i^j. De cette façon il est possible de construire un premier i.st.m* plus classique. Mais le processus d'estimation pondère, donc altère la variable source y_i^j. Comme les π_i^j ne peuvent être calibrés, ils sont construits à partir de connaissances expertes. Par conséquent, un second i.st.m* est proposé, lequel utilise aussi les π_i^j mais cette fois sans altérer ni la valeur ni la nature de y_i^j. Ce second i.st.m* apporte une information différente sur la géographie du PM* d'intérêt et permet de spécifier la robustesse des valeurs prises par le premier.

Principe d'estimation des i.st.m* : approche quantitative et qualitative :
Il en découle deux indicateurs spatiotemporels morbides* (i.st.m).

Les prévalences spatiales pondérées EpiGéoStat converties en Patients-Années :
il s'agit d'un i.st.m* quantitatif, qui modélise la géographie d'un PM* – séquelle - par une prévalence à laquelle sont appliquées des pondérations spatiotemporelles supposées améliorer sa robustesse.

$$z'^{j}_{(\text{U}_k),c} = \hat{f}_c\left(y_i^j ; \pi_i^j \middle| \{\text{tee}_i^j ; \text{V}_{i,1}\}\right)$$

Avec : π_i^j qui représente l'incertitude EpiGeoStat associée à y_i^j, i.e. à la variable séquelle ; Et tee_i^j le temps d'exposition du patient à l'environnement.

Les propensions spatiotemporelles pondérées EpiGéoStat morbides, répétées au prorata du ratio Participation-Exposition :
Il s'agit d'un i.st.m* qualitatif qui modélise la propension qu'ont les individus à développer, dans une U_k, une pathologie – séquelle – en attachant une importance particulière à la qualité spatiotemporelle EpiGéoStat des données utilisées et à la prise en compte conjointe de la connaissance de l'historique médical et du temps d'exposition des individus à l'environnement géographique.

$$z'^{j}_{(\text{U}_k),q} = \hat{f}_q\left(y_i^j \middle| \{\text{tee}_i^j ; \text{tp}_i ; |\pi_i^j| ; \psi_\pi^j ; \text{V}_{i,1}\}\right)$$

Avec : tp_i le temps de participation de l'individu ; Et, ψ_π^j un seuil d'élimination permettant, pour chaque Uk, de sélectionner un sous ensemble de patients, donc de réponses y_i^j, dont la qualité est jugée statistiquement fiable pour être prise en compte. Cette valeur est répétée proportionnellement au rapport entre la durée de connaissance sur l'historique médical et le temps d'exposition des individus à l'environnement géographique. Cette valeur est d'autant plus élevée lorsque la séquelle est développée de façon précoce et que les connaissances temporelles épidémiologiques acquises s'étalent sur une période longue.

Conclusion :

La métrique floue géographique des π_i^j et la prise en compte conjointe *du temps d'exposition à l'environnement* : tee_i^j permettent d'intégrer respectivement : *l'espace - de façon transversale*, et *verticalement - la temporalité* au cœur de la dialectique géographique.

L'injection judicieuse de ces stratégies spatiotemporelles de pondération dans le processus d'estimation des i.st.m* proposés leur donne plus de consistance. Par conséquent, les modélisations spatiotemporelles qui en découlent permettent de parvenir à des représentations plus robustes et plus justes de la réalité géographique des PM.

Cependant toutes ces stratégies d'estimation sont conditionnées par la granularité* des données épidémiologiques disponibles. Afin d'illustrer et de spécifier les méthodes énoncées, elles sont appliquées – dans la section suivante - aux séquelles d'intérêt, i.e. aux informations contenues dans la BD-LEA.

Elles peuvent facilement être étendues à toutes les autres séquelles et à n'importe quelle autre maladie en adaptant, au préalable, la *métrique floue géographique* et l'estimation *du temps d'exposition à l'environnement* à la granularité* des données épidémiologiques disponibles.

SECTION C) SPECIFICATION DE LA METHODE DE MODELISATION - APPLICATION A LEA.

L'estimation des facteurs d'incertitude *EpiGéoStat* ainsi que celle du *temps d'exposition à l'environnement* sont interdépendantes de la *granularité* des données épidémiologiques* et des *informations expertes connexes* disponibles.
L'application des stratégies de pondération aux données de la Cohorte LEA est un excellent moyen pour illustrer les principes des méthodes proposées mais aussi pour spécifier explicitement les processus d'estimation de la *métrique floue géographique* et de celui du *temps d'expositions à l'environnement.*

ESTIMATION DE LA METRIQUE FLOUE GEOGRAPHIQUE

Les objectifs, les hypothèses, les propositions méthodologiques ainsi que les grandes lignes du principe d'estimation de la *métrique floue géographique* ont déjà été énoncés dans la section précédente. Il s'agit ici de les spécifier explicitement et de les appliquer aux séquelles d'intérêt afin de garantir le caractère heuristique des objectifs et des hypothèses, la reproductibilité des propositions, et d'illustrer de façon intelligible leur mise en œuvre.

Spécification des objectifs :
Il s'agit d'attribuer des poids à chacune des informations épidémiologiques morbides – séquelles y_i^j - associées aux patients I_i en fonction des incertitudes Epidémiologiques, Géographiques et Statistiques (EpiGéoStat) qui leur sont associées. Ceci est justement le but de la *métrique floue géographique* qui intervient au niveau de la phase d'agrégation ensembliste des y_i^j dans les U_k. L'intérêt est de parvenir à une représentation géographique plus proche de la réalité géographique des PM* à modéliser (Abramson J.H., Abramson Z.H., 1988). Chaque variable morbide - séquelle y_i^j - permet de savoir si « oui » ou « non » la pathologie a été développée. Or, ces variables n'ont pas la même consistance spatiotemporelle, ce qui est *préjudiciable* au moment de la phase de spatialisation ensembliste. Le rôle de la *métrique floue géographique* est d'injecter, par le biais *de fonctions mathématiques* adaptées, des *connaissances expertes* pour *fusionner de façon harmonieuse* des y_i^j aux *qualités temporelles et spatiales très différentes* (Dubois et Prade, 1994).

Spécification du principe :
l'i.st.m* quantitatif $z'^j_{(U_k),c}$ est une prévalence spatiale pondérée EpiGeoStat convertie en Patients-Années. En expurgeant les pondérations spatiales et temporelles on obtient une *prévalence classique* $z^j_{(U_k),c}$. On conçoit facilement qu'une *prévalence classique* estimée à partir d'une valeur y_i^j unique, sachant que : le patient n'a pas été revu depuis dix ans ; qu'il est spatialisé dans une U_k *gigantesque*, et que son CP est associable à une dizaine d'autres communes, n'offre pas la même qualité EpiGéoStat qu'une $z^j_{(U_k),c}$ supputée dans une U_k : de *petite taille*, dans laquelle un grand nombre de patients est spatialisé, régulièrement revus par les médecins, et que leur CP a permis de les spatialiser *de façon sûre*.
Il convient donc de prendre en compte ces incertitudes EpiGéoStat dans le processus d'estimation des i.st.m* proposés.

Spécification du processus d'estimation :
La stratégie de pondération proposée consiste à faire tendre *raisonnablement* la variable morbide y_i^j vers sa modalité duale, proportionnellement à la valeur d'un facteur de certitude spatiotemporelle π_i^j, i.e. à la qualité EpiGéoStat spatiotemporelle qui lui est associée. En d'autres termes les $\{y_i^j = \text{NON}\}$, une

fois spatialisées, ne sont plus tout à fait des « non », ce sont des « *non un peu oui* » - et vice versa pour ce qui est des $\left\{y_i^j = \text{OUI}\right\}$.

Par conséquent pour que les y_i^j, qui, sous forme codée, sont données par $y_i^j = \{1 \stackrel{\text{def}}{=} \text{OUI}\} \cup \{0 \stackrel{\text{def}}{=} \text{NON}\}$, puissent tendre la modalité duale, les π_i^j sont nécessairement munis d'un signe qui dépend de la valeur morbide, tel que :

$$\xi_i^j = \left(-1 + 2 \cdot \mathbb{1}_{\left\{y_i^j = \text{OUI}\right\}}\right)$$

Les π_i^j font tendre les y_i^j vers la modalité duale mais la pondération doit être *raisonnable,* i.e adaptée à la logique mathématique, et en adéquation avec les hypothèses fondamentales de cette thèse qui postulent que « l'environnement géographique a un effet sur la santé donc sur les PM* étudiés - séquelles ».

<u>Conséquences</u> :
L'incertitude maximale est bornée, i.e. $\left|\pi_i^j\right| \ll 0{,}5$ valeur où on ne distingue plus une séquelle d'une absence de séquelle et au-delà de laquelle l'interprétation mathématique de la stratégie de pondération devient une ineptie.

Aussi s'il existe bien *un effet environnement,* alors la valeur maximale des $\left|\pi_i^j\right|$ doit être supérieure à l'incidence du PM* calculée sur l'intégralité des patients spatialisés, i.e. :

$$\max_{i=\{1,\ldots,n\}} \left|\pi_i^j\right| = \left(k^j * \text{môy}\left(\left\{y_i^j \middle| x_i^{\text{CP}} \neq \phi\right\}\right) \ \wedge 0{,}5\right)$$

En effet, borner les facteurs d'incertitude EpiGéoStat à $\max_{i=\{1,\ldots,n\}}\left|\pi_i^j\right| = \text{môy}\left(\left\{y_i^j \middle| x_i^{\text{CP}} \neq \phi\right\}\right)$ revient à supposer que l'effet environnement est nul ou qu'il est uniforme dans l'espace géographique*. Or si l'effet environnement est discutable, il est *a priori* difficilement niable.

Enfin les facteurs d'incertitude EpiGéoStat embrassent un ensemble d'incertitudes spatiotemporelles, à connotation : épidémiologiques $\pi_i^{j,\text{epid}}$; géographiques : $\pi_i^{j,\text{geo}}$; statistiques $\pi_i^{j,\text{stat}}$, dans une fonction d'agrégation ensembliste simple. La moyenne empirique est utilisée afin de ne pas écraser les réponses y_i^j à spatialiser et de préserver *l'effet information* associé à chacun des poids* :

$$\pi_i^j = f\left(\xi_i^j; \left\{\pi_i^{j,\text{epi}}; \pi_i^{j,\text{geo}}, \pi_i^{j,\text{stat}}\right\}\right) = \frac{\xi_i^j}{3} \cdot \left(\pi_i^{j,\text{epi}} + \pi_i^{j,\text{geo}} + \pi_i^{j,\text{stat}}\right)$$

Les π_i^j constituent la *métrique floue géographique* qui a pour objet de donner plus de consistance aux i.st.m, de modéliser de façon robuste et juste la réalité spatio-temporelle des PM* étudiés.
Il s'agit maintenant de spécifier explicitement les hypothèses et le principe d'estimation des facteurs d'incertitude et de les appliquer aux données de la Cohorte LEA.

LE FACTEUR D'INCERTITUDE GEOGRAPHIQUE

Objectif :

Estimer l'ampleur des incertitudes spatiales liées aux hypothèses de la méthode de spatialisation. Le facteur d'incertitude géographique $\pi_i^{j,geo}$ est défini par la somme de trois *fonctions fragmentaires d'incertitude spatiotemporelle* :

$$\pi_i^{j,geo} = \sum_{l=1}^{3} \left(\text{i}.\, geo_i^j(l) \right) \in [\![0\,;0{,}50]\!]$$

Remarques générales :

i. Dans la pratique, quel que soit le PM* – séquelle - $\pi_i^{j,Géo} \ll 45\%$. les valeurs sporadiquement élevées du facteur d'incertitude géographique concernent les patients spatialisés cumulant plusieurs incertitudes spatiales fragmentaires, i.e. ceux dont : le CP est incertain ; la superficie de l'U_k est grande, et la probabilité de déménager est plus forte.

ii. Les hypothèses et les stratégies d'estimation des *fonctions fragmentaires d'incertitudes spatiotemporelles* sont déclinées subséquemment. Elles sont appliquées à la séquelle : tumeur thyroïdienne (THYR).

1. Incertitudes spatiales inhérentes à la méthode SpaLea :

Hypothèse :

Cette incertitude dépend de : $q_{o,i}$, le nombre de codes INSEE de 2nd ordre attribués à chaque patient. Visuellement l'incertitude de spatialisation est appréciable – pour l'ensemble des patients spatialisés dans une U_k - par le biais de l'indicateur : $\text{SpaLea}^2_{U_{k_o}}$. Or la fonction d'incertitude spatiale fragmentaire se base sur les $q_{o,i}$., dans le cadre d'une approche individus-centrée* préalable à la phase d'agrégation ensembliste dans les U_k.

Remarques liminaires à l'estimation :

i. Certains codes INSEE de 2nd ordre n'existaient pas encore ou plus en 2003. Donc $q_{o,i}$ tend à surestimer l'incertitude associée à la méthode de spatialisation des patients.

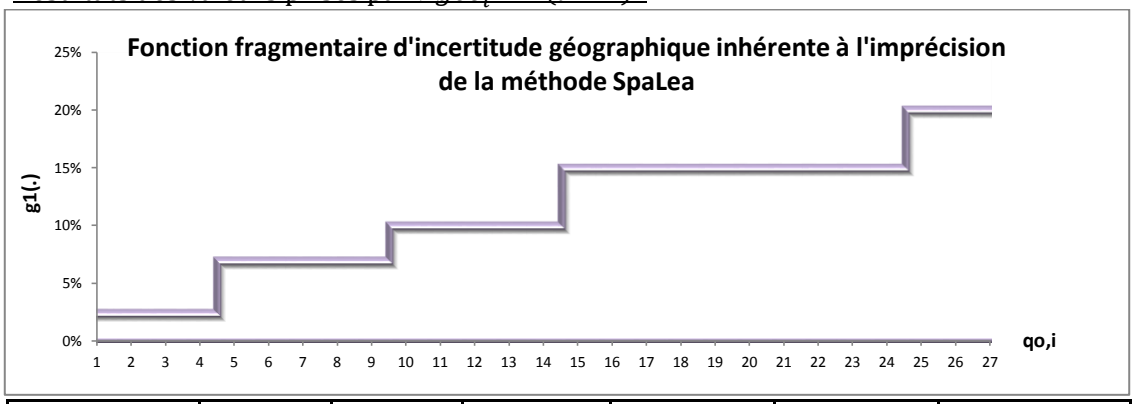
ii. Les codes INSEE de 2nd n'ont pas été pondérés par la surface communale. Cela n'est pas pertinent car les CP sont parfois, même souvent, à cheval sur plusieurs communes (Valetich, 2012).

Stratégie d'estimation :

Les valeurs prises par $\text{i}.\, geo_i^j(l=1)$ sont estimées conditionnellement aux quantiles de la fonction de répartition empirique spatiale des $q_{o,i}$, notée $\hat{F}_n(q_{o,i})$, par une fonction constante par morceaux.

$$\text{i}.\, geo_i^j(l=1) = g_1\left(q_{o,i},\, \hat{F}_n(q_{o,i})\right)$$

Résultats des valeurs prises par i. $geo_i^{\mathrm{THYR}}(l=1)$:



$q_{o,i} \in$	$]\cdot,\cdot]$	$[0,\ 1[$	$[1,\ 5[$	$[5,\ 10[$	$[10,\ 15[$	$[15,\ 25[$	$[25,\ +\infty[$
i. $geo_i^j(1)$		**0%**	**2,5%**	**7%**	**10%**	**15%**	**20%**

Figure 29 : Synthèse des valeurs attribuées par la fonction constante par morceaux pour la séquelle : THYR

La valeur maximale de cette incertitude géographique fragmentaire est bornée conditionnellement aux hypothèses spécifiées auparavant, et ici par :

$$\mathrm{i.geo_i^j}(1) \approx \left(2 * \mathrm{môy}\left(\left\{y_i^{\mathrm{THYR}} \middle| \mathrm{x_i^{CP}} \neq \phi\right\}\right) \wedge 30\%\right)$$

Remarque :

La dernière colonne du tableau ne devrait pas apparaitre. En effet, les patients ayant un nombre de codes INSEE de 2nd ordre supérieur à 15 sont peu nombreux et leur spatialisation a été vérifiée par le biais des ARC. Par conséquent la valeur maximale prise par i. $geo_i^j(1)$ est bornée à 35%.

2. Incertitudes spatiales inhérentes à l'échelle d'investigation

Hypothèse : Les patients sont spatialisés dans des U_k supposées être la commune du lieu de vie. En supposant que cette localisation soit exacte, ces derniers peuvent résider n'importe où, et il est vraisemblable qu'ils se déplacent quotidiennement dans cette enceinte pour des motifs familiaux, sociaux, économiques ou culturels. L'incertitude qui porte sur la variabilité, à micro-échelle, des expositions environnementales subies est intimement liée à l'échelle d'investigation, i.e. à la surface territoriale de l'U_k.

Proposition :

Plus la surface territoriale de l'U_k est grande et plus l'incertitude liée à l'échelle d'investigation est forte.

Remarque liminaire à l'estimation :

Les habitants des petites communes ont tendance à se déplacer sur de plus grandes distances, notamment dans les communes voisines, que ceux qui vivent dans des plus grandes. En, effet les déplacements quotidiens sont fortement conditionnés par les fonctionnalités et les aménités des agglomérations. Par conséquent, la variabilité des expositions environnementales dans les petites communes est plus liée à l'incertitude qui pèse sur la localisation du lieu d'habitation qu'aux mobilités quotidiennes (Hubert, 2009).

Stratégie d'estimation :

l'incertitude géographique inhérente à l'échelle d'investigation dépend de la surface de l'unité géographique dans laquelle l'individu est spatialisé : $S_{i_{(\mathrm{U}_k)}}$. Mais cette relation n'est pas *a priori* linéaire. Il convient de choisir une fonction mathématique avec une forte croissance à l'origine et qui s'atténue ensuite. En l'occurrence une allure logarithmique a été retenue.

$$\mathrm{i.}\, geo_i^j(l=2) = g_2\left(S_{(i|\mathrm{U}_k)};\ \hat{\alpha}, \hat{\beta}\right) = \hat{\alpha} \cdot \ln\left(1 + \hat{\beta} \cdot S_{(i|\mathrm{U}_k)}\right),\ \ \forall i \in \{1, \ldots, n\}$$

Le paramètre de forme $\hat{\beta}$ est basé sur les valeurs de la distribution statistique des $S_{i_{(\mathrm{U}_k)}}$

$$\hat{\beta} = \frac{e^1 - 1}{\hat{Q}_3\left(S_{(i|\mathrm{U}_k)}\right) - \hat{Q}_1\left(S_{(i|\mathrm{U}_k)}\right)} \approx 2{,}27 \times 10^{-5}$$

Tableau des principaux paramètres statistiques des $S_{(i|\mathrm{U}_k)}$

Paramètres	mîn(.)	Q1^(.)	mêd(.)	Q3^(.)	mâx(.)	môy(.)	σ^(.)
S(Uk).i	158	1313	2469	5368	75968	5725,3	8914,39

Tableau 19 : Synthèse statistique des surfaces géographiques communales attribuée aux patients

Résultats, valeurs prises par i. $geo_i^{\mathrm{THYR}}(l = 2)$ - séquelle THYR :

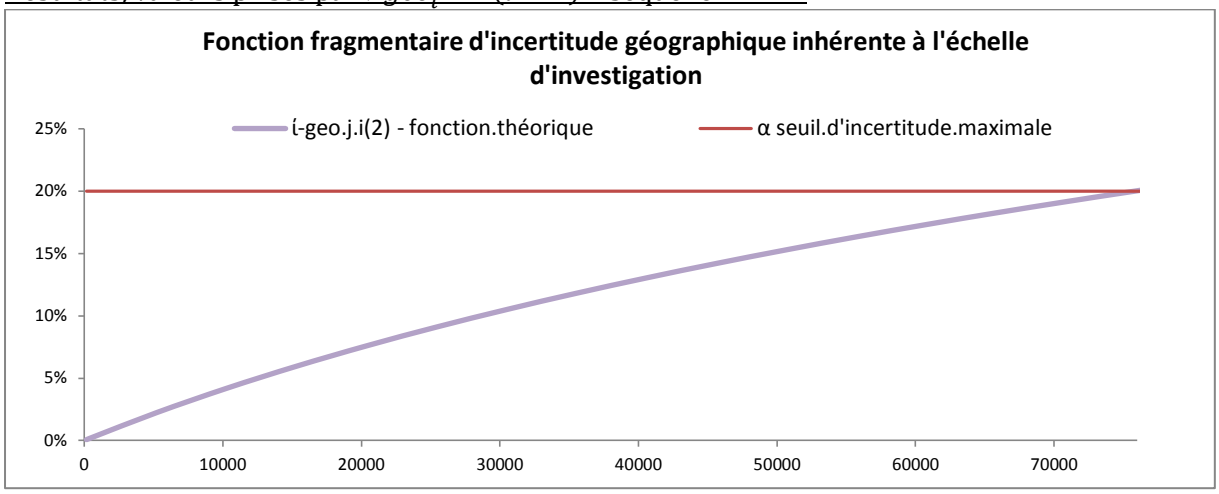


Figure 30 : Synthèse des valeurs attribuées aux patients par la fonction fragmentaire d'incertitude inhérente à l'échelle d'investigations pour la séquelle : THYR

La valeur maximale de cette incertitude fragmentaire géographique d'échelle est bornée afin de ne pas écraser de façon aberrante les valeurs prises par y_i^j *– la valeur de la borne est :*

$$\hat{\alpha} = 20\% \approx \left(\mathbf{2} * \mathrm{môy}\left(\left\{y_i^{\mathrm{THYR}} \middle| x_i^{\mathrm{CP}} \neq \phi\right\}\right) \wedge 20\%\right)$$

Remarque :

Les paramètres de la fonction mathématique associée à i. $\mathrm{geo}_{\mathrm{i}}^{\mathrm{j}}(\mathrm{l} = 2)$ permettent d'atténuer les valeurs prises par i. $geo_i^j(l = 1)$. L'incertitude liée à la localisation résidentielle est estimable précisément par le biais des $S_{(i|\mathrm{U}_k)}$, à l'aune de celle, liée aux déplacements quotidiens qui ne l'est pas.

3. Incertitudes spatiales associées aux mobilités résidentielles

Remarques liminaires générales :

i. Lors de la phase d'agrégation ensembliste des y_i^j dans les U_k, il est particulièrement préjudiciable d'attribuer le même poids aux individus qui y résident depuis toujours et à ceux qui n'y résident plus ou qui viennent juste de s'y installer et qui par conséquent ont côtoyé des environnements différents.

ii. La reconstitution des mobilités résidentielles à partir des données LEA est impossible. De fait, l'incertitude spatiale associée aux trajectoires de vie à moyen et long termes sera estimée à partir de données communautaires.

Hypothèses :

i. On suppose que les patients résident depuis $T_i = \left\|\mathrm{DATE}\left(x_i^{CP}\right) - \mathrm{DATE}\left(x_i^{PPT}\right)\right\|$ dans la commune de première espèce dans laquelle ils sont spatialisés, avec : T_i la distance temps qui sépare la date de prise du premier traitement et celle à laquelle la spatialisation est supposée exacte.

ii. Pour estimer les mobilités extra-communales les données géographiques communautaires les plus fiables sont celles l'INSEE.

Connaissances expertes et données disponibles :

Pour des raisons *techniques, financières, juridiques et sociales*, il n'est pas possible de construire un indicateur géographique permettant d'estimer les mobilités résidentielles réelles et la temporalité, à l'échelle des communes en France.

Les seules données INSEE permettant d'estimer les mobilités résidentielles extra-communales sont celles de l'enquête *Logement 2006*, qui permettent de lier *intention de déménager* et *durée d'occupation du logement* (Couet, 2006).

Intentions de mobilité

Désir de changer de logement selon l'ancienneté d'occupation de la personne de référence (en %)

Ancienneté d'occupation	France Province
Moins d'un an	30
De 1 à moins de 4 ans	35
De 4 à moins de 8 ans	29
De 8 à moins de 12 ans	23
12 ans et plus	10
Ensemble	21

Source : Insee, enquête Logement 2006

Tableau 20 : Intentions de mobilités résidentielles, selon l'enquête Logement 2006
Source : INSEE

Propositions :

i. L'incertitude fragmentaire spatiale inhérente aux trajectoires de vie est estimable à partir des *intentions de déménager* $\mathrm{idd_t}$. Pour cela il convient d'assimiler *la durée d'occupation du logement* $\mathrm{dol_t}$ à T_i et de choisir une fonction mathématique adaptée : i. $\mathrm{geo_i^j}(l = 3)$.

ii. L'enquête fait ressortir qu'au-delà d'une *durée maximale d'occupation du logement* $dol_{\max}^{\sim}$, la *probabilité* de *changer de logement* est à la fois stable et faible.

iii. Afin de ne pas pondérer de façon exagérée les $\mathrm{y_i^j}$ et dans la mesure où les $\mathrm{idd_k}$ surestiment les mobilités résidentielles extra-communales – qui elles, prennent en compte les intra-communales – un *coefficient d'abattement* est introduit : $\tilde{C}$, de sorte que i. $geo_i^j(3) = g_3\left(\{dol_t \equiv T_i\}, \hat{\theta}_m ; \tilde{C}, \left|\hat{C}\right.\right) \leq 15\%$;

Avec : $\widehat{\Theta}_m$ est un vecteur de paramètre estimé.

Stratégie d'estimation :

i. Le modèle choisi est celui de la loi gamma ; Où $\hat{\theta}_m$ est spécifié par le critère des moindres carrés ordinaires (mco) ; (Saporta, 2006).

ii. Les valeurs du vecteur de paramètre $\hat{\theta}_m = \left(\hat{C}; \hat{\theta}; \hat{k}\right)$ sont d'abord estimées numériquement de façon à ajuster $g_3\left(T_i, \hat{\theta}_m \left|\hat{C}\right.\right)$ aux $\mathrm{idd_t}$. Ensuite les valeurs de $\tilde{C}$et de $idd_{\max}^{\sim}$ sont fixées conditionnellement aux propositions énoncées auparavant ; Conséquence :

$$\text{i. } geo_i^j(3) \equiv g_3\left(T_i, \hat{\theta}_m ; \tilde{C}, idd_{\max}^{\sim} \left|\hat{C}\right.\right) = \tilde{C} \cdot (T_i)^{\hat{k}-1} \cdot \frac{e^{\left(-\frac{T_i}{\hat{\theta}}\right)}}{\Gamma\left(\hat{k}\right) \cdot \hat{\theta}^{\hat{k}}} \cdot \mathbb{1}_{\{T_i \leq dol_{\max}^{\sim}\}} + idd_{\max}^{\sim} \cdot \mathbb{1}_{\{T_i \leq dol_{\max}^{\sim}\}}$$

Résultats, valeurs prises par i. $geo_i^{\mathrm{THYR}}(3)$ - séquelle THYR :

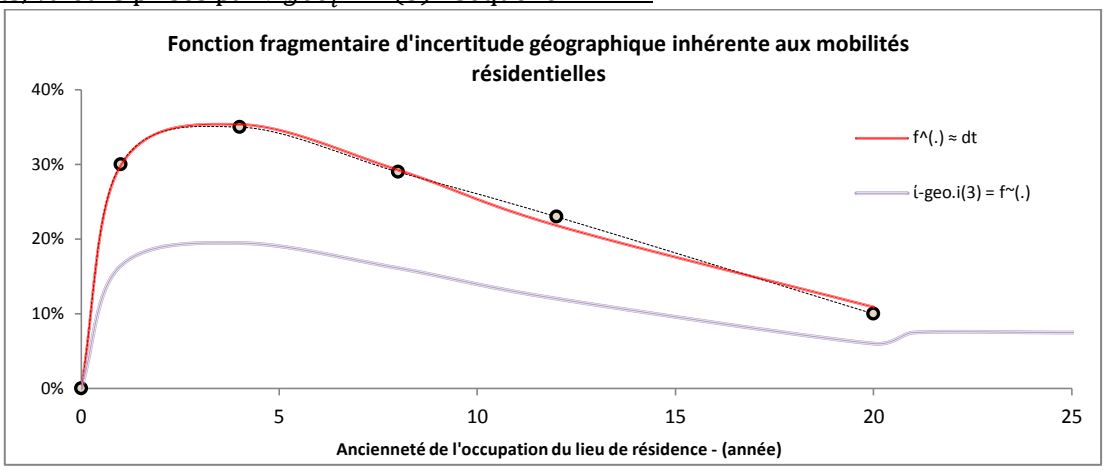


Figure 31 : Valeurs attribuées aux patients par la fonction fragmentaire d'incertitude inhérente aux mobilités résidentielles pour la séquelle : THYR

Les paramètres spécifiés pour cette séquelle sont présentés dans le graphique ci-dessous :

Paramètres	$\hat{k}$	$\hat{\theta}$	$\hat{C}$	$\tilde{C}$	$idd_{\max}^{\sim}$
valeurs	1,36	9,12	5,99	3,30	0,075

Tableau 21 : Vecteur des paramètres utilisés dans la fonction fragmentaire d'incertitude inhérente aux mobilités résidentielles pour la séquelle : THYR

Pour que cette valeur *ait une contrepartie interprétable dans la réalité géographique et en vertu du caractère a-spatial d'éloignement des données INSEE utilisées avec les patients, la valeur maximale de l'incertitude spatiale inhérente aux trajectoires est bornée par :*

$$\max_{\forall i=\{1,\dots,n\}}\{\mathrm{i}.\, geo_i^{THYR}(3)\} \leq 15\% = \left(\mathbf{2} * \mathrm{môy}\left(\{y_i^{\mathrm{THYR}} | x_i^{\mathrm{CP}} \neq \phi\}\right) \wedge 15\%\right)$$

Remarques :

i. Il est important de noter que i. $geo_i^j(l=3)$ est peu fiable, raison pour laquelle son action au sein du système de pondération est la plus faible.

ii. Aussi il est impératif de notifier que par définition, les T_i ne sont pas redondants avec les tee_i^j.

iii. Le concept de Distance a-spatiale morbide* est défini dans cette recherche comme la plausibilité théorique des effets avérés ou suspectés sur la santé des populations des expositions environnementales à des situations à risque ou à des substances physicochimiques

La stratégie d'estimation des $\pi_i^{j,géo}$ est spécifiée, il convient de passer à celle des $\pi_i^{j,épi}$.

LE FACTEUR D'INCERTITUDE SPATIALE EPIDEMIOLOGIQUE

Objectif :

Estimer pour chaque patient, par le biais d'un facteur d'incertitude épidémiologique $\pi_i^{j,épi}$, la qualité spatiotemporelle des données épidémiologiques morbides.

Hypothèse :

Lors de l'agrégation ensembliste des variables morbides y_i^j, dans les unités géographiques, ces dernières ne peuvent pas contribuer la même façon à l'estimation des i.st.m* car elles ont des qualités spatiotemporelles épidémiologiques différentes.

Remarque liminaire :

La stratégie d'estimation du facteur d'incertitude épidémiologique $\pi_i^{j,\text{épi}}$ a été effectuée *sous l'égide d'experts* (Auquier et Michel, 2012).

Proposition :

La stratégie d'estimation des $\pi_i^{j,épi}$ s'opère dans une logique dyadique et se fonde sur un principe de scoring SCORE_i^j prenant en compte plusieurs critères permettant d'estimer l'ampleur des incertitudes spatiotemporelles à partir de données épidémiologiques. On distingue deux cas :

i. Le patient a développé la pathologie $\{y_i^j = 1\}$, cette information est connue avec certitude

ii. Le patient n'a pas développé la pathologie $\{y_i^j = 0\}$, cette information est moins sûre et l'incertitude qui lui est associée peut être estimée à partir de quatre variables LEA.

$$\pi_i^{j,épi} = f_{épi}\left(\text{SCORE}_i^j = \begin{cases} 0 & \text{lorsque: } y_i^j = 1 \\ \left(l_i^j + s_i^j + Q(\text{t}_i) + \text{m}_i\right) & \text{lorsque: } y_i^j = 0 \end{cases}\right)$$

Spécification de la proposition appliquée à LEA :

Lorsque le patient n'a pas développé la séquelle, la qualité spatiotemporelle associée à la variable y_i^j dépend de quatre critères épidémiologiques :

i. <u>La qualité informationnelle de la BD</u>:

Elle diffère d'un patient à l'autre selon que la variable y_i^j est une donnée renseignée ou une lacune - systématiquement comblée « par absence de séquelle ». Les épidémiologistes partent du principe que si un patient développe une séquelle alors l'information finira par « remonter à un moment ou à un autre ». Il n'empêche que lors de l'agrégation ensembliste dans les U_k, une lacune $\{y_i^j = ?\}$ comblée par « non » ne peut pas contribuer de la même façon qu'une absence de séquelle avérée. Dans ce cas le score d'incertitude est augmenté d'un point :

$$l_i^j = \mathbb{1}_{\{y_i^j = ?\}}$$

ii. <u>La qualité informationnelle du suivi :</u>

Deux absences de séquelle n'ont pas le même poids selon que les patients font ou non l'objet d'un dépistage systématique. Les conditions médicales de suivi des séquelles dépendent de l'agressivité du traitement reçu. La recherche de tumeur secondaire est systématique. Celle de tumeur thyroïdienne concerne les patients greffés ou irradiés. Les cataractes sont recherchées pour les mêmes motifs que les thyroïdes et aussi chez les patients qui ont été traités avec une corticothérapie (Michel, Auquier et al., 2007). De fait, un patient non suivi systématiquement prend un point de pénalité :

$$s_i^j = \mathbb{1}_{\{y_i^j = 0\} \cap \{\text{suivi}_i^j = \text{NON}\}}$$

iii. La capacité du patient à développer des séquelles :

Il convient de différencier les patients déclarés en vie et ceux qui sont décédés. Les défunts ne développent plus de séquelle et ne sont plus exposés à l'environnement géographique. Afin d'assurer la reproductibilité de la méthode et de prendre en compte cette spécificité, la variable SCORE_i^j est incrémentée de un, pour les défunts, i.e. :

$$\mathrm{m}_i = \mathbb{1}_{\{y_i^j=0\}\cap\{\text{décédé}_i=\text{OUI}\}}$$

iv. La qualité temporelle de l'information morbide :

Elle est liée à la *temporalité de l'incertitude :* ti_i. Deux patients spatialisés ne peuvent pas contribuer de la même façon sachant que le premier a été revu dans le courant de l'année et que l'autre n'a pas été revu depuis dix ans. Chez ce dernier la séquelle est peut-être latente s'il n'est pas déjà atteint. On supposera que cette incertitude est une fonction linéairement croissante de ti_i, i.e. de la distance temps qui sépare la dernière consultation et la date à laquelle l'information y_i^j est supposée exacte, i.e. la date de la BD-LEA utilisée. Dans ce cas la valeur du score est augmentée proportionnellement à :

$$Q(\mathrm{ti}_i) = \mathrm{ti}_i \Big/ \left(\max_{\forall i=\{1,\ldots,n\}} \{\mathrm{ti}_i\} - \max_{\forall i=\{1,\ldots,n\}} \{\mathrm{ti}_i\} \right)$$

Spécification de la stratégie d'estimation et application à LEA :

Le facteur épidémiologique d'incertitude spatiotemporelle $\pi_i^{j,epi}$ est défini comme une fonction exponentielle croissante du SCORE_i^j, atténuée linéairement, et bornée de sorte à conserver une cohérence mathématique et rester interprétable d'un point de vue épidémiologique.

$$\pi_i^{j,epi} = \left\{ 1 - \hat{\zeta} \cdot \text{SCORE}_i^j + e^{\left(-\hat{\lambda}\cdot \text{SCORE}_i^j\right)} \,\middle|\, \{\hat{\zeta},\hat{\lambda}\} = \underset{\forall\{\zeta,\lambda\}\in\mathbb{R}_*^2}{\operatorname{argmin}} \sum_{l=1}^{5} \left(\pi_l^{j,epid} - \pi_l^{j,exp}\right)^2 \right\}$$

La valeur maximale de $\pi_i^{j,epi}$ ainsi que les poids épidémiologiques $\pi_l^{j,exp}$ à associer aux valeurs entières de SCORE_i^j ont été évalués *sur dire d'experts* (Auquier et Michel, 2012) ; la fonction mathématique a été choisie de la même façon et ses paramètres ont été ajustés aux $\pi_l^{j,exp}$ par le critère des mco (Saporta, 2006).

Résultats des traitements numériques et valeurs prises par $\pi_i^{CATA,épi}$ - séquelle CATA :

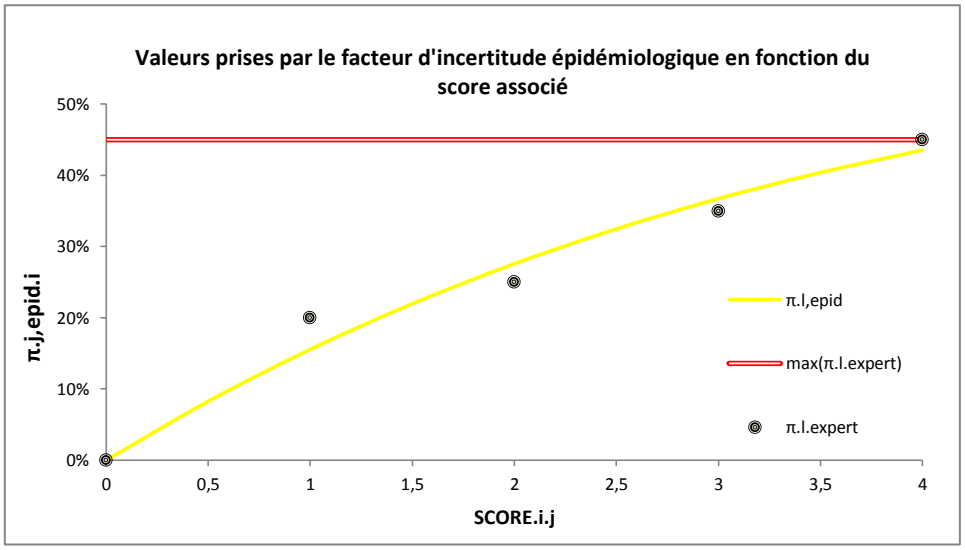


Figure 32 : Valeurs attribuées aux patients par la fonction d'incertitude épidémiologique, pour la séquelle : CATA

Afin de garantir l'adéquation avec l'hypothèse de l'existence d'*un effet environnement* et aussi pour borner la fonction d'incertitude épidémiologique, il a été décidé que :

$$\max_{i=\{1,\dots,n\}}\left\{\pi_i^{j,Epi}\right\} = \left(\left\{k^{CATA} = \mathbf{3}\right\} * môy\left(\left\{y_i^j \middle| x_i^{CP} \neq \phi\right\}\right) \ \wedge 45\%\right)$$

Remarques :

i. Puisque par définition : $ti_i \neq T_i \neq tee_i^j$ alors les variables temporelles ne sont pas redondantes.

ii. Statistiquement il est presque impossible pour les patients d'atteindre un score d'incertitude spatiotemporelle égal à 4. La qualité spatiotemporelle des données épidémiologiques est relativement fiable, par conséquent dans la pratique il a été observé que : $\pi_i^{j,épi} \leq 30\%$.

LE FACTEUR D'INCERTITUDE SPATIALE STATISTIQUE

Objectif :

Estimer pour chaque patient, par le biais d'un facteur statistique d'incertitude spatiotemporelle $\pi_i^{j,stat}$, l'ampleur de l'inconsistance* statistique induite par les i.st.m.

Hypothèse :

Les i.st.m* qui seront proposés ne peuvent pas contribuer de la même façon à l'analyse statistique visant à identifier les DES* géographiques selon qu'un seul patient est spatialisé dans une U_k ou qu'il y en a un grand nombre.

Proposition :

La valeur de $\pi_i^{j,stat}$ est proportionnellement croissante à *l'inconsistance* statistique* associée au processus d'estimation de l'i.st.m, i.e. au nombre de patients $n_{(U_k)}^I$ spatialisés dans chaque U_k. Le facteur statistique d'incertitude spatiotemporelle proposé s'estime de la façon suivante :

$$\pi_i^{j,stat} = f_{Stat}\left(\boldsymbol{n}_{(i|U_k)}^I\right)$$

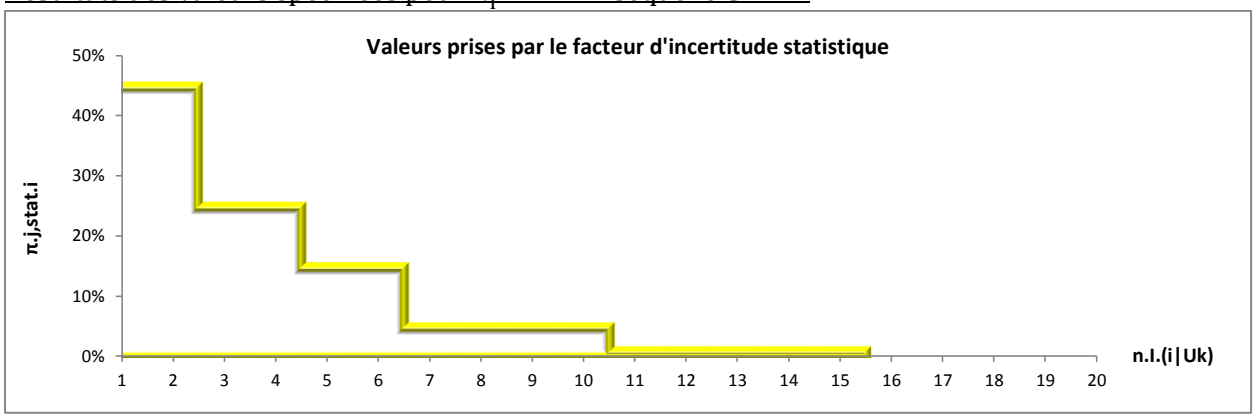
Spécification de la proposition :

Cette incertitude est particulièrement préjudiciable lorsque dans une U_k il n'y a qu'un seul individu spatialisé. *D'un point de vue statistique* l'erreur commise sur l'estimateur est telle qu'il est presque impossible de distinguer une séquelle d'une absence de séquelle (Saporta, 2006).

Stratégie de pondération :

Les valeurs prises par les facteurs statistiques d'incertitude spatiotemporelle $\pi_i^{j,stat}$ sont attribuées par le biais d'une fonction constante par morceaux, doivent être très préjudiciables lorsque $\boldsymbol{n}_{(i|U_k)}^I$ est *petit*.

Résultats des valeurs spécifiées pour $\pi_i^{CATA,stat}$ - séquelle CATA :



$n^{I}_{(i\|U_k)}$	⟦{0}, …, {2}⟧	⟦{2}, …, {4}⟧	⟦{4}, …, {6}⟧	⟦{6}, …, {10}⟧	⟦{10}, …, {15}⟧	⟦{15}, …, {+∞}⟧
$\pi_i^{j,stat}$	45%	25%	15%	5%	1%	0%

Figure 33 : Paramètres spécifiés pour la fonction d'incertitude statistique et valeurs attribuées aux patients, pour la séquelle : CATA

Afin de garantir un caractère préjudiciable il convient d'attacher une importance particulière aux petites valeurs de $n^{I_i}_{(U_k)}$, Donc faire en sorte que k^j soit grand. Pour la séquelle CATA ce paramètre a été spécifié tel que :

$$\pi_i^{j,Stat} \approx \left(\{\mathrm{k^{CATA}} = \mathbf{3,5}\} * \mathrm{môy}\left(\{y_i^j | x_i^{\mathrm{CP}} \neq \phi\}\right) \wedge 45\%\right)$$

Remarque :

Lorsque le nombre d'unités géographiques où $n^{I_i}_{(U_k)} = 1$ est important, le choix de l'échelle d'investigation peut être remis en question. Mais attention, lorsqu'on travaille sur une zone étendue, i.e. un pays, passer à une échelle plus agrégée, i.e. commune vers canton, n'augmente généralement pas la taille des effectifs spatialisés.

En contrepartie, la perte de captation des variabilités environnementales est inéluctable à ce niveau. Quant à l'échelle des départements elle n'est même pas envisageable. Toutefois, lorsque le nombre de U_k où $n^{I_i}_{(U_k)} = 1$ est trop important deux possibilités existent : utiliser une transformation en *Patients-Années* ; Ou basculer dans une logique *individus-centrée* au moment de l'analyse statistique des FE.

La stratégie d'estimation des π_i^j a été spécifiée et appliquée aux données de la Cohorte LEA – et à toutes les séquelles d'intérêt. De fait, la métrique floue géographique a pu être constituée en vue de construire des i.st.m* plus consistants. Les résultats obtenus sont présentés dans cette fin de section.

RESUME, PRESENTATION DES RESULTATS ET REMARQUES

Résumé :

La *métrique floue géographique* est constituée des facteurs d'incertitude spatiotemporelle EpiGéoStat π_i^j qui ont pour but de parvenir à une représentation plus robuste et plus juste de la réalité géographique des PM* – séquelles.

Les facteurs d'incertitude EpiGéoStat π_i^j s'estiment par le biais d'une moyenne empirique des facteurs d'incertitude à connotation : géographique $\pi_i^{j,geo}$, épidémiologique $\pi_i^{j,epi}$ et statistique $\pi_i^{j,stat}$.

Les $\pi_i^j \in [\![-0{,}49;\ 0{,}49]\!]$ sans quoi leurs effets sur les variables épidémiologiques – séquelles - y_i^j seraient aberrants et ininterprétables d'un point de vue épidémiologique. Comme les π_i^j sont dotés d'un signe l'analyse de leurs valeurs est effectuée sur $|\pi_\cdot^j|$.

Présentation des résultats

Le graphique est un diagramme de Tukey multidimensionnel qui représente les valeurs des $|\pi_\cdot^j|$ après application de la stratégie d'estimation des incertitudes EpiGéoStat aux données de la Cohorte LEA, pour chaque séquelle d'intérêt.

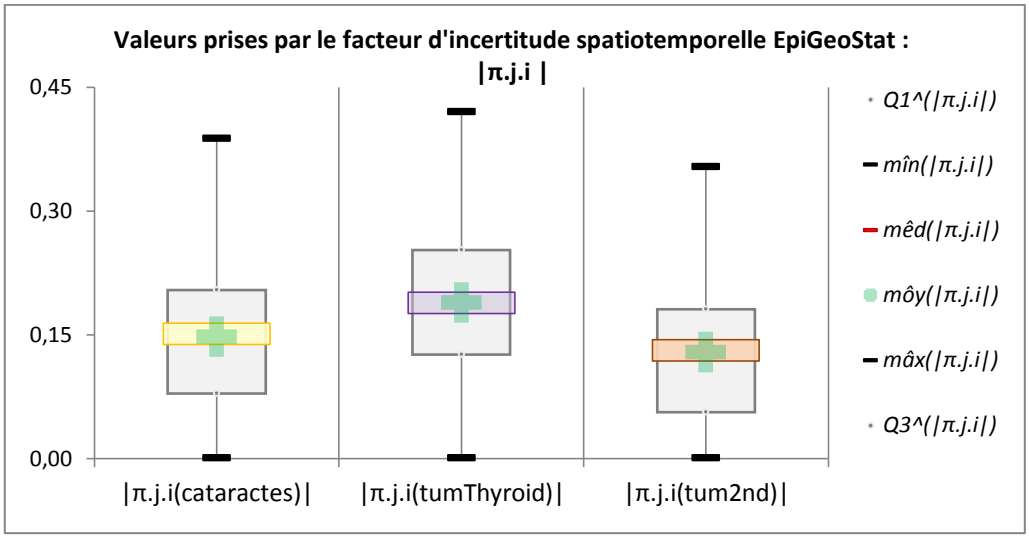


Figure 34 : Diagramme de diagramme de Tukey des valeurs absolues des facteurs d'incertitudes EpiGéoStat, pour les séquelles : CATA, THYR et TUM

Analyses et remarques :

i. Les valeurs prises par les $\pi_i^{j,epi}$ sont beaucoup plus faibles que celles des $\pi_i^{j,stat}$ ou des $\pi_i^{j,geo}$, le processus d'agrégation statistique des poids permet de réguler les valeurs.

ii. Globalement les THYR sont plus incertaines que les CATA qui sont moins sûres que les TUM2.

iii. D'une manière générale 50% des facteurs EpiGéoStat attribués aux patients ont des $|\pi_i^j| \ll 20\%$; Par conséquent, les propositions méthodologiques n'induisent globalement pas de pondérations aberrantes sur les variables morbides y_i^j.

iv. Les valeurs $\max_{i=1,\dots,n}\{|\pi_\cdot^j|\} \in [\![35\%; 43\%]\!]$ ont été attribuées à des patients cumulant les deux incertitudes les plus préjudiciables. i.e. : des I_i spatialisés dans une commune où $n_{(U_k)}^{I_i} = \text{petit}$; Et dont le nombre de codes INSEE de 2nd ordre appariés, $q_{o,i}$ est *grand*.

v. Le nom de la commune de résidence des patients ayant une nombre de codes INSEE de 2nd ordre tel que $q_{o,i} > 15$ a été vérifié par les ARC. De fait, ces patients bénéficient d'une spatialisation exacte, ce qui a permis de diminuer leur $|\pi_\cdot^j|$ en réduisant celle de $\pi_i^{j,géo}$.

vi. Dans la mesure où globalement le nombre de patients spatialisés est faible, il convient de préciser que la méthode de sélection des i.st.e* modélisant la géographie des FE* sera appliquée, d'une part, dans une logique géographique, aux i.st.m* après avoir procédé à une conversion en Patients-Années, et d'autre part, dans le cadre d'une approche *Individus-centrée* afin de corroborer les résultats obtenus par l'approche géographique, et par la même occasion de s'affranchir du problème d'inconsistance* statistique qui lui est associé, ainsi que d'évaluer la robustesse de la *métrique floue géographique* proposée.

L'influence des π_i^j sur le calcul des prévalences spatiales est forte, raison pour laquelle un second indicateur est proposé. Il s'agit des propensions spatiales pondérées, sur lesquelles les π_i^j sont appliqués différemment. Ils n'altèrent pas la nature des variables morbides et attachent une attention particulière à la qualité spatiotemporelles – EpiGéoStat - des données utilisées.

La métrique floue géographique permet d'injecter des *connaissances expertes* épidémiologiques, géographiques et statistiques. Elle grève les i.st.m. d'une notion de qualité spatiotemporelle et par extension améliore leur robustesse spatiale. L'espace est ainsi intégré *transversalement* au cœur de la dialectique géographique. Il est désormais question d'intégrer *verticalement* le temps.

ESTIMATION DES TEMPS D'EXPOSITION A L'ENVIRONNEMENT

Objectif :
Intégrer *verticalement* la notion de temporalité au cœur du processus d'estimation des i.st.m. par le biais d'une stratégie de transformation temporelle qui permet simultanément d'augmenter virtuellement la taille des effectifs spatialisés

Spécification de l'hypothèse :
Lorsqu'on travaille sur des sujets prédisposés il est judicieux d'introduire l'idée de risque temporel. Par exemple chez les individus de la Cohorte LEA le risque de séquelle augmente avec le temps. Quant à la latence, elle diffère selon les caractéristiques environnementales des milieux de vie – du moins c'est ce qui est supposé. Afin de prendre en compte le concept de *risque temporel* des *expositions environnementales potentielles ou intrinsèques* – l'idée est de procéder à une transformation en *Patients-Années*.

Cet artifice permet de dissocier, dans le temps, la *population à risque*, i.e. celle chez qui la pathologie n'a pas encore était décelée des sujets qui l'ont développée. L'avantage de cette transformation est aussi de prendre en compte la vitesse à laquelle les individus sont atteints par les PM* – séquelles. Par conséquent l'intégration du temps d'exposition, dans le processus d'estimation des i.st.m, *améliore l'effet information*-temporel.*
Pragmatiquement, la conversion en *Patients-Années* permet de raisonner non plus sur des patients mais sur des *années d'exposition à l'environnement géographique*. Les effectifs sont donc *virtuellement* augmentés. Le processus d'estimation est simple puisqu'il suffit d'intégrer *le temps d'exposition à l'environnement* (Bernard et Lapointe, 2003).

Spécification du processus d'estimation du temps d'exposition à l'environnement :
On appelle *temps d'exposition à l'environnement* tee_i^j la durée qui sépare l'inclusion dans l'étude épidémiologique et le moment où l'individu développe le PM* étudié. Pour ceux pour qui ne le développent pas, c'est la durée de l'étude épidémiologique – ou *temps de participation* tp_i - qui est considérée - sauf bien sûr si l'individu décède entre temps.
En appliquant cette définition aux données de la cohorte LEA il est possible d'estimer les tp_i et les tee_i^j de tous les patients:

Le temps de participation est la durée qui sépare le diagnostic de la leucémie et la date à laquelle on suppose que les informations $y_i^j = 0$ sont exactes, i.e. celles de la BD-LEA utilisée. Le temps de participation tient compte aussi du fait que le patient est en vie. Autrement dit:

$$\mathrm{tp}_i = \begin{cases} \left(\mathrm{Date.\,BD.\,LEA} - \mathrm{Date.\,deb.\,trait.}\,(\mathrm{I}_i)\right), & \text{lorsque: } \left\{\{y_i^j = 0\} \cap \{\mathrm{m}_i = \mathrm{NON}\}\right\} \\ \left(\mathrm{Date.\,décé.}\,(\mathrm{I}_i) - \mathrm{Date.\,deb.\,trait.}\,(\mathrm{I}_i)\right), & \text{lorsque: } \left\{\{y_i^j = 0\} \cap \{\mathrm{m}_i = \mathrm{OUI}\}\right\} \end{cases}$$

Le temps d'exposition à l'environnement correspond à la durée entre la prise du premier traitement pour la leucémie et le diagnostic d'une séquelle. Quant aux patients n'ayant pas développé cette pathologie, ils constituent l'ensemble des sujets à risque. Dans ce cas le temps d'exposition correspond au

$$\mathrm{tee}_i^j = (\mathrm{Date.\,diag.\,seq.\,j} - \mathrm{Date.\,deb.\,trait.\,I}_i) \cdot \mathbb{1}_{\left\{y_i^j = 1\right\}} + \mathrm{tp}_i \cdot \mathbb{1}_{\left\{y_i^j = 0 \middle| \mathrm{m}_i\right\}}$$

Remarque :
Pour rappel, les patients spatialisés sont ceux de la BD-LEA 2009. Cependant, toutes les variables autres que x_i^{CP} ont été consolidées avec celles de la BD-LEA 2010.

PRESENTATION ET ANALYSE DES RESULTATS

Les résultats des transformations géographiques en Patients-Années sont donnés successivement pour les séquelles : cataractes (CATA), tumeurs thyroïdiennes (THYR) et tumeurs secondaires (TUM2). La variable présentée est *le cumul des Patients-Années spatialisés* dans les U_k, i.e. :

$$n_{(U_k)}^{tee_i^j} = \sum_{i=1}^{n} \left(tee_i^j \cdot \mathbb{1}_{\{V_{i,1}=V_{(U_k)}\}} \right), \qquad \forall k \in \{1, \ldots, q_1\}$$

Les résultats cartographiques :
Ils sont donnés pour les communes de première espèce sises en région PACA et aux alentours pour le Sud-Est de la France ; Et pour les U_k des régions Alsace et Lorraine pour le Nord-Est de la France. Le nom des communes est affiché pour celles qui ont un statut de préfecture, i.e. :

$$label_{(U_k)} = \{Display \quad lorsque: \quad statut_{(U_k)} = \{"prefecteure\ départementale" \cup "préfecture\ de\ régionale"\}$$

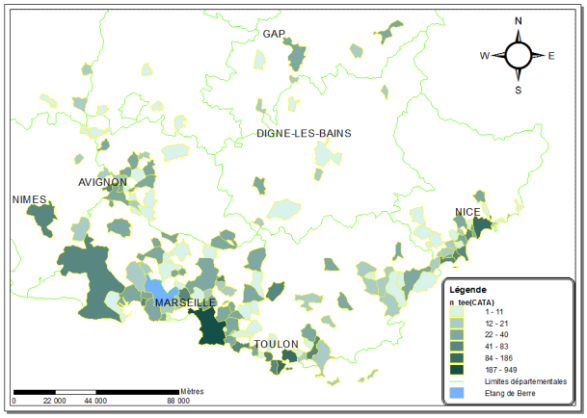
Les résultats statistiques
Ils sont résumés par des histogrammes de la distribution spatiale empirique des cumuls de Patients-Années spatialisés dans l'ensemble des $\mathcal{U}_{\cdot}^{1^{er}}$ – uniquement pour les valeurs discrètes non vides. L'histogramme est complété par un tableau récapitulatif des principaux paramètres statistiques de *position* et de *dispersion* (Saporta, 2006).

Intérêt :
Présenter la distribution spatiale des effectifs spatialisés sur lesquels seront construits les i.st.m* et, en même temps, mesurer l'effet de la conversion en Patients-Années.

RESULTATS OBTENUS POUR LA SEQUELLE : CATA

Présentations cartographiques :

Sud-Est de la France



Nord-Est de la France

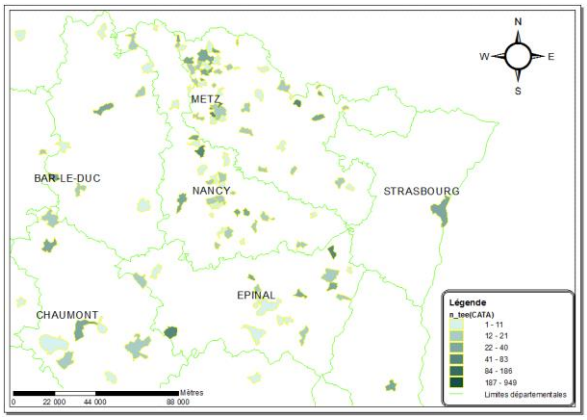


Figure 35 : cartographies des valeurs prises par $n_{\cdot}^{tee\cdot}(CATA)$ dans les U_k

Présentations statistiques :

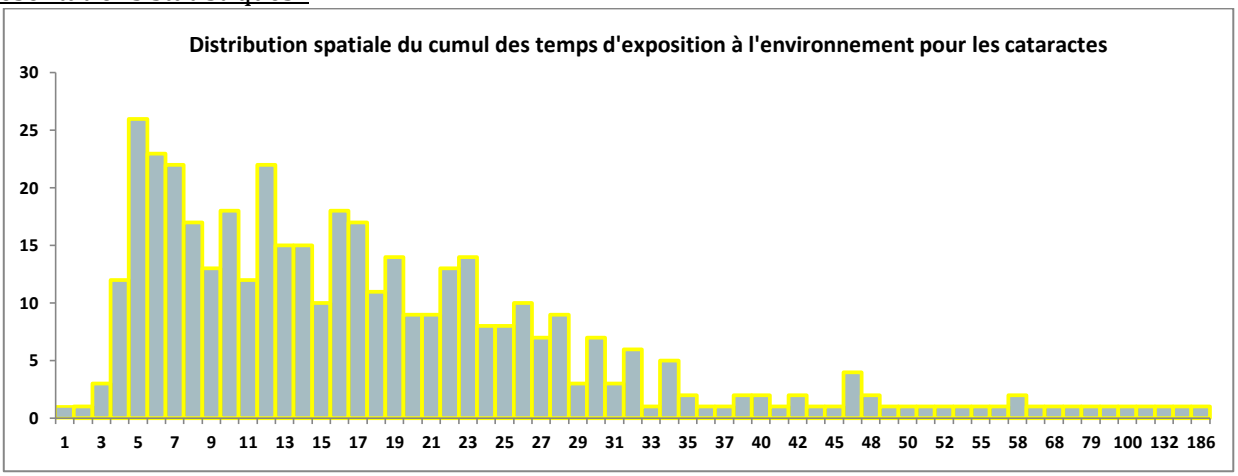


Figure 36 : Histogramme de la distribution spatiale de $n_{\cdot}^{tee\cdot}(CATA)$ pour les U_k sises en France métropolitaine

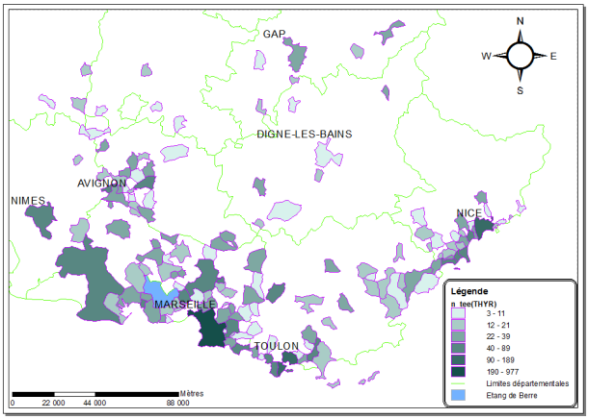
421 - Uk	Paramètres de dispersion & position : **n_tee.i.j-cata_Uk**						
Estimateurs	min(.)	Q1^(.)	mêd(.)	Q3^(.)	max(.)	môy(.)	σ^(.)
Estimation	1	9	16	24	949	21,5	48,84

Tableau 22 : tableau statistique des principaux paramètres associés à la distribution spatiale des $n_{\cdot}^{tee\cdot}(CATA)$

RESULTATS OBTENUS POUR LA SEQUELLE : THYR

Présentations cartographiques :

Sud-Est de la France



Nord-Est de la France

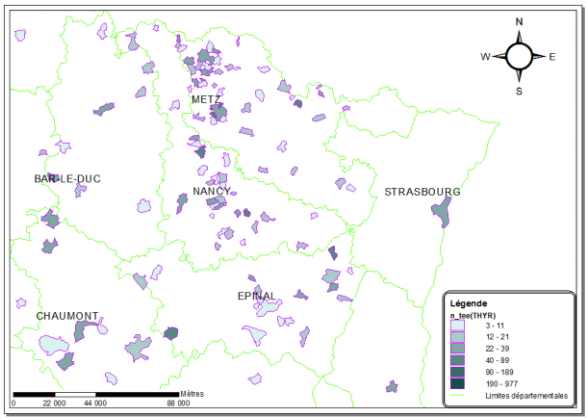


Figure 37 : cartographies des valeurs prises par $n_{.}^{tee.}$(THYR) dans les U_k

Présentations statistiques :

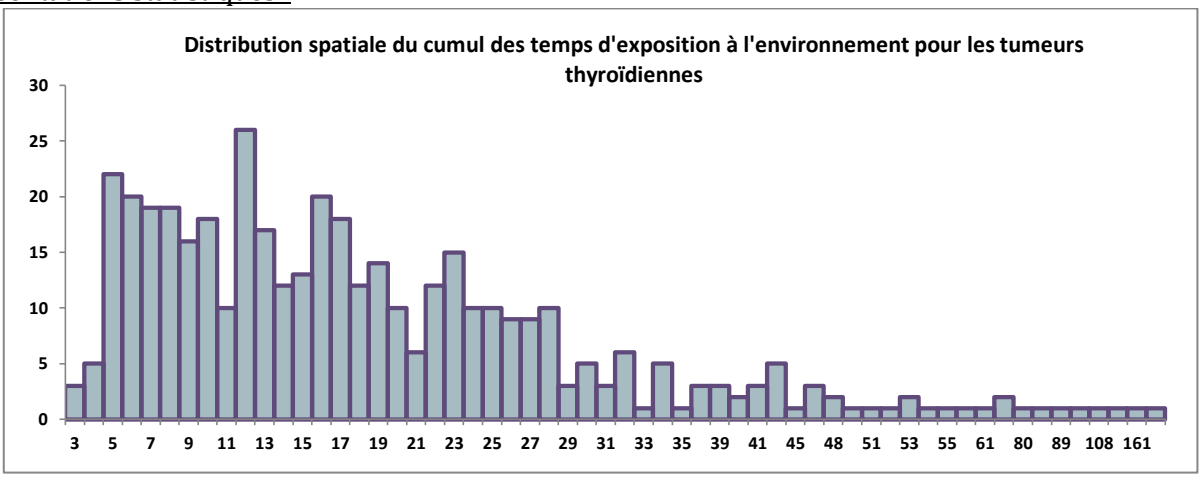


Figure 38 : Histogramme de la distribution spatiale de $n_{.}^{tee.}$(THYR) pour les U_k sises en France métropolitaine

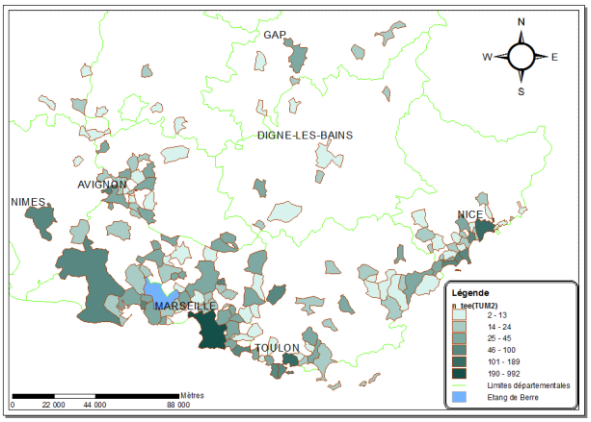
421 - Uk	Paramètres de dispersion & position : **n_tee.i.j-thyr**						
Estimateurs	min(.)	Q1^(.)	mêd(.)	Q3^(.)	max(.)	môy(.)	σ^(.)
Estimation	3	10	16	24	977	22,3	50,13

Tableau 23 : Tableau statistique des principaux paramètres associés à la distribution spatiale des $n_{.}^{tee.}$(THYR)

RESULTATS OBTENUS POUR LA SEQUELLE : TUM2

Présentations cartographiques :

Sud-Est de la France



Nord-Est de la France

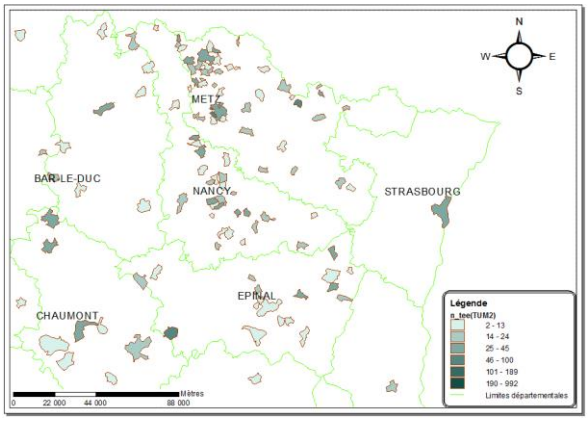


Figure 39 : Cartographies des valeurs prises par $n_{.}^{tee.}(TUM2)$ dans les U_k

Présentations statistiques :

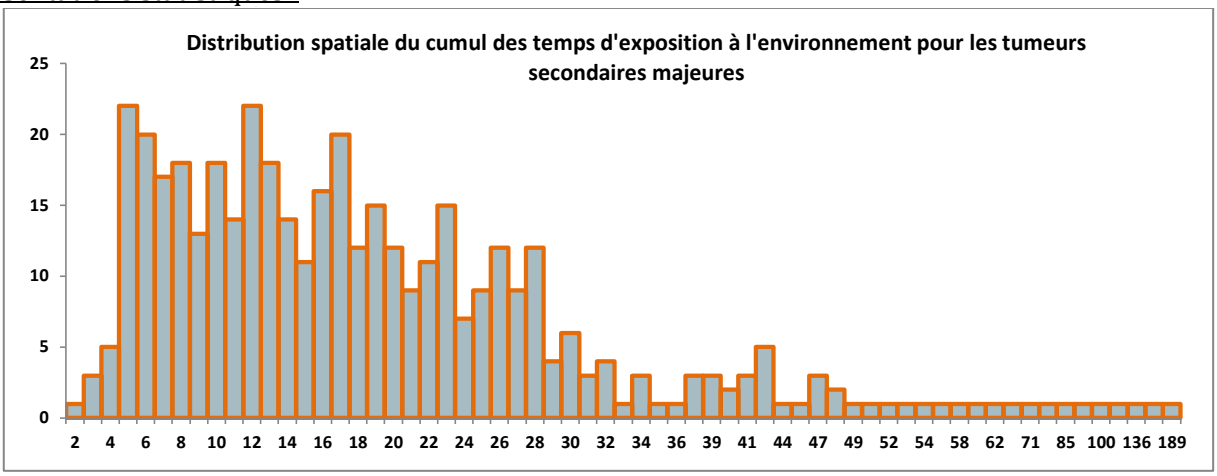


Figure 40 : Histogramme de la distribution spatiale de $n_{.}^{tee.}(TUM2)$ pour les U_k sises en France métropolitaine

421 - Uk	Paramètres de dispersion & position : **n_tee.i.j-tum2**						
Estimateur	min(.)	Q1^(.)	mêd(.)	Q3^(.)	max(.)	môy(.)	σ^(.)
Estimation	2	10	16	25	992	22,5	50,79

Tableau 24 : Tableau statistique des principaux paramètres associés à la distribution spatiale des $n_{.}^{tee.}(TUM2)$

RESUME, ANALYSE ET REMARQUES

Les cartographies présentées et surtout les histogrammes de la distribution spatiale des $n_{.}^{tee}$ montrent que la conversion en *Patients-Années* est un artifice épidémiologique très efficace pour *augmenter virtuellement* les effectifs spatialisés, tout en conservant une échelle parfaitement adaptée à l'analyse géographique des interactions santé-environnement.

Les effectifs spatialisés dans les U_k sont parfois très faibles. L'utilisation d'une échelle plus agrégée, i.e. celle des cantons, a été testée et n'apporte strictement aucune amélioration à cet effet. En contrepartie, elle induit une perte indubitable au niveau de la captation géographique des variabilités environnementales (chapitre.1) ; (chapitre.3). La transformation en Patients-Années permet d'éluder, en partie au moins, l'inconsistance* statistique inhérente à la dialectique géographique. Elle est donc parfaitement adaptée à la modélisation géographique des PM* de petites cohortes, dont LEA fait partie.

D'un point de vue heuristique, la conversion en *Patients-Années* est un moyen d'intégrer conjointement l'effet des expositions environnementales temporelles morbides et la latence des maladies. Elle peut être appliquée à n'importe quel autre PM, et par extension à n'importe quelle autre séquelle développée par les patients de la Cohorte LEA.

En somme, le concept du *temps d'exposition à l'environnement* permet de positionner *verticalement la temporalité* dans la dialectique géographique. L'espace quant à lui est introduit *transversalement* par le biais de *la métrique floue géographique*.

Il s'agit désormais d'énoncer les principes d'intégration de ces systèmes de pondération spatiotemporelle dans l'estimation des deux i.st.m. proposés en vue de modéliser de façon plus robuste et plus juste la géographie de PM* - séquelles. Les propositions heuristiques sont appliquées aux séquelles d'intérêt à partir des données de la Cohorte LEA.

ESTIMATION DES PREVALENCES SPATIALES PONDEREES

Description :
Il s'agit d'un i.st.m* quantitatif, noté $z'^{j}_{(U_k),c}$, qui modélise la géographie d'un PM* – séquelle - par une prévalence à laquelle sont appliquées des pondérations spatiotemporelles pour améliorer sa robustesse.

Hypothèse :
La conversion en patients-années suppose *a priori* qu'il existe un effet environnement. Elle permet de distinguer, dans le processus d'estimation, les patients à risque i.e. ceux qui n'ont pas encore développé la pathologie, de ceux qui l'ont développée mais en tenant compte aussi de la latence, i.e. de la vitesse à laquelle celle-ci s'est manifestée.

Principe d'estimation :
La somme des pathologies - séquelles - développées pondérées par des facteurs d'incertitude spatiotemporelle inhérents à la qualité EpiGéoStat des données et des hypothèses liées à la spatialisation, et réduite par le temps d'exposition à l'environnement de la commune de 1ère espèce.

$$z'^{j}_{(U_k),c} = \frac{1}{n^{\text{tee}^j_i}_{(U_k)}} \cdot \sum_{i=1}^{n} \left(y_i^j + \pi_i^j\right) \cdot \mathbb{1}_{\left\{V_{i,1}=V_{(U_k)}\right\}}$$

Avec : tee_i^j le temps d'exposition à l'environnement ; Et le nombre de Personnes-Années spatialisées, dans chaque U_k, qui est donné par :

$$n^{\text{tee}^j_i}_{(U_k)} = \sum_{i=1}^{n} \left(\text{tee}_i^j \cdot \mathbb{1}_{\left\{V_{i,1}=V_{(U_k)}\right\}}\right)$$

Rappel :
Les $y_i^j = \{\text{"oui "; "non"}\}$ est la variable séquelle.j associée à l'individu i elle vaut 1 lorsque le sujet a développé la pathologie et zéro dans le cas contraire (chapitre.1).

ESTIMATION DES PROPENSIONS SPATIALES PONDEREES.

Description :
Il s'agit d'un i.st.m* qualitatif : $z'^{j}_{(U_k),q}$ qui modélise la propension qu'ont les individus à développer, dans une U_k, une pathologie, en attachant une importance particulière à la qualité spatiotemporelle des données utilisées.

Hypothèse :
Afin de spécifier la qualité spatiotemporelle des informations prises par l'i.st.m* quantitatif : $z'^{j}_{(U_k),c}$, il convient de construire un i.st.m* qualitatif prenant en compte à la fois : La qualité spatiotemporelle EpiGéoStat des données, et conjointement la connaissance temporelle de l'historique médical et du temps d'exposition à l'environnement géographique des patients. Le temps d'exposition à l'environnement présuppose *a priori* un effet environnement et la prise en compte conjointe du temps de participation renforce le poids des pathologies développées de façon précoce.

Principe d'estimation :
Les propensions spatiotemporelles à développer la pathologie – séquelle - sélectionnée EpiGéoStat et répétée au prorata du rapport participation exposition : rpe_i^j attribuent la modalité la plus probable parmi trois : OUI, le nombre de pathologies – séquelles - développées par les individus spatialisés est *curieusement* élevé ; NON ce n'est pas le cas, Ou INCERTAIN, i.e. que la qualité EpiGéoStat associée aux

variables morbides y_i^j spatialisées répétées est trop mauvaise pour que l'on puisse se prononcer à ce sujet dans l'U_k considérée.

$$z'^{j}_{(U_k),q} = \begin{cases} \text{OUI} & \text{lorsque: } \mathbb{P}_{F_n}\left(zq'^{j}_{\{.|U_k\}} = \text{OUI}\right) \geq \kappa^{j}_{\{V_{i,1}\neq\phi\}} \\ \text{INCERTAIN} & \text{lorsque: } \mathbb{P}_{F_n}\left(\left\{zq'^{j}_{\{.|U_k\}} = \mathcal{C}^j\right\}\right) = \phi \\ \text{môd}\left(zq'^{j}_{\{.|U_k\}}\right) & \text{lorsque: Sinon} \end{cases}$$

Proposition :
Maximiser *l'effet information** (Wackernagel, 2003), i.e préserver : la nature de la variable épidémiologique, l'intégration de la notion de qualité EpiGéoStat, l'intégration du temps de d'exposition à l'environnement et de la fiabilité des connaissances médicales disponibles.

Principe de sélection des pathologies dont la qualité est statistiquement fiable :
Les *facteurs d'incertitude spatiotemporelle* π_i^j *ne sont pas directement* appliqués aux variables épidémiologiques, en revanche ils permettent de sélectionner, dans chaque U_k, un sous-ensemble de variables morbides : y_i^j dont la qualité spatiotemporelle *est jugée statistiquement admissible* pour être prise en compte. *Le seuil de sélection* ψ_π^j *de fiabilité EpiGéoStat* retient toutes les variables épidémiologiques dont la valeur absolue du π_i^j n'est pas *anormalement faible.* Il s'estime de la façon suivante :

$$\psi_\pi^j = \left| \text{môy}\left(|\pi_.^j|\right) - t_{(\alpha)} \cdot \frac{\hat{\sigma}\left(|\pi_.^j|\right)}{\sqrt{n}} \right|$$

Avec : $t_{(\alpha)} \xrightarrow[n\to+\infty]{} N\sim\mathcal{N}(0,1),\ \mathbb{P}\left(N \leq t_{(\alpha)}\right) = \{\alpha = 5\%\}$

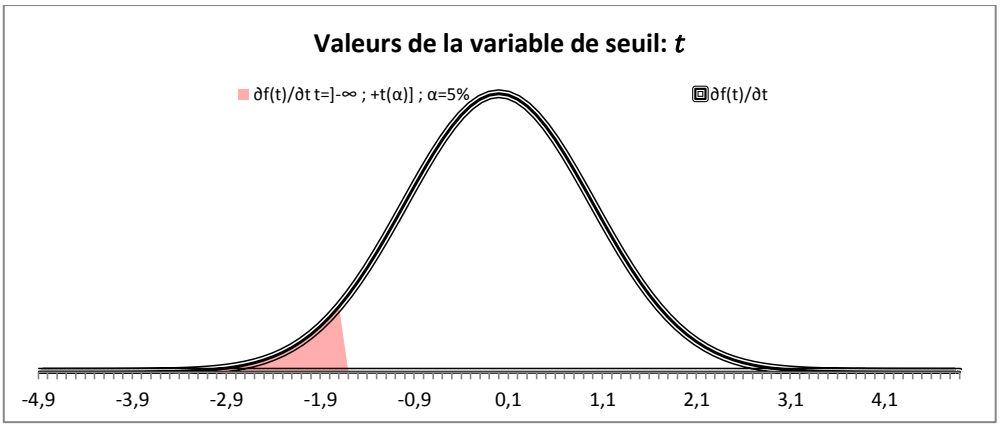


Figure 41 : Représentation unilatérale du niveau de risques statistiquement admis pour l'estimation de ψ_π^j

Principe de répétition temporelle des pathologies au prorata du rapport participation-exposition :
Le ratio de participation-exposition est le rapport, arrondi à l'entier supérieur, entre le temps de participation et le temps d'exposition à l'environnement :

$$rpe_i^j = \left\lceil \frac{tp_i}{tee_i^j} \right\rceil$$

Avec : tp_i et tee_i^j respectivement le temps de participation et le temps d'exposition à l'environnement.

Ce rapport vaut 1 lorsque le patient est déclaré en vie et qu'il n'a pas développé la pathologie - séquelle. En contrepartie, il est supérieur à 1 dès lors qu'il l'a développée ou qu'il est décédé. Cette valeur est d'autant plus élevée que la séquelle est développée de façon précoce et que la durée de participation est importante.

Définition du sous ensemble géographique des pathologies EpiGéoStat statistiquement fiables et répétées au prorata du rapport participation-exposition :

Il s'agit d'un sous-ensemble des pathologies – séquelles - y_i^j, spatialisées dans une U_k, que l'on répète au prorata de la valeur du ratio participation-exposition. Cet ensemble confère un poids important aux pathologies – séquelles - développées rapidement si tant est que la qualité spatiotemporelle EpiGéoStat des données soit suffisamment fiable pour être prise en compte. Le sous-ensemble est défini par :

$$zq'^{j}_{\{.|U_k\}} = \bigcup_{i=1}^{n}\left(\left(\bigcup_{b=1}^{rpe_i^j}\left(y_i^j \cdot \|_{\{b\}} \left| \{V_{i,1} = V_{(U_k)}\} \cap \left\{\left|\pi_i^j\right| > \psi_\pi^j\right\}\right.\right)\right)\right)$$

Proposition du seuil d'identification de la propension géographique à développer la pathologie :

Ce seuil permet de détecter un nombre de pathologies EpiGéoStat statistiquement *fiables*, et que le ratio participation-exposition répète un nombre de fois *curieusement élevé*. L'i.st.m* qualitatif prend la valeur OUI si la probabilité empirique des $\left\{zq'^{j}_{\{.|U_k\}} = \text{OUI}\right\}$ est deux fois plus grande que l'incidence de la pathologie estimée sur l'ensemble des individus spatialisés. Le seuil de détection des propensions géographiques à développer la pathologie –séquelle - est donné par :

$$\kappa^j_{\{V_{i,1}\neq\phi\}} = 2 \cdot \hat{m}oy\left(\bigcup_{i=1}^{n}\left(y_i^j \middle| \{V_{i,1} \neq \phi\}\right)\right)$$

Les valeurs du seuil de détection utilisées :

Séquelle :	CATA	THYR	TUM2
κ.j.{v.i,1≠Φ}	25,97%	20,35%	8,03%

Tableau 25 : Valeurs prises par $\kappa^j_{\{V_{i,1}\neq\phi\}}$ et associées à chaque séquelle

Conclusion :

Les i.st.m* qualitatifs modélisent une sorte de certitude spatiotemporelle sur la susceptibilité qu'ont les prévalences spatiales pondérées $z'^j_{Uk,c}$ à prendre des valeurs plutôt fortes ou faibles. Cependant $z'^j_{(U_k),q}$ n'est pas forcément associé aux extrema $z'^j_{(U_k),c}$. En revanche, en conférant une attention particulière à la qualité EpiGéoStat des données utilisées et en intégrant judicieusement les notions temporelles liées à la latence de la pathologie et à la qualité de l'information épidémiologique disponible sur l'historique médical des individus, il est possible d'une part de détecter les valeurs de $z'^j_{(Uk),c}$ les plus incertaines et donc les Uk sur lesquelles il convient d'éviter de conjecturer d'éventuels liens de causalité entre les expositions environnementales et l'état de santé des individus., et d'autre part, dans les autres, de donner plus de consistance aux valeurs prises par $z'^j_{(U_k),c}$.

Les modélisations géographiques des PM* d'intérêt par les i.st.m* : $z'^j_{(Uk),c}$ et $z'^j_{(Uk),q}$, sont présentées dans la sous-section subséquente. Elles peuvent facilement être étendues à toutes les séquelles développées par les patients et à n'importe quelle autre maladie, en adaptant au préalable *la métrique floue* géographique permettant d'estimer les π_i^j et celle du temps d'exposition à l'environnement tee_i^j.

PRESENTATION ET ANALYSE DES RESULTATS

Les résultats cartographiques et statistiques inhérents à la modélisation géographique sont présentés pour les PM* étudiés, c'est-à-dire les séquelles : cataractes, tumeurs thyroïdiennes et tumeurs secondaires développées par les patients spatialisés. Ensuite, une analyse géographique sommaire est effectuée afin de décrire les éventuelles disparités ou similitudes spatiales ainsi que l'émergence de tendances géographiques morbides plus générales.

Les résultats des modélisations géographiques des PM* d'intérêt sont donnés par les i.st.m* : $z'^{j}_{(Uk),c}$ et $z'^{j}_{(Uk),q}$. Ils sont déclinés successivement pour les séquelles : cataractes (CATA), tumeurs thyroïdiennes (THYR) et tumeurs secondaires (TUM2). Chaque modélisation géographique donne lieu à des cartographies et des synthèses statistiques.

Représentations cartographiques :

Elles sont données pour les communes U_k sises en région PACA et aux alentours dans le Sud-Est de la France ; Et pour les U_k situées dans les régions Alsace et Lorraine pour le Nord-Est de la France.

Cartographies des i.st.m* quantitatifs : $z'^{j}_{(Uk),c}$

Les noms de communes affichés correspondent à celles ayant une *prévalence spatiotemporelle pondérée observée extrême* (Saporta, 2006), i.e. celles pour lesquelles :

$$z'^{j}_{(U_k),c} \geq \left(\widehat{Q}_3\left(z'^{j}_{(U_k),c}\right) + 1{,}5 \times \widehat{IQR}\left(z'^{j}_{(U_k),c}\right)\right)$$

Pour les cartographies des i.st.m* qualitatifs : $z'^{j}_{(Uk),q}$:

Le nom des communes affichées dépend de la pathologie – séquelle – et du nombre de modalités prises par les i.st.m* qualitatifs, soit :

$$\text{label}_{(U_k)} = \begin{cases} \text{Display} \quad \text{lorsque:} \quad \left\{z'^{\text{CATA}}_{(U_k)q} \in \mathcal{C}_j\right\} \cap \{U_k = \mathfrak{U}_{\text{CATA}}\} \\ \text{Display} \quad \text{lorsque:} \quad \left\{\hat{z}'^{\text{THYR}}_{(U_k)q} = \text{OUI}\right\} \\ \text{Display} \quad \text{lorsque:} \quad \left\{z'^{\text{TUM2}}_{(U_k)q} \in \{\text{OUI} \cup \text{INCERTAIN}\}\right\} \end{cases}$$

Avec : $\mathfrak{U}_{\text{CATA}}$: l'ensemble des communes de 1ère espèce dont le nom est : Toulon, Florange, Val-de-Meuse, ou ayant un statut de préfecture ou de préfecture de région, ou encore une *grande* surface territoriale. En d'autres termes :

$$\mathfrak{U}_{\text{CATA}} = \bigcup_{k=1}^{q_1}\left(U_k \middle| \begin{array}{c} \left\{\text{statut}_{(U_k)} = \{\text{Préfecture} \cup \text{Préfecture. de. région}\}\right\} \cup \\ \left\{\text{label}_{(U_k)} = \{\text{Toulon; Florange; Val. De. Meuse}\}\right\} \cup \left\{S_{(U_k)} \geq \widehat{Q}_{10}\left(S_{(U_u)}\right)\right\} \end{array}\right)$$

Les représentations statistiques sont données respectivement pour les i.st.m* quantitatifs : $z'^{j}_{(Uk),c}$ puis pour les *i.st.m* qualitatifs :* $z'^{j}_{(Uk),q}$ par : la distribution spatiale empirique des $z'^{j}_{(Uk),c}$ estimée sur l'ensemble des $\mathcal{U}_{\cdot}^{1^{er}}$ – les valeurs situées à droite de la queue de la distribution spatiale sont généralement agrégées dans la dernière classe – et par un tableau des *principaux paramètres statistiques de position et de dispersion,* et par l'histogramme des fréquences empiriques associées aux modalités des $z'^{j}_{(Uk),q}$ et estimées sur $\mathcal{U}_{\cdot}^{1^{er}}$, aucun tableau n'est associé car le graphique est *correctement documenté* (Saporta, 2006).

Il convient désormais d'aborder la présentation des cartographies et des figures statistiques obtenues pour chacune des séquelles étudiées.

SEQUELLE : CATARACTES

Cartographie des $\mathbf{z'^{CATA}_{(U_k),c}}$ pour le Sud-Est de la France

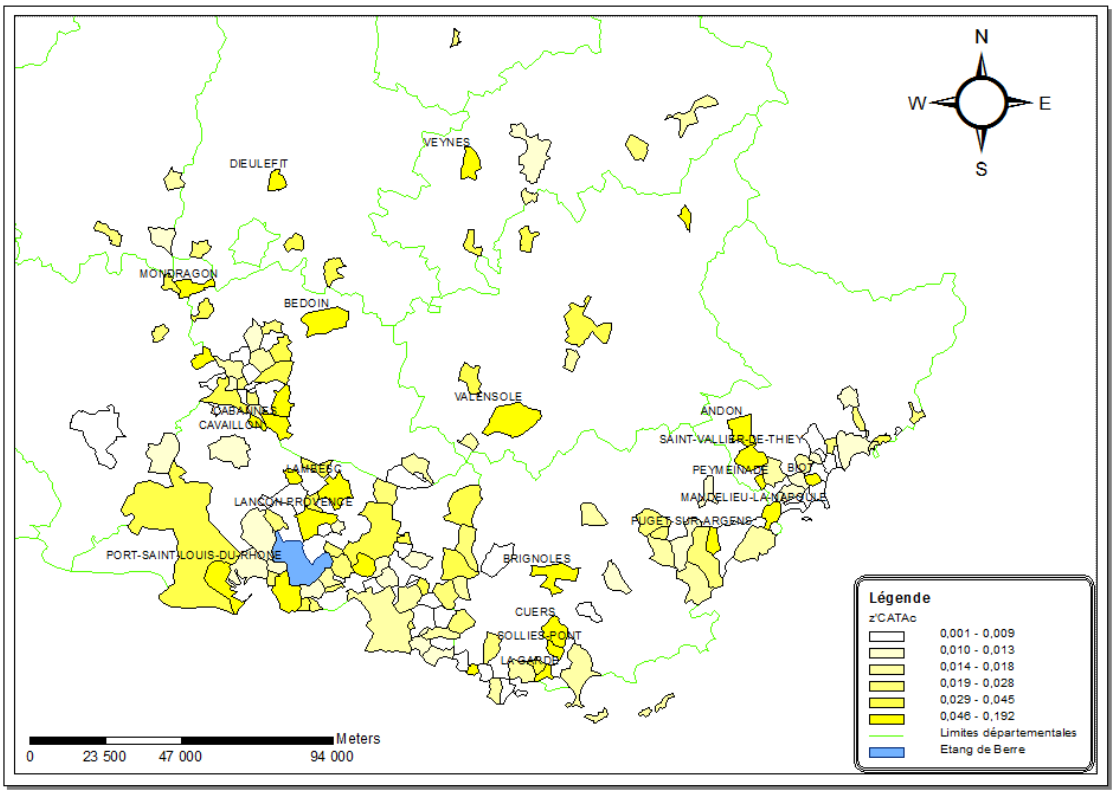


Figure 42 : Valeurs prises par $\mathbf{z'^{CATA}_{(U_k),c}}$ et affichage des $\mathbf{U_k}$ ayant une *prévalence spatiale pondérée observée extrême*

Cartographie des $z'^{CATA}_{(U_k),c}$ pour le Nord-Est de la France

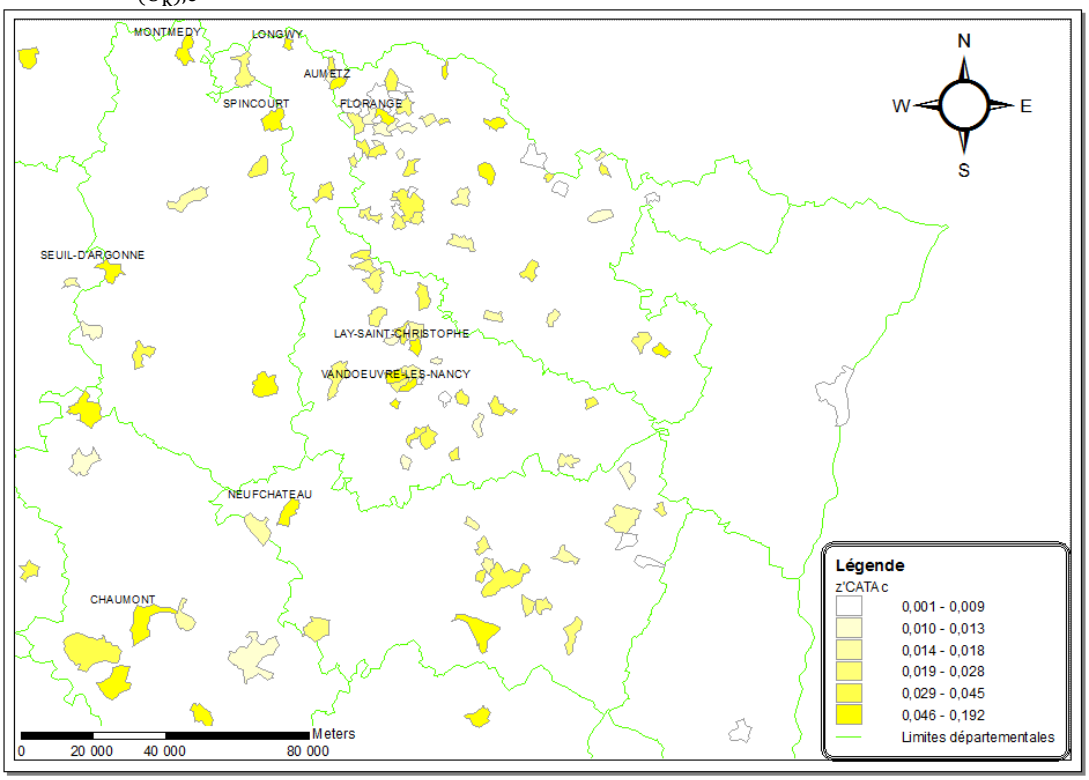


Figure 43 : Valeurs prises par $\mathbf{z'^{CATA}_{(U_k),c}}$ et affichage des $\mathbf{U_k}$ ayant une *prévalence spatiale pondérée observée extrême*

Présentation des résultats statistiques obtenus pour $z'^{CATA}_{(Uk),c}$ sur l'ensemble des $\mathcal{U}^{1er}$

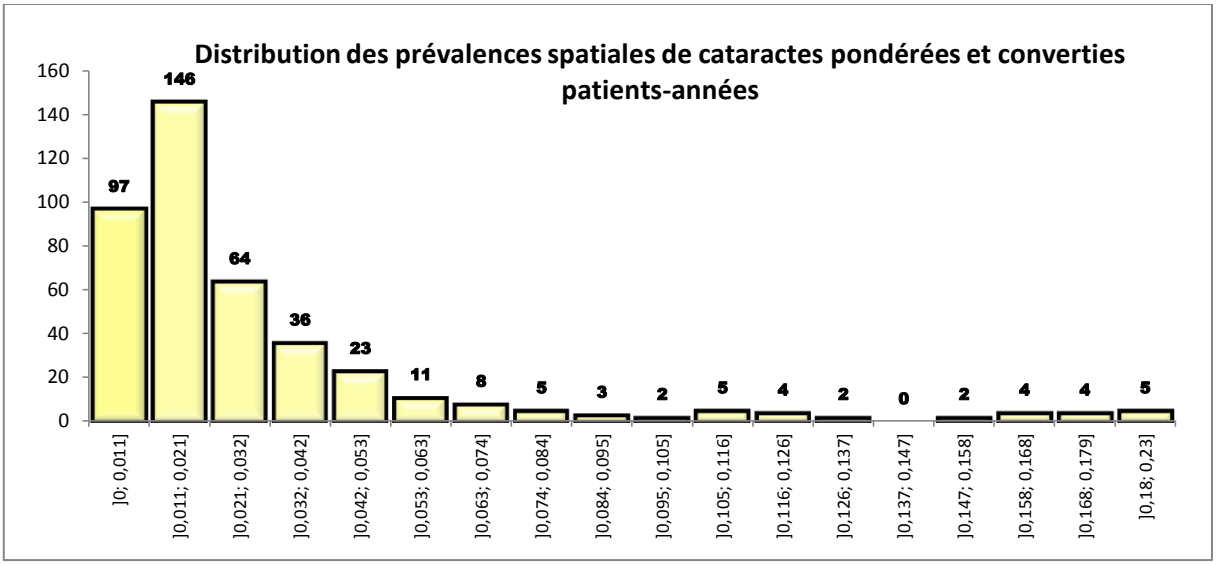


Figure 44 : Histogramme de la distribution spatiale empirique de $z'^{CATA}_{(Uk),c}$

421 - Uk	Paramètres de dispersion et de position : z'CATAc						
Estimateur :	min(.)	Q1^(.)	mêd()	Q3^(.)	max(.)	môy(.)	σ^(.)
Estimation :	**0,001**	**0,011**	**0,018**	**0,034**	**0,192**	**0,0308**	**0,0357**

Tableau 26 : Tableau statistique des principaux paramètres de position et de dispersion de $z'^{CATA}_{(Uk),c}$

Cartographie des $z'^{CATA}_{(U_k),q}$ pour le Sud-Est de la France

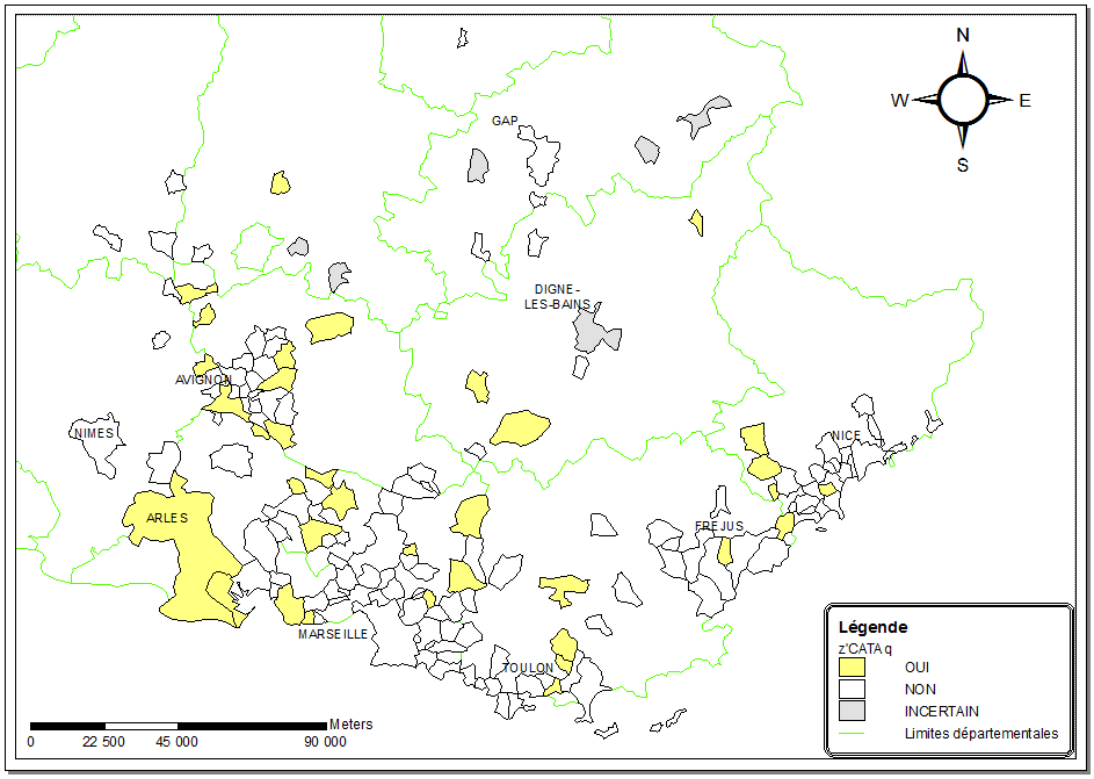


Figure 45 : Valeurs prises par $z'^{CATA}_{(U_k),q}$ et affichage de l'ensemble des communes de : Toulon, Florange, Val-de-Meuse ; Ou ayant un statut de préfecture ou préfecture de région ; Ou encore une *grande* surface territoriale

Cartographie des $\mathbf{z'^{CATA}_{(U_k),q}}$ pour le Nord-Est de la France

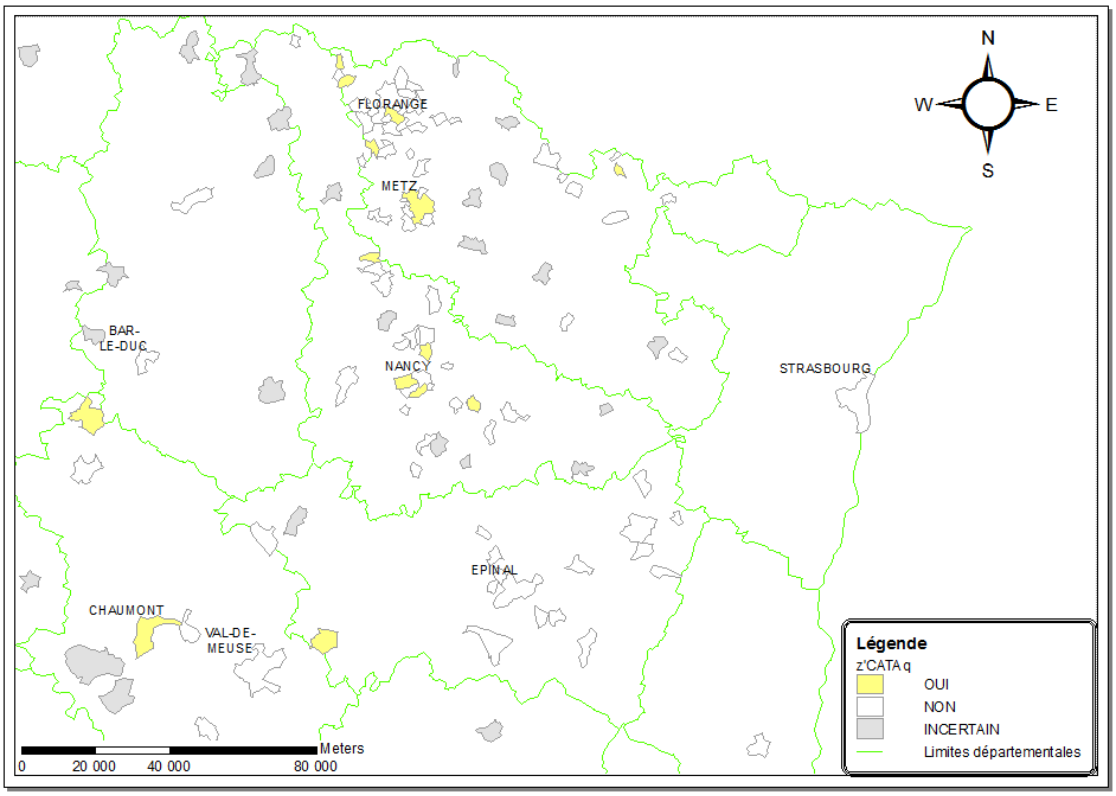


Figure 46 : valeurs prises par $\mathbf{z'^{CATA}_{(U_k),q}}$ et affichage de l'ensemble des communes de : Toulon, Florange, Val-de-Meuse ; Ou ayant un statut de préfecture ou préfecture de région ; Ou encore une *grande* surface territoriale

Présentation des résultats statistiques obtenus pour $z'^{CATA}_{(Uk),q}$ sur l'ensemble des $\mathcal{U}^{1^{er}}$.

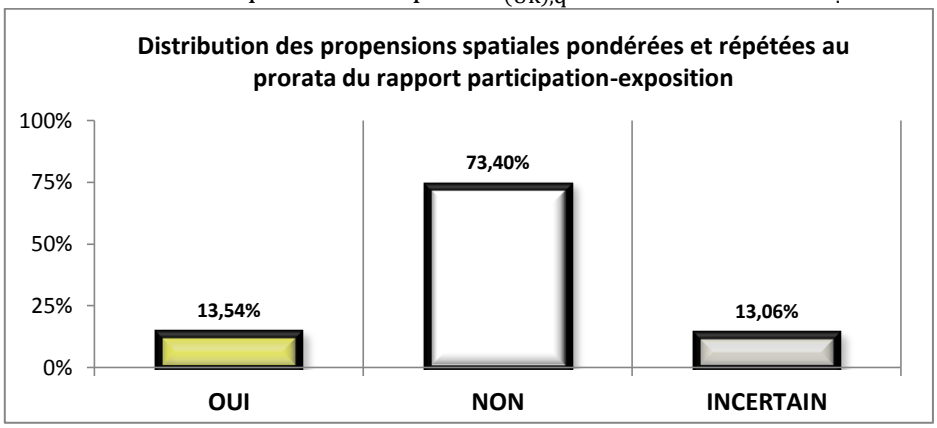


Figure 47 : Histogramme documenté des fréquences empiriques associées aux modalités des $\mathbf{z'^{CATA}_{(Uk),q}}$

SEQUELLE : TUMEURS THYROÏDIENNES

Cartographie des $\mathbf{z'^{THYR}_{(U_k),c}}$ pour le Sud-Est de la France

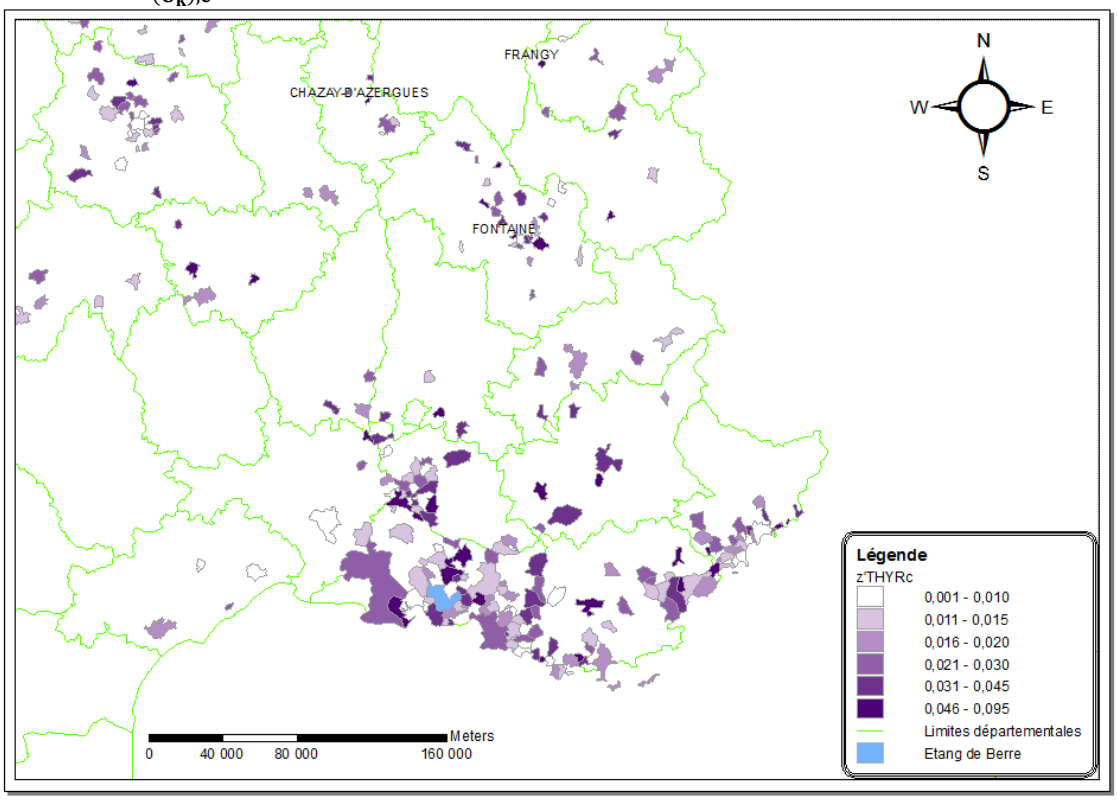


Figure 48 : Valeurs prises par $\mathbf{z'^{THYR}_{(U_k),c}}$ et affichage des $\mathbf{U_k}$ ayant une prévalence spatiale pondérée observée extrême

Cartographie des $z'^{THYR}_{(U_k),c}$ pour le Nord-Est de la France

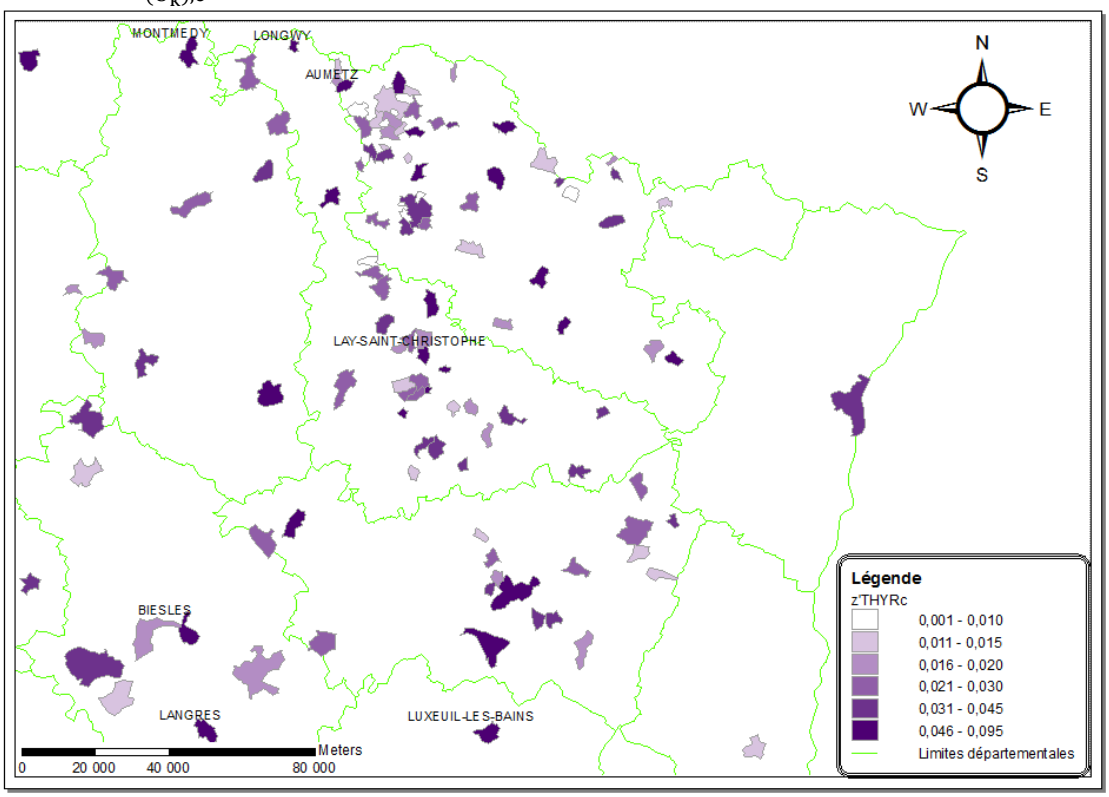


Figure 49 : Valeurs prises par $\mathbf{z'^{THYR}_{(U_k),c}}$ et affichage des $\mathbf{U_k}$ ayant une *prévalence spatiale pondérée observée élevée*

Présentation des résultats statistiques obtenus pour $z'^{THYR}_{(Uk),c}$ sur l'ensemble des $\mathcal{U}^{1er}_{.}$

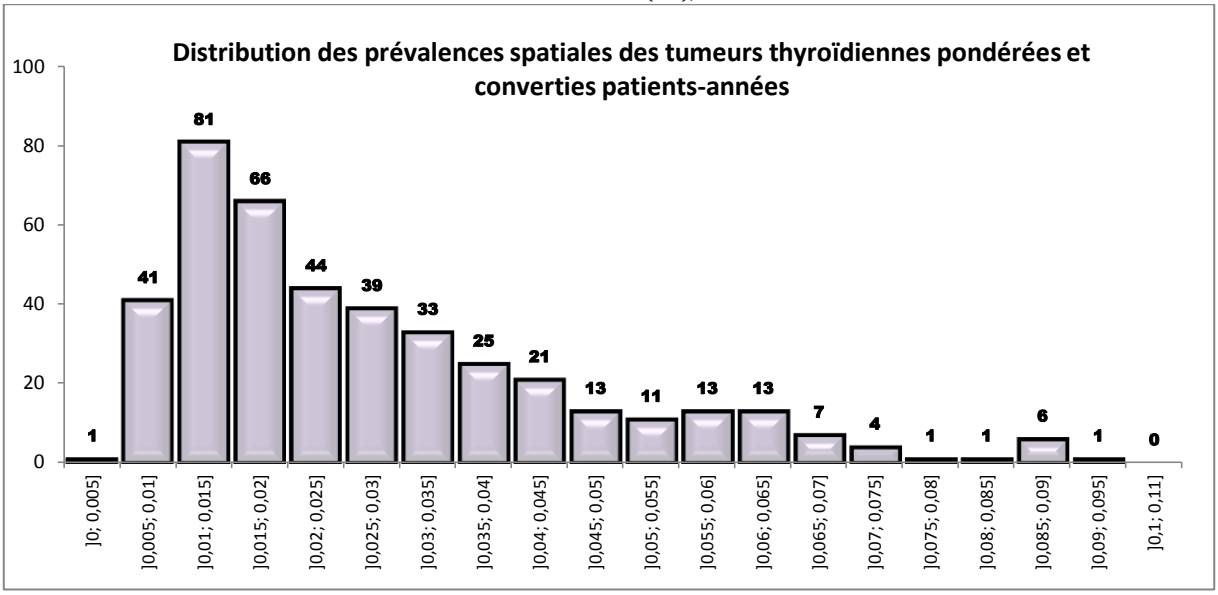


Figure 50 : Histogramme de la distribution spatiale empirique de $z'^{THYR}_{(Uk),c}$

421 - Uk	Paramètres de dispersion et de position : z'THYRc						
Estimateur :	min(.)	Q1^(.)	mêd()	Q3^(.)	max(.)	môy(.)	σ^(.)
Estimation :	**0,005**	**0,014**	**0,022**	**0,037**	**0,091**	**0,0279**	**0,0184**

Tableau 27 : Tableau statistique des principaux paramètres de position et de dispersion de $z'^{THYR}_{(Uk),c}$

Cartographie des $\mathbf{z'^{THYR}_{(U_k),q}}$ pour le Sud-Est de la France

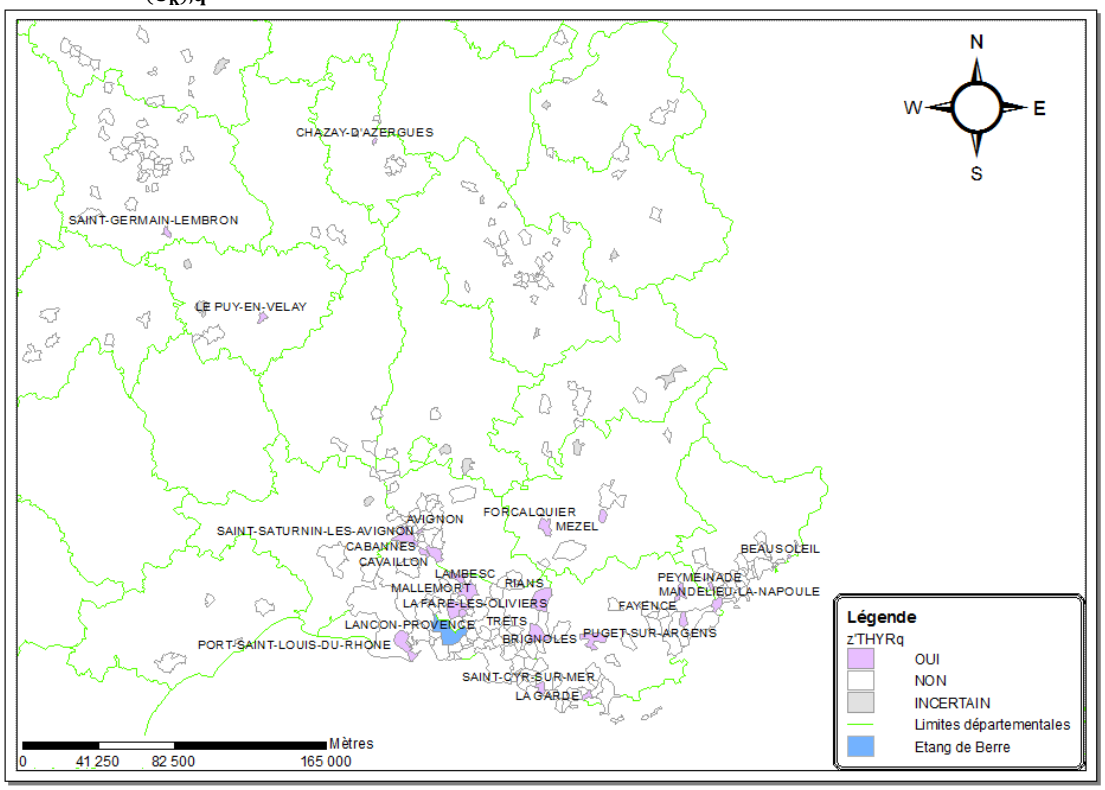


Figure 51 : Valeurs prises par $z'^{THYR}_{(U_k),q}$ et affichage de l'ensemble des communes pour lesquelles $z'^{THYR}_{(U_k),q} = OUI$

Cartographie des $\mathbf{z'^{THYR}_{(U_k),q}}$ pour le Nord-Est de la France

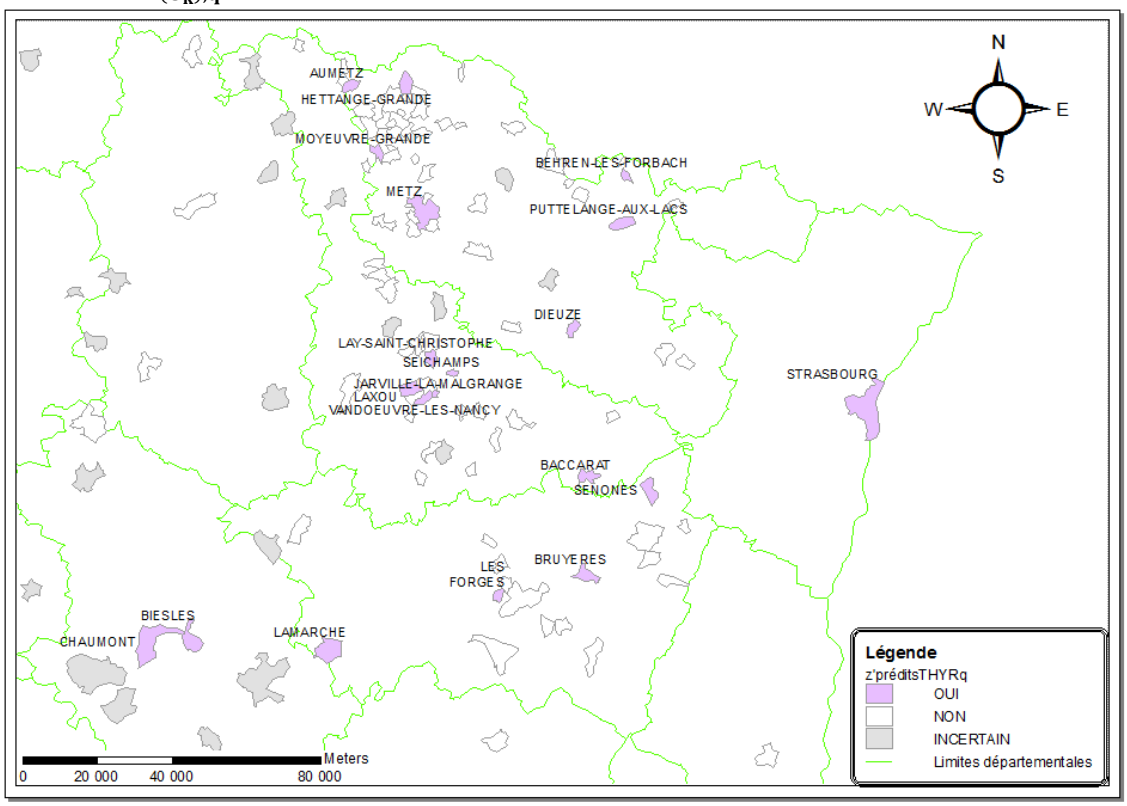


Figure 52 : Valeurs prises par $\mathbf{z'^{THYR}_{(U_k),q}}$ et affichage de l'ensemble des communes pour lesquelles $\mathbf{z'^{THYR}_{(U_k),q} = OUI}$

Présentation des résultats statistiques obtenus pour $z'^{THYR}_{(Uk),q}$ sur l'ensemble des $\mathcal{U}^{1er}$

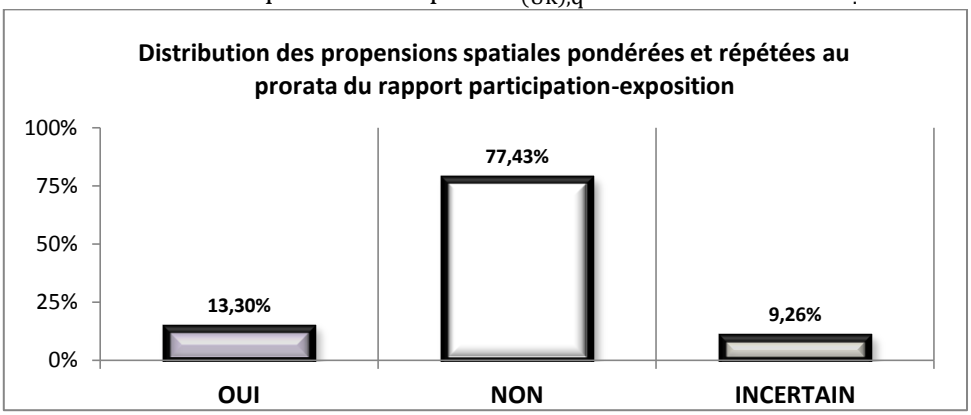


Figure 53 : Histogramme documenté des fréquences empiriques associées aux modalités des $\mathbf{z'^{THYR}_{(Uk),q}}$

SEQUELLE : TUMERURS SECONDAIRES

Cartographie des $\mathbf{z'^{TUM2}_{(U_k),c}}$ pour le Sud-Est de la France

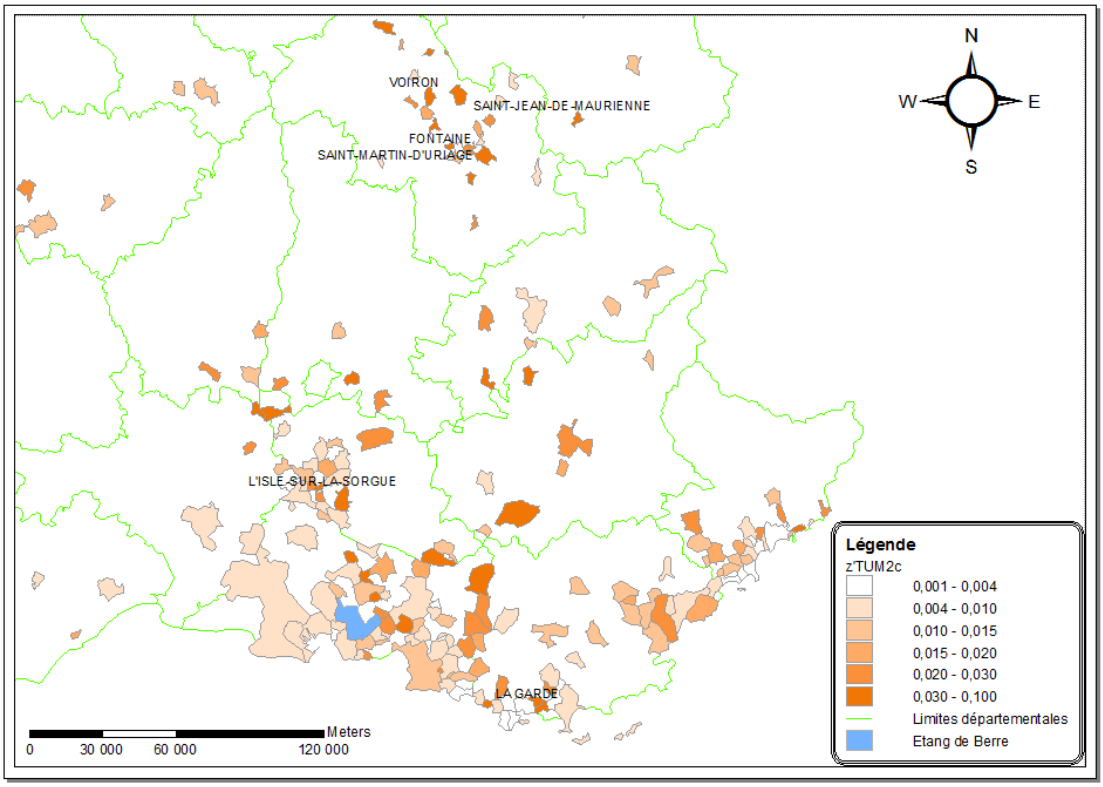


Figure 54 : Valeurs prises par $\mathbf{z'^{TUM2}_{(U_k),c}}$ et affichage des $\mathbf{U_k}$ ayant une *prévalence spatiale pondérée observée extrême*

Cartographie des $z'^{TUM2}_{(U_k),c}$ pour le Nord-Est de la France

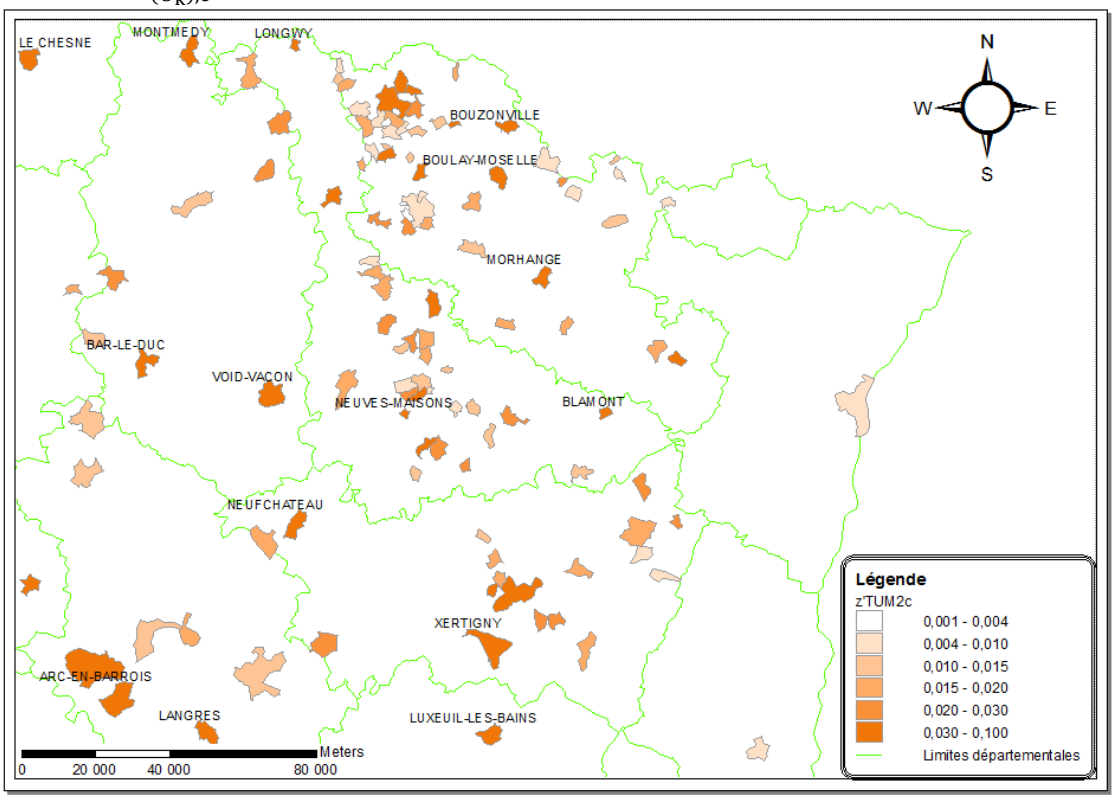


Figure 55 : Valeurs prises par $\mathbf{z'^{TUM2}_{(U_k),c}}$ et affichage des $\mathbf{U_k}$ ayant une *prévalence spatiale pondérée observée élevée*

Présentation des résultats statistiques obtenus pour $z'^{TUM2}_{(Uk),c}$ sur l'ensemble des $\mathcal{U}^{1er}_{.}$

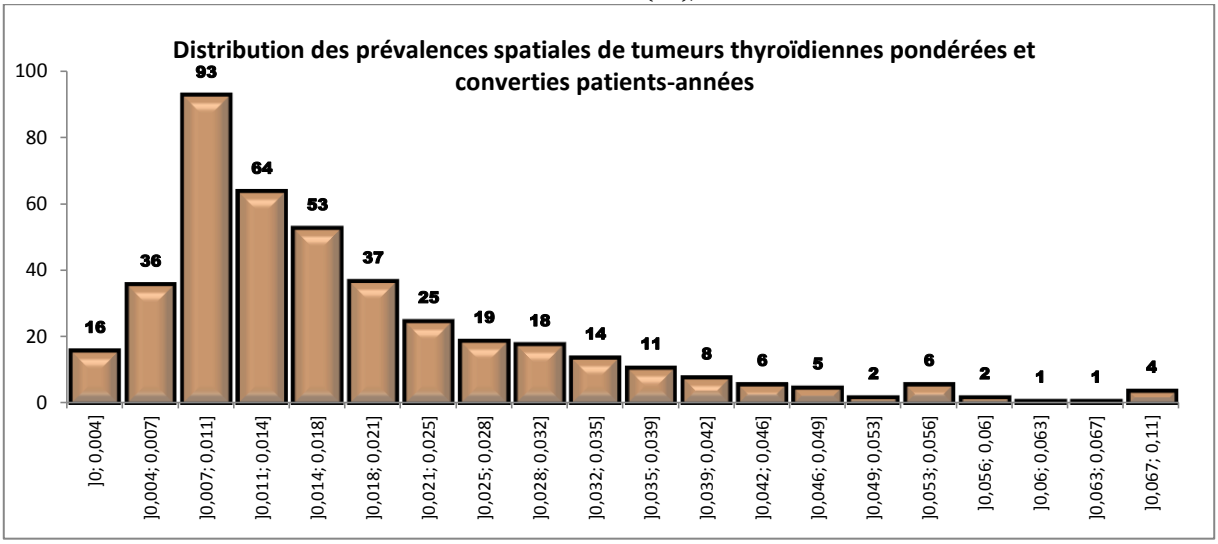


Figure 56 : Histogramme de la distribution spatiale empirique de $z'^{TUM2}_{(Uk),c}$

421 - Uk	Paramètres de dispersion et de position : z'TUM2c						
Estimateur :	min(.)	Q1^(.)	mêd()	Q3^(.)	max(.)	môy(.)	σ^(.)
Estimation :	**0,001**	**0,009**	**0,014**	**0,023**	**0,095**	**0,0183**	**0,0137**

Tableau 28 : Tableau statistique des principaux paramètres de position et de dispersion de $z'^{THYR}_{(Uk),c}$

Cartographie des $\mathbf{z'^{TUM2}_{(U_k),q}}$ pour le Sud-Est de la France

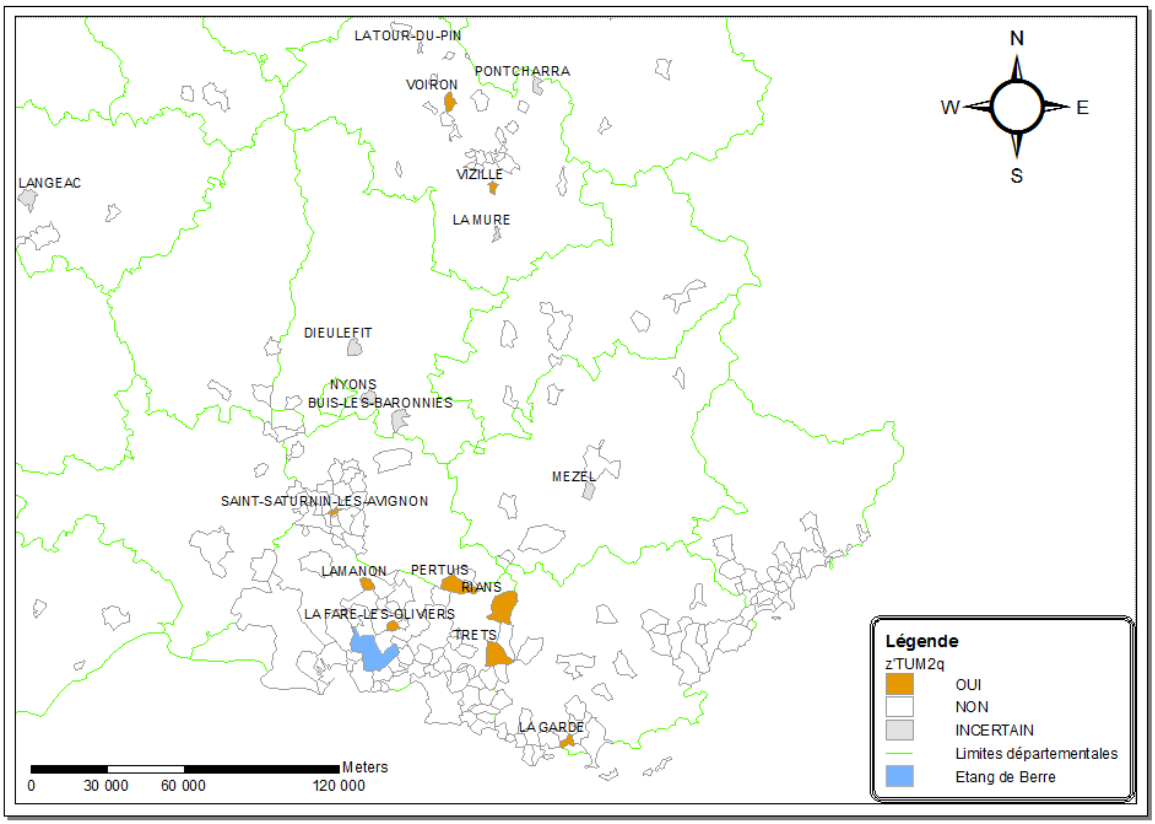


Figure 57 : Valeurs prises par $z'^{TUM2}_{(U_k),q}$ et affichage de l'ensemble des communes pour lesquelles $z'^{TUM2}_{(U_k),q} = \{$OUI ∪ INCERTAIN$\}$

Cartographie des $\mathbf{z'^{TUM2}_{(U_k),q}}$ pour le Nord-Est de la France

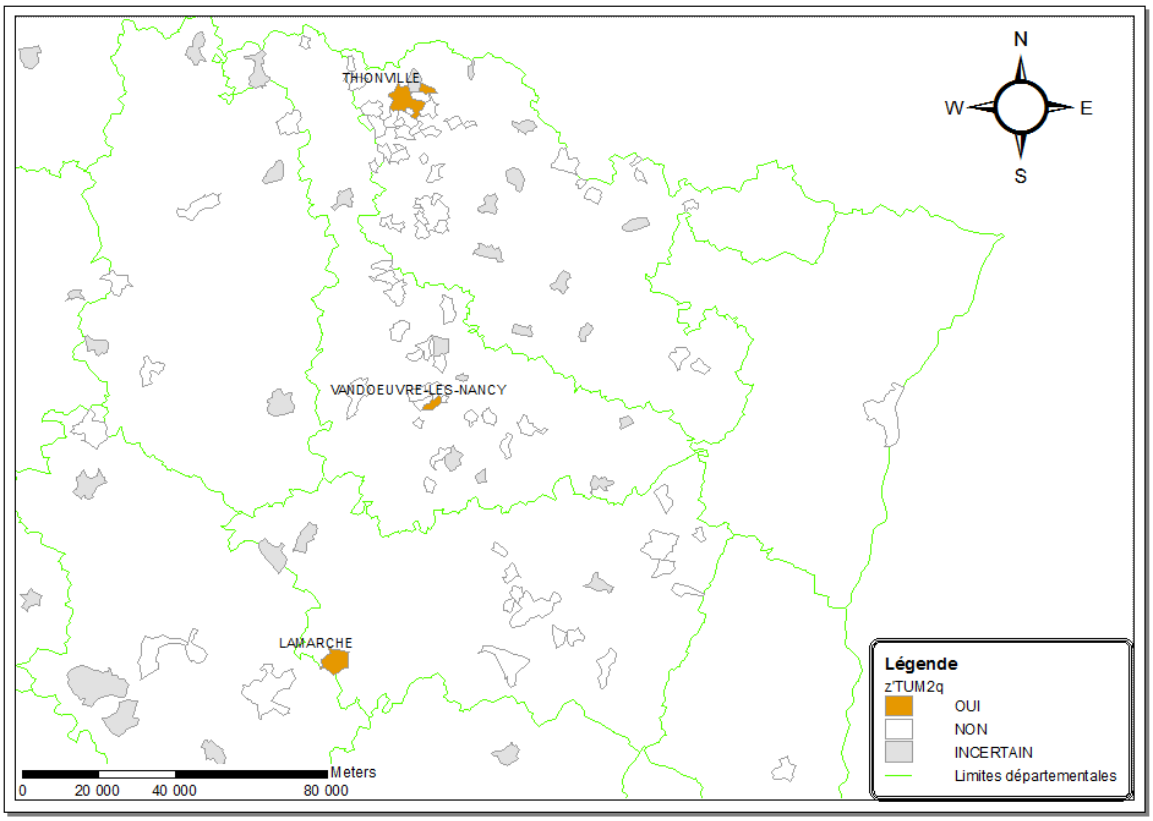


Figure 58 : Valeurs prises par $\mathbf{z'^{TUM2}_{(U_k),q}}$ et affichage de l'ensemble des communes pour lesquelles $\mathbf{z'^{TUM2}_{(U_k),q} = \{OUI \cup INCERTAIN\}}$

Présentation des résultats statistiques obtenus pour $z'^{TUM2}_{(Uk),q}$ sur l'ensemble des $\mathcal{U}^{1^{er}}$

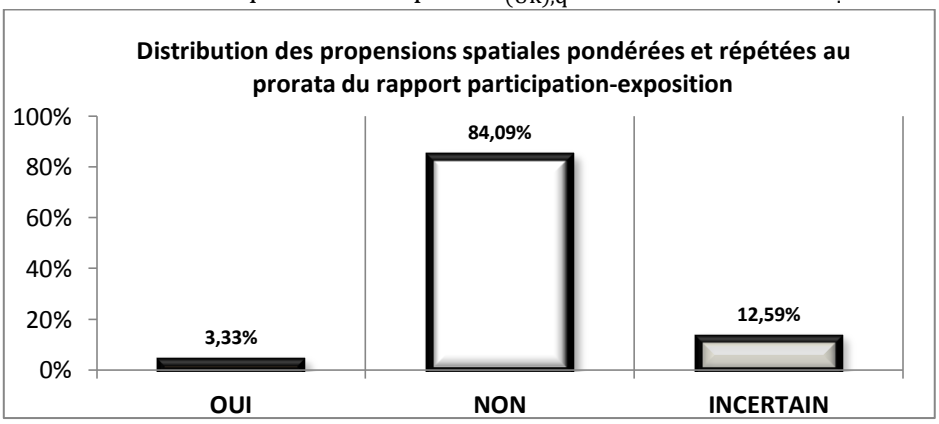


Figure 59 : Histogramme documenté des fréquences empiriques associées aux modalités des $\mathbf{z'^{TUM2}_{(Uk),q}}$

RESUME, ANALYSE ET REMARQUES

Il s'agit maintenant de procéder à une analyse spatiale conjointe des valeurs *extrêmes* prises par $z'^{j}_{(Uk),c}$ et des modalités de $z'^{j}_{(Uk),q}$. Dans un premier temps il est question d'identifier des similitudes ou ressemblances géographiques en étudiant indépendamment chaque séquelle.

Puis, dans un second temps, de faire émerger d'éventuelles tendances ou prédispositions spatiales morbides plus générales. L'analyse conjointe des i.st.m* se focalise plus particulièrement sur les communes situées dans le Sud-Est de la France.

L'analyse des résultats de la modélisation géographique de la séquelle : CATA a permis d'estimer une prévalence spatiotemporelle pondérée observée moyenne de $\bar{z}'^{CATA}_{(Uk),c} = 0{,}0308$. Dans le Sud-Est de la France les communes qui ont à la fois des valeurs *élevées* de $z'^{CATA}_{(Uk),c}$ et de $z'^{CATA}_{(Uk),q} = \text{OUI}$ sont nombreuses. En région PACA, il y a : Port-Saint-Louis-Du-Rhône, Lambesc, Lançon en Provence, Cavaillon, Cabanes, Bédouin, Mondragon, Andon, Biot, Saint-Vallier-De-Thiey. Dans ces communes, il semblerait que les patients spatialisés aient tendance à développer des cataractes.

La commune de Fontaine est située en région Rhône-Alpes. Elle n'est pas affichée sur les cartes mais comme elle est impliquée dans l'analyse des deux autres séquelles, il convient de remarquer qu'il s'agit d'une valeur de $z'^{CATA}_{(Uk),c}$. Extrême. Cependant, à l'instar de la commune de Veynes dans l'arrière-pays de la région PACA, la qualité des modélisations géographiques de Fontaine ne permet pas d'attacher trop d'importance à cette prévalence élevée dans la mesure où les propensions spatiotemporelles pondérées à développer des cataractes ne sont pas fiables, i.e. que $z'^{CATA}_{(Uk),q} = \text{INCERTAIN}$.

L'analyse des résultats de la modélisation géographique de la séquelle : THYR a permis d'estimer une prévalence spatiotemporelle pondérée observée moyenne de $\bar{z}'^{THYR}_{(Uk),c} = 0{,}0279$. Dans le Sud-Est de la France les communes qui ont des $z'^{THYR}_{(Uk),c}$ *extrêmes* sont rares. Il n'y en a aucune en PACA et seulement trois en région Rhône-Alpes, celles de : Frangy, Fontaine, Chazay-d'Azergues.
Or, seule la commune de Chazay-d'Azergues a une $z'^{CATA}_{(Uk),q} = \text{OUI}$; Ce qui signifie que sa qualité spatiotemporelle de modélisation permet de conjecturer que les patients spatialisés ont probablement tendance à développer, dans cette Uk, des tumeurs thyroïdiennes.

Quant à la commune rurale de Frangy, $z'^{THYR}_{(Uk),q} = \text{INCERTAIN}$ et par conséquent, la qualité spatiotemporelle des modélisations géographiques rend l'interprétation de la valeur *extrême :* $z'^{THYR}_{(Uk),c}$ peu fiable. Enfin, pour la commune de Fontaine, $z'^{THYR}_{(Uk),q} = \text{NON}$, il s'agit d'un antagonisme géographique qui rend difficile l'interprétation conjointe des deux i.st.m.

L'analyse des résultats statistiques et cartographiques de la modélisation de la séquelle : TUM2 a permis d'estimer une prévalence spatiotemporelle pondérée observée moyenne de $\bar{z}'^{TUM2}_{(Uk),c} = 0{,}0183$. Il s'agit de la plus faible de toutes les séquelles étudiées. Dans le Sud-Est de la France les communes qui ont des valeurs de $z'^{TUM2}_{(Uk),c}$ *extrêmes* sont rares, il n'y en a que deux, celles de La Garde et de L'Isle-Sur-Sorgue. On peut aussi noter la présence de Fontaine en région Rhône-Alpes que l'on retrouve aussi pour les $z'^{THYR}_{(Uk),c}$ *extrêmes.*

Mais pour la seule commune de La Garde, $z'^{TUM2}_{(Uk),q} = \text{OUI}$, donc une qualité spatiotemporelle suffisante pour conjecturer que les patients spatialisés ont tendance à développer des tumeurs secondaires. En revanche, pour les communes de L'Isle-Sur-Sorgue et de Fontaine la valeur de la propension

spatiotemporelle pondérée vaut : $z'^{TUM2}_{(Uk),q} = NON$, ce qui engendre un antagonisme géographique et qui rend le croisement des i.st.m, *a priori* impossible, ou du moins difficilement interprétable.

L'analyse conjointe des $z'^{j}_{(U_k),c}$ *extrêmes* et des $z'^{j}_{(U_k),q}$ met en évidence la grande attention qu'il convient de porter à l'interprétation des $z'^{j}_{(U_k),c}$. En effet, elles doivent nécessairement être croisées avec les $z'^{j}_{(U_k),q}$ - qui permettent d'identifier les communes où la qualité spatiotemporelle EpiGéoStat rend les $z'^{j}_{(U_k),c}$ significatives. Mais l'analyse conjointe des $z'^{j}_{(U_k),q}$ engendre aussi des *antagonismes géographiques* qui rendent l'interprétation des $z'^{j}_{(U_k),c}$ *incertaine*, difficile, voire impossible.

L'analyse des valeurs de $z'^{j}_{(U_k),c}$ *extrêmes* et des modalités de $z'^{j}_{(U_k),q}$, fait ressortir des *incertitudes* et des *antagonismes*, i.e. des *conflits spatiotemporels* d'interprétation liés à la qualité des hypothèses, des méthodes et des données utilisées. Ces ambivalences sont encore plus courantes dans les communes du Nord-Est de la France où la spatialisation des patients est beaucoup plus incertaine, et où la proportion d'individus spatialisés est deux fois inférieure à celle localisée dans le Sud-Est. Raisons pour lesquelles l'analyse descriptive s'est limitée aux communes sises en région PACA et aux alentours.

Il est difficile de croiser de la sorte les valeurs $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$ et il convient de se demander si la stratégie utilisée est adaptée.

Toutefois, cette analyse spatiale croisée sommaire - menée séquelle par séquelle - met en évidence quelques similitudes et des disparités géographiques, donc une tendance à développer des séquelles dans certaines communes et pas dans d'autres.

Plus globalement, on remarque que la prévalence spatiotemporelle pondérée observée moyenne estimée sur les CATA est proche de celle des THYR qui est beaucoup plus faible que pour les TUM2.
Ce qui correspond à ce qui était attendu, et par conséquent conjecture que la *métrique floue* et les *conversions en Patients-Années* ne bruitent pas les modélisations géographiques. Au contraire, elles leur confèrent une plus grande robustesse.

Il semble *a priori* très difficile de suspecter des tendances accrues à développer des PM, i.e. systématiquement des séquelles. Cela renforce l'hypothèse qu'en dépit de l'influence sur l'état de santé des traitements reçus pour la leucémie, il y a autre chose. Et que cette autre chose est sans doute l'*effet d'exposition à des Facteurs Environnementaux * combinés.*

Il n'y a pas de commune où systématiquement $z'^{j}_{(U_k),c}$ prend une valeur extrême et où $z'^{j}_{(U_k),q} = OUI$ lorsque l'on considère toutes les séquelles. Cependant, cette combinaison particulière – caractérisant des prédispositions géographiques morbides - s'observe bien simultanément mais sur des communes proches, qui se trouvent dans des zones géographiques relativement réduites, quand elles ne sont pas situées en contiguïté spatiale directe. Par conséquent un *effet environnement* suggéré précédemment est presque évident.

L'analyse des $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$ telle qu'elle a été menée met en évidence *un effet environnement* mais reste très limitée dans l'interprétation des modélisations géographiques morbides. Néanmoins, les $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$ apportent indubitablement des informations différentes qui sont *a priori* robustes, intéressantes et adaptées à la dialectique géographique.

Il convient donc de proposer une façon plus *judicieuse* de croiser ces deux i.st.m* en vue de procéder à des analyses spatiales plus fines.

Et pour cause, le positionnement scientifique du Géographe de la Santé exhorte à ne pas se limiter seulement à des modélisations spatiotemporelles mais bien de caractériser les espaces géographiques en fonction du type de risque morbide auquel sont assujetties les populations *in situ.* Et ce dans le but bien sûr d'assurer, grâce à des mesures politiques *le contrôle territorial* d'une bonne santé environnementale* (Salem, 1995).

Perspective nécessaire dans un contexte communautaire qui incite à élaborer *des outils géographiques permettant d'évaluer les risques environnementaux auxquels sont assujetties les populations locales* (Commission des Communautés Européennes, 2003).

SECTION D) CARACTERISATION DES RISQUES D'EXPOSITIONS GEOGRAPHIQUES

L'idée :
Combiner *judicieusement* l'information contenue dans les i.st.m* $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$ afin de les agréger en un seul i.st.m, noté $z^{REG.j}_{(U_k)}$, susceptible de caractériser dans chaque U_k, le Risque d'Exposition Géographique (REG) à un Phénomène Morbide* (PM) particulier – e.g. une maladie, ou une séquelle nommée j.

Objectif :
Il s'agit d'une proposition méthodologique vouée à modéliser le REG à des phénomènes morbides par le biais d'une stratégie de fusion *par seuillage géographique*.

Considération théorique liminaire :
« *Seule l'éventuelle existence d'un effet de seuil [...] peut être interprétable dans la réalité géographique.* [Cependant] *il y a un réel problème d'identification de la valeur mathématique d'un paramètre avec sa signification géographique* (L. Sanders , 1992-2).

Proposition :
Déterminer un seuil *d'exposition géographique élastique*, noté φ^*_j, à partir duquel la prévalence spatiale pondérée EpiGéoStat convertie en patients-années est *anormalement* élevée, et combiner ces valeurs avec l'information qualitative spatiotemporelle des propensions spatiales pondérées.
L'i.st.m* $z^{REG.j}_{(U_k)}$ résultant est de nature qualitative et doit permettre de discrétiser les U_k selon un type de Risque d'Exposition Géographique (REG) morbides.
En guise d'illustration la proposition est appliquée aux trois séquelles d'intérêt : CATA, THYR, TUM2.

Proposition d'une typologie caractérisant le REG :
Associer à chaque U_k un épithète au risque que les patients encourent du simple fait d'y résider, au vu des connaissances épidémiologiques observées disponibles. Le Risque d'Exposition Géographique à la pathologie.j (REG.j) peut être :

1. **PROBABLE :** les communes dont la prévalence spatiale pondérée EpiGéoStat convertie en patients-années est supérieure au seuil *d'exposition géographique élastique* : $\left\{z'^{j}_{(U_k),c} \geq \varphi^*_{(\cdot)}\right\}$ et où la propension spatiale pondérée par les incertitudes spatiotemporelles est à la fois estimable et forte, i.e. : $\left\{z'^{j}_{(U_k),q} = \text{"OUI"}\right\}$.

2. **POSSIBLE :** les communes dont la prévalence spatiale pondérée est supérieure au seuil d'exposition géographique élastique, i.e. :$\left\{z'^{j}_{(U_k),c} \geq \varphi^*_{(\cdot)}\right\}$ et où la qualité de l'information est suffisamment bonne pour que la propension spatiale pondérée répétée au prorata du rpe^j_i permette d'estimer l'i.st. qualitatif : $\left\{z'^{j}_{(U_k),q} = \text{NON}\right\}$.

3. **INDEMONTRABLE :** toutes les communes, quelle que soit la valeur de la prévalence spatiale pondérée convertie en *patients-années* : $\left\{z'^{j}_{(U_k),c} \in \mathbb{R}\right\}$, dès lors que la qualité spatiotemporelle associée à la méthode de spatialisation et aux informations utilisées, n'est pas fiable, i.e. $\left\{z'^{j}_{(U_k),q} = \text{"INCERTAIN"}\right\}$.

4. **FAIBLE :** les communes où la prévalence spatiale pondérée convertie en années d'exposition à l'environnement est strictement inférieure au seuil d'exposition $\left\{z'^{j}_{(U_k),c} < \varphi^{*}_{(\cdot)}\right\}$ et dont les incertitudes spatiotemporelles à connotation géographique, statistique et épidémiologique permettent de supputer une propension spatiale pondérée, telle que $\left\{z'^{j}_{(U_k),q} = \text{"NON"}\right\}$.

Principe d'estimation :
L'i.st.m* d'Exposition Géographique s'évalue donc selon la règle suivante :

$$z^{REG.j}_{(U_k)} = \begin{cases} \text{PROBABLE} \quad \text{Lorsque:} \quad \left\{\left\{z'^{j}_{(U_k),c} \geq \hat{\varphi}^{*}_{j}\right\} \cap \left\{z'^{j}_{(U_k),q} = \text{"OUI"}\right\}\right\} \\ \text{POSSIBLE} \quad \text{Lorsque:} \quad \left\{\left\{z'^{j}_{(U_k),c} \geq \hat{\varphi}^{*}_{j}\right\} \cap \left\{z'^{j}_{(U_k),q} = \text{"NON"}\right\}\right\} \\ \text{INDEMONTRABLE} \quad \text{Lorsque:} \quad \left\{\left\{z'^{j}_{(U_k),c} \in \mathbb{R}_{+}\right\} \cap \left\{z'^{j}_{(U_k),q} = \text{"INCERTAIN"}\right\}\right\} \\ \text{FAIBLE} \quad \text{Lorsque:} \quad \left\{\left\{z'^{j}_{(U_k),c} < \hat{\varphi}^{*}_{j}\right\} \cap \left\{z'^{j}_{(U_k),q} = \text{"NON"}\right\}\right\} \end{cases}$$

Avec $\varphi^{*}_{(\cdot)}$ ledit seuil *d'exposition géographique élastique* qui permet de fusionner *judicieusement* l'information contenue dans les deux i.st.m* proposés et qu'il convient à présent de définir.

SEUIL D'EXPOSITION GEOGRAPHIQUE ELASTIQUE

Remarque liminaire :
Si le phénomène morbide est complètement aléatoire, i.e. indépendant de toute composante géographique, alors les prévalences spatiotemporelles pondérées EpiGéoStat converties en patients-années estimées dans chaque U_k doivent, en vertu *la loi des grands nombres*, tendre vers l'incidence spatiotemporelle pondérée estimée sur la toute population spatialisée (Bernoulli, 1713) - or ce n'est pas le cas

Hypothèse :
Les disparités géographiques observées sur les $z'^{j}_{(U_k),c}$ permettent de conjecturer que l'environnement a *probablement* un effet sur l'incidence des PM* étudiés. Par conséquent la modélisation du REG sera fondée sur une stratégie statistique d'identification par seuillage des $z'^{j}_{(U_k),c}$ *anormalement* élevés à partir de l'estimation de la moyenne globale pondérée et en tenant compte aussi des distorsions géographiques inhérentes aux *pondérations EpiGéoStat* et à la transformation en *patients-années*.

Proposition :
Le seuil d'exposition géographique φ^{*}_{j} représente une borne supérieure de la moyenne géographique pondérée qui ne peut pas être dépassée du simple fait du hasard. Il est estimé modulo un niveau de risque admis α^{j} - en s'affranchissant de toute hypothèse sur la loi de probabilité sous-jacente - auquel on grève un *coefficient d'élasticité* géographique $\xi^{j}_{géo}$ estimé par le biais des connaissances expertes sur la nature et la géographie du PM* ainsi que sur les distorsions induites par les π^{j}_{i} de la métrique floue et celles engendrées par l'intégration des temps d'exposition à l'environnement tee^{j}_{i}, sur la modélisation des variabilités géographiques PM* par $z'^{j}_{(U_k),c}$.

Remarques :
La métrique floue et la conversion en *patients-années* augmentent virtuellement la taille de la cohorte en faisant tendre, pour chaque patient, l'absence de séquelle vers le OUI proportionnellement aux incertitudes

associées, et vice-versa pour le NON. Cette stratégie surestime la réalité morbide mais représente de façon plus robuste et plus juste les variabilités géographiques.

Conséquence : l'effet conjoint des π_i^j et des tee_i^j a des répercussions sur les $z'^j_{(U_k),c}$, donc sur les caractéristiques statistiques de la distribution spatiale des $z'^j_{(U_k),c}$, et il convient de les exploiter

Remarque experte :
Si le seuil est fixé au niveau de la borne supérieure des prévalences spatiales pondérées le nombre de $\left\{z_{(U_k)}^{REG.j} = \text{PROBABLE}\right\}$ sera environ égal à celui des $\left\{z'^j_{(U_k),q} = \text{OUI}\right\}$; Quant au nombre de $\left\{z_{(U_k)}^{REG.j} = \text{POSSIBLE}\right\}$ il sera aussi grand qu'aberrant, d'abord parce que les $z'^j_{(U_k),q}$ renseignent autant sur la propension à développer la séquelle que sur la qualité des données utilisées, et ensuite car les stratégies mathématiques utilisées pour calculer les $z'^j_{(U_k),c}$ et les $z'^j_{(U_k),q}$ ne garantissent pas que les communes où $\left\{z'^j_{(U_k),q} = \text{OUI}\right\}$ sont associées systématiquement aux valeurs de $z'^j_{(U_k),c}$ les plus fortes.
Or, le nombre de modalités tel que $\left\{z_{(U_k)}^{REG.j} = \text{PROBABLE} \cup \text{POSSIBLE}\right\}$ ne doit pas être surestimé, surtout que l'un comme l'autre de ces deux épithètes est préjudiciable pour les communes concernées. L'objectif des φ_j^* est de juguler le nombre de $\left\{z_{(U_k)}^{REG.j} = \text{POSSIBLE}\right\}$ et de rationaliser le nombre $\left\{z_{(U_k)}^{REG.j} = \text{PROBABLE}\right\}$ de sorte qu'ils soient représentatifs de la réalité géographique modélisée.

Proposition experte :
C'est là qu'entre en scène le coefficient d'élasticité géographique $\xi_{géo}^j$ dont l'objet est de d'affiner l'estimation de φ_j^*. L'idée est de corriger la borne supérieure de la moyenne boostrap géographique pondérée de façon rationnelle, i.e. sans amplifier le rôle correcteur de la métrique floue et de la transformation en *patients-années.*

Remarque :
Une stratégie de seuillage classique directement basée sur les écarts-types n'est pas pertinente car elle engendre une amplification des distorsions apportées par les π_i^j et et les tee_i^j.

Stratégie d'estimation :
Le seuil d'élasticité géographique associé au PM* - séquelle « j » - proposé s'estime de la façon suivante :

$$\varphi_j^* = \left\lceil \xi_{géo}^j \cdot \left(\bar{z}'^j_{(U_k),c} + b_{1-\alpha}^{*,j} \cdot \frac{\hat{\sigma}\left(z'^j_{(U_k),c}\right)}{\sqrt{N(U_k)}} \right) \right\rceil$$

Avec : $\lceil \cdot \rceil = \text{"la fonction arrondie à l'entier supérieur"}$; $b_{1-\alpha}^{*,j}$ est un paramètre statistique bootstrap qui dépend du niveau de risque admis α^j ; Et $\xi_{géo}^j$ est le coefficient d'élasticité géographique. Ces deux paramètres sont évalués au vu des caractéristiques statistiques de $z'^j_{(U_k),c}$ ou à partir de connaissances expertes.

ESTIMATION DES PARAMETRES DU SEUIL

L'estimation du seuil d'élasticité géographique φ_j^* fait intervenir deux paramètres.
Le premier est $b_{1-\alpha}^{*,j}$ qui s'estime – par bootstrap - conditionnellement à *un niveau de risque* α^j. Il permet d'obtenir une borne supérieure de la moyenne géographique des $z'^j_{(U_k),c}$;
Le second est le *coefficient d'élasticité géographique* $\xi_{géo}^j$ qui permet de pondérer judicieusement cette borne supérieure de la moyenne en tenant compte des connaissances expertes disponibles sur la nature du

phénomène morbide et les distorsions géographiques induites par *la métrique floue géographique* et la transformation en *patients-années*.

ESTIMATION DE $B_{1-A}^{*,J}$

Remarque liminaire : l'estimation de la variable $b_{1-\alpha}^{*,j}$ fait intervenir le paramètre α^j. Ensuite le processus d'estimation de la statistique boostrap $b_{1-\alpha}^{*,j}$ est normalisé. Par conséquent l'intérêt porte sur α^j qui correspond au niveau de risque admis dans l'estimation d'une borne supérieure de la moyenne spatiale pondérée qui ne devrait pas être dépassée du simple fait du hasard.

Hypothèse : Il n'y a aucune raison pour que la valeur de α^j soit constante, elle doit varier en fonction de la nature du PM* d'intérêt - séquelle.

Proposition théorique : en théorie, plus α^j est grand et plus $b_{1-\alpha}^{*,j}$ sera statistiquement petit. Aussi les valeurs prises par α^j doivent faire allégeance aux prénotions statistiques empiriquement établies, inhérentes au choix d'un niveau de risque (Saporta, 2006).
Conséquence : les valeurs prises par α^j sont comprises entre :

$$\alpha^j \in [\![5\% \; ; \; 10\%]\!]$$

Proposition pragmatique : la valeur de α^j peut être choisie en fonction des caractéristiques statistiques de la variable $z'^j_{(U_k),c}$. qui sont influencées par l'action combinée des facteurs d'incertitude spatiotemporelle π_i^j et de l'injection des temps d'exposition des patients à l'environnement tee_i^j. En particulier les valeurs prises par l'estimateur de la moyenne et l'écart-type.

Séquelle	CATA	THYR	TUM2
$\bar{z}'^j_{(U_k),c}$	0,03082	0,02790	0,01826
$\hat{\sigma}\left(z'^j_{(U_k),c}\right)$	0,03567	0,01838	0,01372

Tableau 29 : Rappel de la moyenne et de l'écart-type estimés sur $z'^j_{(U_k),c}$ pour chaque séquelle

Conséquence : afin de ne pas amplifier les distorsions géographiques engendrées par le rôle conjoint des π_i^j et des tee_i^j la valeur spécifiée *a priori* pour α^j peut être choisie inversement proportionnelle$\hat{\sigma}\left(z'^j_{(U_k),c}\right)$.

Application aux données LEA : Les valeurs utilisées pour la modélisation du REG aux séquelles appliquées aux données de la Cohorte LEA sont :

Séquelle :	CATA	THYR	TUM2
α^j	10%	5%	5%

Tableau 30 : Valeurs de α^j spécifiées pour chaque séquelle

Remarque :
Le paramètre α^j permet de moduler le nombre de $\left\{z_{(U_k)}^{REG.j} = \text{PROBABLE} \cup \text{POSSIBLE}\right\}$ sans quoi il est soit surévalué, soit sous-évalué.

Ensuite, afin de profiter pleinement de l'information contenue par $z'^j_{(U_k),c}$ l'estimation de *cette moyenne spatiale pondérée anormalement élevée* est effectuée par bootstrap.

Processus d'estimation bootstrap :
L'estimation de la variable $b_{1-\alpha}^{*,j}$ s'effectue par bootstrap. Elle est dite *semi-paramétrique* puisque conditionnée par la valeur choisie pour α^j.

Remarque théorique :
L'idée du bootstrap est de parvenir à une estimation plus robuste des paramètres statistiques, uniquement à partir des informations disponibles, i.e. en s'affranchissant de toute hypothèse sur la loi suivie par la variable d'intérêt (Chernick, 1999).

Principe d'estimation:
Le bootstrap consiste à générer B ré-échantillonnages bootstrap conditionnellement aux données disponibles.

$$z'^{j,*,\cdot}_{(U_k),c} = \left(z'^{j,*,1}_{(U_k),c}, \dots, z'^{j,*,B}_{(U_k),c}\right)$$

De telle sorte qu'ils soient tirés aléatoirement dans des multinomiales uniformes discrètes :

$$z'^{j,*,b}_{(U_k),c} \sim \mathcal{M}\left(\left(z'^{j}_{(U_1),c}, \dots, z'^{j}_{(U_{q_1}),c}\right); \frac{1}{\mathrm{N(U_k)}}\right), \quad \forall\, b = \{1, \dots, B\}$$

Une fois les échantillons bootstrap générés, on peut, par agrégation ensembliste, estimer la moyenne et l'écart-type bootstrap :

$$\mathbb{E}^*\left[z'^{j,*,\cdot}_{(U_k),c}\right] \approx \bar{z}'^{j,*}_{(U_k),c} \overset{\text{def}}{=} \frac{1}{B}\sum_{b=1}^{B} \mathrm{m\hat{o}y}_b\left(z'^{j,*,b}_{(U_k),c}\right)$$

$$\sigma^*\left[z'^{j,*,\cdot}_{(U_k),c}\right] \approx \hat{\sigma}^B\left(z'^{j,*,\cdot}_{(U_k),c}\right) \overset{\text{def}}{=} \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\mathrm{m\hat{o}y}_b\left(z'^{j,*,b}_{(U_k),c}\right) - \bar{z}'^{j,*}_{(U_k),c}\right)^2}$$

Ensuite une statistique bootstrap particulière, qui assure une convergence Gaussienne asymptotique, est estimée :

$$\mathrm{T}_j^{*,b} = \frac{\bar{z}'^{j,*,b}_{(U_k),c} - \bar{z}'^{j,*}_{(U_k),c}}{\hat{\sigma}^B\left(z'^{j,*}_{(U_k),c}\right)} \xrightarrow[\mathrm{n}\to+\infty]{\text{approximativement}} \mathrm{N} \sim \mathcal{N}(0,1); \quad \forall\, \mathrm{b} = \{1, \dots, \mathrm{B}\}$$

Enfin pour obtenir le paramètre $\mathrm{b}^{*,\mathrm{j}}_{1-\alpha}$ il suffit de procéder à l'estimation du percentile bootstrap à partir de la valeur spécifiée pour α^{j} :

$$b^{*,j}_{1-\alpha} = \mathbb{P}_{F_n^{*,b}}\left(\mathrm{T}_j^{*,b} \le t_{(1-\alpha)}\right), \qquad \text{avec: } \alpha^j \in [\![5\%;\ 10\%]\!]$$

Aussi, il est possible de construire une fonction de répartition empirique bootstrap à partir de cette statistique et de vérifier son caractère approximativement Gaussien :

$$F_n^{*,b}(t_j) = \mathbb{P}_{F_n^{*,b}}\left(\mathrm{T}_j^* \le t\right) = \frac{1}{\mathrm{B}}\sum_{b=1}^{B} \mathbb{1}_{\left\{\mathrm{T}_j^{*,b} \le t \,\middle|\, \left(\bar{z}'^{j,*,b}_{(U_k),c}\right)_{1\le k\le q_1}\right\}}$$

Hypothèse théorique:
Si la distribution des $\bar{\mathrm{z}}'^{\mathrm{j},*,\mathrm{b}}_{(\mathrm{U_k}),\mathrm{c}}$ suit *approximativement* une $\mathcal{N}(0,1)$, cela peut être on peut vérifier visuellement. L'hypothèse d'une convergence asymptotique permet de conjecture l'écart et le gain de précision apporté par le bootstrap.

Application aux données LEA : le processus a été appliqué aux $\mathrm{z}'^{\mathrm{j}}_{(\mathrm{U_k}),\mathrm{c}}$ pour chacune des séquelles. Les distributions bootstrap ont été mises en perspective d'une Gaussienne centrée réduite :

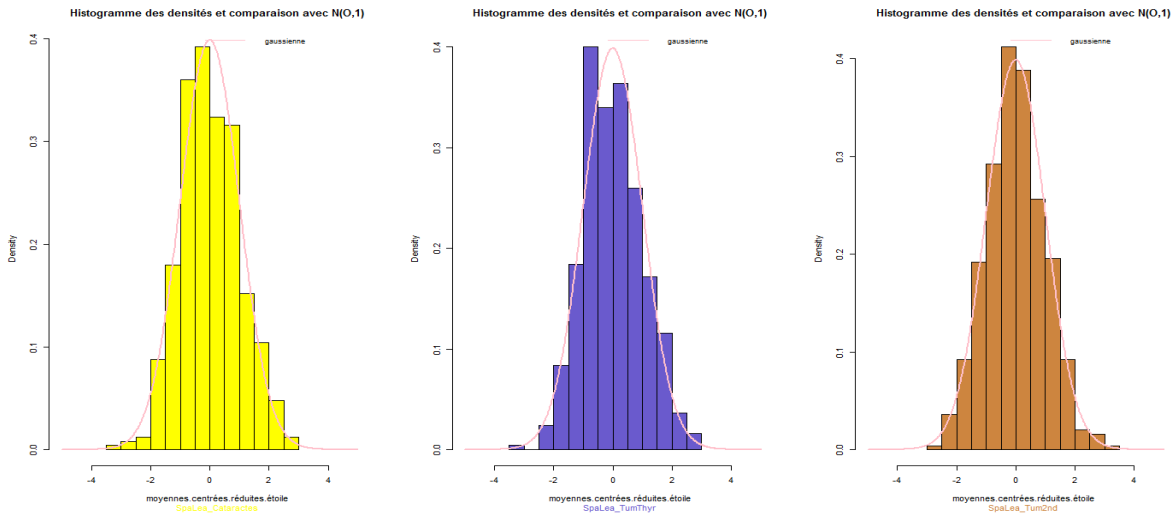


Figure 60 : Comparaison de la distribution boostrap des $T_j^{*,b}$ et de celle de $\mathcal{N}(0,1)$ pour chaque séquelle.

Le calcul des $b_{1-\alpha}^{*,j}$ a été effectué avec le logiciel R (Institute for Statistics and Mathematics, 1997).

ESTIMATION DE ξ_{geo}^{j}

Remarque liminaire :
L'action du paramètre α^j est très modérée. Elle ne permet pas d'obtenir un seuil suffisamment robuste pour juguler le nombre de valeurs $\left\{z_{(U_k)}^{REG.j} = \text{POSSIBLE}\right\}$ et assigner correctement les modalités $\left\{z_{(U_k)}^{REG.j} = \text{PROBABLE}\right\}$.

Proposition :
Pour pallier cette carence, le coefficient d'élasticité géographique $\xi_{géo}^{j}$ est introduit. Au vu de la stratégie d'estimation proposée pour le seuil φ_j^*, la valeur de $\xi_{géo}^{j}$ est spécifiée - à l'aune des α^j - conditionnellement *aux connaissances expertes* acquises sur la nature et la géographie du PM* et proportionnellement aux distorsions géographiques induites par l'action combinée des π_i^j et des tee_i^j sur la variabilité des $z'^{j}_{(U_k),c}$.

Remarque :
Par ailleurs, puisque la géographie des risques d'exposition aux PM* - séquelles n'est pas connue *a priori* les $\xi_{géo}^{j}$ ne peuvent pas être calibrées

Hypothèse subsidiaire :
Si l'estimation de la borne supérieure de la moyenne géographique pondérée est suffisamment robuste alors les valeurs prises par $\xi_{géo}^{j}$ ne sont pas aléatoires.

Proposition subsidiaire :
Il est vraisemblable de penser que le coefficient d'élasticité géographique peut prendre une gamme bornée de valeurs. L'expérience a montré que cette assertion était vraie, et que sur les données de la Cohorte LEA :

$$\xi_{géo}^{j} \in \Im_{\xi} = [\![1\ ;\ 2]\!]$$

Considérations expertes d'estimation :

L'action des π_i^j est plus forte sur les THYR que sur toutes les autres séquelles. En contrepartie l'incidence des THYR observées (10,2%) sur les patients spatialisés est proche de celle estimée sur les CATA observées (13%) et par conséquent l'effet particulièrement important des tee_i^j vient tempérer celle des π_i^j. Ainsi, l'action conjointe des π_i^j et des tee_i^j sur la modélisation géographique des PM* – séquelles pour les patients spatialisés est importante sur les CATA, plus modérée sur les THYR et moindre sur les TUM2.

Conséquence :

Il en découle la hiérarchie suivante :

$$\xi_{géo}^{CATA} \leq \xi_{géo}^{THYR} \leq \xi_{géo}^{TUM2}$$

Application aux données LEA :

Séquelle :	CATA	THYR	TUM2
$\xi_{géo}^{j}$	1.10	1.50%	1.90%

Tableau 31 : Valeurs de $\xi_{géo}^{j}$ spécifiées pour chaque séquelle

En somme, le niveau de risque α^j amplifié par le coefficient d'élasticité géographique $\xi_{géo}^{j}$ jugule globalement le nombre de $\left\{z_{(U_k)}^{REG.j} = \text{PROBABLE} \cup \text{POSSIBLE}\right\}$ conditionnellement à l'impact conjoint de la *métrique floue géographique* et de celle la conversion *en patients-années*.

De fait, un *seuil d'élasticité géographique* φ_j^* robuste peut être estimé afin de caractériser les espaces en fonction des Risques d'Exposition Géographique à des PM* - séquelles.

Ils permettent de tempérer le nombre $\left\{z_{(U_k)}^{REG.j} = \text{PROBABLE}\right\}$ et parallèlement de contrebalancer judicieusement cette perte par des $\left\{z_{(U_k)}^{REG.j} = \text{POSSIBLE}\right\}$ – afin d'obtenir des représentations plus précises et plus adaptées à la réalité géographique des niveaux de Risques d'Expositions aux Phénomènes Morbides.

Cette stratégie de caractérisation des espaces géographiques en fonction d'une typologie de REG a été appliquée aux données de la Cohorte LEA. Les résultats statistiques et cartographiques obtenus sont présentés conséquemment.

PRESENTATION ET ANALYSE DES RESULTATS

Les résultats de la caractérisation des unités géographiques communales $U_{\boldsymbol{k}}$ en fonction du Risque d'Exposition Géographique sont donnés successivement pour les séquelles : cataractes (CATA), tumeurs thyroïdiennes (THYR) et tumeurs secondaires (TUM2).
Les paramètres utilisés pour l'estimation du seuil d'élasticité géographique φ_j^* sont ceux spécifiés auparavant.

Les résultats cartographiques :
Les communes de première espèce concernées sont celles sises en région PACA et aux alentours pour le Sud-Est de la France, et pour les U_k de la région Alsace-Lorraine pour le Nord-Est de la France. Les cartes sont documentées avec le nom des communes qui interviennent dans l'analyse en fonction du type de séquelle considéré et de telle sorte que les cartes restent intelligibles :

$$\text{label}_{(U_k)} = \begin{cases} \text{Display} \quad \text{lorsque:} \quad \left\{ z_{(U_k)}^{REG(CATA)} = \text{PROBABLE} \right\} \\ \text{Display} \quad \text{lorsque:} \quad \left\{ z_{(U_k)}^{REG(THYR)} = \{\text{PROBABLE}\} \right\} \\ \text{Display} \quad \text{lorsque:} \quad \left\{ z_{(U_k)}^{REG(TUM2)} = \{\text{PROBABLE} \cup \text{POSSIBLE}\} \right\} \end{cases}$$

Représentations statistiques
Elles sont données par l'histogramme des fréquences empiriques associées aux modalités des $z_{(Uk)}^{REG.j}$ – estimées sur l'ensemble des $\mathcal{U}_{\cdot}^{1^{er}}$. Les tableaux statistiques ne sont pas donnés car les graphiques sont *correctement documentés* (Saporta, 2006).

SEQUELLE : CATARACTES

Cartographie des $z_{(U_k)}^{REG(CATA)}$ pour le Sud-Est de la France

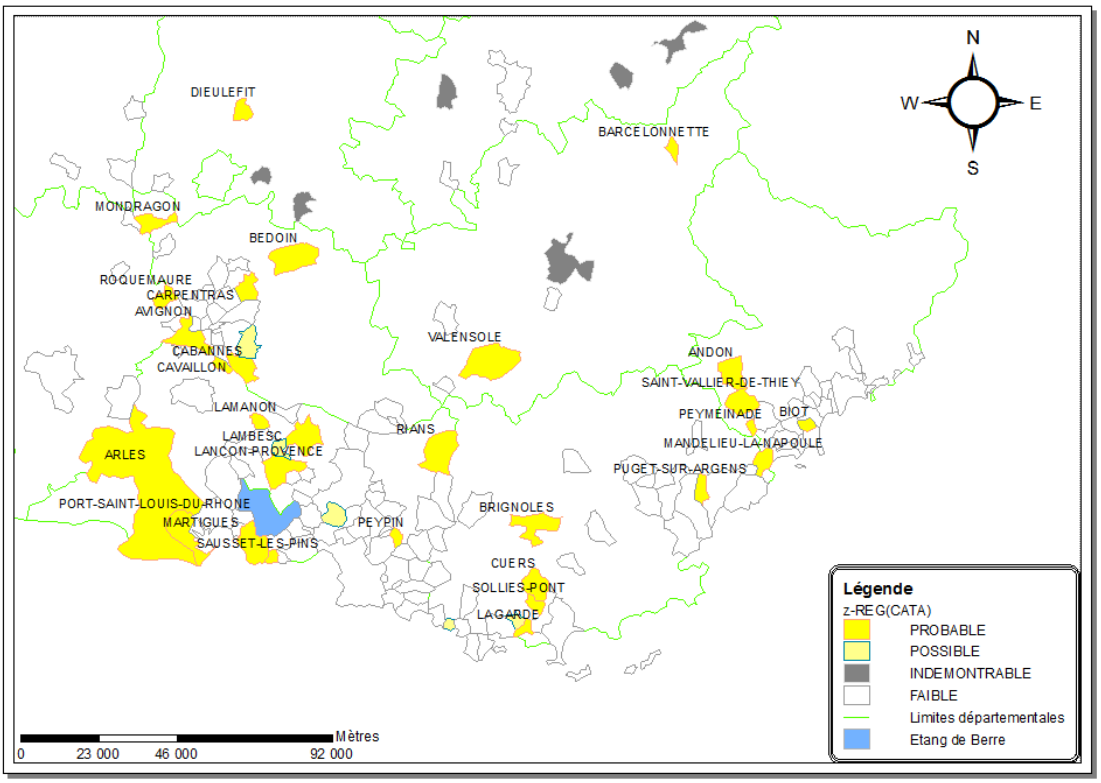


Figure 61 : Valeurs prises par $z_{(U_k)}^{REG(CATA)}$ et affichage des noms de communes où le REG est *probable*

Cartographie des $z_{(U_k)}^{REG(CATA)}$ pour le Nord-Est de la France

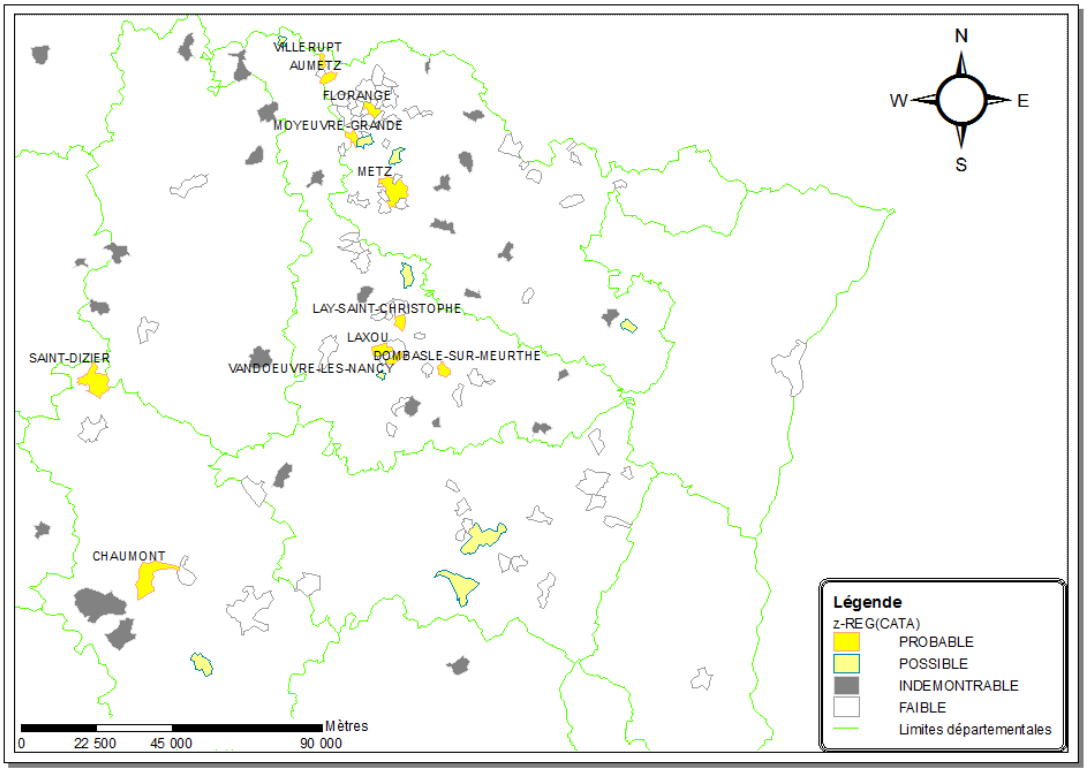


Figure 62 : Valeurs prises par $z_{(U_k)}^{REG(CATA)}$ et affichage des noms de communes où le REG est *probable*

Présentation des résultats statistiques obtenus pour $z_{(U_k)}^{REG(CATA)}$ sur l'ensemble des $\mathcal{U}_{\cdot}^{1^{er}}$

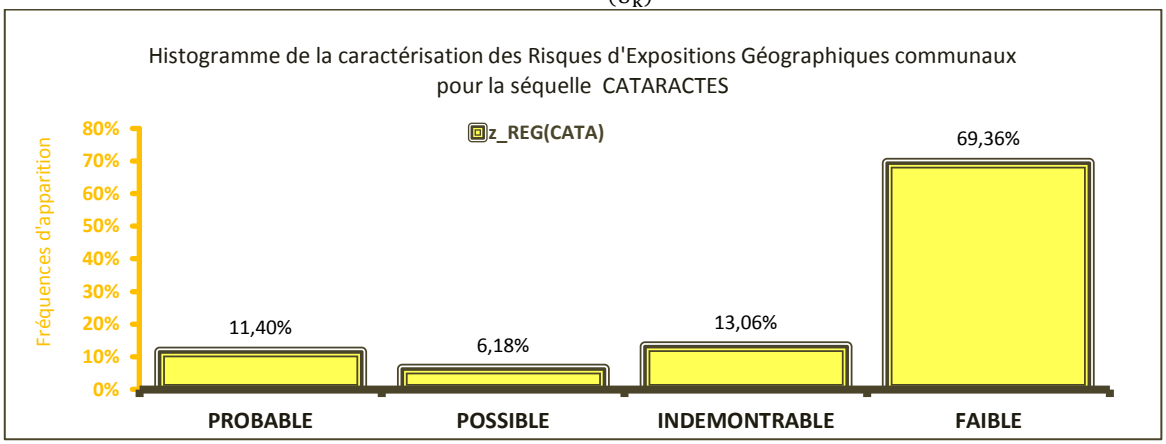


Figure 63 : Histogramme documenté des fréquences empiriques associées aux modalités des $z_{(U_k)}^{REG(CATA)}$

SEQUELLE : TUMEURS THYROÏDIENNES

Cartographie des $z_{(U_k)}^{REG(THYR)}$ pour le Sud-Est de la France

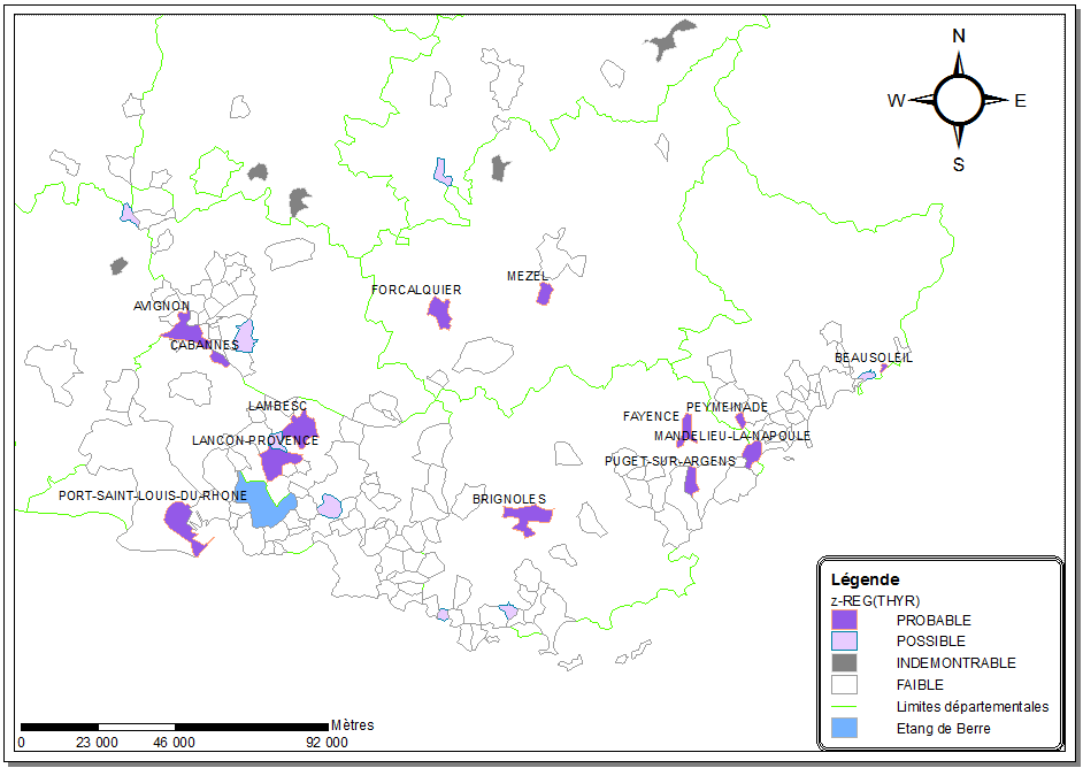


Figure 64 : Valeurs prises par $z_{(U_k)}^{REG(THYR)}$ et affichage des noms de communes où le REG est *probable*

Cartographie des $z_{(U_k)}^{REG(THYR)}$ pour le Nord-Est de la France

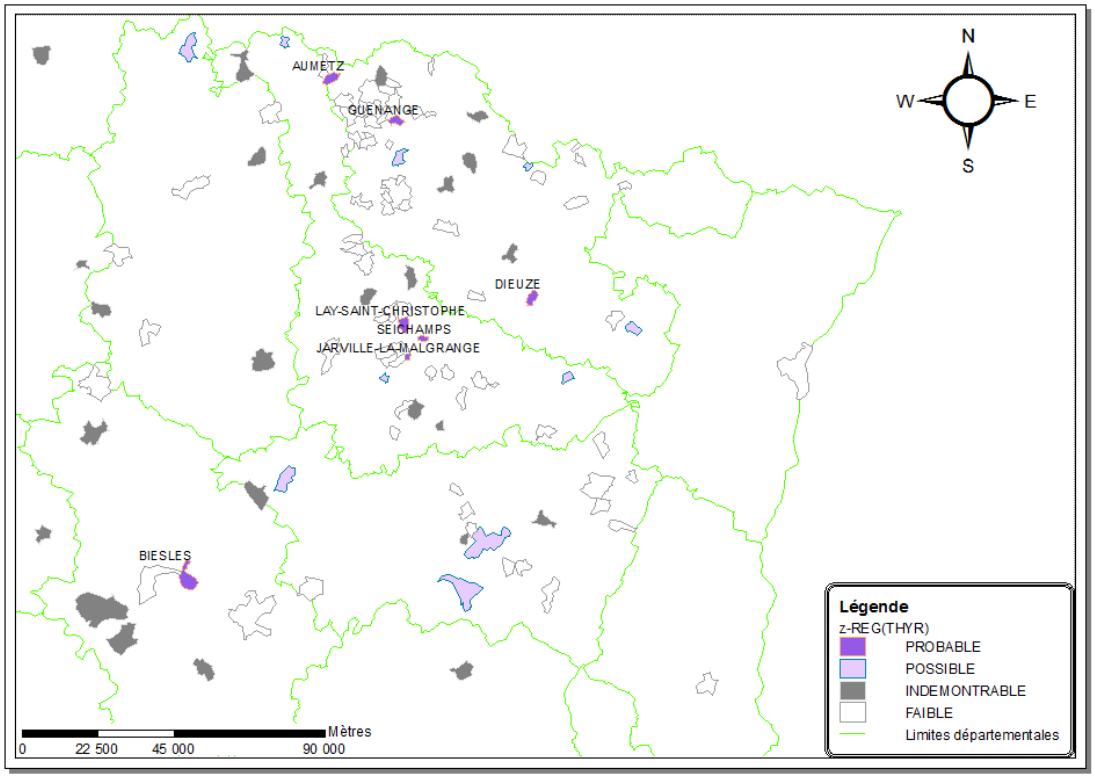


Figure 65 : Valeurs prises par $z_{(U_k)}^{REG(THYR)}$ et affichage des noms de communes où le REG est *probable*

Présentation des résultats statistiques obtenus pour $z_{(U_k)}^{REG(THYR)}$ sur l'ensemble des $\mathcal{U}_{.}^{1^{er}}$

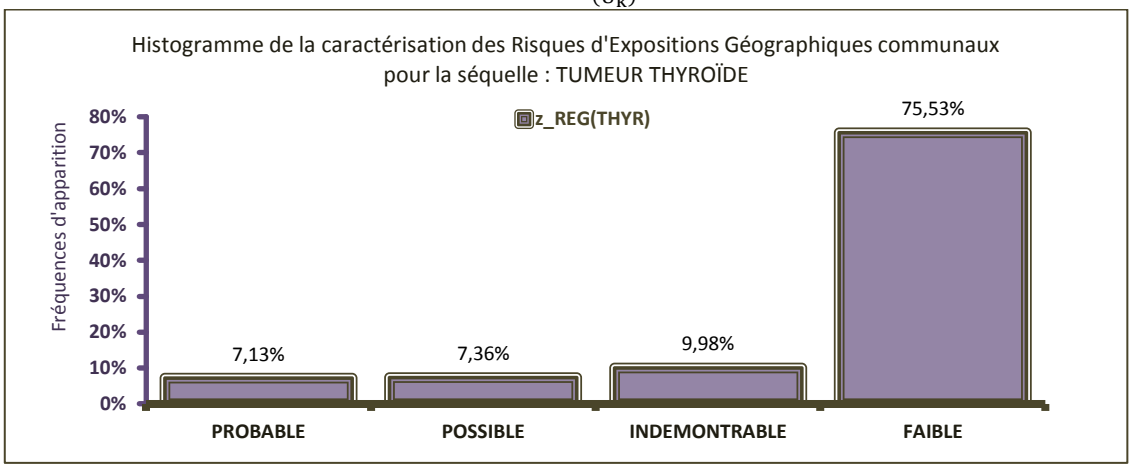


Figure 66 : Histogramme documenté des fréquences empiriques associées aux modalités des $z_{(U_k)}^{REG(THYR)}$

SEQUELLE : TUMEURS SECONDAIRES

Cartographie des $z_{(U_k)}^{REG(TUM2)}$ pour le Sud-Est de la France

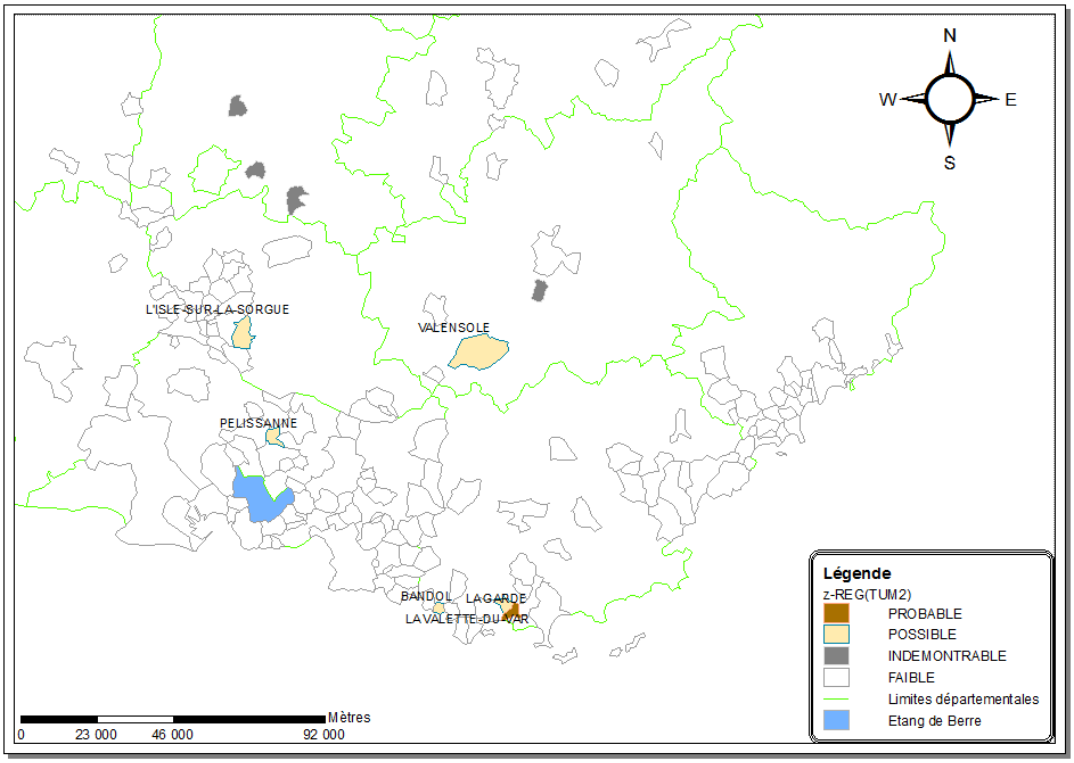


Figure 67 : Valeurs prises par $z_{(U_k)}^{REG(TUM2)}$ et affichage des noms de communes où le REG est *probable ou possible*

Cartographie des $z_{(U_k)}^{REG(TUM2)}$ pour le Nord-Est de la France

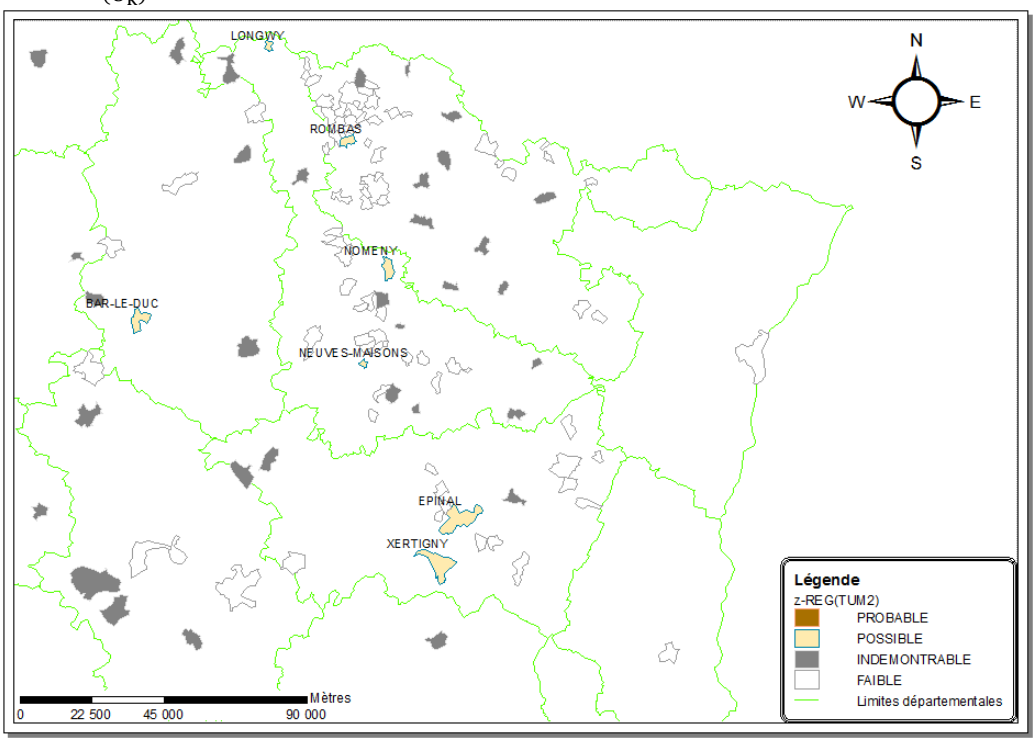


Figure 68 : Valeurs prises par $z_{(U_k)}^{REG(TUM2)}$ et affichage des noms de communes où le REG est *probable ou possible*

Présentation des résultats statistiques obtenus pour $z_{(U_k)}^{REG(TUM2)}$ sur l'ensemble des $\mathcal{U}_{\cdot}^{1^{er}}$

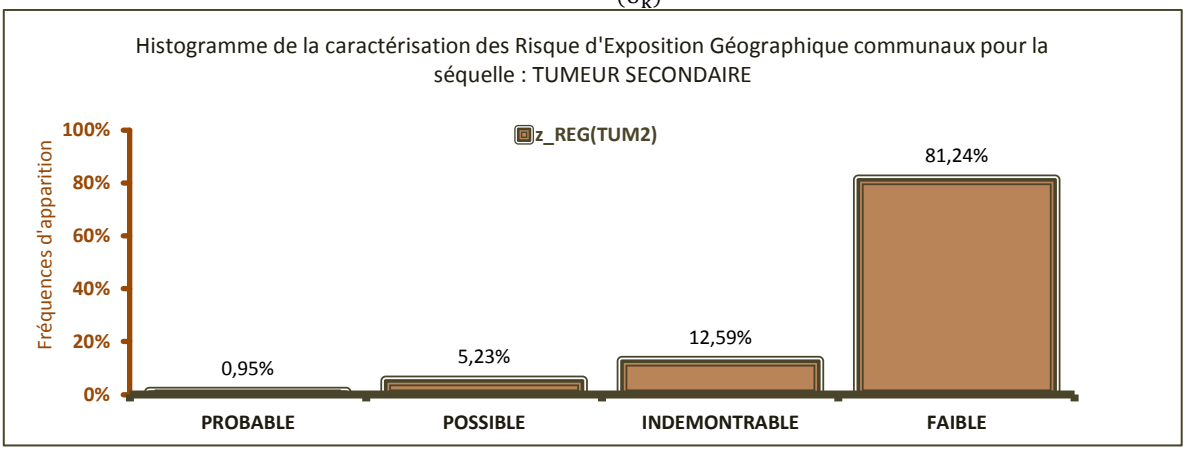


Figure 69 : Histogramme documenté des fréquences empiriques associées aux modalités des $z_{(U_k)}^{REG(TUM2)}$

RESUME, ANALYSE ET REMARQUES

L'analyse spatiale descriptive des Risques d'Exposition Géographique (REG) aux Phénomènes Morbides (PM) d'intérêt - séquelles – caractérisant les $\mathcal{U}^{1er}$, a pour objet d'identifier des disparités et des similitudes (ou géométries) géographiques, donc de faire émerger des tendances globales en matière de REG.

Analyse spatiale :
La modélisation du Risque d'Exposition Géographique (REG) appliquée à la séquelle CATA caractérise 11,4% des communes de : PROBABLE. S'agissant de la séquelle THYR, le REG : PROBABLE caractérise 7,1% des communes. Et cette proportion atteint 14,5% des U_k lorsqu'on considère aussi les REG : POSSIBLE. Enfin, pour les TUM2 cette proportion est plus faible, les REG : POSSIBLE ou PROBABLE concernent seulement 6,2% des communes. Pour toutes les séquelles confondues les modalités PROBABLE sont essentiellement affectées aux U_k situées dans le Sud-Est de la France.

Pour 13,1% des communes le REG aux CATA est qualifié de INDEMONTRABLE. Cette impossibilité à caractériser le REG touche 10,0% des U_k pour la séquelle THYR et cette proportion atteint 12,5% pour les TUM2.

La caractérisation des espaces par un REG est feutrée par le spectre d'incertitudes multiples. L'affection des modalités du REG : PROBABLE et aussi du REG : POSSIBLE sont *préjudiciables* pour les communes concernées. Par conséquent, mieux vaut être circonspect plutôt qu'affecter à tort une modalité contraire à la réalité géographique lorsque les incertitudes sont trop fortes. Cependant, il convient de ne pas considérer non plus ces communes comme des U_k où aucun patient ne serait spatialisé puisque des informations spatialisées morbides sont néanmoins disponibles. Les U_k où le REG est INDEMONTRABLE doivent être traitées avec beaucoup d'égard.

La stratégie d'estimation proposée affecte les $z_{(U_k)}^{REG.j} = $ INDEMONTRABLE associés aux U_k où peu de patients résident et dont le CP matérialise des zones géographiques étendues ou dont la qualité spatiotemporelle des données épidémiologiques est peu fiable. Quelle que soit la séquelle considérée, comme le montrent les résultats cartographiques, les U_k concernées se situent majoritairement dans le Nord la France. Par conséquent, l'analyse spatiale descriptive des $z_{(U_k)}^{REG.j}$ se limitera aux communes sises en région PACA. Compte tenu : des incertitudes multiples qui maculent la modélisation des PM*, de l'impossibilité de valider la méthode par la géographie réelle des REG qui est inconue, et du fait que cette recherche a pour finalité l'identification de DES, alors l'analyse spatiale des $z_{(U_k)}^{REG.j}$ sera sommaire.

En région PACA, $z_{(U_k)}^{REG.TUM2} = $ PROBABLE caractérise uniquement la commune de La-Garde ; Et $z_{(U_k)}^{REG.TUM2} = $ POSSIBLE concerne cinq communes : La-Valette-du-Var, Bandol, Valensole, L'île-sur-la-Sorgue et Pelissanne. Le REG de THYR est PROBABLE pour 13 communes, dont : Beausoleil, Peymeinade, Mandelieu-la-Napoule, Fayence, Puget-sur-Argens, Brignoles, Lançon-en-Provence, Avignon, Cabannes, Port-Saint-Louis-du-Rhône, Lambesc. Quant à, $z_{(U_k)}^{REG.THYR} = $ POSSIBLE il concerne les communes de : La-Valette-du-Var, Bandol, L'île-sur-la-Sorgue et Pelissanne, ainsi que Eze et Cabries. Enfin, les $z_{(Uk)}^{REG.CATA} = $ PROBABLE touchent 27 communes en PACA, avec entre autres : Valensole, Biot, Port-Saint-Louis-Du-Rhône, Mandelieu-La-Napoule, Saint-Vallier-De-Thiey, Cabannes, Brignoles, Puget-Sur-Argens, Avignon, Carpentras, Cavaillon Mondragon, Lançon-en-Provence, Sausset-Les-Pins, Martigues, La-Garde, Sollies-Pont, Lamanon, Arles, Lambesc. Et cinq sont caractérisées par un REG de CATA : POSSIBLE, celles de : Bandol, Pelissanne, L'île-sur-la-sorgue, La-Valette-du-Var et Cabries.

Il existe donc des disparités géographiques évidentes dans l'affectation des REG mais aussi des régularités géographiques morbides. En l'occurrence, la commune de La-Garde a un $z_{(Uk)}^{REG.j} = $ PROBABLE

pour les CATA et les TUM2 ; Et il en est de même concernant les THYR et les CATA pour les sept communes suivantes : Mandelieu-la-Napoule, Puget-Sur-Argens, Brignoles, Lançon-en-Provence, Avignon, Cabannes et Lambesc.

D'une manière plus générale, les communes de La-Valette-du-Var, Bandol, L'île-sur-la-Sorgue, Pelissanne sont caractérisées $z_{(U_k)}^{REG.j} = \{PROBABLE \cup POSSIBLE\}$ pour toutes les séquelles.

Interprétation et remarques heuristiques :
Les régularités géographiques observées dénotent la présence d'un effet *environnement*. Il convient cependant de se demander si les prédispositions morbides mises en exergue par les $z_{(Uk)}^{REG.j}$ sont induites plutôt par les FIM, i.e. des agrégations des patients ayant reçu des protocoles de traitement lourds, ou s'il s'agit de communes favorisant les expositions combinées à des FE* délétères, ou encore s'il cela n'est pas simplement le fruit d'un hasard quelconque.

Les quatre U_k dans lesquelles les REG suggèrent une propension à développer toutes les séquelles renforcent l'hypothèse selon laquelle la géographie des CIM* a un effet sur celle des PM. Manifestement, la situation est celle d'une agrégation de patients ayant reçu des protocoles de traitement lourds qui ont des effets secondaires au long cours. D'où l'importance d'intégrer les FIM* qui ont indubitablement une influence sur l'état de santé des patients.

En contrepartie, le fait que ces singularités soient anecdotiques et qu'elles s'observent sur des modalités de REG différentes renforce aussi l'hypothèse de l'effet, au moins contributif, d'expositions à des FE* particuliers ou combinées. D'où l'intérêt de modéliser aussi l'environnement géographique dans toute sa plénitude, i.e. ses composantes SAN, SOCIO.ECO et PHY.CHIM, dont la variabilité diffère avec des gradients spatiotemporels forts pour certains, et qui est stationnaire pour d'autres.

A l'instar de l'analyse spatiale conjointe des $z'^{j}_{(U_k),c}$ *extrêmes* et des $z'^{j}_{(U_k),q}$, celle menée sur les $z_{(U_k)}^{REG.j}$ met en exergue, avec une évidence encore plus forte, le fait que la géographie des PM* n'est pas aléatoire. Les régularités observées en sont la preuve. Et elles sont d'autant plus prégnantes lorsqu'on s'attache à la *géométrie des contiguïtés spatiales*. En effet, les cartographies présentées montrent que pour toutes les séquelles, il est possible de former des *blocs de communes caractérisées par un même REG*, avec des U_k proches les unes des autres lorsqu'elles ne se trouvent pas en contiguïté spatiale directe.

Ce constat renforce l'hypothèse du rôle, *au moins contributif*, *des FE* sur l'état de santé des patients* - jusqu'à le rendre même évident. Sans pour autant infléchir l'effet inéluctable des CIM* sur la géographie des PM* d'intérêt modélisée par les i.st. . $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$, puisque les $z_{(U_k)}^{REG.j}$ ne constituent qu'une stratégie de fusion d'informations vouée à l'interprétation géographique de ces derniers.

SYNTHESE DU CHAPITRE 2

Le premier i.st.m* proposé est quantitatif, il est noté $z'^{j}_{(Uk),c}$. Il représente des prévalences de pathologies – séquelles - *pondérées EpiGéoStat* et tient compte du *temps d'exposition à l'environnement géographique*.

Le second i.st.m* est qualitatif, il est noté $z'^{j}_{(Uk),q}$ et exprime la *propension qu'ont les individus à développer une pathologie* – séquelle – en accordant une attention particulière à la *qualité spatiale EpiGéoStat* des données utilisées et conjointement à la *connaissance temporelle médicale disponible* pour chaque individu réduit par son temps d'exposition à l'environnement géographique.

Enfin un troisième i.st.m* est proposé. Il est noté $z^{REG.j}_{(U_k)}$ et résulte d'une *fusion de l'information* contenue dans les deux premiers. La stratégie d'agrégation se fonde sur l'estimation d'un *seuil d'élasticité géographique* φ^*_j mêlant *judicieusement* : Les caractéristiques statistiques des deux premiers i.st.m* ; Et des *connaissances expertes* sur les *processus de modélisation* et la *nature des PM** - séquelles. Les $z^{REG.j}_{(U_k)}$ caractérisent chaque U_k par un type de Risque d'Exposition Géographique (REG) morbide, qui peut être : PROBABLE ; POSSIBLE ; INDEMONTRABLE ; FAIBLE.

Ces propositions heuristiques ont été appliquées aux séquelles d'intérêt à partir des données de la Cohorte LEA. Elles peuvent cependant être étendues facilement à toutes les autres séquelles développées par les patients et à n'importe quelle maladie en adaptant au préalable : La *métrique floue* proposée et l'estimation des *temps d'exposition à l'environnement*, inhérents à la modélisation géographique des PM* ; Ainsi que la valeur des paramètres du *seuil d'élasticité géographique*, pour ce qui est de la caractérisation des REG morbides.

Les $z'^{j}_{(U_k),c}$ sont particulièrement affectés par les *facteurs EpiGéoStat* d'incertitude spatiotemporelle : π_i^j. Leurs valeurs sont donc difficiles à interpréter. Cependant l'utilisation conjointe des $z'^{j}_{(U_k),q}$, permet d'estimer la qualité spatiotemporelle associée aux hypothèses, aux méthodes et aux données utilisées, et donc de spécifier l'information apportée par les prévalences spatiales pondérées *EpiGéoStat* et converties en *Patients-Années*. Aussi les π_i^j permettent d'augmenter la robustesse spatiale et donc *d'intégrer l'espace au cœur du processus de modélisation géographique*.

La conversion en *Patients-Années* peut être étendue à n'importe quel autre PM. Cette transformation permet *d'augmenter virtuellement la taille de la cohorte*. Par conséquent, elle est adaptée à la modélisation géographique des PM* de petites cohortes, dont LEA fait partie. De plus, les *temps d'exposition à l'environnement* permettent d'intégrer *la temporalité au cœur du processus de modélisation,* ce qui donne plus de sens et de consistance à la dialectique géographique des interactions santé-environnement.

L'analyse des $z'^{j}_{(U_k),c}$ *extrêmes* et $z'^{j}_{(U_k),q}$ a mis en évidence, de façon très pragmatique, que la *métrique floue géographique* couplée à une transformation en *Patients-Années* ne bruitait pas la modélisation géographique des PM, mais qu'au contraire, les pondérations semblaient améliorer significativement la précision et la robustesse des i.st.m* proposés

L'analyse conjointe des $z'^{j}_{(U_k),c}$ *extrêmes* et $z'^{j}_{(U_k),q}$ met en évidence des disparités géographiques morbides, donc une tendance à développer certaines séquelles dans des communes et pas dans d'autres – ce qui suggère *un effet environnement*. Mais surtout l'observation de ces similitudes géographiques

étendues à l'ensemble des pathologies – séquelles – permet d'identifier des *blocs* de U_k situées dans des zones géographiques peu étendues lorsqu'elles ne se trouvent pas en *contiguïté spatiale directe*. Par conséquent, l'*effet de l'environnement géographique* est renforcé sans pour autant renier l'influence évidente des CIM, e.g. des traitements reçus, sur l'état de santé des patients.

L'analyse conjointe des $z'^{j}_{(U_k),c}$ *extrêmes* et $z'^{j}_{(U_k),q}$ est cependant limitée par la présence d'antagonismes et d'incertitudes qui rendent difficile l'interprétation de la géographie des PM* qu'ils modélisent. Puisque le travail du géographe de la santé ne saurait se limiter à des modélisations spatiotemporelles et qu'il a aussi pour devoir de caractériser les territoires, les $z^{REG.j}_{(U_k)}$ constituent une stratégie de fusion d'informations judicieuse vouée à la discrétisation des espaces en fonction des Risques d'Exposition Géographique (REG) à des phénomènes morbides particuliers.

Les incertitudes qui maculent la modélisation des PM* peuvent avoir des répercussions sur l'affectation des REG. Or, il convient de ne pas caractériser à tort les Unités Géographiques : U_k de risquées alors qu'elles ne le sont pas.
Mieux vaut être circonspect. En effet, des REG : PROBABLE ou POSSIBLE sont particulièrement préjudiciables pour les communes visées. Raison pour laquelle la stratégie de caractérisation permet d'intégrer des *connaissances expertes à connotation statistique, géographique et épidémiologique* – fondées sur des prénotions bibliographiques et méthodologiques, de sorte à ne pas surévaluer, ni sous-évaluer, les REG morbides auxquels sont assujettis les patients, et par extension les populations.

L'analyse des $z^{REG}_{(U_k)}$ a permis de montrer avec une grande acuité qu'il existe des disparités géographiques en matière de REG aux phénomènes morbides. Toutefois le nombre de communes caractérisées par un REG PROBABLE reste conditionné par la proportion de patients spatialisés qui ont développé le PM* étudié.

Par ailleurs, les similitudes géographiques observées sur les U_k sont systématiquement qualifiées, pour toutes séquelles, par un REG. Et parallèlement, les U_k caractérisées par des REG *préjudiciables* qui forment des *blocs de communes* parce qu'elles sont proches les unes des autres lorsqu'elles ne se trouvent contiguïté spatiale directe - constat qui avait déjà été établi par l'analyse conjointe des $z^{j}_{(U_k),c}$ *extrêmes* et $z'^{j}_{(U_k),q}$ mais qui est encore plus prégnant sur les $z^{REG.j}_{(U_k)}$ - permettent de conjecturer un *effet de l'environnement géographique évident*.

CONCLUSION DU CHAPITRE 2

Les analyses spatiales des i.st.m* proposés montrent donc qu'il existe un effet, au moins contributif, des FE* combinés sur l'état de santé des patients. Cette hypothèse est particulièrement prégnante.
En effet, si les risques morbides étaient uniquement dus aux FIM, i.e. à la géographie des CIM* - dont les traitements reçus font partie – alors les mêmes U_k devraient systématiquement être qualifiées par un même type de REG. Or ce n'est pas exactement ce qui est observé sur les tumeurs secondaires (TUM2) et les tumeurs thyroïdiennes (THYR), qui sont des séquelles pourtant proches, ni *a fortiori* sur les cataractes (CATA).

CONCLUSION DE LA PARTIE I

La modélisation des Risques d'Exposition Géographique (REG) aux Phénomènes Morbides (PM) développés par les patients de la Cohorte LEA a permis de montrer qu'en agrégeant des U_k situées en contiguïté spatiale directe ou partielle, il était possible de former des *blocs* caractérisés par un même REG.
Ces *cluster* ou *agrégats spatiotemporels morbides* mettent en exergue le fait que la géographie des séquelles n'est pas aléatoire et que l'*environnement*, tel qu'il a été défini, a sa part de responsabilité (Tillaut, 2005).

Dès lors deux hypothèses se présentent : (i) *Les séquelles développées par les patients sont déterminées par leur CIM, en l'occurrence par les effets au long cours de traitements agressifs, et rien de plus* ; (ii) *La géographie des CIM, en particulier l'historique médical des patients, est déterminante sur l'incidence des séquelles mais ne suffit pas à les expliquer, ni à conjecturer les prédispositions géographiques morbides observées – qui sont conditionnées aussi par l'action contributive de FE* combinés*.

En vertu de la première loi de la Géographie - qui postule que *ce qui est proche se ressemble et ce qui est éloigné dissemble* (Tobler, 1970) – on peut supposer que la variabilité spatiale des FE, dans *les clusters* caractérisés par un même REG, est faible et que par conséquent la seconde éventualité est la plus probante.

La géographie des séquelles n'est donc pas uniquement conditionnée par celle des CIM. Les FE* : SAN, SOCIO.ECO et PHY.CHIM ont aussi une influence contributive dont il convient d'évaluer l'impact. Et, par hypothèse, l'effet combiné des expositions environnementales peut être étendu à l'état de santé général des populations.

Conformément au domaine scientifique dans lequel cette recherche s'inscrit l'*environnement géographique* est considéré dans toute sa plénitude. Et le rôle du Géographe de la Santé est notamment de construire des i.st.m* et des i.st.e, *pertinents*, fiables* et *adaptés à la gestion des territoires.* Par suite, son devoir est aussi d'identifier des DES* afin de réduire les Risques d'Exposition Géographique morbides, en donnant aux praticiens de santé et aux acteurs politiques des moyens opérationnels pour *protéger la santé individuelle et promouvoir la santé publique* (Salem, 1995).

Pour les épidémiologistes les FE-SAN* sont *a priori* les DES* les plus adéquats pour expliquer l'état de santé des populations – en particulier lorsqu'on travaille sur des séquelles (Barnett, Wrigley et al., 2002). Par contre, les Géographes de la Santé privilégient plutôt la piste des expositions géographiques combinées à des FE-PHY.CHIM* (Brucker-Davis, 1998).

Les points de vue des deux disciplines convergent sur le caractère déterministe des FIM* et l'influence potentielle et indirecte des FE-SOCIO.ECO* (Chaix, Merlo et al., 2005).

Les FE-PHY.CHIM* sont controversés du fait de la difficulté qu'il y a à établir des relations causales significatives entre les *expositions environnementales* à de faibles doses de substances toxiques et l'incidence géographique des maladies. Généralement, les *cluster*s *morbides* identifiés dans les études géographiques sont maculés d'incertitudes et l'effet des FE-PHY.CHIM* étudiés a autant *de chances d'être corrélé que parfaitement indépendant du PM** (Dejour-Salamancad, Gomes-Do-Espirito-Santo et al., 2005).

A cela s'ajoutent des problèmes méthodologiques liés à l'utilisation de modèles statistiques de *puissance faible* et *de données environnementales de mauvaise qualité.* Or, avec l'émergence exacerbée de BD

contenant des variables aux caractéristiques granulaires de plus en plus fiables, *les possibilités des modélisations géographiques environnementales ne semblent plus être un problème* (Zeitouni, 2006).
Conséquence, on assiste actuellement à un regain d'attention pour les expositions géographiques chroniques à des doses environnementales de substances toxiques, ce qui constitue un véritable changement de paradigme pour les épidémiologistes (Inserm - expertise collective, 2005).

Les maladies sont multifactorielles. L'environnement géographique est multidimensionnel. Par conséquent l'approche retenue est pluridisciplinaire. Des connaissances en géographie de la santé et en épidémiologie spatiale sont nécessaires pour répondre à la problématique de santé publique posée. Mais la contribution des sciences statistiques, littéralement *sciences des états* - ou encore *physique sociale* (Quételet, 1969) – est essentielle pour analyser conjointement la complexité de la géométrie géographique des FE/FIM* *pertinents** et des PM* *d'intérêt*. Actuellement, les méthodes d'analyse multidimensionnelle non paramétriques offrent en ce sens des *perspectives prometteuses* (Cartier, Villani et al., 2012).

Dans le cadre de cette recherche l'identification des DES* géographiques s'opèrera grâce des *machines learning* et à une stratégie de *sélection de variables* robuste, innovante et adaptée aux jeux de données de *grande dimension* (Ben Ishak et Ghattas, 2005), dont celui constitué par les i.st.e* voués à la modélisation des FE/FIM* et des i.st.m* : $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$ proposés pour celle des PM* fait partie. Les $z^{REG.j}_{(U_k)}$, par contre, n'interviendront pas directement dans l'identification des DES* puisqu'ils sont voués uniquement à l'interprétation géographique de ces derniers.

La modélisation géographique des PM* est maculée d'incertitudes spatiotemporelles protéiformes qui rendent l'identification des DES* complexe. En effet, les i.st.m* proposés à cet effet sont sous le joug d'un système de pondérations spatiotemporelles fondé sur des *connaissances expertes* et une *transformation temporelle*.

Ici, c'est la significativité des i.st.m* proposés qui se pose, i.e. leur capacité à modéliser la réalité géographique des PM* étudiés. Et pour cause, *il est courant que les indicateurs spatiaux ne modélisent pas du tout la géographie des phénomènes pour lesquels ils ont été proposés*, mais qu'ils modélisent insidieusement quelque chose de radicalement différent (Brook, Lohr et al., 1984).

Par conséquent, dans la pratique l'identification des DES* à partir de la géographie des PM* d'intérêt doit inéluctablement être validée, ou au moins corroborée, par une approche *individus centrée*, i.e. une analyse menée directement sur les variables LEA représentatives des séquelles développées par les patients. Cet angle d'attaque permettra de donner plus de consistance aux résultats obtenus et par la même occasion de valider la robustesse des i.st.m* proposés.

Cependant, le problème de la significativité des indicateurs spatiaux concerne aussi la modélisation géographique des FE/FIM* *pertinents**. Pour pallier cet écueil, les BD environnementales et les méthodes les plus fiables ont été déclinées dans l'état de l'art, et par suite, le concept de FE/FIM* *Curieux* de test* a été introduit.

Enfin, il convient de rappeler qu'il s'agit d'une thèse en Géographie et que quelle que soit la robustesse des propositions méthodologiques et des données sources utilisées, *l'exposition aux risques dans l'espace et le temps [ne peut être] exprimée que sous la forme d'une relation possible entre le milieu et l'homme* (Bailly et Beguin, 2005). Donc, par extension, que les liens de *cause à effet* identifiés ne pourront pas atteindre le niveau de rigueur *des relations causales épidémiologiques* (Hill, 1965).
Le positionnement scientifique, les objectifs, les hypothèses fondamentales, les méthodes de modélisation et de sélection les plus robustes à l'heure actuelle ainsi que les FE/FIM* pertinents* et les BD accessibles les plus fiables permettant d'en modéliser la géographie ont été déclinés.

La géographie des PM* a été modélisée par des i.st.m* robustes - peu ou prou – qui ont permis de conjecturer un *effet environnement évident*. Il convient désormais de proposer des i.st.e* adaptés à la modélisation géographique des FE/FIM* *pertinents** et de tenter d'identifier des DES, ce qui est l'objet de la Partie II.

PARTIE.II : MODELISATIONS GEOGRAPHIQUES ENVIRONNEMENTALES ET IDENTIFICATION DES DETERMINANTS DE SANTE

Le chapitre 1 a permis de dresser un état des connaissances en santé environnementale*, de déterminer des outils de modélisation géographique robustes et de passer en revue les bases de données disponibles utilisées afin de proposer des méthodes de modélisation et de sélection des FE* géographiques qui *a priori* déterminent l'état de santé des populations. Le chapitre 2 propose des méthodes dédiées à la modélisation géographique des Phénomènes Morbides* (PM) d'intérêt par le biais d'indicateurs spatiotemporels morbides* (i.st.m) quantitatifs $z'^{j}_{(U_k),c}$ et qualitatifs $z'^{j}_{(U_k),q}$. Une stratégie de fusion de l'information contenue dans ces indicateurs spatiaux a permis d'en construire un troisième : $z^{REG.j}_{(U_k)}$, utile à l'analyse spatiale de la disparité des Risques d'Exposition Géographique (REG) morbides. Ce dernier a montré avec une grande acuité que certaines Unités Géographiques : U_k pouvaient être caractérisées par un même REG, et que des blocs d'U_k pouvaient être formés par contigüité spatiale. La mise en évidence de *ces clusters* - ou *agrégats spatiotemporels morbides* (Tillaut, 2005) laisse présumer que la géographie des PM* d'intérêt (séquelles) n'est pas aléatoire et qu'il existe un effet *environnement évident, au moins contributif*.

L'objectif de la partie II est, dans un premier temps (chapitre 3), de modéliser la géographie des FE/FIM, par des indicateurs spatiotemporels environnementaux* (i.st.e) *consistants* notés $x^{l}_{(U_k)}$, en combinant, par des stratégies d'intégration, des outils et des méthodes de modélisation appropriées à la prise en compte de toute l'information contenue dans les données mobilisées. L'idée est d'intégrer *horizontalement* l'espace et *verticalement* la temporalité au cœur du processus d'estimation des i.st.e* (Peguy, 1996). Puis, dans un second temps (chapitre 4), il s'agira de proposer un instrument et une méthode d'analyse statistique multivariée suffisamment puissante – quitte à en adapter les contours – pour caractériser de façon fiable les interactions spatiales multidimensionnelles entre les i.st.e* et les i.st.m* constituant les jeux de données géographiques confectionnés. Le but étant d'identifier parmi tous les FE/FIM* intégrés ceux qui conditionnent les états de santé étudiés.

En épidémiologie spatiale comme en géographie de la santé les paradigmes convergent désormais et il convient de raisonner en termes d'expositions environnementales combinées (Leux et Guénel, 2010). Par conséquent la définition de l'environnement qui a été retenue est celle de l'OMS – soit tous les facteurs possibles susceptibles de caractériser l'environnement géographique des milieux de vie et la population cible à l'exception des facteurs génétiques (World Health Organization, 2009).
L'espace géographique* environnemental considéré est multidimensionnel et les facteurs qui le caractérisent sont discrétisés en quatre composantes : Les Facteurs Individuels et Médicaux* (FIM), et les Facteurs Environnementaux* (FE) à connotation : Sanitaire (FE-SAN) ; Socio-économique (FE-SOCIO.ECO) ; Et physicochimique (FE-PHY.CIM).

Les FE-SAN* et les FE-SOCIO.ECO* caractérisent des expositions à des situations contextuelles à risque. Les FE-PHY.CHIM* décrivent la présence de substances toxiques dans les *milieux environnementaux*, i.e. l'eau, l'air, le sol et les matrices biologiques, ou dans *les milieux d'exposition*, i.e. les éléments des milieux environnementaux susceptibles de rentrer en contact avec les organismes humains (Caudeville J., Boudet C. et al., 2012).

Enfin, les FIM* représentent la géographie des Caractéristiques Individuelles et Médicales*(CIM) des patients qui peuvent influencer la morbidité. Les CIM* sont accessibles par le biais de données épidémiologiques et ont un impact évident sur l'état de santé (chapitre.1).
Il est nécessaire de les intégrer mais ils ne suffisent pas à expliquer totalement l'état de santé des populations. En effet, les études menées sur les séquelles suggèrent un effet environnement contributif probant (Michel, Auquier et al., 2007). Certains facteurs génétiques sont indubitablement des déterminants de santé (Stein, 2012). Ils ne sont pourtant pas intégrés bien qu'ils soient particulièrement *pertinents**. (Salem, 1995). Cela constitue un biais dans l'analyse des interactions santé-environnement (Abramson J.H., Abramson Z.H., 1988). Mais il s'agit d'une position contrainte et non pas choisie puisque la BD-LEA n'intègre pas encore la dimension génétique des patients (chapitre.1).

L'accessibilité aux données est la première limite à la modélisation des FE/FIM. Elle est, en outre, conditionnée par la qualité de la granularité* (nature, échelle, temporalité, précisions, lacunes…) des informations disponibles dans les BD existantes (Zeitouni, 2006).

Les FE/FIM pertinents** représentent tous *des expositions environnementales*, *internes* ou *externes*, *avérées* ou simplement *suspectées*, à des *substances toxiques qui ont des effets déterministes ou stochastiques*. Ou encore, les expositions *à des situations conjoncturelles à risque* induisant des *prédispositions morbides collectives et individuelles* (Haddad, 1992) - décrites dans la l'état de l'art. Dans la mesure où il n'existe pas de littérature sur les séquelles étudiées, *les facteurs pertinents** sont ceux documentés pour avoir des effets sur les *sujets sains*, qui sont supposés délétères, *a fortiori*, sur ceux prédisposés (chapitre.1).
Les FE/FIM* pertinents* ont été discrétisés selon deux types d'expositions. Les *expositions environnementales géographiques potentielles* qui sont modélisées à partir de mesures effectuées dans les milieux environnementaux ou d'indicateurs communautaires et qui constituent la majorité des i.st.e* proposés. Et les *expositions environnementales géographiques intrinsèques* qui sont modélisées partir de variables CIM* ou doses d'exposition à des substances délétères présentes dans les *milieux de contact*. Cette différence est fixée par la granularité* des inputs (chapitre.1).

Nonobstant, rien ne permet de conjecturer que les i.st.e* modélisant des *expositions intrinsèques* sont plus robustes que ceux qui caractérisent *des expositions potentielles*. Et pour cause, les indicateurs spatiaux des risques de contamination humaine sont associés à des incertitudes multiformes qui les rendent rarement *significatifs* (Mercat-Rommens, Chojnacki et al., 2008). Il est impossible de reconstituer de façon fiable les *parcours de vie* des populations (Couet, 2006) ; D'estimer les sensibilités génétiques de chacun (Sobol, 2004) ; De prendre en compte les variabilités spatiotemporelles de ces facteurs (Zaidi, Bhatnargar, et al., 2000). A tel point que l'idéal pour caractériser les interactions santé-environnement est souvent de raisonner de façon plus globale et à des échelles agrégées de façon à réduire *l'effet exacerbé des bruits de fond environnementaux à micro-échelle* (Pistocchi A. Sarigiannis DA., Vizcanio P., 2010).

L'échelle des communes s'est imposée comme la plus adéquate pour modéliser la géographie des PM* et celle des FE/FIM* *pertinents**. Ce choix repose sur un ensemble de considérations théoriques et pratiques : La Géographie est une Science Humaine et Sociale qui doit avoir des répercussions sur la *vie sociale,* et l'étude des interactions complexes entre *l'homme et son milieu doit être abordée* de façon *expérimentale* et à un niveau *macroscopique* (Merleau-Ponty, 1945) .

Aussi, la géographie de la santé a pour dessein de caractériser les *états de santé* par des indicateurs spatiaux qui justifient *socialement* les mesures politiques prises en santé environnementale*. Et, ces derniers *ne trouvent leur pleine intelligibilité qu'à l'échelle des territoires* (Salem, 1995) . Par ailleurs, il convient de considérer la granularité* de l'ensemble des données utilisées de sorte à *minimiser le*

*concept du biais conditionnel** (Marcotte, 2008) ; en outre, l'échelle doit permettre de capter *la variabilité géographique des* expositions *potentielles* ou *intrinsèques* (Chaix, Merlo et al., 2005).

Cette thèse est fondée sur l'hypothèse que si les i.st.m* et i.st.e* sont suffisamment fiables pour modéliser la géographie des PM* et des FE/FIM* alors il est possible d'identifier des DES* (chapitre.1). Mais, il est courant que les indicateurs spatiaux modélisent insidieusement des phénomènes radicalement différents de ceux qu'ils sont sensés caractériser (Brook, Lohr et al., 1984). À cet écueil s'ajoute le fait que *la procédure de sélection des variables* proposée – VSURF (Genuer, 2010)- n'est pas prouvée de façon théorique et qu'en dépit de sa puissance pratique incontestable, rien ne garantit que les FE/FIM* géographiques qui expliquent statistiquement les PM* soient nécessairement les vrais Déterminants Environnementaux de Santé* (DES). Ces constats conduisent à proposer :

*Le concept de Distance a-spatiale morbide**, est définie comme la plausibilité théorique des effets avérés ou suspectés sur l'état de santé des populations de FE* représentatifs d'expositions environnementales à des situations à risque ou à des substances physicochimiques particulières (chapitre.1).

Les FE/FIM Curieux* de test*, qui représentent *des expositions géographiques à des faits de santé* dont la Distance a-spatiale morbide* est éloignée des *états de santé* étudiés ; soit parce que l'état des connaissances est controversé à ce sujet ; soit parce qu'il s'agit de FE/FIM* correctement documentés mais dont la granularité* des données mobilisées est médiocre et induit une *distorsion spatiale ou temporelle dans le processus* d'estimation des i.st.e.

Dans chaque composante environnementale des *FE/FIM* Curieux** seront introduits afin d'estimer la capacité de VSURF à détecter et à disjoindre *les variables de bruit des variables explicatives,* et en même temps de comparer les niveaux d'importance conférés aux i.st.e* représentatifs des Déterminants Environnementaux de Santé* (DES), et des Facteurs Environnementaux à Risques : Contributifs (FREC) ou Potentiellement Aggravant (FREPA). Les FE/FIM* curieux* n'ont *a priori* aucun effet sur les PM* étudiés. Et dans le cas contraire, il convient de s'interroger sur leur contrepartie dans la réalité géographique, i.e. sur les phénomènes géographiques que les i.st.e* proposés, pour modéliser les FE/FIM* curieux*, représentent réellement.

Présentation du chapitre 3 :

L'objectif est de proposer, à partir des variables mobilisées, pour tous les FE/FIM* *pertinents* ou Curieux** intégrables, des stratégies de modélisation spatiotemporelle robustes, des méthodes et des outils adaptés à leur mise en œuvre – au regard des considérations théoriques déclinées dans l'état de l'art.

La géographie des FE/FIM* *pertinents** ou *Curieux** intégrables est modélisée par le biais d'i.st.e, notés $x^{l}_{(U_k)}$ dans les communes U_k où des patients sont spatialisés. Ces i.st.e* se fondent sur des méthodes mathématiques et des outils de traitement permettant de mettre en œuvre les stratégies d'intégration de l'espace et de la temporalité au cœur du processus de modélisation géographique.

La géographie de la santé *considère l'espace comme une entrée dans la compréhension des processus entre la santé et l'environnement*. Les indicateurs spatiaux utilisés décrivent *les sociétés* et permettent de montrer, avec une grande acuité, que *des populations confrontées à des milieux environnementaux analogues* sont exposées à des risques morbides différents *du fait de leur gestion particulière des espaces* (Salem, Rican S et al., 2006).

La géographie de la santé est *mal-aimée* de l'épidémiologie spatiale et de la santé publique car considérée à tort comme une *science de l'inventaire* qui utilise des méthodes archaïques discutables d'analyse et de représentation (Salem, 1995). Mais ces considérations sont dépassées. En dépit des quelques indigences méthodologiques suggérées dans l'état de l'art, liées à l'impossibilité pour les manuels d'enseignement de suivre l'évolution incessante des outils implémentés dans les Systèmes d'Informations Géographiques et à la simplicité des modélisations proposées, les indicateurs géographiques sont généralement assez représentatifs de la réalité (Lahousse et Piédanna, 1998).

Les Systèmes d'Informations Géographiques (SIG) sont des logiciels adaptés à la modélisation géographique. Les SIG actuels sont très puissants et permettent d'intégrer et de représenter dans l'espace et le temps des phénomènes protéiformes à partir de données géographiques multidimensionnelles, incomplètes, imprécises ou incertaines, de nature qualitative ou quantitative. Le SIG utilisé dans le cadre de cette thèse est ArcGis.10 et les outils de traitement numérique implémentés suffisent à supputer la quasi intégralité des i.st.e* $x^{l}_{(U_k)}$ inhérents aux stratégies d'intégration proposées (ESRI, 2013).
Quelques méthodes de modélisation probabilistes et géostatistiques qui sont encore l'apanage d'instruments mathématiques particuliers, ont soit requis l'utilisation du logiciel R (Institute for Statistics and Mathematics, 1997), soit été programmées en Visual Basic pour Application (Microsoft, 2013).
Les stratégies d'intégration proposées sont destinées à intégrer des données mobilisées par des processus d'harmonisation : *verticaux,* i.e. temporels ; Et *horizontaux*, i.e. spatial, conditionnellement à leurs granularité*, et enfin adaptées au *support* - i.e. à l'échelle d'investigation retenue, celle des communes : U_k. - (Peguy, 1996).
Il s'agit de proposer des transformations mathématiques fondées sur le concept de *minimisation* du *biais conditionnel** (Marcotte, 2008). En d'autres termes les stratégies d'intégration doivent d'une part , maximiser *l'effet support**, i.e les caractéristiques spatiales et temporelles des données d'échantillonnage, et d'autre part, optimiser l'*effet information**, i.e. pallier les incertitudes induites par d'éventuelles lacunes, harmoniser leurs échelles, et fusionner des informations géographiques de nature et de précision diverses afin de les adapter aux phénomènes environnementaux à modéliser (Baillargeon, 2005). Les stratégies d'intégration proposées suggèrent des processus numériques d'estimation de sorte que les $x^{l}_{(U_k)}$ représentent *au mieux* la réalité géographique des FE/FIM.

Les stratégies d'agrégation temporelle des données spatiales utilisées en géographie sont simples, l'estimateur statistique est choisi subjectivement. Néanmoins il représente généralement assez bien la réalité géographique (Pumain et Saint-Julien, 1997). Cependant, il est préférable que ce choix prenne en compte les caractéristiques statistiques des *séries temporelles* lorsque les données spatiales sont *disponibles à différentes dates* (Lütkepohl, 1991). De fait, les stratégies d'intégration *verticales* proposées déterminent l'estimateur statistique conditionnellement à *une analyse probabiliste qui s'appuie sur le concept de stabilité temporelle apparente* (Hamilton, 1994).

Les stratégies d'intégration horizontales utilisées *en analyse spatiale* sont basées sur des *opérateurs* ensemblistes d'agrégation ou de désagrégation et la fusion d'informations géographiques connexes *socio-économiques, démographiques ou administratives* (Pumain et Saint-Julien, 1997), qui peuvent facilement être couplées à *des modèles statistiques paramétriques* (Charre, 1995). Les techniques d'analyse spatiale classiques sont particulièrement adaptées à la modélisation des phénomènes géographiques. La quasi-totalité d'entre elles sont disponibles dans l'extension d'ArcGis : *Spatial Analyst* (ESRI, 2013).
Un intérêt particulier est porté aux instruments SIG récemment implémentés ou améliorés, en l'occurrence, aux possibilités offertes par le module *ModelBuilder* (ESRI, 2013) qui permet de procéder à des calibrations itératives afin d'améliorer la précision spatiale des i.st.e* construits par capture de données raster, c'est le cas par exemple de la variabilité *des expositions géographiques potentielles* à des substances physicochimiques délétères combinées et estimées à partir de l'occupation biophysique des sol (données CLC). Ce module permet également d'effectuer des modélisations spatiales dynamiques pour estimer les expositions géographiques potentielles à des radionucléides artificiels diffusés dans les milieux environnementaux – à partir de la proximité des INB* (données ASN). Une autre extension, particulièrement appropriée pour la modélisation géographique, est *Geostatistical Analyst* (ESRI, 2013). Elle permet de construire *des couches de surfaces continues* à partir des données géo-localisées. Elle est utilisée pour la modélisation de la variabilité des expositions *à la radioactivité environnementale* (données RNM) ou celle *des expositions potentielles aux paramètres météorologiques* (données Météo-France). Parmi *les nombreuses méthodes de reconstitution spatiale* implémentées, une revue des forces et des faiblesses de chacune permettra de légitimer l'intérêt porté aux *géostatistiques uni-variables* (Matheron, 1965) et *multi-variables* (Wackernagel, 2003), et plus particulièrement aux techniques de

krigeage qui permettent d'évaluer en tous points de l'espace les valeurs de *variables régionalisées** (Marcotte, 2008).

Présentation du chapitre4 :
L'ultime étape de cette recherche a pour dessein de caractériser les interactions statistiques entre les i.st.m* et les i.st.e* proposés afin d'identifier les Déterminants Environnementaux de Santé* (DES) géographiques. La géographie de la santé, à l'instar de l'épidémiologie spatiale, souffre d'une sorte de retard méthodologique sur ce point, qui touche à des problématiques de recherches contemporaines dans le domaine des statistiques et des probabilités.
En effet, en géographie de la santé, les méthodes multidimensionnelles d'analyse statistique paramétriques sont encore largement utilisées (Rémy, Handschumacher et al., 2011) - en dépit de leur faiblesse à caractériser la complexité des interactions statistiques des variables contenues dans des jeux de données géographiques. Ce constat peut être étendu à la géographie en général, (Groupe CHADULE, 1997) en particulier à sa variation quantitative : l'analyse spatiale (Pumain et Saint-Julien, 1997). Les modèles paramétriques ne sont pas assez puissants au sens statistique du terme pour caractériser de façon fiable les interactions multidimensionnelles constituées de variables de nature et de précision diverses. Pourtant, les modèles statistiques sont *a priori* les instruments les plus adaptés *pour l'analyse, la synthèse et l'exploration de la complexité spatiale* (Charre, 1995). Les phénomènes spatiotemporels sont complexes car ils découlent, entre autres, de processus stochastiques, i.e. qui se produisent dans un milieu déterminé *dont la nature et les mécanismes sont connus* mais dont l'issue ne peut être conjecturée par la seule observation *des états antérieurs de ce milieu* (Peguy, 1996). Les outils statistiques intègrent l'idée du hasard en se fondant sur des théories probabilistes et *le calcul des probabilités est généralement défini comme étant l'étude des lois du hasard* (Borel, 1967). Ils sont donc bien adaptés. D'ailleurs, le datamining* *couplé à des méthodes d'ensemble* fondées sur des processus stochastiques d'apprentissage et à des modèles non paramétriques offre désormais la puissance nécessaire pour expliquer – à sa façon – un grand nombre de phénomènes jusqu'alors sans solution (Cartier, Villani et al., 2012). Les méthodes d'analyse multidimensionnelle contemporaines permettent de résoudre des problèmes de *sélection de variables dans des jeux en grandes dimensions* (Tuleau-Malot, 2005). Les SIG ne sont pas en mesure d'appréhender ces jeux multidimensionnels protéiformes particuliers qui représentent aujourd'hui la majorité des jeux de données géographiques. Le champ *des théories de l'apprentissage* par des *machines learning* offre des *perspectives explicatives et prédictives* indubitables, toute la difficulté étant de choisir l'algorithme et la méthode de sélection adaptés à la problématique et aux sources d'informations disponibles (Han et Kamber, 2006). Ces procédures de sélection restent cependant l'apanage de logiciels statistiques particuliers et ne sont pas, ou seulement partiellement, programmées dans les *packages* proposés (Institute for Statistics and Mathematics, 1997).

Dans le chapitre 4, il sera question de spécifier l'algorithme le plus adapté au regard des connaissances disponibles et des caractéristiques du jeu de données contenant les i.st.m* et les i.st.e* proposés pour modéliser la géographie des PM* d'intérêt et des FE-FIM* pertinents* et Curieux* (Ghattas et Ben Ishak, 2008). L'algorithme retenu est *randomForest** - il permet, entre autres, de caractériser, par le biais de scores, l'importance des variables, i.e. des i.st.m* et des i.st.e* (Breiman, 2001). Cependant, l'importance des variables ne présume pas de la manière dont elles interagissent entre elles, i.e. de dissocier les i.st.e* qui représentent des bruits de fond environnementaux de ceux qui conditionnent la variabilité des i.st .m. Une stratégie de sélection par seuillage – nommée VSURF - a récemment été proposée. Elle permet justement de disjoindre les variables de bruit des variables explicatives (Genuer, 2010).

La méthode VSURF sera appliquée à cette fin. Il conviendra cependant d'en modifier légèrement les contours – pour l'adapter à la dialectique géographique, afin d'une part, de caractériser les interactions spatiales entre les i.st.m, modélisant les PM* d'intérêt, et les i.st.e* proposées, et d'autre part, dans le chapitre suivant, de modéliser la géographie des FE/FIM* intégrés.

CHAPITRE 3 : MODELISATIONS GEOGRAPHIQUES ENVIRONNEMENTALES

Ce chapitre propose des stratégies spatiotemporelles d'intégration et des instruments statistiques permettant de modéliser, de façon précise et fiable, la variabilité géographique des FE/FIM* *pertinents** ou *Curieux*.* Les propositions heuristiques permettent de supputer des i.st.e* robustes par une minimisation du concept de *biais conditionnel** (Marcotte, 2008)des sources d'informations disponibles - de sorte qu'ils soient adaptés à la finalité recherchée, i.e. l'identification les DES, des FREC et des FREPA.
Les stratégies d'estimation des i.st.e* sont déclinées successivement par composante environnementale : Individuelle et Médicale, Sanitaire, Socio-économique et physicochimique, après un bref rappel des éléments théoriques et des données disponibles qui leur confèrent un caractère intégrable.
L'analyse des résultats de la modélisation des FE/FIM* *intégrés* sera concise puisqu'elle ne permet pas d'identifier des DES. Le seul moyen pour parvenir à démêler l'écheveau de cette question complexe est d'utiliser un instrument de datamining* et une procédure de sélection permettant de mettre simultanément en concurrence les i.st.m* et tous les i.st.e.

SECTION A) FACTEURS INDIVIDUELS ET MEDICAUX

Les Facteurs Individuels et Médicaux* (FIM) modélisent la géographie des Caractéristiques Individuelles et Médicales*(CIM) de la population d'intérêt, i.e. les patients spatialisés de la Cohorte LEA. Les CIM* sont des informations à caractère individuel, comportemental ou relatives à l'historique médical.
En géographie de la santé les CIM* sont parfois négligées. Or le positionnement scientifique du géographe consiste à prendre en compte tous *les facteurs ayant au moins une influence suspectée* et à raisonner en terme *d'expositions combinées* (Leux et Guénel, 2010) L'omission des CIM* revient à négliger les expositions environnementales des patients, i.e. liées au phénotype (âge, sexe), au style de vie (pratique d'activité physique) et à l'historique médical (type de LEUC, agressivité de traitement) de l'échantillon de la population. Ne pas considérer les FIM* biaise l'analyse géographique des interactions santé-environnement, *a fortiori* lorsque ces informations sont accessibles (Abramson J.H., Abramson Z.H., 1988).

Les CIM* sont des variables de la BD-LEA, notées $x_i^{l:CIM}$: la Caractéristique Individuelle ou Médicale « l » associée au patient « i ». Les études menées sur LEA, dans une logique *individus-centrée**, ont montré que certaines CIM* conditionnent *l'état de santé* des patients. Mais elles ne suffisent pas à les expliquer - et l'environnement joue probablement un rôle important (Michel, Auquier et al., 2007).
Le Professeur Michel et le Professeur Auquier, à partir d'un compromis entre plausibilité des interactions avec les séquelles et fiabilité de la granularité* des données LEA, ont préconisé de modéliser la géographie de certaines CIM. Les FIM* jugés *pertinents** sont :
Le genre : $x_{\mathrm{i}}^{\mathrm{SEXE}} = \{\mathrm{GARÇON, FILLE}\}$; Le type de LEUC traitée $\mathrm{x_i^{TYPLEUC}} = \{\mathrm{LAL} *; \mathrm{LAM}\}$; L'âge au moment du diagnostic de la LEUC $x_{\mathrm{i}}^{\mathrm{AGE\_DIAG}} \in \{\mathbb{N}|\mathrm{unité: année}\}$; La durée du suivi : $x_{\mathrm{i}}^{\mathrm{DSUIVI}} \in \{\mathbb{N}|\mathrm{unité: année}\}$; Le type de traitement reçu $x_i^{\mathrm{PROTOC}} = \{11.\,\mathrm{protocoles.\,possibles}\}$; Les rechutes $x_{\mathrm{i}}^{\mathrm{RECHUT}} = \{\mathrm{OUI; NON}\}$; Le recours à une greffe de moelle : $x_{\mathrm{i}}^{\mathrm{GREF}} = \{\mathrm{OUI; NON}\}$; L'(es) irradiation(s) corporelle(s) totale(s) : $x_{\mathrm{i}}^{\mathrm{IRCAT}} = \{\mathrm{OUI; NON}\}$.
Le Professeur Auquier a notifié *a posteriori* une ambiguïté quant à l'interprétation médicale de $\mathrm{x_i^{DSUIVI}}$ qui pouvait être connotée à la fois positivement, comme relevant de l'accessibilité aux soins et négativement, comme un effet lié à la durée des expositions environnementales délétères.

Quant au FIM* Curieux* introduit, il s'agit de l'intensité de l'activité physique, notée x_i^{ACPHY}. La pratique d'une activité sportive a des effets positifs multiples documentés sur l'état de santé (Ghanbari-Niaki, Saghebjoo, et al., 2009). Cependant cette variable LEA est entachée de lacunes. De plus, son caractère comportemental crée une *distorsion morbide*, elle est plus éloignée des patients que celle touchant au génotype ou à l'historique médical et par conséquent, remplit à merveille son rôle de FIM* *Curieux**.

La géographie des FIM* représente des *expositions environnementales intrinsèques* dans la mesure où la Distance a-spatiale morbide* avec les patients est beaucoup plus courte que celle des autres FE* : SAN, SOCIO.ECO ou PHY.CHIM. En effet, ces derniers sont modélisés à partir de mesures environnementales ou d'indicateurs communautaires caractérisant les milieux de vie – i.e. de variables plus incertaines, imprécises et empreintes de bruits de fond environnementaux – ce qui crée une distorsion morbide sur l'effet des expositions environnementales géographiques modélisées. Il convient de prendre en compte cette spécificité dans le processus d'estimation des i.st.e* des $x_{(\mathrm{U_k})}^{\mathrm{l:FIM}}$.

PROPOSITIONS HEURISTIQUES ET STRATEGIE D'INTEGRATION DES CARACTERISTIQUES INDIVIDUELLES ET MEDICALES

Objectif :
Modéliser dans les communes de 1ère espèce la géographie des FIM* par le biais d'i.st.e* robustes et fiables. Ces i.st.e* sont notés $x'^{\mathrm{l:FIM}}_{(U_k)}$. Il doivent être adaptés à la fois à : La granularité* des variables LEA $x_i^{l:CIM}$, i.e. prendre en compte leur qualité spatiotemporelle intrinsèque lors du processus d'agrégation ensembliste des $x_i^{l:CIM}$ dans les $\mathrm{U_k}$, et à l'identification des DES, i.e. induire une distorsion a-spatiale morbide afin que les $x_i^{l:CIM}$ soient plus en harmonie avec les variables environnementales utilisées pour caractériser les FE* : SAN, SOCIO.ECO et PHY.CHIM.

Remarques liminaires :
Les $x_{\mathrm{i}}^{\mathrm{CIM}}$ peuvent être de nature quantitative discrète : $x_{\mathrm{i}}^{\mathrm{AGE\_DIAG}}$; $\mathrm{x_i^{DSUIVI}}$, qualitative booléenne : $x_{\mathrm{i}}^{\mathrm{SEXE}}$; $x_{\mathrm{i}}^{\mathrm{GREF}}$, ou qualitative multi-classes : $x_{\mathrm{i}}^{\mathrm{ACPHY}}$; $\mathrm{x_i^{PROTOC}}$…
Les $x_{\mathrm{i}}^{\mathrm{CIM}}$ ont des qualités spatiotemporelles à connotation géographique, épidémiologique et statistique qui varient pour chaque patient (Abramson J.H., Abramson Z.H., 1988).

Proposition principale :
A l'instar de la géographie des PM* modélisée par les i.st.m* $z'^{\mathrm{j}}_{(U_k),q}$ et $z'^{\mathrm{j}}_{(U_k),c}$ via des stratégies de *fusion d'informations expertes*, dans l'idée de la théorie *des ensembles flous,* à partir des variables séquelles $y_{\mathrm{i}}^{\mathrm{j}}$, la stratégie d'intégration des FIM* se base sur une agrégation ensembliste spatiotemporelle pondérée EpiGéoStat. Elle est caractérisée par les i.st.e* $x'^{\mathrm{l:FIM}}_{(U_k)}$ auxquels il s'agit de conférer une consistance *horizontale* (i.e. spatiale) et *verticale* (i.e. temporelle) adaptée (Peguy, 1996).

Hypothèse principale :
La théorie *des ensembles flous* permet d'améliorer la qualité du processus de fusion de données de nature, de précision, d'incertitude, d'échelle spatiale et temporelle différentes. L'idée consiste à injecter de la *connaissance experte* par le biais de *fonctions mathématiques*. La théorie *des ensembles flous* est applicable à toutes les problématiques de fusion si tant est qu'elle soit spécifiée pour *le domaine scientifique*, *la finalité*, *les connaissances expertes et les données auxiliaires intégrables* susceptibles d'améliorer la qualité du processus d'agrégation ensembliste (Dubois Didier, Prade Henri, 2004).

Spécification de la proposition :
Afin de parvenir à une représentation plus robuste et plus juste de la réalité géographique des CIM, par le biais d'i.st.e, une *métrique floue géographique* – constituée de facteurs de certitude spatiotemporelle EpiGéoStat : υ_i^l, est proposée. La *métrique floue géographique* se compose de trois facteurs thématiques

de certitude spatiotemporelle, à connotation : géographique : $\upsilon_i^{l,geo}$, épidémiologique : $\upsilon_i^{l,epid}$, et statistique $\upsilon_i^{l,stat}$.

Remarques auxiliaires :
La métrique floue géographique utilisée pour les FIM* se compose *de facteurs de certitude EpiGéoStat :* υ_i^l., alors que celle proposée pour les PM* est constituée de facteurs d'incertitude : π_i^j.
La métrique floue des υ_i^l proposée pour les FIM, à l'instar de celle des π_i^j pour les PM, ne peut pas être être calibrée puisque la réalité géographique des phénomènes n'est pas connue *a priori.* La stratégie d'estimation est fondée sur *des connaissances expertes.*
L'injection d'informations *expertes* se fait par le biais de fonctions mathématiques *adaptées* à la *granularité* des* $x_i^{l:CIM}$ et à la dialectique de modélisation géographique.
Les υ_i^l ne doivent pas altérer la nature ni avoir un effet exagéré sur les $x_i^{l:CIM}$ de façon à ne pas induire des incertitudes plus fortes que le gain de précision spatiotemporel apporté.
A l'aune de la modélisation géographique des PM* par les i.st.m* : $z'^j_{(U_k),.}$ il n'y a pas lieu d'intégrer le *temps d'exposition à l'environnement* tee_i^j (Bernard et Lapointe, 2003). Et pour cause, les variables $x_i^{l:CIM}$, qu'il s'agisse du sexe, de la LEUC traitée, du protocole de traitement reçu, ou encore de l'âge au diagnostic, sont parfaitement *stables dans le temps* donc indépendantes du tee_i^j.

a. Seule exception x_i^{APHY} qui peut varier dans le temps. L'incertitude temporelle quant à la pratique et à l'intensité des activités physiques est prise en compte dans le processus d'estimation des $\upsilon_i^{l,épi}$.
b. Autre exception, qui n'en est pas une. x_i^{RECHUT} n'est pas stable et peut varier au fil du temps et des expositions environnementales donc de : tee_i^j. Cependant, x_i^{RECHUT} n'est pas considérée pour son caractère morbide déclaré ou latent. Elle est intégrée en tant que variable potentiellement explicative pour l'effet combiné des rechutes et des traitements - qui ont un impact avéré sur le risque de séquelle. Le traitement est prescrit après le diagnostic de la rechute. Donc, il est nécessairement renseigné dans LEA, alors cette information est à la fois stable et sûre.

Les variables x_i^{PROTOC} et x_i^{APHY} contiennent des lacunes, respectivement 2 et 270 sur les 943 patients de la Base LEA 2009 consolidée.

La stratégie d'estimation des υ_i^l est à la fois analogue à celle des π_i^j et plus simple. La synoptique d'estimation est donc décrite dans ses grandes lignes de sorte à garantir son intelligibilité et sa reproductibilité.

PRINCIPE DE LA STRATEGIE D'INTEGRATION SPATIOTEMPORELLE DES DONNEES LEA

Certaines CIM* sont entachées des lacunes. Or pour obtenir des i.st.e* robustes il convient, entre autres, d'optimiser *l'effet information** des sources (Baillargeon, 2005). Cela commence, en palliant cette défaillance avec une règle statistique adaptée (Saporta, 2006).

COMBLEMENT DES LACUNES

Remarque liminaire :
La stratégie *bouche-trou* proposée doit, dans la mesure du possible, éviter de réduire les effectifs spatialisés. Cette phase est nécessaire et préalable à la modélisation géographique des CIM.

Objectif :
Eviter de perdre en puissance statistique en supputant des i.st.e* contenant les lacunes ou en diminuant les effectifs spatialisés.

Proposition méthodologique :
La stratégie *de comblement* des lacunes doit être adaptée aux variables qualitatives ou quantitatives par le biais d'un estimateur statistique adapté et *consistant* (Liaw et Wiener, 2006).
Principe d'estimation de la stratégie de comblement :
Les individus sont éliminés lorsque la proportion de lacunes, estimée sur la population spatialisée, est supérieure à 30%. Sinon les lacunes sont comblées par le mode ou par la médiane selon la nature de la variable, i.e. que :

$$\left\{x_{\mathrm{i}}^{\mathrm{l}}=\text{lacune}\right\}=\begin{cases}\mathrm{mêd}\left(x_{\{\mathrm{i}|\mathrm{abs.lacune}\}}^{\mathrm{l}}\right) & \text{lorsque:} \quad \left\{\mathrm{N}\left(\mathrm{x}_{\{\mathrm{i}|\mathrm{abs.lacune}\}}^{\mathrm{l}}\right)\leq 30\%\right\}\cap\left\{x_{\cdot}^{\mathrm{l}}\in\mathbb{R}^{\mathrm{n}}\right\}\\ \mathrm{môd}\left(x_{\{\mathrm{i}|\mathrm{abs.lacune}\}}^{\mathrm{l}}\right) & \text{lorsque:} \quad \left\{\mathrm{N}\left(\mathrm{x}_{\{\mathrm{i}|\mathrm{abs.lacune}\}}^{\mathrm{l}}\right)\leq 30\%\right\}\cap\left\{x_{\cdot}^{\mathrm{l}}\in\left\{\mathcal{C}^{\mathrm{j}}\in\mathbb{N}^{\mathrm{c_j}}\right\}\right\}\\ \phi\ ;\ \mathrm{I}_{\left\{\mathrm{i}\middle|\mathrm{x_i^l}\neq\mathrm{lacune}\right\}}^{\mathrm{l}} & \text{lorsque :} \quad \left\{\mathrm{card}\left(\mathrm{x}_{\{\mathrm{i}|\mathrm{abs.lacune}\}}^{\mathrm{l}}\right)/\mathrm{n}>30\%\right\}\end{cases}$$

Avec : $\mathrm{N}\left(x_{\{i|abs.lacune\}}^{l}\right)=\left\{\mathrm{card}\left(x_{\{i|abs.lacune\}}^{l}\right)/\mathrm{n}\leq 30\%\right\}$; $x_{\{i|abs.lacune\}}^{l}$ l'ensemble des variables non lacunaires ; Et $\mathrm{I}_{\left\{\mathrm{i}\middle|\mathrm{x_i^l}\neq\mathrm{lacune}\right\}}^{\mathrm{l}}$ l'ensemble des individus spatialisés dont x_i^l ne contient pas de lacune

$$x_{\{\mathrm{i}|\mathrm{abs.lacune}\}}^{\mathrm{l}}=\bigcup_{i=1}^{n}\left(x_{\mathrm{i}}^{\mathrm{l}}\middle|x_{\mathrm{i}}^{\mathrm{l}}\neq\ \text{»lacune »}\right)$$

Remarque :
Cette stratégie de comblement des lacunes utilise un estimateur statistique tant que celui-ci reste consistant *en pratique*. Sinon les variables lacunaires sont *rendues inactives* (Saporta, 2006).

ESTIMATION DE LA METRIQUE FLOUE

Objectif :
Construire une *métrique floue* en s'appuyant sur des *connaissances expertes* afin d'aboutir à une modélisation plus robuste et plus juste de la réalité géographique des FIM* lors de la phase d'agrégation ensembliste des CIM* : $x_{\mathrm{i}}^{\mathrm{l}}$ dans les $\mathrm{U_k}$. L'optique est de rationaliser l'impact spatiotemporel des $x_{\mathrm{i}}^{\mathrm{l}}$ dans le processus d'agrégation au regard de sa granularité*.

Spécification de l'hypothèse :
Les facteurs de certitude spatiotemporelle $\upsilon_{\mathrm{i}}^{\mathrm{j}}$ permettent de donner plus de robustesse au processus de fusion des $x_{\mathrm{i}}^{\mathrm{l}}$ en se fondant sur la théorie *des ensembles flous* (Dubois Didier, Prade Henri, 2004). En l'occurrence, l'injection des *connaissances expertes* s'effectue par le biais des *fonctions mathématiques* permettant d'évaluer la certitude spatiotemporelle par une combinaison d'informations à connotation : Epidémiologique, Géographique et Statistique (EpiGéoStat).

Principe d'estimation :
Les facteurs composites de certitude EpiGéoStat : $\upsilon_{\mathrm{i}}^{\mathrm{l}}$ sont obtenus par la moyenne empirique des facteurs thématiques de certitude spatiotemporelle, tel que :

$$v_i^l=\frac{1}{n_w}\cdot\sum_{w=1}^{n_w}v_i^{l,w}\in[\![0,1]\!]$$

En l'occurrence $n_{\upsilon} = 3$ puisque les facteurs thématiques de certitude spatiotemporelle considérés sont : $\upsilon_i^{l,geo}$, $\upsilon_i^{l,epid}$, $\upsilon_i^{l,stat}$.

Remarque :
Le principe d'agrégation ensembliste *des poids thématiques de certitude spatiotemporelle* est simple, ce qui présuppose que le niveau de contribution de chacun d'eux, i.e. de leur fiabilité, doit être pris en compte indépendamment dans le processus d'estimation.

SYNOPTIQUE D'ESTIMATION DU FACTEUR GEOGRAPHIQUE DE CERTITUDE SPATIOTEMPORELLE

Remarques liminaires :
A l'aune de ce qui avait été proposé pour $\pi_i^{j,geo}$, le facteur géographique de certitude spatiotemporelle $\upsilon_i^{l,geo}$ ne peut pas être considéré comme une fonction décroissante de la superficie territoriale des U_k. En effet, les x_i^l, sont temporellement stables, i.e. que le sexe, le type de LEUC ou encore l'agressivité du traitement reçu sont indépendants des expositions environnementales, donc de celle induite par la variabilité spatiale de l'incertitude liée à l'échelle d'investigation.
L'incertitude spatiotemporelle de la localisation des variables CIM* dans l'espace géographique* est la plus préjudiciable de toutes les incertitudes thématiques. Par conséquent le facteur géographique de certitude spatiotemporelle peut être nul, i.e. $\upsilon_i^{l,égo} = 0$, si l'incertitude thématique associée est trop forte.

Spécification de l'hypothèse :
La certitude géographique est inversement proportionnelle à l'imprécision spatiale liée aux hypothèses de la méthode SpaLea et à l'incertitude spatiale des mobilités résidentielles temporelles associées aux trajectoires de vie à moyen et long termes.

Proposition d'une stratégie d'estimation :
Les poids géographiques de certitude $\upsilon_i^{j,geo}$ sont spécifiés par une formule simple permettant d'inverser les valeurs des fonctions géographiques fragmentaires d'incertitude spatiotemporelle qui avaient été proposées pour $\pi_i^{j,geo}$ – (chapitre.2) – tel que :

$$\upsilon_i^{j,geo} = 1 - \left(g_1^l(q_{o,i}) + g_2^l(T_i, \widehat{\Theta}_m; \widehat{C})\right) \in [\![0;\ 1]\!]$$

Remarque préalable à l'estimation :
Les hypothèses ainsi que les stratégies d'estimation des fonctions d'incertitude spatiotemporelle $g_\cdot^l(\cdot)$ sont identiques à celles proposées pour les $\pi_i^{j,geo}$. En revanche, leurs paramètres ont été réadaptés à l'agrégation des CIM* dans les U_k au vu des propositions heuristiques énoncées auparavant et de la qualité des connaissances expertes injectées. En l'occurrence, la fiabilité de la fonction géographique fragmentaire d'incertitude utilisant les informations issues de LEA et de SpaLea $g_1^l(q_{o,i})$ est bien supérieure à celle de $g_2^l(T_i, \widehat{\Theta}_m; \widehat{C})$ qui est basée sur des données communales socio-économiques.

Stratégie d'estimation de l'incertitude fragmentaire liée à la méthode de spatialisation

Rappel des hypothèses émises:

L'incertitude liée à la spatialisation des patients s'apprécie visuellement par le biais de l'indicateur $SpaLea^{2}_{U_{k_o}}$ représenté en vert dans les communes de seconde espèce (chapitre 2). Mais comme cette incertitude spatiale fragmentaire est attribuée dans une logique *individus-centrée**, préalable à la phase d'agrégation ensembliste des CIM* dans les U_k, elle dépend plutôt de : $q_{o,i}$, i.e. le nombre de codes INSEE de 2nd ordre attribués à chaque patient.

Fonction fragmentaire d'incertitude inhérente à la méthode de spatialisation :

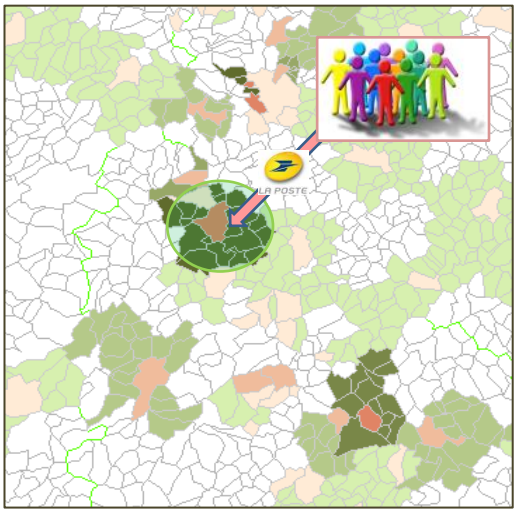
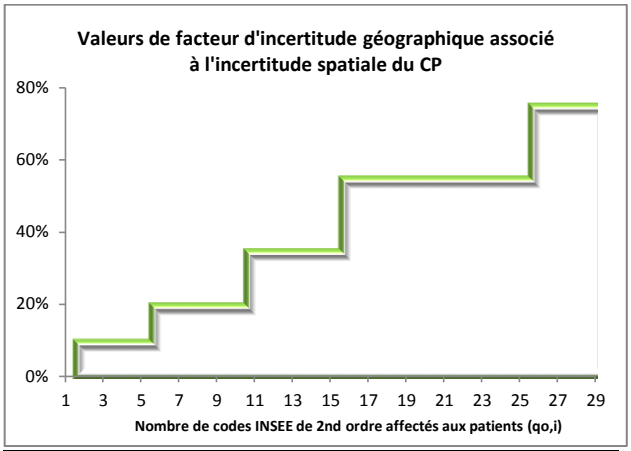


Figure 70 : Synoptique d'estimation de l'incertitude fragmentaire de localisation spatiale



$q_{o,i} \in$	$[\![0,\ 1[\![$	$[\![1,\ 5[\![$	$[\![5,\ 10[\![$	$[\![10,\ 15[\![$	$[\![15,\ 25[\![$	$[\![25,$
$g^{l}_{1}(q_{o,i})$	0%	10%	20%	35%	55%	75%

Figure 71 : Allure de la fonction fragmentaire d'incertitude de localisation spatiale et paramètres spécifiés

Spécification des paramètres :

Les valeurs prises par $g^{l}_{1}(q_{o,i})$ sont affectées par le biais d'une fonction constante par morceaux paramétrée conditionnellement aux caractéristiques de la fonction de répartition empirique spatiale des $q_{o,i}$, notée $\hat{F}_n(q_{o,i})$, à la stratégie d'estimation proposée pour $\upsilon_i^{j,geo}$, et à la fiabilité des informations liées à cette incertitude géographique fragmentaire (i.e. au CP des patients et aux hypothèses du concept de spatialisation de 1ère espèce) ; ce qui permet de borner les valeurs prises par la première fonction d'incertitude fragmentaire à $g^{l}_{1}(\cdot) \in [\![0;\ 0{,}75]\!]$

Remarque :

Les patients dont le nombre de codes INSEE de 2nd ordre était supérieur à 15 ont été vérifiés par les ARC. Conséquence, pour ces derniers, la valeur de $g^{l}_{1}(q_{o,i})$ est nulle. Mais la mise au point d'une méthode nécessite de prendre en compte toutes les éventualités.

Stratégie d'estimation de l'incertitude fragmentaire liée aux mobilités résidentielles à moyen et long terme.

Rappel des hypothèses émises :

Lors de la phase d'agrégation ensembliste des x_i^l dans les U_k il est préjudiciable d'attribuer la même importance aux CIM* des individus qui y résident qu'à celles des individus qui ont probablement déménagé dans une autre commune.

L'incertitude spatiale fragmentaire inhérente *aux trajectoires de vie temporelles* peut être estimée uniquement à partir des *intentions de déménager* idd_t en fonction de *la durée d'occupation du logement* dol_t . Il s'agit des seules données communales disponibles permettant de lier les mobilités spatiales résidentielles et la temporalité, à l'échelle des communes en France. Elles sont extraites de l'enquête INSEE *logement 2006* (Couet, 2006).

La durée d'occupation du logement dol_t est assimilée à T_i, i.e. la distance temps qui sépare la date du premier traitement reçu et celle à laquelle la variable x_i^l est supposée correctement spatialisée, i.e. la dernière date à laquelle le patient a décliné le CP de son lieu de résidence.

Fonction fragmentaire d'incertitude inhérente aux mobilités résidentielles :

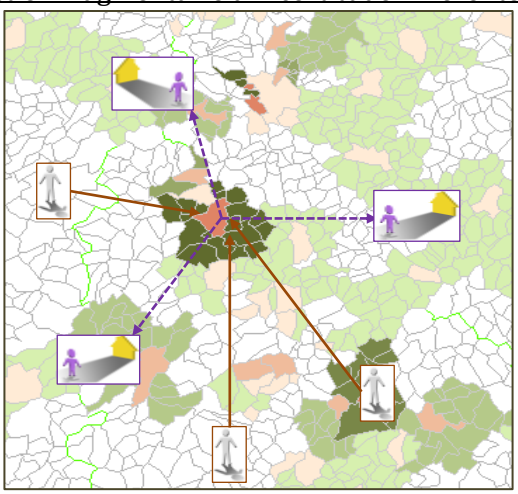
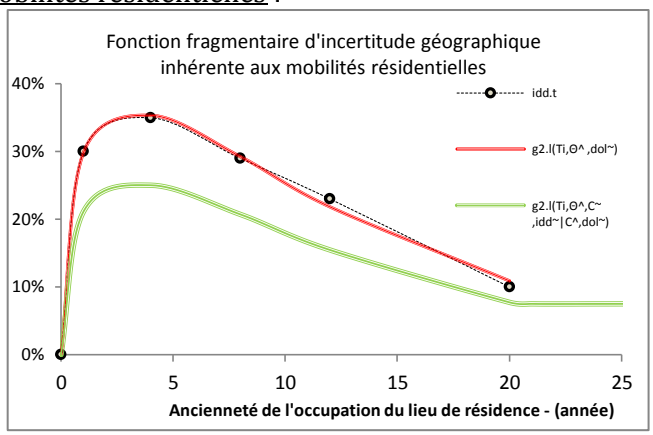


Figure 72 : Synoptique d'estimation de l'incertitude fragmentaire liée aux mobilités résidentielles temporelles



Paramètres	$\hat{k}$	$\hat{\theta}$	$\hat{C}$	$\tilde{C}$	$\text{idd}_{\max}^{\sim}$
valeurs	1,36	9,12	5,99	4,24	0,075

Figure 73 : Allure de la fonction d'incertitude des mobilités résidentielles temporelles et paramètres spécifiés

Spécification des paramètres :

Le modèle mathématique choisi est la loi gamma. Il est identique à celui utilisé pour $\text{i.geo}_i^j(w = 1)$ puisque le phénomène et les données utilisées idd_t sont les mêmes. L'enquête logement 2006 affirme qu'au-delà d'une durée maximale d'occupation du logement $\text{dol}_{\max}^{\sim}$, la probabilité de changer de logement est à la fois stable et faible : $\text{idd}_{\cdot}^{\sim}$ (Couet, 2006) ;

L'ajustement du modèle $g_2^l(T_i, \hat{\Theta}_m, \text{dol}_{\max}^{\sim})$ aux $(t, \text{idd}_t \,|\text{dol}_{\max}^{\sim})$ où $\hat{\Theta}_m = (\hat{C}; \hat{\theta}; \hat{k})$ est un vecteur, de paramètres, qui est spécifié dans un premier temps par le critère des moindres carrés ordinaires (mco) (Saporta, 2006).

La fonction d'incertitude géographique fragmentaire des mobilités résidentielles temporelles est :

$$g_2^l(T_i, \hat{\theta}_m; \tilde{C}, idd_{\max}^{\sim} \,|\hat{C}, dol_{\max}^{\sim}) = \tilde{C} \cdot (T_i)^{\hat{k}-1} \cdot \frac{e^{\left(-\frac{T_i}{\hat{\theta}}\right)}}{\Gamma(\hat{k}) \cdot \hat{\theta}^{\hat{k}}} \cdot \mathbb{1}_{\{T_i \leq dol_{\max}^{\sim}\}} + idd_{\cdot}^{\sim} \cdot \mathbb{1}_{\{T_i > dol_{\max}^{\sim}\}}$$

Les valeurs des *paramètres experts* : $\tilde{C}$ et $idd_{\max}^{\sim}$ sont affinées au regard des conclusions de l'enquête *logement 2006*, de la stratégie d'estimation des $\upsilon_i^{j,geo} \in [\![0,1]\!]$ et du poids dévolu à la fiabilité *des* idd_t qui bornent cette seconde fonction d'incertitude fragmentaire à $g_2^l(\cdot) \in [\![0;\ 0{,}25]\!]$.

Remarques :

Les variables de la BD-LEA ne permettent pas de reconstituer les mobilités résidentielles des patients et aucune reconstitution diachronique de x_i^{CP} n'est financièrement envisageable bien qu'elle permettrait d'estimer de façon plus précise les mobilités extra-communales des patients ; c'est la raison pour laquelle elles sont estimées à partir de données INSEE.

L'incertitude spatiale fragmentaire basée sur les mobilités extra-communales résidentielles temporelles : $g_2^l(\cdot)$ est moins fiable que celle inhérente à la localisation spatiale des patients, utilisée pour $g_1^l(\cdot)$, raison pour laquelle son influence sur $\upsilon_i^{j,geo}$ est beaucoup plus modérée.

SYNOPTIQUE D'ESTIMATION DU FACTEUR EPIDEMIOLOGIQUE DE CERTITUDE SPATIOTEMPORELLE

Remarques liminaires :
Les variables CIM* ont des qualités spatiotemporelles différentes qui doivent être prises en compte lors de leur agrégation dans les U_k (Abramson J.H., Abramson Z.H., 1988).

Les valeurs des x_i^l sont des variables LEA *historiques* particulièrement sûres et bien renseignées dans la BD. La majorité d'entre elles est temporellement stable. Le sexe, qui est déterminé à la naissance, le protocole de traitement, qui est fixé par une Réunion de Concertation Pluridisciplinaire (RCP), et la leucémie traitée, sont des données invariantes dans le temps. Par conséquent les valeurs doivent être $\upsilon_i^{j,épi} \gg 0$. La stratégie d'estimation du facteur épidémiologique de certitude spatiotemporelle : $\upsilon_i^{l,épi}$ a été construite à l'appui de *recommandations expertes* (Auquier et Michel, 2012).

Spécification de l'hypothèse
La qualité spatiotemporelle des variables CIM* dépend : du taux de lacunes l_i^l qu'elles contiennent, et du temps d'incertitude : ti_i lorsqu'elles ne sont pas temporellement stables.

Proposition d'une stratégie d'estimation :
Les poids épidémiologiques de certitude $\upsilon_i^{l,épi}$ sont spécifiés inversement aux valeurs des fonctions épidémiologiques fragmentaires d'incertitude spatiotemporelle, dans une logique analogue à celle proposée pour les $\upsilon_i^{l,géo}$, tel que :

$$\upsilon_i^{l,épi} = 1 - \left(e_1(l_i^l) + e_2(ti_i)\right) \in [\![0{,}4;\ 1]\!]$$

Remarque préalable à l'estimation :
Les fonctions épidémiologiques fragmentaires d'incertitude sont des fonctions mathématiques.

Schéma et principe d'estimation du facteur épidémiologique de certitude spatiotemporelle

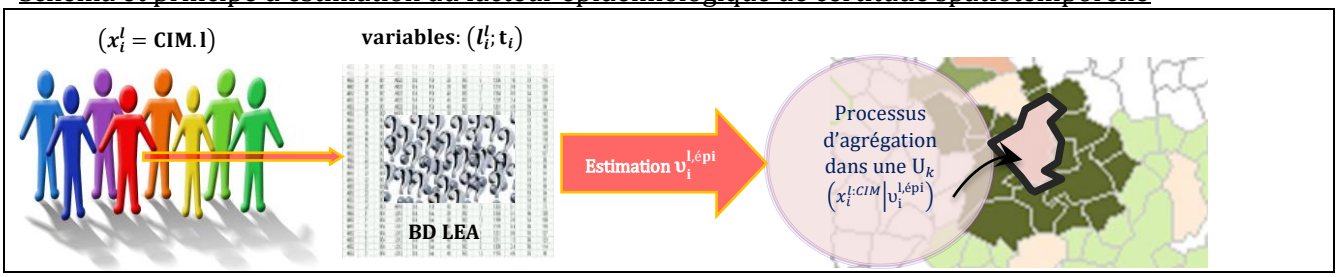


Figure 74 : Incertitudes fragmentaires induites par le niveau de lacune et par la variabilité temporelle de la variable CIM

Spécification des hypothèses :
Les fonctions épidémiologiques fragmentaires d'incertitude spatiotemporelle sont

$e_1(l_i^l)$	$\mathbf{e_2(ti_i)}$
Incertitude liée à qualité informationnelle de x_i^l	Incertitude liée à la variabilité temporelle des x_i^l
Elle s'estime proportionnellement au niveau de lacunes de la variable CIM. $\mathbf{l_i^l}$ vaut 1 lorsque x_i^lest une valeur comblée par un estimateur statistique consistant (Saporta, 2006). Dans le cas contraire la valeur renvoyée par la variable $\mathbf{l_i^l}$ vaut 0.	Elle est proportionnellement croissante à $\mathbf{ti_i}$: la distance temps entre la date à laquelle le patient a été interrogé pour la dernière fois au sujet de x_i^l et celle à laquelle cette information est supposée exacte et correctement spatialisée – celle de la BD-LEA utilisée
Lorsque x_i^l est une valeur comblée le poids d'incertitude dévolu au risque que l'estimateur statistique utilisé soit une information erronée est préjudiciable et comme $e_1(\cdot) \leq 0{,}6$, alors : $e_1(l_i^l) \in [\![0;\ 0{,}5]\!]$	Les x_i^l instables sont généralement des caractéristiques comportementales, donc moins bien renseignées dans la BD-LEA. Par hypothèse, leurs effets sur les séquelles sont plus modérés, plus incertains, conséquence : $e_2(ti_i) \in [\![0;\ 0{,}1]\!]$

Stratégie d'estimation et spécification des paramètres utilisés :
Les paramètres utilisés ont été spécifiés au regard des hypothèses, de la stratégie et des remarques déclinées

$$\upsilon_i^{l,epid} = 1 - \left(\left| \frac{1}{n} \sum_{i=1}^{n} \left(\left\{ l_i^l = \mathbb{1}_{\{x_i^l = lacune\}} \right\} \right) \right| \wedge 50\% \right) - \frac{0{,}1 \cdot ti_i}{\left(\max_{\forall i=\{1,\ldots,n\}} \{ti_i\} - \max_{\forall i=\{1,\ldots,n\}} \{ti_i\} \right)}$$

Tableau 32 : Principe d'estimation du facteur de certitude spatiotemporelle épidémiologique

Remarques :
Par définition : $ti_i \neq T_i$, par conséquent, les facteurs thématiques de certitude temporelle $\upsilon_i^{l,épi}$ et $\upsilon_i^{l,géo}$ne sont pas redondants.

L'incertitude liée au taux de lacune $e_1(l_i^l)$ est prise en compte uniquement s'il s'agit d'une donnée comblée. Dans l'éventualité où l'estimateur statistique est inconsistant et où la modélisation géographique du CIM* est faite à partir d'un sous-échantillon de patients spatialisés réduit : $I^l_{\{i | x_i^l \neq lacune\}}$ les poids épidémiologiques de certitude sont sûrs : $\upsilon_i^{l,épi} = 1$. Or, cette éventualité est très préjudiciable et elle est prise en compte dans la stratégie d'estimation des facteurs statistiques de certitude spatiotemporelle fragmentaire $\upsilon_i^{l,stat}$.

SYNOPTIQUE D'ESTIMATION DES FACTEURS D'INCERTITUDE A CONNOTATION STATISTIQUE

Remarques liminaires :
L'incertitude liée au taux de lacune $e_1(l_i^l)$ affecte fortement les valeurs prises par $\upsilon_i^{l,épi}$ mais uniquement lorsque les $x_i^l \neq$ lacune parce qu'elles ont été comblées par une statistique consistante (Saporta, 2006).
Dans le cas contraire la modélisation des FIM* est effectuée à partir d'un sous-échantillon de la population spatialisée. Cette éventualité particulièrement préjudiciable en termes d'inconsistance* statistique doit être prise en compte dans l'estimation des facteurs statistiques de certitude spatiotemporelle. Toutefois, même si le nombre de CIM* utilisées est petit, l'U_k ne peut pas être considérée comme une commune où aucune information ne serait disponible, par conséquent : $\upsilon_i^{l,épi} \geq 0{,}2$.

Spécification de l'hypothèse :
La consistance statistique de la valeur affectée à $x'^{l:FIM}_{(U_k)}$ dépend du nombre $x_i^{l:CIM}$ utilisé, donc du nombre de patients spatialisés dans l'U_k considérée.

Principe de la stratégie d'estimation :
Les facteurs statistiques de certitude spatiotemporelle sont attribués aux patients spatialisés : I_i en fonction du nombre de patients interférant dans le calcul des $x'^{l:FIM}_{(U_k)}$, i.e. sachant $x_i^l \neq$ lacune. La valeur de $\upsilon_i^{l,stat}$ s'estime par le biais d'une fonction statistique d'incertitude spatiotemporelle et du nombre d'individus spatialisés utilisables.

$$\upsilon_i^{l,stat} = 1 - s_1^l\left(n^I_{\left(i\middle|U_k \cap x_i^l\right)}\right) \in [\![0{,}2;\ 1]\!]$$

Remarque préalable à l'estimation :
Le nombre d'individus spatialisés utilisables pour l'estimation du FIM* s'estime de la façon suivante :

$$n^I_{\left(i\middle|U_k \cap x_i^l\right)} = \sum_{i=1}^{n}\left(\mathbb{1}_{\{x_i^l \neq \text{lacune}\} \cap \{V_{i,1}=V_{(U_k)}\}}\right),\quad \forall\ k \in \{1,\ \ldots,\ q_1\}$$

Avec : $V_{i,1}$ le code INSEE de 1er ordre et q_1 le nombre total de communes de 1ère espèce (chapitre 2).
Spécification de l'hypothèse émise:
Cette incertitude est particulièrement préjudiciable lorsque, dans une U_k, le nombre de CIM* utilisables pour l'estimation du FIM* *est petit*. Mais il convient aussi de ne pas considérer les U_k où peu d'individus sont spatialisés comme des communes où aucune information épidémiologique ne serait disponible. Par conséquent, les valeurs de $s_1^l(\cdot)$ doivent être grandes, préjudiciables, et correctement spécifiées pour les petites valeurs de $n^I_{\left(i\middle|U_k \cap l_i^l\right)}$.

Schéma et principe d'estimation de la fonction statistique d'incertitude spatiotemporelle

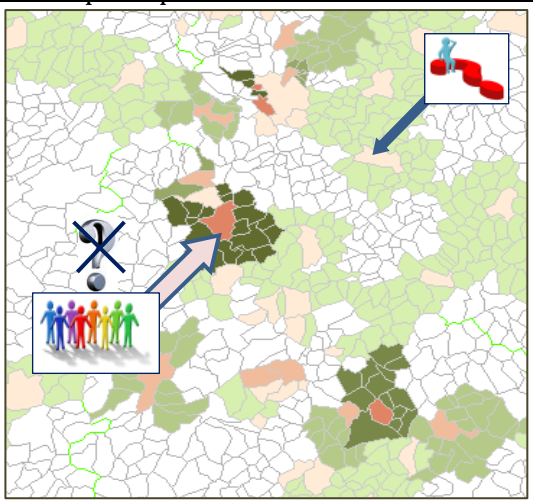
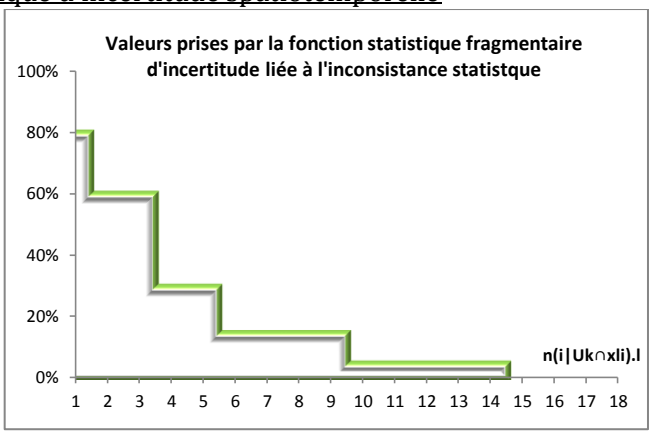


Figure 75 :Synoptique d'estimation de l'incertitude incertitude fragmentaire d'inconsistance* statistique



$n^I_{(i\|U_k \cap x_i^l)}$	[1, 2[	[2, 4[	[4, 6[	[6, 10[	[10, 15[	[15, +∞[
$s_1^l(\cdot)$	80%	60%	30%	15%	5%	0%

Figure 76 : Allure de la fonction d'incertitude fragmentaire d'inconsistance* statistique et paramètres spécifiés

Principe d'estimation :
Les valeurs de $s_1^l(\cdot)$ sont attribuées par une fonction constante par morceaux. Ses paramètres ont été fixés au vu des hypothèses, des remarques et de la stratégie d'estimation des $\upsilon_i^{l,stat}$ en se basant sur les caractéristiques de la distribution spatiale du nombre de CIM* utilisables pour l'estimation du FIM

Remarque :
Il est particulièrement préjudiciable de travailler avec des $x'^{l:FIM}_{(U_k)}$ estimés à partir de peu de CIM* et d'autant plus, lorsque les x_i^l lacunaires ne peuvent pas être comblées par une statistique consistante. Aussi, les U_k où très peu de patients sont spatialisés ont des valeurs de $x'^{l:FIM}_{(U_k)}$ et des distances *a-spatiales morbides* particulièrement proches des i.st.m* : $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$. Cependant, le biais induit par l'inconsistance* statistique est pris en charge par π_i^j qui, dans cette configuration particulière, engendre un bruit particulièrement fort sur $z'^{j}_{(U_k),c}$ et porte généralement la valeur de $z'^{j}_{(U_k),c}$ à INCERTAIN.

ANALYSE ET REMARQUES

Commentaires sur les valeurs prises :
La moyenne des facteurs composites de certitude EpiGéoStat est estimée à : $\bar{v}^l_{\cdot} \approx 70\%$.
Plus de 75% des υ_i^l ont une valeur supérieure à 60%.
Aucune valeur de v_i^l n'est inférieure à 35% de certitude.
Le minimum vaut : $v_i^l = 35\%$, il est atteint pour la variable x_i^{ACPHY}, pour un patient cumulant plusieurs incertitudes spatiotemporelles EpiGéoStat.
Les valeurs maximales des v_i^l appartiennent à l'intervalle : $[\![94\%; 96\%]\!]$.

Remarques :
Les valeurs prises par les v_i^l sont relativement élevées et n'ont pas d'effet direct sur les x_i^l – dans la stratégie d'estimation des $x'^{l:FIM}_{(U_k)}$ proposée subséquemment– ils ne pondèrent pas de façon aberrante les CIM* et par conséquent améliorent la capacité des $x'^{l:FIM}_{(U_k)}$ à modéliser les FIM.
La métrique floue géographique de certitude des υ_i^j est une façon d'intégrer l'espace et le temps au cœur du processus d'estimation des i.st.e. En dépit du fait que le tee_i^l n'est pas intégré puisqu'il n'a généralement pas d'influence sur les CIM, la prise en compte des T_i dans les $\upsilon_i^{l,\text{géo}}$ et de ti_i dans les $\upsilon_i^{l,\text{épi}}$ permet bien d'introduire la temporalité au cœur du processus de modélisation des FIM.

ESTIMATION DES INDICATEURS SPATIOTEMPORELS

Objectif :
Modéliser, d'une part, de façon précise, la réalité géographique des FIM* intégrés, et d'autre part, concevoir une stratégie permettant d'induire un bruit de fond spatiotemporel de façon à augmenter la distance a-spatiale morbide* des CIM. Autrement dit, il s'agit de randomiser les informations extraites de la base LEA : $x_i^{l:CIM}$ pour diminuer le niveau de précision avec lequel elles décrivent l'environnement médical, physiologique ou comportemental inhérent à chaque patient. L'optique est d'harmoniser les expositions géographiques représentées par les FIM* et celles consécutives aux FE* : SAN, SOCIO.ECO et PHY.CHIM. Ces dernières sont estimées à partir de données plus *éloignées* des patients et grevées de bruits de fond environnementaux (Mandin, 2004). Lesquels sont induits par des incertitudes engendrées, entre autres, par des systèmes de collecte mesurant des paramètres disparates, avec des niveaux de précision hétéroclites, des défaillances impliquant des lacunes, et enfin des données recueillies parfois à des temporalités et en des lieux différents.

Hypothèse heuristique :
Dans chaque commune, les variables caractérisant les CIM* des patients utilisées pour modéliser la géographie des FIM* doivent être agrégées différemment selon leur nature. Aussi les x_i^l ont des qualités spatiotemporelles hétérogènes qui sont évaluées par des facteurs de certitude EpiGéoStat υ_i^l. Ces derniers peuvent être utilisés pour conditionner la stratégie de modélisation afin de *d'optimiser l'effet information** (Marcotte, 2008), de sorte à ce que les i.st.e* $x'^{l:FIM}_{(U_k)}$ soient représentatifs de la variabilité spatiale des FIM* et aussi adaptés à l'identification de DES.

Spécification de l'hypothèse heuristique :
Les facteurs EpiGéoStat de certitude spatiotemporelle υ_i^l, sont utilisés pour donner plus de consistance à la modélisation des FIM. Mais en contrepartie ils doivent aussi bruiter les i.st.e* de façon à induire une *distorsion géographique morbide* avec les patients. Ils ont aussi pour rôle d'augmenter la Distance a-spatiale morbide* *de l'exposition aux FIM** qui sont calculées par le biais de variables *individus-centrée**.

Proposition méthodologique :
Concevoir des i.st.e* $x'^{l:FIM}_{(\mathrm{U_k})}$ adaptés à la nature des variables épidémiologiques utilisées, et proposer une stratégie d'intégration de l'information contenue dans les υ_i^l afin de prendre en compte à la fois la qualité spatiotemporelle des données sources utilisées x_i^l- de façon à modéliser les FIM* d'une façon fiable en ramenant la Distance a-spatiale morbide* des expositions aux FIM* à celle des autres FE* à connotation SAN, SOCIO.ECO et PHY.CHIM.

Principe d'estimation des i.st.e* quantitatifs:
Lorsque x_i^l est quantitative comme l'âge au moment du diagnostic ou la durée du suivi, l'i.st.e* représentatif de la géographie du CIM* est une moyenne des valeurs des x_i^l répétées au prorata du facteur de certitude EpiGéoStat υ_i^l ·associé à chaque patient spatialisé dans une même $\mathrm{U}_{\mathbf{k}}$.

$$x'^{l:FIM}_{(\mathrm{U_k})} = \frac{1}{n^{\mathrm{I}}_{(U_{\boldsymbol{k}}|\upsilon)}} \cdot \sum_{i=1}^{n} \left(\bigcup_{\mathrm{b}=1}^{[\upsilon_i^l \cdot 10]} \left(x_i^l \cdot \mathbb{1}_{\{\mathrm{b}\}} \right) \right) \cdot \mathbb{1}_{\left\{\mathrm{V_{i,1}}=\mathrm{V}_{(\mathrm{U_k})}\right\}}$$

Avec : $n^{\mathrm{I}}_{(U_{\boldsymbol{k}}|\upsilon)}$ le nombre de patients spatialisés dans $\mathrm{U}_{\mathbf{k}}$ répété au prorata de la valeur arrondie à l'entier le plus proche du υ_i^l et multipliée par dix.

$$n^{\mathrm{I}}_{(\mathrm{U_k}|\upsilon)} = \mathrm{card}\left(\bigcup_{\mathrm{b}=1}^{[\upsilon_i^l \cdot 10]} \left(x_i^l \cdot \mathbb{1}_{\{\mathrm{b}\}} \,\middle|\, \left\{\mathrm{V_{i,1}} = \mathrm{V}_{(\mathrm{U_k})}\right\} \right) \right)$$

Remarque heuristique :
La stratégie d'agrégation permet de créer un bruit de fond puisqu'il s'agit d'une moyenne spatiale pondérée EpiGéoStat, tout en attribuant néanmoins un poids plus important aux données épidémiologiques les plus consistantes d'un point de vue spatiotemporel.

Principe d'estimation des i.st.e* qualitatifs :
Lorsque x_i^l est qualitative, soit booléenne – i.e. le type de LEUC traitée, le sexe du patient, soit multi-classes – i.e. le protocole de traitement reçu, le type d'activité physique pratiquée - la géographie des CIM* est modélisée par l'estimateur du mode à partir des x_i^l répétées au prorata du facteur EpiGéoStat de certitude $\upsilon_{\mathrm{i}}^{\mathrm{l}}$, associé aux patients d'une même $\mathrm{U_k}$, mais uniquement pour ceux dont la valeur de $\upsilon_{\mathrm{i}}^{\mathrm{l}}$ est suffisamment robuste, tel que:

$$x'^{l:FIM}_{(U_k)} = \begin{cases} \underset{\mathcal{C}^{\mathrm{l}}=\{1,\ldots,\mathrm{c_l}\}}{\mathrm{argmax}} \left\{ \mathbb{P}_{\mathrm{F_n}}\left(\dot{\mathrm{x}}^{\mathrm{l}}_{(\cdot|\cdot)} = \mathcal{C}^{\mathrm{l}} \right) \right\} & \text{lorsque:} \quad \mathbb{P}_{\mathrm{F_n}}\left(\left\{ \dot{\mathrm{x}}^{\mathrm{l}}_{(\cdot|\cdot)} = \mathcal{C}^{\mathrm{l}} \right\} \right) \neq \dfrac{1}{\mathrm{c_l}} \\ \mathrm{môd}\left(\mathrm{x}^{\mathrm{l}}_{\cdot} \right) & \text{lorsque:} \quad \mathbb{P}_{\mathrm{F_n}}\left(\left\{ \dot{\mathrm{x}}^{\mathrm{l}}_{(\cdot|\cdot)} = \mathcal{C}^{\mathrm{l}} \right\} \right) = \left\{ \dfrac{1}{\mathrm{c_l}} \cup \phi \right\} \end{cases}$$

Avec : $\mathrm{c_l}$ le nombre de modalités de la variable x_i^l ; $x_\cdot^l$ les variables associées à l'intégralité des patients spatialisés ; $\dot{x}^l_{(I_i \subset U_k | \upsilon_i^l > \bar{\psi}'_\upsilon)}$ *les* x_i^l des patients spatialisés dans l'$\mathrm{U_k}$ et répétées au prorata du produit $\left(\upsilon_{\mathrm{i}}^{\mathrm{l}} \cdot 10\right)$, arrondi à l'entier le plus proche, lorsque la certitude associée à la CIM* est strictement supérieure à un seuil d'élimination des CIM* incertains $\bar{\psi}'_\upsilon$, tel que :

$$\left\{\dot{x}^{l}_{(\cdot|\cdot)} \overset{\text{def}}{\equiv} \dot{x}^{l}_{(I_i \subset U_k | \upsilon^{l}_i > \bar{\psi}'_{\upsilon})}\right\} = \bigcup_{i=1}^{n}\left(\bigcup_{b=1}^{[\upsilon^{l}_i \cdot 10]} \left(x^{l}_i \cdot \mathbb{1}_{\{b\}} \left| \{V_{i,1} = V_{(U_k)}\} \cap \{\upsilon^{l}_i > \bar{\psi}'_{\upsilon}\}\right.\right)\right)$$

Lorsque les valeurs υ^{l}_{i} sont inférieures ou égales à un seuil moyen d'élimination des variables CIM* qui ne sont pas sûres d'un point de vue EpiGéoStat, alors $\{x^{l}_{i} = \phi\}$. Le seuil d'élimination des x^{l}_{i} dont la qualité spatiotemporelle *n'est pas a priori statistiquement admissible*, i.e. associée à un facteur EpiGéoStat de certitude spatiotemporelle *anormalement* faible, est défini par :

$$\bar{\psi}'_{\upsilon} = \left\lfloor \text{môy}\left(\bigcup_{l=1}^{n^{CIM}_{l}} \bigcup_{i=1}^{n} (\upsilon^{l}_{i})\right) - t_{(1-\alpha)} \cdot \frac{\hat{\sigma}\left(\bigcup_{l=1}^{n^{CIM}_{l}} \bigcup_{i=1}^{n} (\upsilon^{l}_{i})\right)}{\sqrt{n^{CIM}_{l} \cdot n}} \right\rfloor$$

Avec : n^{CIM}_{l} le nombre de variables CIM* intégrées ; $t_{(\cdot)} \xrightarrow[n\to+\infty]{} Z \sim \mathcal{N}(0,1)$ une variable gaussienne telle que $\mathbb{P}\left(\mathcal{N}(0,1) \leq t_{(\alpha)}\right) = \{\alpha = 2{,}5\%\}$; môy$(\cdot)$ l'estimateur de la moyenne ; Et $\hat{\sigma}(\cdot)$ l'estimateur biaisé de l'écart-type (Saporta, 2006).

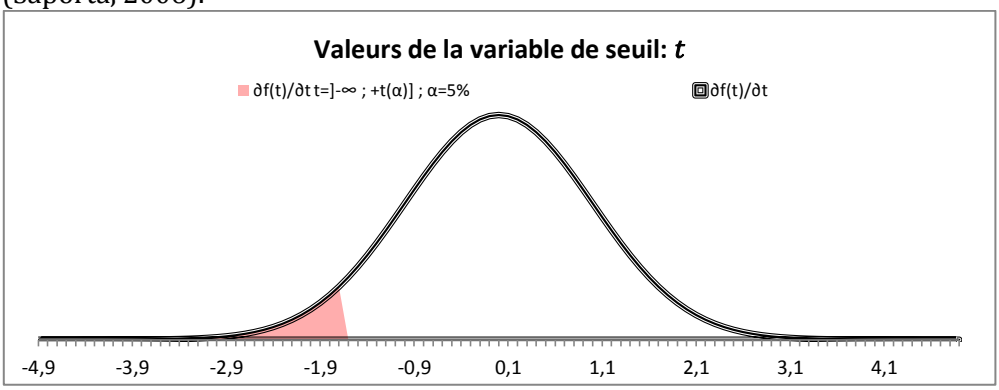


Figure 77 : Représentation d'un seuil gaussien unilatéral avec le niveau de risques admis pour l'estimation de $\bar{\psi}'_{\upsilon}$

Remarque sur l'estimation : le nombre total de x^{l}_{i} éliminées est de 24. Il est identique pour toutes les variables CIM, à l'exception de x^{ACPHY}_{i} pour laquelle 37 valeurs sont supprimées - parmi les 747 associées à l'intégralité des patients spatialisés.

Remarques heuristiques :

La stratégie de sélection des $\dot{x}^{l}_{\left(I_i \subset U_k \left| \upsilon^{l}_i > \bar{\psi}'_{\upsilon}\right.\right)}$ permet bien de retenir un sous-ensemble de patients spatialisés dont la qualité EpiGéoStat *des* x^{l}_{i} est plus représentative et plus adéquate à la modélisation spatiotemporelle de la réalité géographique des FIM.

La règle spécifiée pour l'estimation des i.st.e* des FIM* qualitatif permet de bruiter les variables $x'^{l:FIM}_{(Uk)}$ en ne retenant que la modalité la plus probable estimée sur le sous-ensemble $\dot{x}^{l}_{(\cdot|\cdot)}$, ou en cas d'équiprobabilité sur $x^{l}_{\cdot}$. Cela permet de fait, d'obtenir systématiquement une valeur pour $x'^{l:FIM}_{(U_k)}$ et d'induire une *distorsion géographique* de sorte que la Distance a-spatiale morbide* des expositions aux FIM* soit plus semblable à celle des autres FE* empreints des bruits de fond environnementaux divers.

Les stratégies d'estimation proposées ont été appliquées aux $x^{l:CIM}_{i}$ et ont permis d'estimer les i.st.e* $x'^{l:FIM}_{(Uk)}$ et par conséquent de modéliser la variabilité spatiale des FIM* *pertinents* et Curieux** intégrés.

PRESENTATION DES RESULTATS ET REMARQUES

Les résultats cartographiques de la modélisation géographique des Caractéristiques Individuelles et Médicales*(CIM) par le biais des i.st.e* $x'^{l:FIM}_{(Uk)}$ proposés sont déclinés pour chacun des FIM* *pertinents* et Curieux** intégrés. Ils représentent la variabilité spatiale des expositions environnementales microcosmiques inhérentes aux CIM* des patients spatialisés.

Les résultats cartographiques sont présentés dans les U_k sises en région PACA et aux alentours. Leur analyse visuelle ne permet pas d'identifier des DES*. Le seul moyen de parvenir à caractériser la complexité des régularités spatiales entre les i.st.e* et les i.st.m* est de les mettre simultanément en perspective et d'utiliser une procédure mathématique adaptée (chapitre.4).
Aussi l'indicateur $\mathrm{SpaLea}^2_{\mathrm{U_{k_o}}}$ représentant l'incertitude spatiale associée à l'identification des U_k induite par les hypothèses de la méthode SpaLea et à la granularité* des données n'est pas présenté pour ne pas surcharger l'affichage.

La documentation des cartes décline l'i.st.e* $x'^{FIM}_{(U_k)}$ présenté et le type de variabilité spatiale modélisée par le FIM. Un tableau agrémente chacune d'elles. Celui-ci contient deux indicateurs statistiques qui diffèrent selon la nature qualitative ou quantitative de la variable. Le premier est estimé dans une logique individus-centrée* sur : x_i^l, et le second dans logique spatiale sur l'en : $x'^{FIM}_{(U_k)}$.

<u>Indicateurs statistiques présentés pour les CIM* et les FIM* de nature quantitative :</u>
La valeur moyenne (vm) de la variable CIM* $x^{\mathrm{l}}_{\mathrm{i}}$, estimée sur les patients spatialisés (ps) :

$$\mathrm{vm.\,ps}\left(x^{\mathrm{l}}_{\mathrm{i}}\middle|\mathrm{I}_{(\mathrm{i}|V_{i,1}\neq\phi)}\right) \overset{\text{def}}{=} \bar{x}^{\mathrm{l}}_{\mathrm{i}}$$

La valeur moyenne pondérée (vmp) au prorata de la certitude EpiGéoStat estimée sur l'ensemble des i.st.e* $x'^{\mathrm{l:FIM}}_{(\mathrm{U}_k)}$ associés aux communes de première espèce (cpe) :

$$\mathrm{vmp.\,cpe}\left(x'^{\mathrm{l:FIM}}_{(\mathrm{U}_k)}\middle|\mathrm{I}_{(\mathrm{i}|V_{i,1}\neq\phi)}\right) \overset{\text{def}}{=} \bar{x}'^{\mathrm{l}}_{(\mathrm{U}_k)}$$

<u>Indicateurs statistiques présentés pour les CIM* et les FIM* de nature qualitative :</u>
La proportion d'individus (pi) dont le CIM* : $x^{\mathrm{l}}_{\mathrm{i}}$ est caractérisé par la modalité c^l_k, estimée sur l'ensemble des patients spatialisés (ps).

$$\mathrm{pi.\,ps}\left(x^{\mathrm{l:CIM}}_{\mathrm{i}} = \mathrm{c}^l_k\right) \overset{\text{def}}{=} \frac{1}{n}\cdot\mathrm{card}\left(x^{\mathrm{l}}_{\mathrm{i}} = \mathrm{c}^l_k\middle|\mathrm{I}_{(\mathrm{i}|V_{i,1}\neq\phi)}\right)$$

La proportion communale (pc) où les CIM* du sous-ensemble de patients spatialisés de façon fiable (psf), au sens EpiGéoStat du terme, peuvent majoritairement être décrits par la modalité $\mathrm{c}^{\mathrm{l}}_{\mathrm{k}}$ - i.e. la proportion des $x'^{\mathrm{l:FIM}}_{(\mathrm{U}_k)}$ associés aux $\mathcal{U}^{1^{\mathrm{er}}}_{\cdot}$ prenant majoritairement la modalité $\mathrm{c}^{\mathrm{l}}_{\mathrm{k}}$:

$$\mathrm{pc.\,psf}\left(x'^{\mathrm{l}}_{(\mathrm{U}_k)} = \mathrm{c}^l_k\right) \overset{\text{def}}{=} \frac{1}{\mathrm{N}(\mathrm{U}_k)}\cdot\mathrm{card}\left(x'^{\mathrm{l}}_{(\mathrm{U}_k)} = \mathrm{c}^l_k\middle|\mathrm{I}_{(\mathrm{i}|V_{i,1}\neq\phi)}\right)$$

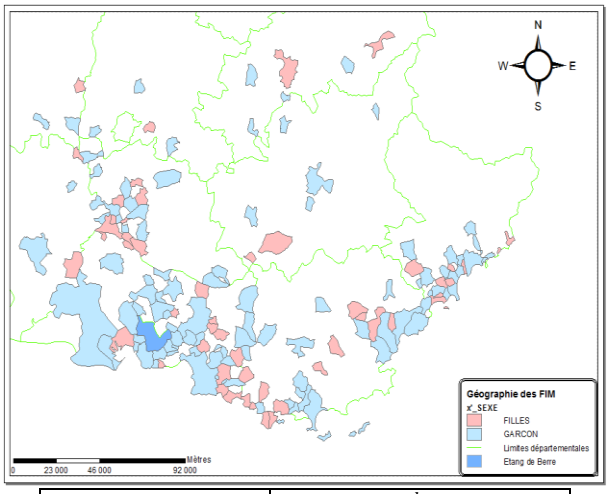
Aussi les modalités des $\mathrm{x}'^{\mathrm{FIM}}_{(\mathrm{U_k})}$ de nature qualitative sont affichées en lettres majuscules pour garantir la lisibilité de la légende.

Les remarques effectuées sur la modélisation géographique des FIM* sont déclinées uniquement lorsque des singularités spatiales sont observables.

CARTOGRAPHIES ET STATISTIQUES DE LA GEOGRAPHIE DES FACTEURS INDIVIDUELS ET MEDICAUX

Géographie des Caractéristiques Individuelles et Médicales*de la population LEA

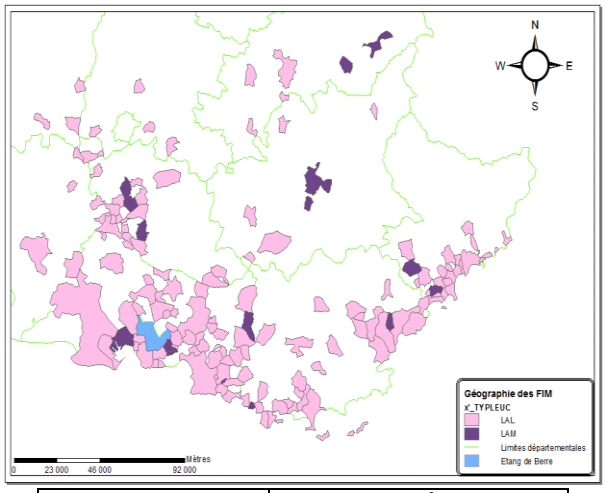
Variabilités spatiotemporelles des genres



l: SEXE	Modalité : c_k^l = GARCON
pi. ps$\left(x_i^{l:CIM} = c_k^l\right)$	56%
pc. psf$\left(x'^{l:FIM}_{(U_k)} = c_k^l\right)$	63%

Figure 78 : i.st.e : $x'^{SEXE}_{(U_k)}$: genre majoritaire EpiGéoStat des patients spatialisés

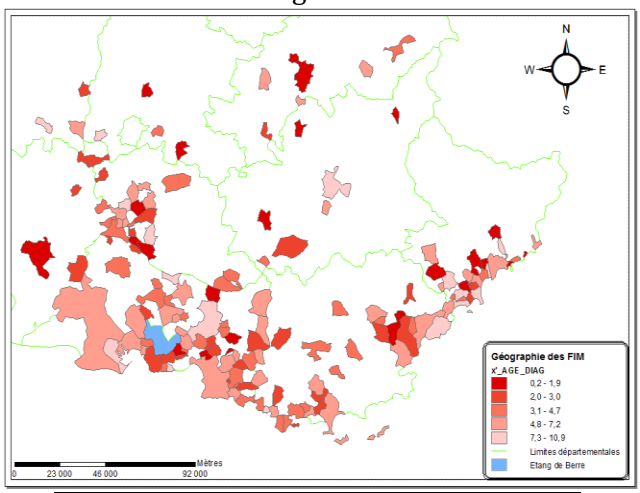
Variabilités spatiotemporelles des leucémies diagnostiquées



l: TYPELEUC	Modalité : c_k^l = LAL *
pi. ps$\left(x_i^{l:CIM} = c_k^l\right)$	86,3%
pc. psf$\left(x'^{l:FIM}_{(U_k)} = c_k^l\right)$	89,0%

Figure 79 : i.st.e : $x'^{TYPLEUC}_{(U_k)}$: types de LA majoritairement traitées EpiGéoStat

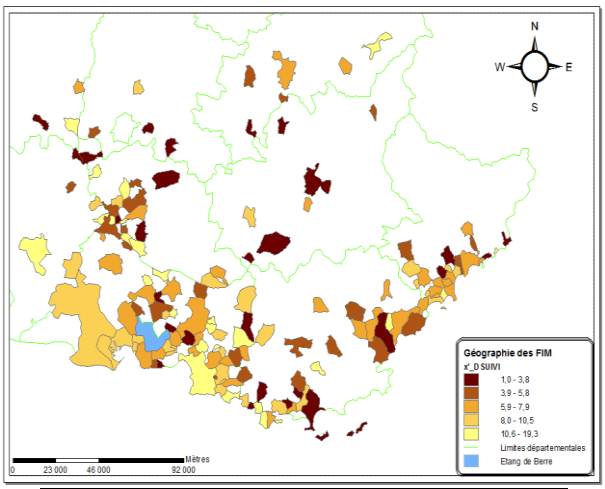
Variabilités spatiotemporelles des âges au diagnostic



l: AGE_DIAG	Nature : quantitative
$\bar{x}_i^l$	6,2 ans
$\bar{x}'^{l}_{(U_k)}$	5,1 ans

Figure 80 : i.st.e : $x'^{AGE\_DIAG}_{(U_k)}$: âge moyen EpiGéoStat au moment du diagnostic de la leucémie

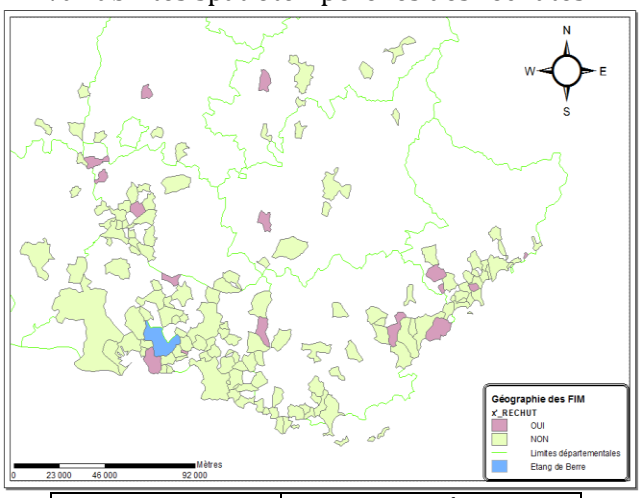
Variabilités spatiotemporelles des durées de suivi



l: DSUIVI	Nature : quantitative
$\bar{x}_i^l$	19,1 ans
$\bar{x}'^{l}_{(U_k)}$	17,3 ans

Figure 81 : i.st.e : $x'^{DSUIVI}_{(U_k)}$: durée moyenne EpiGéoStat du suivi des patients

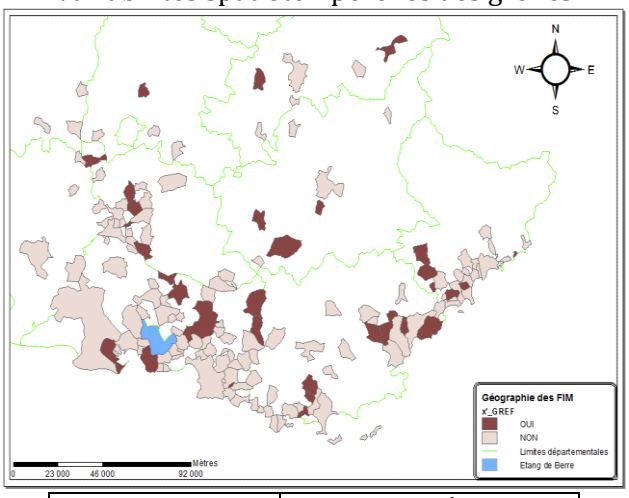
Variabilités spatiotemporelles des rechutes



l: RECHUT	Modalité : c_k^l = OUI
pi. ps$\left(x_i^{l:CIM} = c_k^l\right)$	16,3%
pc. psf$\left(x'^{l:FIM}_{(U_k)} = c_k^l\right)$	11,9%

Figure 82 : i.st.e : $x'^{RECHUT}_{(U_k)}$: proportion EpiGéoStat des patients qui ont reçu des traitements complémentaires liés à une rechute.

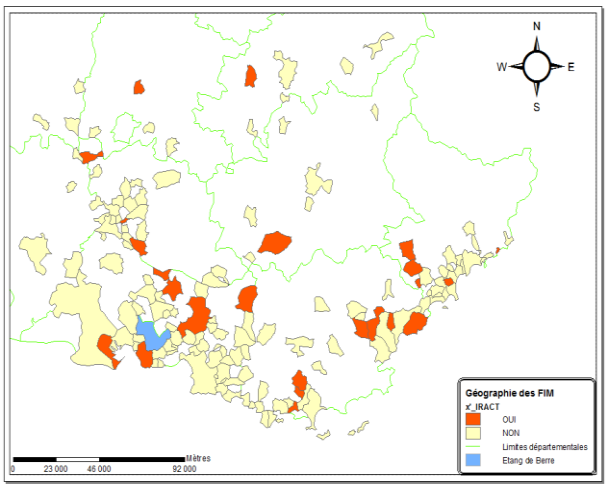
Variabilités spatiotemporelles des greffes



l: GREF	Modalité : c_k^l = OUI
pi. ps$\left(x_i^{l:CIM} = c_k^l\right)$	26,9%
pc. psf$\left(x'^{l:FIM}_{(U_k)} = c_k^l\right)$	21,0%

Figure 83 : i.st.e : $x'^{GREF}_{(U_k)}$: proportion EpiGéoStat des patients greffés dans le cadre du traitement de leur leucémie.

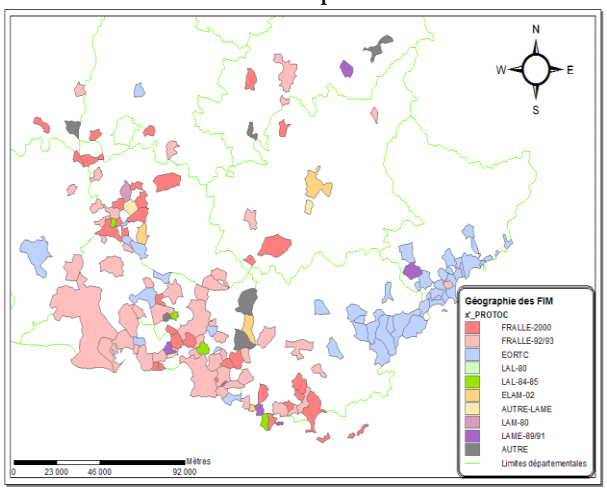
Variabilités spatiotemporelles des irradiations thérapeutiques



l: IRACT	Modalité : c_k^l = OUI
pi. ps$\left(x_i^{l:CIM} = c_k^l\right)$	19,4%
pc. psf$\left(x'^{l:FIM}_{(U_k)} = c_k^l\right)$	15,9%

Figure 84 : i.st.e : $x'^{IRCAT}_{(U_k)}$: proportion EpiGéoStat des patients qui ont subi une irradiation corporelle totale.

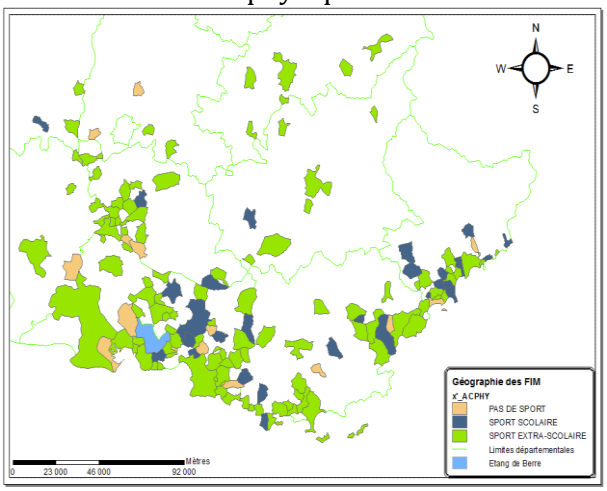
Variabilités spatiotemporelles des protocoles de traitement prescrits



l: PROTOC	COMMENTAIRES
Protocoles les plus utilisés sur l'ensemble des patients spatialisés	21% ont été traités avec EORTC ; 27% avec Fralle-92/93 et 19% avec Fralles-2000 ; 32% restant ont reçu d'autres protocoles
Protocoles EpiGéoStat majoritairement prescrits dans les U_k	l'EORTC est dans 20% des U_k ; Fralle-92/93 dans 36% des U_k ; Fralle-2000 dans 21% des U_k ; D'autres protocoles sont prescrits dans les 23% des U_k restantes

Figure 85 : i.st.e : $x'^{PROTOC}_{(U_k)}$ protocole EpiGéoStat majoritairement utilisé pour traiter les leucémies parmi 11 possibles.

Variabilités spatiotemporelles de l'activité physique



l: ACPHY	COMMENTAIRES
Sur l'ensemble des patients spatialisés	19,3% déclarent ne pratiquer aucun sport ; 27% pratiquent une activité sportive scolaire 35,8% ont une activité sportive extra-scolaire 17,9% n'ont pas répondu
Observation EpiGéoStat majoritaire sur l'ensemble des communes de 1ère espèce	Aucune activité physique dans 17% des U_k Activité physique scolaire dans 20% des U_k Discipline sportive extra-scolaire 63% des U_k 0% de lacunes puisque comblées à la source

Figure 86 : i.st.e : $x'^{ACPHY}_{(U_k)}$ nature et intensité EpiGéoStat de l'activité sportive majoritairement pratiquée.

REMARQUES

Remarques particulières :

Variabilité spatiale des greffes et des irradiations corporelles totales :
les cartographies des niveaux géographiques de patients greffés, modélisés par $x'^{GREF}_{(U_k)}$, et celles des irradiations corporelles totales, modélisées par $x'^{IRACT}_{(U_k)}$, présentent des ressemblances visuelles troublantes dans les communes situées en PACA et aux alentours. En dépit du fait que la proportion d'U_k où les patients sont majoritairement greffés soit plus élevée que celle où ils sont majoritairement irradiés – il semblerait que ces deux i.st.e* soient spatialement corrélés.

Variabilité spatiale des protocoles de traitements majoritairement utilisés :
Les cartographies mettent en exergue une composante territoriale forte quant au type de protocole prescrit dans les $\mathcal{U}^{1^{er}}_{.}$. Dans toutes les communes situées dans le sud-est de la région PACA, les patients sont toujours majoritairement traités avec l'EROTC. Et dans celles situées dans le sud-ouest de la région PACA, ils reçoivent presque systématiquement les protocoles : Fralle.92-93 et Fralle.2000. Il existe donc des disparités géographiques évidentes au niveau du protocole de traitement prescrit. Cependant, il ne

s'agit pas d'une inégalité géographique *d'accès au traitement de référence* - (Klein, 1989). Du moins pas exactement. Les enfants traités pour une LAL* reçoivent en réalité le protocole européen auquel a adhéré le centre de référence par lequel ils ont été suivis. En PACA, il n'y en a que deux : Le CHU de Nice qui utilise l'EORT et le CHU de Marseille qui a adhéré aux protocoles de type Fralle. Toutefois, *il n'y a pas de différence significative entre EORT et les Fralles.* Ils contiennent exactement les mêmes principes actifs, *seuls les excipients changent*. Et, *a priori* ce sont les principes actifs qui sont toxiques. Quant au différentiel de risque induit par la synergie entre les deux – *rien n'est démontré* (Auquier, 2013a).

Variabilité spatiale de l'intensité des activités physiques pratiquées :
Dans la majorité des communes fortement urbanisées, plus de 50% des individus pratiquent des activités sportives extra-scolaires, i.e. une activité physique plutôt régulière et intensive. Aussi l'i.st.e* $x'^{\mathrm{ACPHY}}_{(\mathrm{U_k})}$ ne contient pas de lacunes puisqu'elles sont comblées à la source sur les x_i^{ACPHY} par la stratégie *bouche-trou* proposée, afin de ne pas perdre en puissance statistique lors de l'identification des DES, des FREC et des FREPAS (chapitre 4).

Remarque générale :
Globalement, le nombre de patients spatialisés dans les $\mathrm{U_k}$ *est petit*. Bien que les stratégies de pondération spatiotemporelle appliquées par les υ_i^j sur les i.st.e* caractérisant les FIM, et conjointement par les tee_i^j et les π_i^j sur les i.st.m, permettent d'augmenter virtuellement les effectifs spatialisés et *a priori* de représenter de façon plus juste la réalité géographique des PM, il n'en demeure pas moins que ces stratégies de pondération sont subjectives puisque fondées sur des connaissances expertes.
Par conséquent, l'identification des DES* sera effectuée – par le biais de VSURF (Genuer, Poggi et al., 2013) – dans un premier temps dans une logique géographique, i.e. sur les i.st.m* $z'^j_{(\mathrm{U_k}),c}$ et $z'^j_{(\mathrm{U_k}),q}$ et les i.st.e* modélisant la géographie des FE/FIM* proposés. Puis dans un second temps, une approche *individus-centrée** sera menée. C'est-à-dire directement sur les variables séquelles y_i^j et en appariant, à chaque patient, les caractéristiques environnementales de sa commune de résidence – i.e. les i.st.e* $x^{\mathrm{l:FE}}_{(\mathrm{U_k})}$ et les $x'^{\mathrm{l:FIM}}_{(\mathrm{U_k})}$. Cette seconde application permettra d'une part de corroborer - ou non - les résultats obtenus par l'approche géographique, et dans l'affirmative, de valider, par la même occasion les stratégies de pondération proposées pour la modélisation géographique des PM* et des FIM.

La géographie des FIM* se passe de commentaires approfondis puisque l'analyse de la *complexité* des interactions spatiales - en vue d'identifier des DES, des FREC et des FREPA – par le biais des i.st.m* modélisant la géographie des PM* n'est envisageable qu'en mettant en perspective simultanément tous les i.st.e. destinés à la modélisation géographique des FE/FIM* - dont ceux utilisés pour modéliser la composante environnementale SAN sont décrits dans la section suivante

SECTION B) FACTEURS ENVIRONNEMENTAUX SANITAIRES

Les Facteurs Environnementaux* (FE) à connotation Sanitaire (FE-SAN) sont intégrés dans l'optique de modéliser les variabilités géographiques de *l'accès aux soins qui* est défini, en géographie de la santé, comme *la capacité matérielle d'accéder à des ressources sanitaires et aux services de santé et qui est désormais considérée comme un déterminant de santé ou un éventuel facteur de risque* (Picheral, 2001). L'accès aux soins est un concept multidisciplinaire et les FE-SAN* caractérisent *l'accessibilité physique à l'offre de soins* qui en constitue sa dimension *socio-spatiale* (Penchansky et Thomas, 1981).
Les disparités géographiques en matière d'accès aux soins dépendent de la *qualité des tissus sanitaires* territoriaux*. Leurs effets *directs* ou *indirects* sur l'état de santé des populations sont une thématique de santé publique particulièrement documentée en géographie de la santé (Salem, Rican S et al., 2006) et en épidémiologie spatiale (Penchansky et Thomas, 1981). En l'occurrence, *l'effet des FE-SAN* sur l'incidence et la gravité des séquelles étudiées : CATA, THYR, TUM2 est particulièrement probant* (Auquier, 2010).
Les géographes de la santé sont des orfèvres pour quantifier *l'accessibilité géographique aux soins*. En France, les indicateurs spatiaux sont construits dans une logique territoriale, à partir de données étatiques, *et l'échelle des communes* est *a priori* la plus adaptée pour *capter les variabilités spatiales* de recours aux soins (Chaix, Merlo et al., 2005). La fiabilité de ces indicateurs est conditionnée par leur capacité à prendre en compte les dimensions *sociales*, *économiques* et temporelles *de l'accès aux soins et de leur capacité à s'affranchir au mieux* des *barrières spatiales virtuelles insidieusement induites par* les limites administratives qui circonviennent les sources d'information disponibles (Harrouin, Aligon et al., 2012).

En France, l'Institut de recherche et de documentation en économie de la santé (Irdes) est l'organisme chargé de mettre au point les indicateurs spatiaux caractérisant l'accès géographique aux soins en fonction de la qualité de tissus sanitaires* territoriaux (Irdes, 2012). Ces indicateurs sont expertisés par la Direction de la Recherche et des Etudes, de l'Evaluation et des Statistiques (DREES) et par les Agences Régionales de Santé (ARS). Lorsqu'ils sont validés, des KIT sont développés afin d'estimer ces indicateurs spatiaux, à différentes dates, à l'échelle des cantons ou des communes (ARS, 2012). A l'heure actuelle les variables spatiotemporelles les plus robustes d'un point de vue géographique sont *les distances temporelles d'accès* (DTA) et récemment, *l'Accès Potentiel Localisé** (APL).
Les seules données publiques disponibles à l'échelle des communes : U_u sur l'ensemble de la France métropolitaine sont les DTA par la route aux items sanitaires* en 2007 (DREES, 2012), et depuis peu, les indicateurs APL pour deux types de professionnels de santé libéraux en 2010 (DREES, 2012).
Les items sanitaires décrivent* l'offre de soins territoriale, i.e les spatialités médicales des praticiens de santé libéraux, ainsi que le *plateau technique des établissements de santé* : services et Equipements et Matériels Lourds (EML*).
Les DTA estiment une distance temporelle routière moyenne qui sépare les populations des *items sanitaires** les plus proches. Les DTA sont notées $x^{DTA}_{(U_u),\{t'=2007\}}$, elle sont exprimées en minutes - de nature quantitative discrète. Par hypothèse, le patient se déplace en voiture et emprunte l'itinéraire le plus court. La demande, i.e. la population *in situ*, permet de pondérer les temps d'accès aux items sanitaires*. Mais Les DTA sont biaisées car l'offre et le réseau routier sont circonscrits par des barrières spatiales virtuelles et spécieuses que représentent les limites administratives. De plus les DTA sont nulles lorsque *l'item sanitaire* considéré se trouve dans la commune en question (Coldefy, Lucas-Gabrielli et al., 2011).
Pour pallier cette carence qui affecte principalement la modélisation de l'accès aux praticiens de santé libéraux, l'indicateur APL a été développé. Il embrasse simultanément l'offre de soins, la demande et le temps d'accès aux items sanitaires* en s'affranchissant des limites administratives, i.e. qu'il est estimé sur des *secteurs flottants* - il s'appuie sur *la méthode Two-step floating catchment area*. Les indicateurs sont notés APL $x^{APL}_{(U_u),\{t'=2010\}}$. Il s'agit d'une densité qui s'exprime en nombre de médecins pour mille

habitants mais il est uniquement disponible pour les généralistes hors Mode d'Exercice Particulier* (MEP) i.e. ceux ayant des compétences connexes en acupuncture, homéopathie, angiologie, ainsi que pour les ophtalmologues (Barlet, Lucas-Gabrielli et al., 2012).

A l'heure actuelle APL est l'indicateur le plus consistant d'un point de vue spatiotemporel. Cependant, l'attractivité potentielle sur des zones flottantes *représentatives des bassins de vie* est controversée par un *système de pondération expert* qui ne peut pas être calibré, et qui de fait induit de la subjectivité dans le processus de modélisation. Le système de pondération en question influence fortement les valeurs prises par l'indicateur (Harrouin, Aligon et al., 2012).

Les FE-SAN pertinents**, modélisés par le biais des i.st.e* $x'^{l:SAN}_{(U_k)}$ décrivent les variabilités spatiales de l'accès géographique aux items sanitaires* potentiellement liées aux PM* étudiés, séquelles :

- partir des $x^{DTA}_{(U_u),\{t'=2007\}}$ pour les praticiens libéraux : généralistes (GENE), généralistes MEP et Omnipraticiens, oto-Rhino Laryngologues (ORL), ophtalmologues (OPHT), radiologues (RADIO), pédiatres (PEDI)
- Et par les $x^{APL}_{(U_u),\{t'=2010\}}$ pour les généralistes hors MEP et aux ophtalmologues (OPHT). Les DTA permettent aussi de modéliser l'accès aux plateaux techniques des établissements de santé - aux services hospitaliers de : pédiatrie (PEDIA), hématologie (HEMA.s), oto-Rhino-Laryngologie (ORL.s), ophtalmologie (OPHT.s), endocrinologie (ENDO.s), neurologie médicale et neurochirurgie (NEUR.s) ainsi qu'aux EML* de type appareils d'Imagerie par Résonance Magnétique* (IRM), Scanner* (SCAN), caméra à scintillation (CAME), appareil de Tomographie par Emission de Positons (TEP).

Les *FE-SAN* Curieux* de test* constituent les i.st.e* sanitaires : $x'^{l:SAN}_{(U_k)}$ - construits à partir des DTA qui n'ont pas *a priori* de lien avec les PM* d'intérêt. Par exemple les i.st.e* $x'^{l:OPTH}_{(U_k)}$ seront introduits pour l'identification des DES* lors de l'analyse de THYR à partir des i.st.m* : $z'^{THYR}_{(U_k),c}$ et $z'^{THYR}_{(U_k),q}$.

PROPOSITIONS HEURISTIQUES ET STRATEGIE D'INTEGRATION DES FE-SAN

Objectif :
Modéliser dans les communes de 1ère espèce U_k, par le biais d'i.st.e* robustes et fiables notés $x'^{l:SAN}_{(U_k)}$, le plus précisément possible, la réalité géographique des FE-SAN* intégrés. Il s'agit de prendre en compte la granularité* des variables géographiques mobilisées, i.e. les DTA $x^{DTA}_{(U_u),\{t'\}}$ aux items sanitaires* pertinents* et les $x^{APL}_{(U_u),\{t'\}}$. Bien qu'il s'agisse déjà d'indicateurs *spatiaux modélisant l'accès aux soins,* ils ne sont pas directement intégrables. Leur granularité* intrinsèque n'est pas adaptée à la finalité recherchée, i.e. l'identification des DES* à partir des i.st.m* $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$.

Remarques liminaires :
Les types d'items sanitaires* jugés pertinents* ont été retenus en adéquation avec les PM* d'intérêt. Cependant certaines variables mobilisées sont très proches et induisent des redondances. LES DTA ne sont pas toujours robustes pour estimer l'accès aux praticiens de santé et les indicateurs APL sont disponibles uniquement pour les médecins généralistes (GENE) et les ophtalmologues (OPHT). Il convient de remarquer aussi qu'aucune de ces variables ne contient de lacunes.

Les APL et les DTA mobilisées sont déclinés à l'échelle des communes sur l'intégralité de la France métropolitaine. Cependant, dans les grandes villes, les disparités géographiques en matière d'accessibilité aux soins sont parfois importantes et dépendent de la localisation du lieu de résidence (Coldefy, Lucas-Gabrielli et al., 2011). Pour prendre en compte cette spécificité, les indicateurs spatiaux

d'accès aux soins sont parfois estimés à des échelles plus fines - celle des arrondissements (Ar_u). Trois communes de première espèce sont concernées : Marseille, Lyon et Paris.

Les APL $x^{APL}_{(U_u),\{t'=2010\}}$ et les DTA $x^{DTA}_{(U_u),\{t'=2007\}}$ ne sont disponibles qu'à une temporalité spécifique qui n'est pas représentative de la période recouverte par LEA, i.e. 1980 à 2010. Or, l'accès spatial aux soins varie dans le temps, ce qui pose un problème de consistance temporelle.

En outre, comme les temporalités des indicateurs spatiaux sont différentes de celles de la base SIG utilisée, où les Unités Géographiques repensent les limites territoriales des communes en 2003, un conflit inter-sources est induit au moment de l'appariement des $x^{APL}_{(U_u),t'}$ et des $x^{DTA}_{(U_u),t'}$ aux U_k.

Hypothèse principale :
La granularité* intrinsèque des indicateurs spatiaux APL et des DTA n'est pas celle requise par des i.st.e* adaptés à la modélisation géographique des FE-SAN* pertinents*. Des traitements spatiotemporels doivent être appliqués afin de minimiser le concept de *biais conditionnel**. Il s'agit de proposer des stratégies spatiotemporelles d'intégration de façon à ce que les i.st.e* proposés soient en adéquation avec les spécificités des phénomènes environnementaux SAN retenus et la finalité recherchée (Marcotte, 2008).
Les stratégies d'intégration spatiotemporelles proposées se fondent sur des fonctions mathématiques et sur l'incorporation de connaissances géographiques expertes. Il s'agit, d'une part, d'optimiser *l'effet information** en tentant d'amoindrir les incertitudes des données brutes utilisées, et d'autre part, de maximiser *l'effet de support*,* i.e. de leur grever les caractéristiques spatiotemporelles les plus fiables possible afin que les i.st.e* proposés représentent au mieux la géographie des FE-SAN* intégrables (Baillargeon, 2005).

Proposition méthodologique principale :
La stratégie d'estimation des i.st.e* destinés à modéliser la géographie des FE-SAN* vise à *maximiser l'effet information** par le biais d'une fusion statistique réduisant *les redondances typologiques* (Saporta, 2006)et par des processus *d'agrégation spatiale pondérée géographiquement* afin d'harmoniser les indicateurs déclinés à l'échelle des arrondissements (Ar_u) à celle des U_k (Pumain et Saint-Julien, 1997). Quant à *l'effet de support**, il est optimisé par le biais d'une stratégie de randomisation géographique destinée à pallier *l'inconsistance* temporelle* des données disponibles qui surestiment l'accès aux soins sur la période d'investigation – en introduisant *des bruits blancs* territoriaux à partir des informations disponibles (Saporta, 2006) , et en proposant une stratégie d'appariement aux U_k, conçue dans une logique diachronique territoriale cohérente.

La stratégie d'intégration des variables de la Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques (DREES) se décline dans un processus en deux étapes. Il s'agit d'abord d'optimiser *l'effet information* ;* puis de maximiser *l'effet de support**.

PRINCIPE DE LA STRATEGIE D'INTEGRATION SPATIOTEMPORELLE DES VARIABLES SANITAIRES

Afin de minimiser le concept de *biais conditionnel** des APL $x^{APL}_{(U_u),\{t'=2010\}}$ et les DTA $x^{DTA}_{(U_u),\{t'=2007\}}$, il s'agit, dans un premier temps, *d'optimiser l'effet information** par fusion statistique des types d'items sanitaires* redondants, et en agrégeant les indicateurs déclinés à micro-échelles, pui, dans un second temps, d'une part, *de maximiser l'effet de support** par un processus stochastique géographique permettant de réduire une accessibilité surestimée par la granularité* temporelle des inputs et, d'autre part, d'apparier aux U_k - dans une logique territoriale et diachronique cohérente - les i.st.e* SAN ainsi confectionnés.

OPTIMISATION DE L'EFFET INFORMATION

La phase d'optimisation de l'effet information* des i.st.e* voués à modéliser la géographie des FE-SAN* s'opère d'abord par un processus de fusion des redondances et ensuite, par une procédure d'harmonisation d'échelle, des données mobilisées, à celle des U_k.

FUSION STATISTIQUE D'INFORMATIONS REDONDANTES

Remarques liminaires :

L'intégration conjointe de toutes les DTA caractérisant l'accès aux généralistes est exprimée en fonction du mode d'exercice : simple, MEP, ou omnipraticien et du secteur d'activité du praticien de santé (Coldefy, Lucas-Gabrielli et al., 2011).

Le prix des consultations des médecins exerçant en secteur 1 est fixé par la Caisse d'Assurance Maladie (CAM). Ceux exerçant en secteur 2 sont libres de fixer leurs honoraires mais la CAM impose de le faire avec *tact et mesure*. Enfin les 500 médecins exerçant en secteur 3 sont libres de pratiquer les tarifs qu'ils souhaitent et la CAM ne rembourse qu'un euro par consultation (Caisse nationale de l'assurance maladie des travailleurs salariés, 2012).

Hypothèse :

Les patients de la cohorte LEA qui désirent s'orienter vers un généraliste privilégient le besoin de consulter aux spécialités auxiliaires et au secteur d'activité du généraliste. *De plus, ces spécificités n'ont aucune incidence sur la qualité des soins reçus* (Auquier, 2013a).

Proposition d'une stratégie d'estimation **:**

Il s'agit de fusionner les redondances de granularité* (Saporta, 2006). Les DTA destinées à la modélisation géographique de l'accès à un généraliste sont obtenues par minimisation de toutes les DTA disponibles :

$$x_{(U_u),t}^{GEN} = \min_{k=\{1;2;3\}} \left\{ x_{(U_u),t}^{\{GENE|Secteur=k\}};\ x_{(U_u),t}^{\{GENE.MEP|Secteur=k\}};\ x_{(U_u),t}^{\{OMNIPRATICIEN|Secteur=k\}} \right\}$$

Une fois les redondances éliminées par fusion il convient d'harmoniser les échelles des indicateurs et l'échelle d'investigation retenue.

UNIFORMISATION D'ECHELLES : AGREGATION

Remarques liminaires :

Certaines DTA et APL sont déclinés à micro-échelles dans l'optique de décrire, de façon plus précise, l'accès aux soins dans les grandes villes françaises. C'est le cas des communes de **:** Marseille, découpée en 16 (Ar_u) ; Lyon - en 9 (Ar_u) ; Paris - en 20 (Ar_u).

Hypothèse :

En *analyse spatiale* les processus d'agrégation ou de désagrégation d'échelle se fondent sur des *opérateurs mathématiques ensemblistes* et incorporent un système de pondération par le biais de variables auxiliaires sociales, économiques, démographiques ou administratives (Pumain et Saint-Julien, 1997)

Principe d'estimation :

Afin d'uniformiser l'échelle des indicateurs spatiaux mobilisés et de faire en sorte qu'elle soit en adéquation avec l'échelle d'investigation retenue, les DTA et APL à l'échelle des Ar_u sont fusionnées par le biais d'une stratégie d'agrégation pondérée adaptée. Les APL et les DTA intègrent déjà des systèmes de pondération plus ou moins complexes fondés sur des dimensions géographiques : sociales, sanitaires et temporelles. Par conséquent la stratégie d'harmonisation d'échelles utilise une variable

administrative : $x_{(.)}^{\mathrm{SG}}$ la surface géographique des arrondissements $\mathrm{Ar_u}$ inclus dans la commune considérée.

$$x_{(\mathrm{U_u}),t'}^{\mathrm{l:SAN}} = \frac{\left(\mathrm{n}_{(\mathrm{Ar}_{.}|\mathrm{U_u})}\right)^{-1}}{x_{(\mathrm{U}_u)}^{\mathrm{SG}}} \cdot \sum_{\mathrm{u}=1}^{\mathrm{n}_{(\mathrm{Ar}_{.}|\mathrm{U_u})}} \left(x_{(\mathrm{Ar_u}),\mathrm{t}'}^{\mathrm{SG}} \cdot x_{(\mathrm{Ar_u}),\mathrm{t}'}^{l} \middle| \mathrm{Ar_u} \subseteq \mathrm{U_u}\right)$$

Désormais l'effet information* est optimisé. L'harmonisation de la granularité* inputs a permis d'obtenir des variables spatiotemporelles plus précises, plus adaptées à la modélisation des FE-SAN. Il est désormais question de maximiser l'effet de support*.

MAXIMISATION DE L'EFFET DE SUPPORT

La phase d'optimisation de l'effet de support* inhérent à l'estimation des i.st.e* SAN s'opère en deux temps : Le premier, est un processus d'harmonisation probabiliste des temporalités ; le second, est un appariement diachronique des i.st.e* au $\mathrm{U_k}$, dans une logique territoriale cohérente.

PROCESSUS D'HARMONISATION TEMPORELLE PROBABILISTE

Remarques liminaires :

Les DTA et APL sont déclinés pour des temporalités particulièrement récentes, respectivement 2007 et 2010, donc peu représentatives de l'accessibilité géographique dans la période recouverte par l'étude LEA, i.e. 1980 à 2010. Ces indicateurs surestiment *l'accès géographique à l'offre et à la qualité des soins sur cette période – tant il elle vrai que l'accès aux soins s'améliore avec le temps. En revanche, les disparités géographiques tendent à se pérenniser, bien que certaines disparaissent et d'autres se dessinent*. Par conséquent les DTA et les APL sont adaptés à la modélisation des variabilités spatiales de l'accès géographique aux *items sanitaires** (Lucas-Gabrielli, 2012).

En particulier, les DTA surestiment fortement l'accessibilité réelle aux soins dans la mesure où elles ont été calculées *au moment où le niveau de praticiens de santé, en France, atteignait son paroxysme*. (Coldefy, Lucas-Gabrielli et al., 2011).

Hypothèse :

L'accès à des soins de qualité s'améliore avec le temps grâce aux progrès médicaux, à la modernisation des tissus sanitaires* et à l'évolution des modes de transport, les disparités géographiques tendent à se pérenniser, lorsqu'elles ne se creusent pas (Brunet, Théry et al., 2009). Les politiques visant à les amoindrir sont fondées sur des Schémas Régionaux d'Organisation des Soins (SROS), i.e. dans une logique régionale (DILA, 2002).

Principe d'estimation :

Les indicateurs APL et a *fortiori les* DTA surestiment l'accès géographique à l'offre de soins. Afin de maximiser *l'effet du support* temporel et dans la mesure où aucune stratégie d'agrégation *verticale* n'est envisageable car les indicateurs sont déclinés à une date unique, l'intégration de la dimension temporelle est effectuée par le biais d'un processus *de randomisation Gaussien à effet modéré*, adapté à la logique spatiale des politiques sanitaires menées. Le but est de diminuer les valeurs prises par les indicateurs spatiaux (Saporta, 2006).

$$x'^{l}_{(U_u)} = x^{l}_{(U_u),t'} + \mathbb{t}_{\mathrm{u}(\alpha)} \cdot \frac{\hat{\sigma}\left(x^{l}_{(\mathrm{U}_{.},t'|U_u \subseteq \mathrm{RE}_r)}\right)}{\sqrt{card(\mathrm{U}_{.}, t'|U_u \subseteq \mathrm{RE}_r)}}$$

Avec : $x^{l}_{(U_{.},t'|U_u \subseteq \mathrm{RE}_r)}$ l'ensemble des communes de France métropolitaine contenues dans la région RE_r sachant que U_u y est incluse à la date t' ;

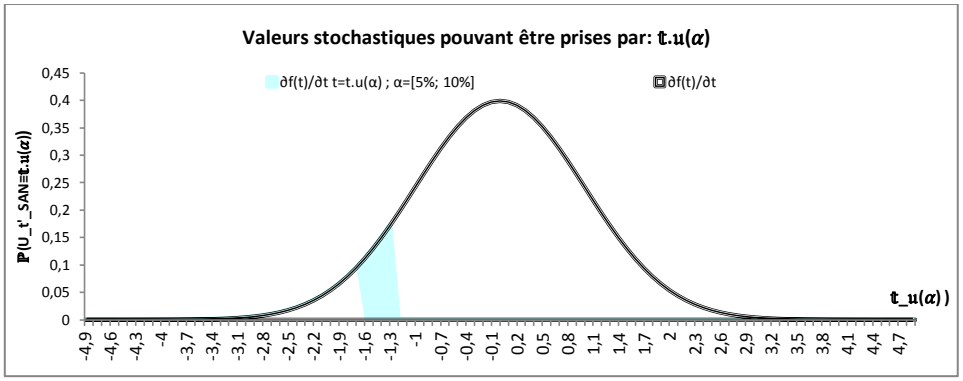


Figure 87 : Amplitude des valeurs stochastiques prises par $\mathbb{t}_{u(\alpha)}$ en fonction de $u(\alpha)$

Et avec : $\mathbb{t}_{u(\alpha)} \xrightarrow[n\to+\infty]{} U_{\mathbb{t}}^{SAN} \sim \mathcal{N}(0,1)$ tel que : $\mathbb{P}\left(U_{\mathbb{t}}^{SAN} \equiv \mathbb{t}_{u(\alpha)}\right) = \{u(\alpha) \sim \mathbb{U}[\![5\%; 10\%]\!]\}$.

Remarque particulière :

APL est déjà assujetti à un système de pondération complexe et *controversé* aux dimensions sociale, sanitaire et temporelle permettant d'estimer des secteurs flottants i.e. *des zones à cheval* sur plusieurs communes pour s'affranchir des barrières spatiales fixées par les limites administratives et qui ont une influence forte sur les valeurs prises (Harrouin, Aligon et al., 2012). Il convient donc de ne pas le complexifier encore et par extension de prendre le risque de le rendre inconsistant en le bruitant, de fait :

$$x_{(U_u)}^{l:APL} \equiv x'^{l:APL}_{(U_u),t'}$$

PROCESSUS D'HARMONISATION SPATIALE DIACHRONIQUE

Les indicateurs spatiaux ont été harmonisés à l'échelle des communes U_u, et randomisés par un processus stochastique Gaussien qui tient compte de la logique spatiale des politiques sanitaires et qui vise à les grever d'une composante temporelle adéquate avec celle de l'étude LEA.

Remarques liminaires :

Les U_u de la BD SIG géofla utilisée représentent les surfaces géographiques communales telles qu'elles étaient en 2003 (IGN, 2004).

Il y a unicité entre les codes INSEE et les unités géographiques communales mais les valeurs des codes géographiques changent au fil du temps. *Elles peuvent être appariées grâce aux tables des correspondances diachroniques des codes INSEE historiques* (INSEE, 2012b)

Hypothèse :

Les codes INSEE évoluent dans le temps et cette variabilité est conditionnée par des motifs sociaux, économiques, démographiques et politiques. Chaque année les découpages administratifs des communes françaises évolue au gré des processus de création/suppression ou de fusion/division de communes (Bellin, Morin et al., 2011).

Principe d'appariement :

Le découpage administratif des U_u évolue dans le temps, ce qui pose un problème d'interopérabilité des sources. Les $x'^{l}_{(U_u)}$ construits à partir des stratégies proposées sont appariés, dans une logique territoriale cohérente sur le plan diachronique, aux U_k afin d'obtenir les i.st.e* définitifs, ce qui s'opère de la façon suivante :

$$x'^{l:SAN}_{(U_k)} = \left[\text{môy}\left(\bigcup_{U_u=1}^{n_{U_u}} \left(x'^{l}_{(U_u)} \middle| (U_u, t|t = t') \subseteq (U_k, t|t = 2003) \right) \right) \right]$$

REMARQUES

Les i.st.e* destinés à modéliser les FE-SAN, dans les communes de 1ère espèce, sont notés $x'^{\mathrm{l:SAN}}_{(\mathrm{U}_k)}$ lorsqu'ils ne sont pas assujettis au processus de randomisation temporelle, c'est le cas de ceux obtenus à partir des DTA 2007. En revanche, ils sont notés $x^{\mathrm{l:SAN}}_{(\mathrm{U}_k)}$ lorsque ce n'est pas le cas, comme pour ceux obtenus à partir des APL 2010.

D'une manière générale, les i.st.e* construits à partir de variables quantitatives discrètes sont arrondis à l'entier le plus proche. Et ceux construits avec des variables quantitatives continues sont arrondis en fonction de la précision granulaire des inputs utilisés – i.e. les i.st.e* issus des APL 2010 sont arrondis à trois chiffres après la virgule.

Les stratégies pour minimiser le bais conditionnel des indicateurs spatiaux inputs utilisés par des i.st.e* $x'^{\mathrm{l:SAN}}_{(\mathrm{U}_k)}$ représentatifs de la variabilité géographique des *FE-SAN* pertinents* et Curieux** et à l'identification de DES, ont été appliquées de façon à garantir leur caractère reproductible.

PRESENTATION DES RESULTATS ET REMARQUES

Les résultats cartographiques de la modélisation géographique des FE-SAN* *pertinents* et Curieux**, par le biais des i.st.e* $x'^{\mathrm{l:SAN}}_{(\mathrm{U}_k)}$ et $x^{\mathrm{l:APL_SAN}}_{(\mathrm{U}_k)}$proposés représentent, la variabilité spatiale des expositions environnementales géographiques potentielles liée à l'accès géographique aux soins.

Ces expositions sont déclinées par des distances temps en minutes d'accès par la route à tous les items sanitaires* susceptibles d'avoir une interaction avec les PM* – séquelles. Et pour les médecins généralistes et les ophtalmologues, elles sont aussi caractérisées en termes d'Accès Potentiel Localisé* – exprimé pour mille habitants.

Les résultats cartographiques sont présentés dans les $\mathrm{U_k}$ situées en région PACA et aux alentours car leur analyse ne permet pas d'identifier des DES. Le seul moyen pour y parvenir est de mettre simultanément en perspective tous les i.st.e* et les i.st.m* proposés, et d'utiliser une procédure mathématique adaptée (chapitre.4).

L'indicateur $\mathrm{SpaLea}^2_{\mathrm{U_{k_o}}}$ représentant l'incertitude spatiale associée à l'identification des U_k inhérente aux hypothèses de la méthode SpaLea et à la granularité* des données utilisées n'est pas représenté pour ne pas surcharger l'affichage.

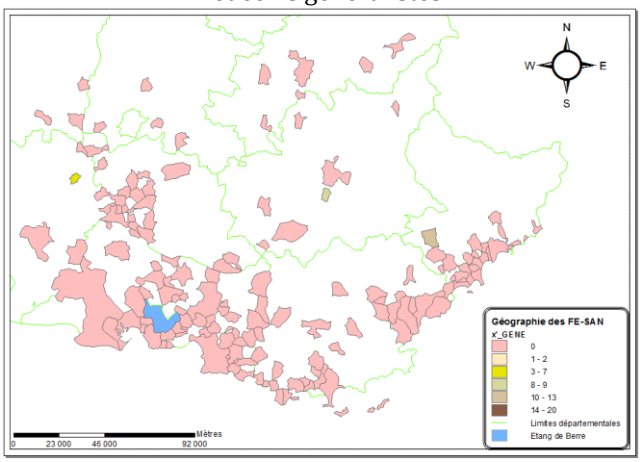
La documentation des cartes décline l'i.st.e* $x'^{\mathrm{l}}_{(\mathrm{U}_k)}$ présenté et le type de variabilité spatiale modélisée. Un tableau statistique accompagne chacune d'entre elles. Il décline les principaux paramètres *de position et de dispersion*, i.e. la moyenne :$\mathrm{môy}\left(x'^{\mathrm{l}}_{(\mathrm{U}_k)}\right)$, l'estimateur biaisé de l'écart-type : $\widehat{\sigma}\left(x'^{\mathrm{l}}_{(\mathrm{U}_k)}\right)$; ainsi que les trois premiers quartiles : $\widehat{\mathrm{Q}}_1\left(x'^{\mathrm{l}}_{(\mathrm{U}_k)}\right)$, $\mathrm{mêd}\left(x'^{\mathrm{l}}_{(\mathrm{U}_k)}\right)$; $\widehat{\mathrm{Q}}_3\left(x'^{\mathrm{l}}_{(\mathrm{U}_k)}\right)$. Ils sont estimés sur l'ensemble des U_k.

Des remarques sont portées uniquement lorsque des singularités spatiales sont observables.

CARTOGRAPHIES ET STATISTIQUES DE LA GEOGRAPHIE DES FE-SAN

Géographie de l'accessibilité aux items sanitaires* territoriaux

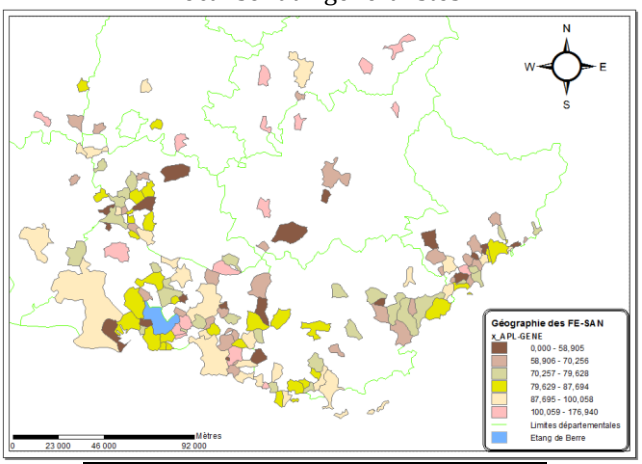
Variabilités spatiotemporelles de l'accès temporel aux médecins généralistes



Statistique	i.st.e $:x'^{\mathrm{GENE}}_{(U_k)}$ (min)
môy(·)	0,38
$\widehat{\sigma}(\cdot)$	2,052
$\widehat{Q}_1(\cdot)$	0
mêd(·)	0
$\widehat{Q}_3(\cdot)$	0

Figure 88 : i.st.e* : $x'^{\mathrm{GENE}}_{(U_k)}$: Distance temps moyenne d'accès par la route à un médecin généraliste

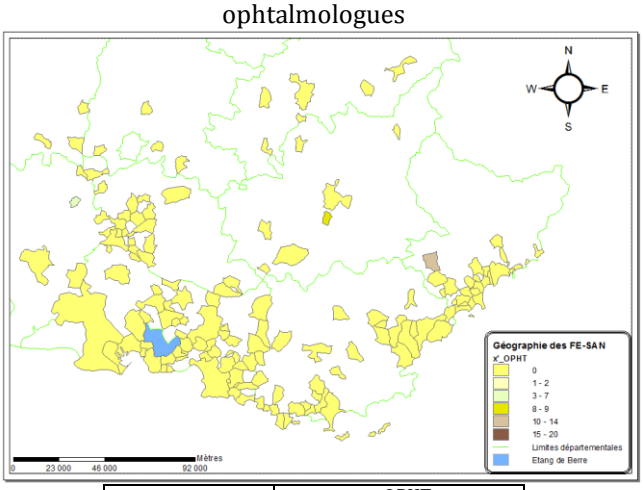
Variabilités spatiotemporelles de l'Accès Potentiel Localisé* aux généralistes



Statistique	i.st.e $:x^{\mathrm{APL\_GENE}}_{(U_k)}$(s.u.)
môy(·)	79,96
$\widehat{\sigma}(\cdot)$	24,346
$\widehat{Q}_1(\cdot)$	65,00
mêd(·)	79,63
$\widehat{Q}_3(\cdot)$	92,68

Figure 89 : i.st.e* : $x^{\mathrm{APL\_GEN}}_{(U_k)}$: APL moyenne communale à un médecin généraliste, non MEP, pour 1000 habitants

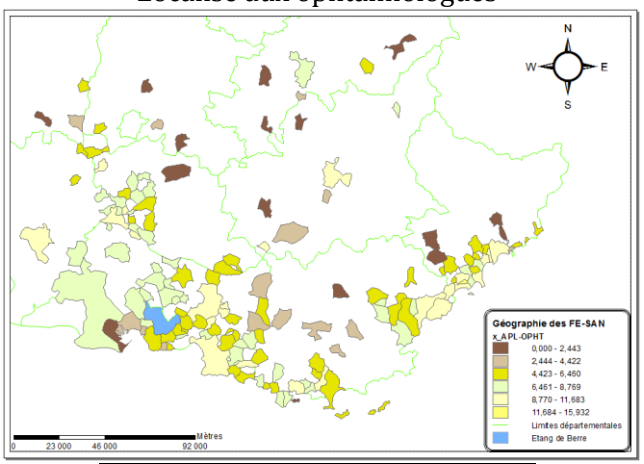
Variabilités spatiotemporelles de l'accès temporel aux ophtalmologues



Statistique	i.st.e $:x'^{\mathrm{OPHT}}_{(U_k)}$ (min)
môy(·)	0,41
$\widehat{\sigma}(\cdot)$	2,067
$\widehat{Q}_1(\cdot)$	0
mêd(·)	0
$\widehat{Q}_3(\cdot)$	0

Figure 90 : i.st.e* : $x'^{\mathrm{ORL}}_{(U_k)}$: Distance temps moyenne d'accès par la route à un ophtalmologue

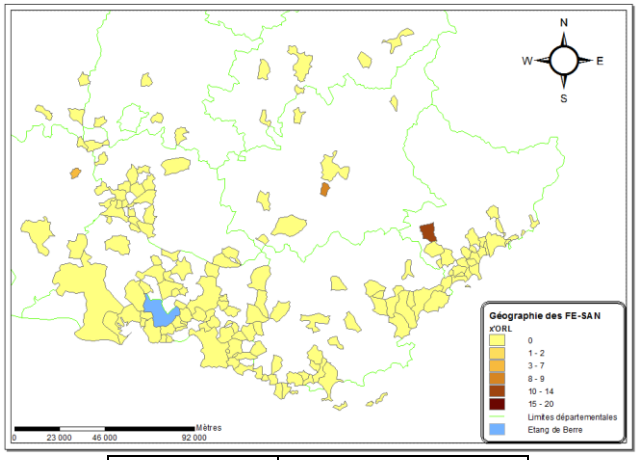
Variabilités spatiotemporelles de l'Accès Potentiel Localisé aux ophtalmologues



Statistique	i.st.e $:x^{\mathrm{APL\_OPHT}}_{(U_k)}$ (s.u.)
môy(·)	5,65
$\widehat{\sigma}(\cdot)$	3,182
$\widehat{Q}_1(\cdot)$	3,20
mêd(·)	5,37
$\widehat{Q}_3(\cdot)$	7,53

Figure 91 : i.st.e* : $x^{\mathrm{APL\_ORL}}_{(U_k)}$: APL moyenne communale à un d'otorhino laryngologue pour 1000 habitants

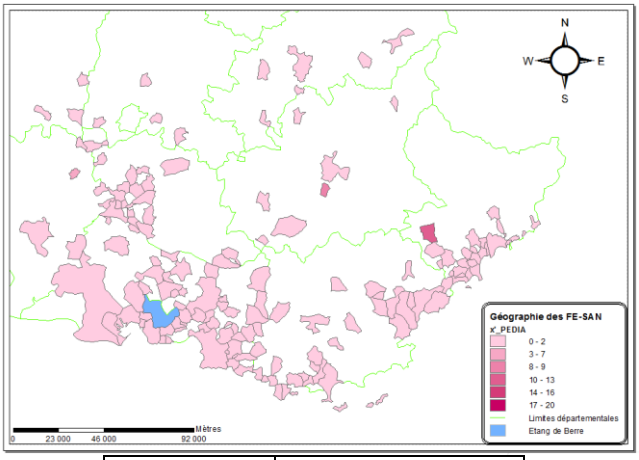
Variabilités spatiotemporelles de l'accès temporel aux ORL



Statistique	i.st.e $:x'^{\mathrm{ORL}}_{(U_k)}$ (min)
môy(·)	0,41
$\widehat{\sigma}(\cdot)$	2,067
$\widehat{Q}_1(\cdot)$	0
mêd(·)	0
$\widehat{Q}_3(\cdot)$	0

Figure 92 : i.st.e* : $x'^{\mathrm{ORL}}_{(U_k)}$: Distance temps moyenne d'accès par la route à un otorhino laryngologue

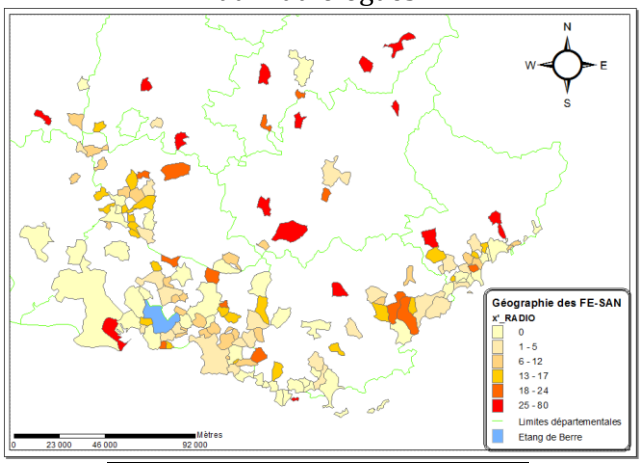
Variabilités spatiotemporelles de l'accès temporel aux pédiatres



Statistique	i.st.e $:x'^{\mathrm{PEDIA}}_{(U_k)}$ (min)
môy(·)	0,47
$\widehat{\sigma}(\cdot)$	2,403
$\widehat{Q}_1(\cdot)$	0,00
mêd(·)	0,00
$\widehat{Q}_3(\cdot)$	0,00

Figure 93 : i.st.e* : $x'^{\mathrm{PEDIA}}_{(U_k)}$: Distance temps moyenne d'accès par la route à un pédiatre

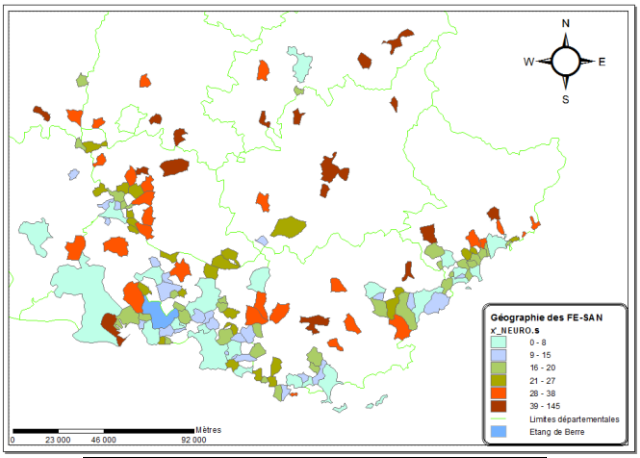
Variabilités spatiotemporelles de l'accès temporel aux radiologues



Statistique	i.st.e $:x'^{\mathrm{RADIO}}_{(U_k)}$ (min)
môy(·)	12,05
$\widehat{\sigma}(\cdot)$	12,624
$\widehat{Q}_1(\cdot)$	1,00
mêd(·)	11,00
$\widehat{Q}_3(\cdot)$	18,00

Figure 94 : i.st.e* : $x'^{\mathrm{RADIO}}_{(U_k)}$: Distance temps moyenne d'accès par la route à un radiologue

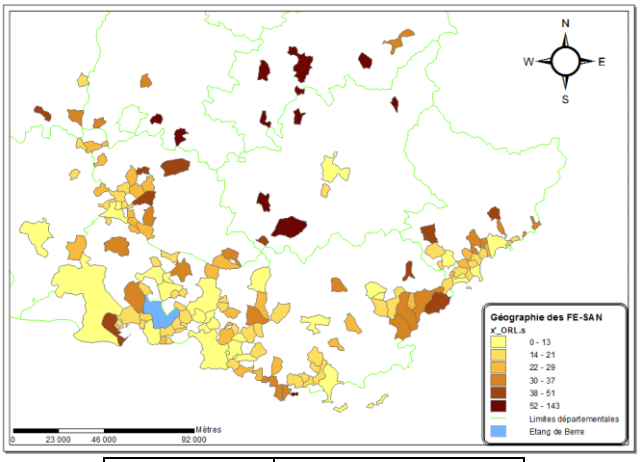
Variabilités spatiotemporelles de l'accès temporel à un service de neurologie



Statistique	i.st.e $:x'^{\mathrm{NEUROs}}_{(U_k)}$ (min)
môy(·)	23,03
$\widehat{\sigma}(\cdot)$	17,739
$\widehat{Q}_1(\cdot)$	12,50
mêd(·)	20,00
$\widehat{Q}_3(\cdot)$	31,00

Figure 95 : i.st.e* : $x'^{\mathrm{NEUROs}}_{(U_k)}$: Distance temps moyenne d'accès par la route à un service hospitalier de neurologie

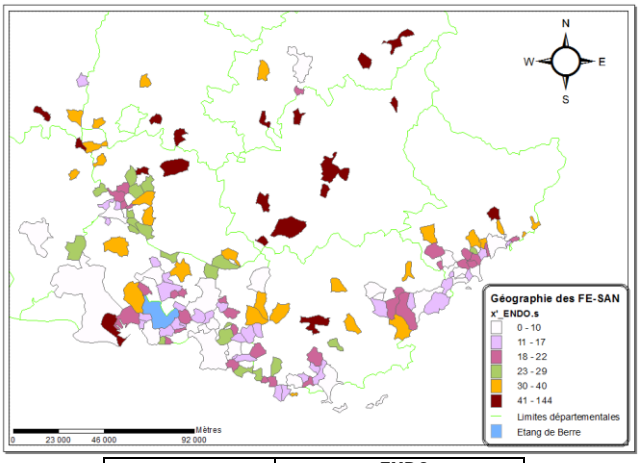
Variabilités spatiotemporelles de l'accès temporel à un service ORL



Statistique	i.st.e $:x'^{\mathrm{ORLs}}_{(U_k)}$ (min)
môy(·)	30,87
$\hat{\sigma}(\cdot)$	20,218
$\hat{Q}_1(\cdot)$	17,00
mêd(·)	29,00
$\hat{Q}_3(\cdot)$	41,00

Figure 96 : i.st.e* : $x'^{\mathrm{ORLs}}_{(U_k)}$: Distance temps moyenne d'accès par la route à un service hospitalier d'otorhinolaryngologie

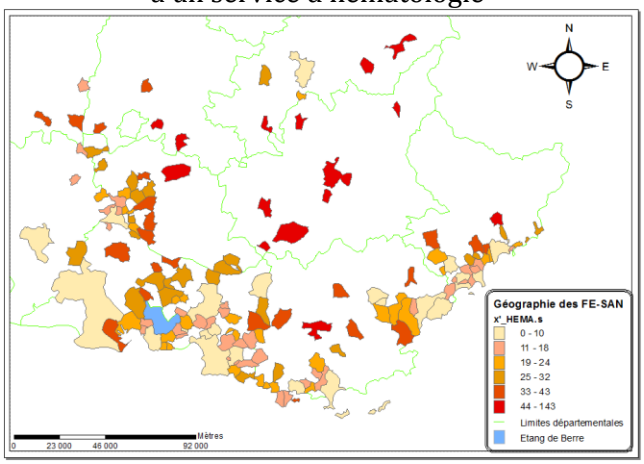
Variabilités spatiotemporelles de l'accès temporel à un service d'endocrinologie



Statistique	i.st.e $:x'^{\mathrm{ENDOs}}_{(U_k)}$ (min)
môy(·)	24,71
$\hat{\sigma}(\cdot)$	18,378
$\hat{Q}_1(\cdot)$	14,00
mêd(·)	21,00
$\hat{Q}_3(\cdot)$	33,00

Figure 97 : i.st.e* : $x'^{\mathrm{NEUROs}}_{(U_k)}$: Distance temps moyenne d'accès par la route à un service hospitalier d'endocrinologie

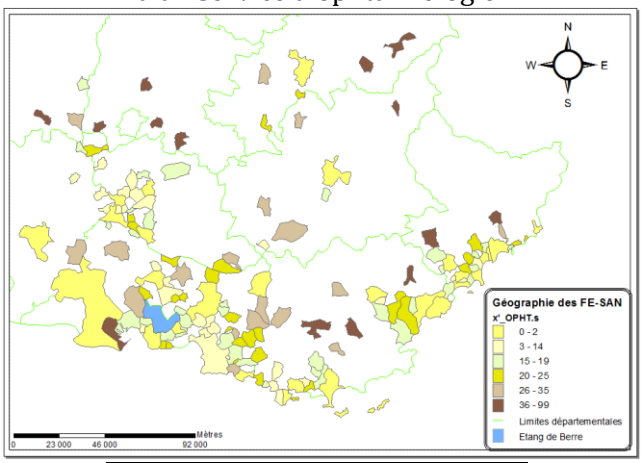
Variabilités spatiotemporelles de l'accès temporel à un service d'hématologie



Statistique	i.st.e $:x'^{\mathrm{HEMAs}}_{(U_k)}$ (min)
môy(·)	25,65
$\hat{\sigma}(\cdot)$	18,351
$\hat{Q}_1(\cdot)$	15,00
mêd(·)	23,00
$\hat{Q}_3(\cdot)$	35,00

Figure 98 : i.st.e* : $x'^{\mathrm{HEMAs}}_{(U_k)}$: Distance temps moyenne d'accès par la route à un service hospitalier d'hématologie

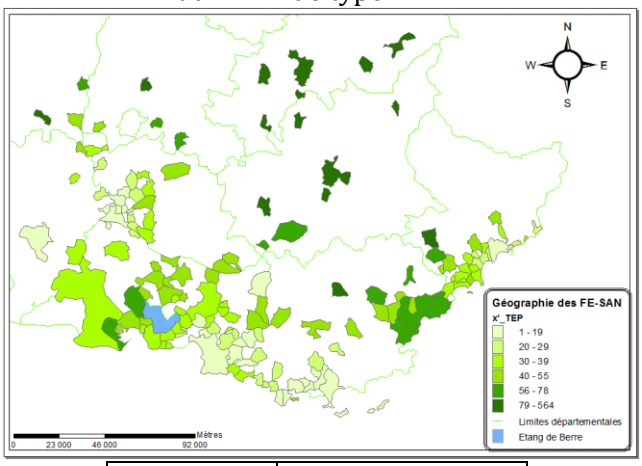
Variabilités spatiotemporelles de l'accès temporel à un service d'ophtalmologie



Statistique	i.st.e $:x'^{\mathrm{OPTHs}}_{(U_k)}$ (min)
môy(·)	20,30
$\hat{\sigma}(\cdot)$	15,776
$\hat{Q}_1(\cdot)$	10,00
mêd(·)	19,00
$\hat{Q}_3(\cdot)$	27,00

Figure 99 : i.st.e* : $x'^{\mathrm{NEUROs}}_{(U_k)}$: Distance temps moyenne d'accès par la route à un service hospitalier d'ophtalmologie

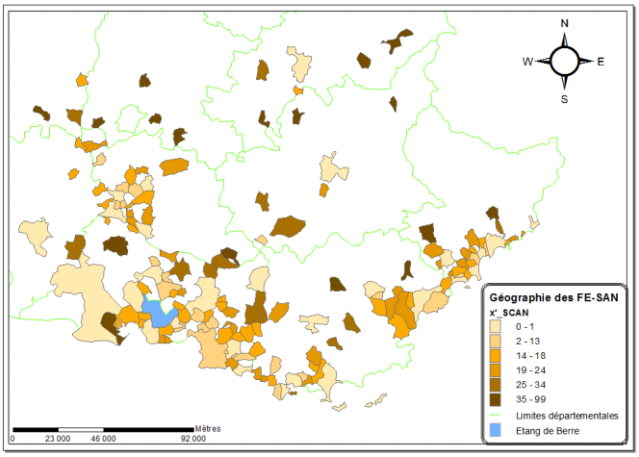
Variabilités spatiotemporelles de l'accès temporel aux ELM de type : TEP



Statistique	i.st.e $:x'^{\mathrm{TEP}}_{(\mathrm{U}_k)}$ (min)
môy(·)	60,00
$\widehat{\sigma}(\cdot)$	86,650
$\widehat{\mathrm{Q}}_1(\cdot)$	23,50
mêd(·)	37,00
$\widehat{\mathrm{Q}}_3(\cdot)$	65,00

Figure 100 : i.st.e* : $\mathrm{x}'^{\mathrm{TEP}}_{(\mathrm{U_k})}$: Distance temps moyenne d'accès par la route à un tomographe par émission de positons

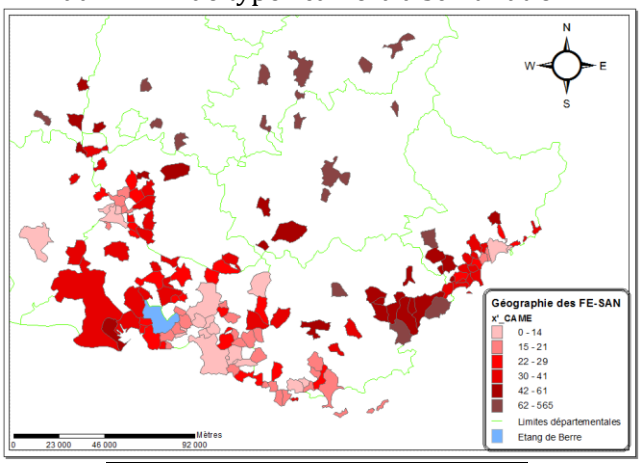
Variabilités spatiotemporelles de l'accès temporel aux EML* de type : Scanner*



Statistique	i.st.e $:x'^{\mathrm{SCAN}}_{(\mathrm{U}_k)}$ (min)
môy(·)	19,16
$\widehat{\sigma}(\cdot)$	14,866
$\widehat{\mathrm{Q}}_1(\cdot)$	10,00
mêd(·)	18,00
$\widehat{\mathrm{Q}}_3(\cdot)$	26,00

Figure 101 : i.st.e* : $x'^{SCAN}_{(U_k)}$: Distance temps moyenne d'accès par la route à un Scanner*

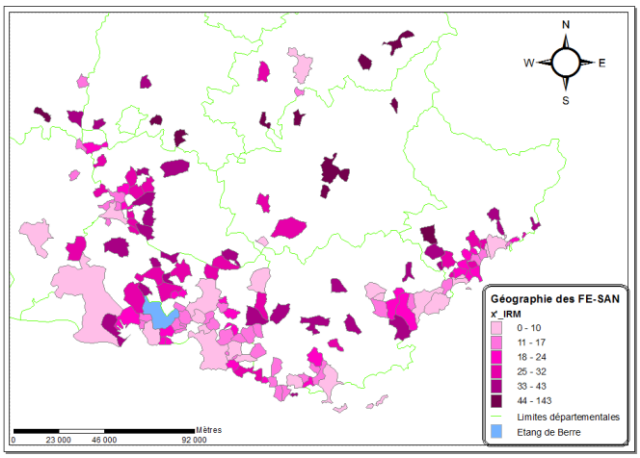
Variabilités spatiotemporelles de l'accès temporel aux EML* de type : caméra à scintillation



Statistique	i.st.e $:x'^{\mathrm{CAME}}_{(\mathrm{U}_k)}$ (min)
môy(·)	49,26
$\widehat{\sigma}(\cdot)$	86,897
$\widehat{\mathrm{Q}}_1(\cdot)$	18,00
mêd(·)	28,00
$\widehat{\mathrm{Q}}_3(\cdot)$	50,50

Figure 102 : i.st.e* : $x'^{CAME}_{(U_k)}$: Distance temps moyenne d'accès par la route à un caméra à scintillation

Variabilités spatiotemporelles de l'accès temporel aux EML* de type : IRM



Statistique	i.st.e $:x'^{\mathrm{IRM}}_{(\mathrm{U}_k)}$ (min)
môy(·)	24,54
$\widehat{\sigma}(\cdot)$	18,61
$\widehat{\mathrm{Q}}_1(\cdot)$	14,00
mêd(·)	23,00
$\widehat{\mathrm{Q}}_3(\cdot)$	34,00

Figure 103 : i.st.e* : $x'^{IRM}_{(U_k)}$: Distance temps moyenne d'accès par la route à un a appareil d''imagerie par résonance magnétique

REMARQUES

Les résultats cartographiques ainsi que les tableaux statistiques qui les complètent permettent de montrer que :
Les $x'^{l:SAN}_{(U_k)}$ estimés à partir des DTA modélisent des disparités géographiques en matière d'accès aux services hospitaliers et aux EML*. En revanche, ce n'est pas le cas pour ceux caractérisant l'accès aux praticiens libéraux, qui semble être à la fois uniforme et excellent sur l'intégralité du territoire, à l'exception des radiologues (RADIO) – pour lesquels les $x'^{l:SAN}_{(U_k)}$ sont *a priori* adéquates.

Ce constat va à l'encontre des prénotions documentées en géographie de la santé et en épidémiologie spatiale. Or les communes situées dans l'arrière-pays de la région PACA sont pourtant de *vrais déserts médicaux*. Le principe d'estimation des DTA porte à zéro la valeur de celle-ci dès lors que la spécialité libérale est repérée dans la commune. Cette stratégie pouvait être utilisée dans le passé mais elle n'est plus d'actualité. Et de fait, les $x'^{l:SAN}_{(U_k)}$ modélisant la géographie de l'accès aux praticiens libéraux, estimées à partir des DTA, sont inconsistantes. De plus, les DTA ne tiennent pas compte non plus des spécificités géo-démographiques et géo-sociales des territoires telles que l'influence de la quantité de personnes âgées sur les délais d'obtention d'un rendez-vous, les horaires d'ouverture des cabinets ou les déplacements extra-communaux.

Les $x^{l:APL\_SAN}_{(U_k)}$ estimés à partir des APL permettent de pallier les lacunes énoncées. En dépit des controverses soulevées au sujet des poids de l'indicateur APL (chapitre 1), ils modélisent avec une grande acuité les disparités communales d'accès aux praticiens libéraux. Mais ils ne sont disponibles que pour les Médecins généralistes et les ophtalmologues

Les i.st.e* représentatifs des FE-SAN* surestiment l'accès géographique à l'offre de soins territoriale de 1980 à 2010, la période couverte par LEA. Cependant, même si les valeurs sont un temps surévaluées, ils restent représentatifs des disparités géographiques de l'accès aux soins.
Le principal reproche que l'on peut faire aux i.st.e, qu'ils soient estimés à partir des DTA ou des APL, est de supposer *une accessibilité équivalente à l'intérieur d'une même zone géographique* (Talen et Anselin, 1998). Mais cette remarque touche autant à la granularité* d'échelle des données sources qu'à celle de l'échelle d'investigation retenue.

L'intégration conjointe des $x'^{\mathrm{l:SAN}}_{(U_k)}$ pour les généralistes et les ophtalmologues estimés à partir des DTA et des $x^{\mathrm{l:SAN}}_{(U_k)}$ estimés à partir des APL, dans la procédure d'identification des DES, n'induit *a priori aucune redondance puisque* $x'^{\mathrm{l:SAN}}_{(U_k)}$ sont inconsistants d'un point de vue statistique (Saporta, 2006).
Il s'agit désormais de décliner les stratégies de caractérisation spatiotemporelle des expositions géographiques environnementales potentielles aux FE-SOCIO.ECO* pertinents* et Curieux* intégrables.

SECTION C) FACTEURS ENVIRONNEMENTAUX SOCIO-ECONOMIQUES

Les Facteurs Environnementaux* (FE) socio-économiques (FE-SOCIO.ECO) constituent la dimension *a-spatiale* de l'accès géographique aux soins. Il s'agit d'un concept interdisciplinaire où s'enchevêtrent des dimensions sociales, économiques et géographiques (Litva et Eyles, 1995). Le comportement *vis-à-vis du recours aux soins* ne dépend pas uniquement des caractéristiques comportementales, financières et médicales des individus. L*'effet de contexte représente* la dimension *collective* du milieu de vie et de ses répercussions *directes ou indirectes* sur les croyances et les conduites individuels vis-à-vis du recours aux soins. (Chaix, Merlo et al., 2005). La dimension *collective* de conjonctures socio-économiques *défavorables* influence l'état de santé général des populations et conditionne *les distances individuelles perçues*, *économiques, sociales, culturelles* d'accès aux soins. La *spécialisation socio-économique des territoires* peut induire *des conduites à risques*, *du stress contextuel* ou des expositions à des substances toxiques liées e.g. à l'industrie ou l'agriculture – on parle alors *d'espaces géographiques prédisposant aux phénomènes morbides* (Haddad, 1992).
Il existe une pléthore d'indicateurs spatiaux dans littérature. Les plus classiques sont simples, robustes et caractérisent des conjonctures géographiques globales d'ordre culturel, social et économique (Powell, 1995). D'autres sont plus spécifiques et quantifient la variabilité spatiale des niveaux de vie ou de la répartition des richesses (Carstairs et Morris, 1989). Enfin, des indices atypiques proposent d'évaluer des distances a-spatiales *psychologiques d'accès aux soins* (Benach et Yasui, 1999) ou des niveaux *de défaveur sociale* à partir de transformations topologiques de données géographiques socio-économiques (Townsend, 1987)

En géographie de la santé et en épidémiologie spatiale, le principe est le même, la variabilité spatiale de *l'effet de contexte* s'évalue par le biais d'indicateurs spatiaux conjoncturels construits à partir de données étatiques.
La modélisation géographique est conditionnée par la qualité des indicateurs géographiques conjoncturels disponibles. Les BD retenues pour modéliser la géographie des FE-SOCIO.ECO* pertinents* et Curieux* contiennent des variables disponibles sur l'intégralité du territoire français métropolitain et dont la granularité*, i.e. l'échelle, les temporalités, la précision, le niveau de lacunes, sont en adéquation avec la problématique, le positionnement scientifique et les hypothèses inhérentes à cette recherche.

En l'occurrence il s'agit de la BD de l'Institut National de la Statistique et des Etudes Economiques (INSEE) et plus particulièrement de la sous-base : *Données Locales*. Les indicateurs géographiques sont disponibles à différentes dates et décrivent le panorama démographique, social et économique des communes de France métropolitaine (INSEE, 2012c), à l'exception de quelques rares variables contenant des lacunes, dans des zones couvertes par le secret statistique (INSEE, 2013). Les indicateurs INSEE sont des données spatiotemporelles étatiques particulièrement fiables. Elle ont été mobilisées pour construire des i.st.e* permettant de modéliser la géographie des *FE-SOCIO.ECO* pertinents** représentatifs : des comportements spatiaux vis-à-vis du recours aux soins, des efforts politiques de durabilité et de leurs répercussions sur les attraits sociaux, économiques et sanitaires des territoires, des expositions potentielles à des substances toxiques induites par la spécialisation socio-professionnelle ou économique des territoires, et des prédispositions contextuelles aux phénomènes morbides.
Quant à la BD de l'Observatoire National de la Délinquance et des Réponses Pénales (ONDRP), elle contient des indicateurs géographiques composés d'index spatiaux *thématiques représentatifs des infractions* territoriales disponibles à différentes dates et avec le pas de temps est régulier, sur l'intégralité de la France Métropolitaine (ONDRP, 2011). Seul bémol, l'échelle la plus fine est celle des départements français. Les variables ONDRP sont utilisées pour modéliser la géographie des *FE-SOCIO.ECO* Curieux** par les niveaux de stress potentiellement perçu et induit par l'insécurité territoriale contextuelle. Le sentiment de stress potentiellement perçu ou subi par la population peut

être d'origine socio-contextuelle et ses effets physiologiques sont particulièrement bien documentés, e.g. les troubles de l'humeur et de l'anxiété (Godin, Kittel, et al., 2005), ils ont une influence néfaste sur l'état de santé général des populations. Cependant la nature et la granularité* d'échelle des données disponibles confèrent aux variables ONDRP une Distance a-spatiale morbide* assez éloignée des PM* d'intérêt.

Les variabilités spatiales des FE-SOCIO.ECO* *pertinents* et Curieux* de test* seront modélisées par des i.st.e* notés : $x_{(U_k)}^{l:SOCIO.ECO}$, et les stratégies d'intégration spatiotemporelle des données mobilisées sont déclinées subséquemment.

PROPOSITIONS HEURISTIQUES D'INTEGRATION DES FE-SOCIO.ECO

Objectif :
Modéliser dans les communes de 1ère espèce U_k la géographie des FE-SOCIO.ECO* à partir des données mobilisées en utilisant toutes leurs caractéristiques granulaires. Les i.st.e* $x_{(U_k)}^{SOCI.ECO}$ proposés doivent être représentatifs de la réalité géographique des FE-SOCIO.ECO, i.e. adaptés à la finalité cherchée qui est l'identification des DES* à partir d'une procédure de sélection de variables multidimensionnelle au regard des i.st.m* $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$ proposés pour modéliser la géographie des PM* d'intérêt.

Remarques liminaires :
Parmi la pléthore de BD existantes ce sont celles de l'INSEE et de l'ONDRP qui ont été retenues pour la fiabilité des indicateurs géographiques qu'elles contiennent, leur capacité à modéliser les FE-SOCIO.ECO* pertinents* et Curieux* et leur disponibilité sur l'intégralité de la France métropolitaine. Toutefois, ces indicateurs ne sont pas directement intégrables. Leur granularité* doit être optimisée, certains doivent être fusionnés afin de réduire leurs biais géographiques, leurs redondances…
Certaines variables INSEE contiennent des lacunes CNIL. C'est le cas par exemple des revenus médians ou des indices de GINI qui ne sont pas disponibles dans certaines unités géographiques couvertes par le secret statistique.
Les indicateurs géographiques INSEE, à l'instar des variables DREES, sont disponibles à l'échelle des communes sauf pour trois U_k - Marseille Lyon et Paris - où ils sont déclinés à l'échelle des arrondissements(Ar_u). Par contre, les index temporels spatialisés ONDRP sont disponibles uniquement à l'échelle des départements(De_u).
S'agissant des temporalités : Les $x_{(U_u),\{t'=t.INSEE\}}^{l:INSE}$ mobilisés sont des chroniques temporelles dont le pas de temps est soit annuel, soit variable – avec des *sauts* - et les périodes recouvertes sont variables. Quant aux $x_{(De_u),\{t'=t.ONDRP\}}^{l:ONDRP}$ mobilisés, ils sont déclinés annuellement sur la période 1996 à 2011.
Enfin, les variables INSEE déclinées à l'échelle des communes U_u ne peuvent pas être directement appariées aux U_k. de la base SIG géofla 2003 puisque des limites territoriales varient au cours du temps

Hypothèse principale :
Les variables INSEE et ONDRP utilisées pour confectionner les i.st.e* $x_{(U_k)}^{SOCI.ECO}$ doivent être soumises à des traitements spatiotemporels minimisant le concept de *biais conditionnel** afin de modéliser le mieux possible la réalité géographique des expositions environnementales aux FE-SOCIO.ECO* (Marcotte, 2008). Les stratégies d'intégration spatiotemporelles proposées se fondent sur des fonctions mathématiques permettant d'effectuer des transformations topologiques et statistiques *et qui incorporent des connaissances géographiques expertes*. Il s'agit d'un processus en deux étapes qui consiste d'abord à optimiser *l'effet information**, i.e. amoindrir les incertitudes et les imprécisions des données sources utilisées, et ensuite, à *maximiser l'effet de support**, i.e. maximiser leurs caractéristiques granulaires spatiotemporelles afin qu'elles soient adaptées à la finalité recherchée (Baillargeon, 2005).

Proposition méthodologique principale :
La stratégie d'estimation des i.st.e* destinés à modéliser la géographie des FE-SOCIO.ECO* comporte une *d'optimisation de l'effet information** – qui se compose : *d'une stratégie comblement des lacunes* par des statistiques consistantes (Liaw et Wiener, 2006) adaptées à logique territoriale (Charre, 1995), des processus de *fusion statistique visant à éliminer les redondances typologiques* (Saporta, 2006)*, supprimer les biais de masse* en ramenant les variables de type dénombrement à la nature et à l'échelle des objets géographiques ciblés (Lahousse et Piédanna, 1998) *et procéder à des agrégations topologiques, dans l'idée de la théorie des ensembles flous, pour construire des variables en adéquation avec les FE-SOCIO.ECO* retenus* (Dubois Didier, Prade Henri, 2004). Par ailleurs, il s'agit de proposer des techniques *d'agrégation ou de désagrégation spatiale, pondérées géographiquement,* afin d'harmoniser les échelles (Pumain et Saint-Julien, 1997).
Quant à la phase de maximisation de *l'effet de support*, elle* repose sur une stratégie d'agrégation *verticale* des *variables spatiotemporelles optimisées* (Peguy, 1996), qui se fonde sur le concept de *stationnarité temporelle apparente* (Lütkepohl, 1991) avec une statistique représentative choisie à partir d'un processus de décision probabiliste (Saporta, 2006) *;* et sur une procédure d'appariement, conçue dans une logique diachronique et territoriale cohérente.

Les stratégies d'intégration des variables INSEE et ONDRP ont pour but de proposer des i.st.e* $x_{(U_k)}^{l:SOCIO.ECO}$ robustes. Il s'agit d'un processus en deux étapes qui consiste d'abord à *optimiser l'effet information**.

PRINCIPE DES STRATEGIES D'INTEGRATION DES VARIABLES SPATIOTEMPORELLES

Afin de minimiser le biais conditionnel* des $x_{(U_k)}^{l:SOCIO.ECO}$ les stratégies proposées consistent à optimiser l'effet information* des variables INSEE et ONDRP mobilisées en commençant par combler les lacunes, puis en améliorant par des processus de fusion leurs caractéristiques granulaires informationnelles, et enfin les uniformiser à l'échelle des U_k.

OPTIMISATION DE L'EFFET INFORMATION

La première étape est une stratégie de comblement des lacunes adaptée et applicable à toutes les variables spatiotemporelles mobilisées, i.e. à celles utilisées pour les FE-SOCIO.ECO* mais aussi à celles utilisées pour la modélisation des FE-PHY.CHIM.

STRATEGIE TERRITORIALE STATISTIQUE DE COMBLEMENT DES LACUNES

Remarques liminaires :
Les variables ONDRP ne contiennent pas de lacune. Et d'une manière générale, toutes les variables mobilisées pour la modélisation des FE, dont celles de l'INSEE font partie, ont été choisies car elles ont un niveau de lacunes spatiales inférieur à 30%. De plus, toutes les variables lacunaires mobilisées dans le cadre de cette recherche sont de nature quantitative.
La loi n°78-17 du 6 janvier 1978 modifiée interdit à l'INSEE de diffuser la valeur de certaines variables $x_{(U_u),t.INSEE}^{l:INSEE}$ dans les unités géographiques couvertes par le secret statistique – pour ne pas porter atteinte à la vie privée des populations *in situ* (INSEE, 2013).
Parmi les variables mobilisées, celles concernées sont le revenu fiscal médian déclaré par les ménages $x_{(U_u),t}^{RevMed.m}$ et le revenu fiscal médian déclaré par personne $x_{(U_u),t}^{RevMed.p}$ dans U_u comptant moins de 50 ménages, ainsi que les indices de Gini estimés par ménage $x_{(U_u),t}^{Gini.m}$ et par personne $x_{(U_u),t}^{Gini.p}$ dans les U_u de moins de 2000 habitants, au moment du recensement de 1999 (INSEE, 2010).

Spécification de l'hypothèse :

La stratégie *de comblement* des lacunes doit être adaptée à la nature des variables et doit utiliser une statistique *consistante* (Liaw et Wiener, 2006). De plus, comme la dialectique est géographique, elle doit être en adéquation avec l'échelle d'investigation, i.e. pensée dans logique statistique territoriale (Charre, 1995) et par extension – pour ce qui est de la modélisation de FE-SOCIO.ECO* - intégrer le fait que les variabilités spatiales conjoncturelles sont intimement liées aux tendances et aux actions des politiques municipales, lesquelles sont généralement indépendantes de celles menées dans les communes voisines, mais néanmoins orientées, à des échelles départementales, par les cycles gouvernementaux et leurs directives déconcentrées (Bailly et Beguin, 2005).

Proposition d'une stratégie d'estimation :

On choisit l'estimateur de la moyenne des valeurs communales d'un même département De_u si la *variabilité spatiale apparente* des $x^{l}_{(U_u),t}$ ne biaise pas l'estimateur de l'opérateur espérance. Autrement dit, tant que le nombre d'extrema spatiaux n'influence pas de façon exagérée la statistique proposée, soit encore que le nombre de valeurs caractérisées d'*anormalement* erratiques, avec un niveau statistique de risque α bilatère et élevé, est inférieur à $\{\mathbb{S} = 20\%\}$.

Dans le cas contraire, et puisque par hypothèse les conjonctures socio-économiques sont conditionnées par les politiques municipales, qu'elles sont indépendantes de celles des communes voisines, mais néanmoins influencées par un contexte plus général – l'alternative consiste à utiliser l'estimateur de la médiane que lorsque le nombre d'extrema est très grand - supérieur à 50% des effectifs (Saporta, 2006) – et qu'il convient de toute façon de prendre en compte dans cette configuration particulière de choses.

$$\left\{x^{l}_{(U_u|De_u),t} \neq ?\right\}\begin{cases} \text{môy}\left(\left\{x^{l}_{(U_u\subseteq De_u),t} \neq ?\right\}\right) & \text{si :} \quad \text{card}\left(\bigcup_{u=1}^{n^{l}_{(U_.|De_u,t)}} \left(\left\{x^{l}_{(U_u\subseteq De_u),t} \neq ?\right\}\right) \notin {}^{(\pm)}_{\mathbb{S}}\psi^{l}_{(De_u),t}\right) \leq N^{l}_{\mathbb{S},t} \\ \text{mêd}\left(\left\{x^{l}_{(U_u\subseteq DEP_d),t} \neq ?\right\}\right) & \text{si:} \quad \text{card}\left(\bigcup_{u=1}^{n^{l}_{(U_.|De_u,t)}} \left(\left\{x^{l}_{(U_u\subseteq De_u),t} \neq ?\right\}\right) \notin {}^{(\pm)}_{\mathbb{S}}\psi^{l}_{(De_u),t}\right) > N^{l}_{\mathbb{S},t} \end{cases}$$

Avec ; $\left\{x^{l}_{(U_u|De_u),t} \neq ?\right\}$ une lacune ; $n^{l}_{(U_.|De_u,t)}$ le nombre de communes U_u dans le département u à l'année t ; $N^{l}_{\mathbb{S},t}$ le nombre d'extrema spatiaux caractérisant une *variabilité spatiale apparente anormalement erratique ;* ${}^{(\pm)}_{\mathbb{S}}\psi^{l}_{(De_u)}$ un intervalle de confiance, de niveau α, caractérisant les extrema spatiaux, tel que :.

$${}^{(\pm)}_{\mathbb{S}}\psi^{l}_{(De_u),t} = \left[\left[\text{môy}\left(\left\{x^{l}_{(U_u\subseteq De_u),t} \neq ?\right\}\right) + \mathbb{t}_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\left(\left\{x^{l}_{(U_u\subseteq De_u),t} \neq ?\right\}\right)}\right.\right.; \left.\left.\text{môy}\left(\left\{x^{l}_{(U_u\subseteq De_u),t} \neq ?\right\}\right) - \mathbb{t}_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\left(\left\{x^{l}_{(U_u\subseteq De_u),t} \neq ?\right\}\right)}\right)\right]$$

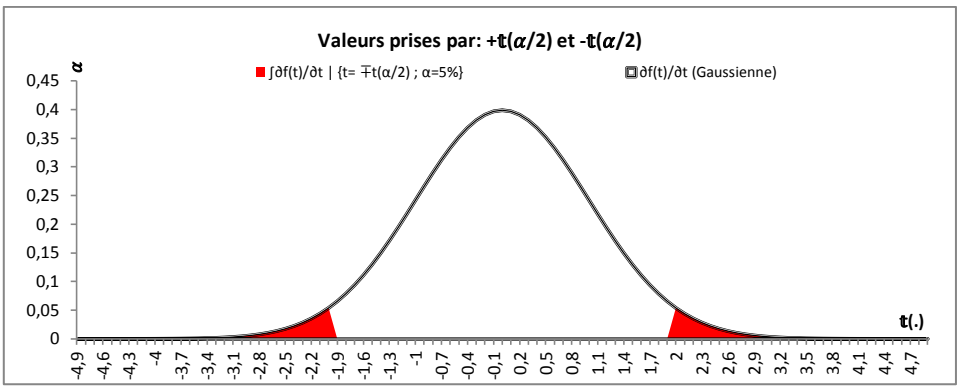


Figure 104 : Valeurs déterministes prises par $\mathbb{t}_{(\frac{\alpha}{2})}$ lorsque $\alpha = 5\%$

Les paramètres associés à la caractérisation de la variabilité spatiale apparente sont:

$$\mathbb{t}_{(\cdot)} \xrightarrow[n\to+\infty]{} Z\sim\mathcal{N}(0,1),\ \ 2\cdot\mathbb{P}\left(\mathcal{N}(0,1)\leq \mathbb{t}_{\left(\frac{\alpha}{2}\right)}\right)=\{\alpha=5\%\}\ ;\ N^{l}_{\mathbb{S},t}=\left\lceil \mathbb{S}\cdot n^{l}_{(U_{.}|De_{u},t)}\right\rceil ;\{\mathbb{S}=20\%\}$$

Une fois les lacunes comblées, il est possible de procéder à la fusion statistique et topologique des données mobilisées.

FUSION STATISTIQUE ET TOPOLOGIQUE DE L'INFORMATION GEOGRAPHIQUE DISPONIBLE

Remarques liminaires :

La stratégie de fusion des données spatiales disponibles a pour dessein de proposer une des procédures topologiques d'agrégation d'informations – par le biais d'indicateurs géographiques auxiliaires - à connotation : socio-économique, statistique et géographique. Il s'agit d'obtenir des variables localisées à différentes dates, et dont la précision informationnelle a été adaptée à la modélisation des FE-SOCIO.ECO* intégrables.

Les caractéristiques granulaires temporelles et les unités des indicateurs géographiques INSEE et ONDRP utilisés ont été déclinées (chapitre.1)

Hypothèses principales :

Les indicateurs géographiques INSEE et ONDRP sont presque systématiquement des dénombrements spatiaux déclinés à différentes dates. Ils comportent donc un biais *d'échelle* ou *de masse* (Lahousse et Piédanna, 1998). Par conséquent, il convient de les rapporter à l'objet géographique et à la nature de la cible – la taille de la population ou à la surface géographique – i.e. de *privilégier les densités spatiales temporelles* (Pumain et Saint-Julien, 1997). Les combinaisons topologiques de variables, fondées sur la *théorie des ensembles flous* par des *fusions topologiques,* permettent d'améliorer la qualité informationnelle (Dubois Didier, Prade Henri, 2004) lorsque les *informations statistiques agrégées* ne sont pas *redondantes* (Saporta, 2006). Il s'agit de confectionner des variables temporelles localisées grevées d'une information assez précise pour modéliser les contextes territoriaux induisant *des prédispositions géographiques morbides* (Haddad, 1992)

Les FE-SOCIO.ECO* retenus sont modélisés à partir d'i.st.e* $x^{l}_{(U_k)}$. Chacun caractérise des variabilités spatiotemporelles contextuelles et plusieurs i.st. peuvent décrire des spécificités géographiques d'un même FE. Ainsi, pour chacune des variables temporelles localisées dont la précision informationnelle est adaptée par fusion, les hypothèses, le processus d'estimation et parfois un complément granulaire sont précisés, en complément des caractéristiques spécifiées (chapitre.1).

APPLICATION AUX VARIABLES INSEE

Géographie des comportements vis-à-vis du recours à l'offre de soins territoriale

Spécification de l'hypothèse :

La variabilité spatiale des niveaux de vie, du pouvoir d'achat et de la répartition des richesses permet d'estimer indirectement les comportements vis-à-vis du recours aux soins et de la consommation de soins de santé (Chaix, Merlo et al., 2005), et plus spécifiquement, les disparités spatiales des niveaux de précarité ou de pauvreté socio-professionnelle sont de bons moyens pour évaluer le climat d'aversion économique à la consommation de soins, une sorte de défaveur territoriale sanitaire qui va de pair avec la création de déserts médicaux et une diminution de la qualité des soins reçus (Penchansky et Thomas, 1981).

Stratégie d'estimation : Variabilités spatiales et temporelles des niveaux de vie
La proportion de foyers fiscaux imposables : le nombre de foyers fiscaux imposables divisé par le nombre total de foyers fiscaux

$$X^{tx.FoyFisc}_{(U_u),t} = \frac{X^{FoyFisImp}_{(U_u),t}}{X^{FoyFisTot}_{(U_u),t}}$$

Stratégie d'estimation : Variabilités spatiales des niveaux de revenus
Le revenu fiscal net médian déclaré par les ménages ou par les personnes

$$\left\{X^{RevMed.m}_{(U_u),t} \cup X^{RevMed.p}_{(U_u),t}\right\}$$

Précision granulaire : Lacunes sont comblées par la stratégie proposée dans les U_u de moins de 2000 habitants au sens du recensement de 1999 ; les revenus fiscaux déficitaires sont portés à zéro (INSEE, 2010).

Stratégie d'estimation : Variabilités spatiales et temporelles de la répartition des richesses
L'indice INSEE de Gini est estimé pour les ménages et pour les personnes.

$$\left\{X^{GINI.m}_{(U_u),t} \cup X^{GINI.p}_{(U_u),t}\right\}$$

Précision granulométrique : Les lacunes sont comblées par la stratégie proposée dans les U_u de moins 50 ménages. L'indice de GINI est un indicateur géographique reflétant la concentration des revenus fiscaux, et à valeur dans $[\![0; 1]\!]$, 0 représente une égalité parfaite dans la capacité financière des entités fiscales, et 1 lorsqu'une entité fiscale concentre, à elle seule, la totalité des revenus de la zone (INSEE, 2012c)

Stratégie d'estimation Variabilités spatiales des niveaux de précarité
Le taux de chômage correspond au nombre de chômeurs de 15 à 64 ans divisé par le nombre d'actifs pour la même tranche d'âge :

$$X^{tx.CHOM}_{(U_u),t} = \frac{X^{CHOM}_{(U_u),t}}{X^{POP.ACTIVE}_{(U_u),t}}$$

Stratégie d'estimation : Variabilités spatiales des niveaux de richesse socio-professionnelle
Le taux d'ouvriers est le nombre d'ouvriers de 15 à 64 ans divisé par le nombre d'actifs appartenant à la même tranche d'âge et exerçant une activité

$$x^{l:\,tx.OUV}_{(U_u),t} = \frac{x^{OUV}_{(U_u),t}}{x^{POP}_{(U_u),t} - x^{CHOM}_{(U_u),t}}$$

Géographie de la qualité des politiques menées en matière de durabilité et de leurs répercussions sur les attraits sociaux, économiques et sanitaires des territoires
Spécification des hypothèses :
L'analyse conjointe des taux de mortalité, de l'accroissement démographique global et de l'accroissement naturel permet d'estimer indirectement les aménités sanitaires, socio-économiques, culturelles et environnementales des territoires, inhérentes aux mesures politiques instituées en matière de durabilité des espaces. Lorsque les politiques menées sont efficaces, de l'attractivité est suscitée, la qualité de vie s'améliore ; l'espérance de vie augmente, l'attractivité suscite des migrations. La population croit et s'investit à son tour dans les processus de développement durable, facilitant entre autres la consommation de soins (Zorman. Michel, 2001).

Stratégie d'estimation : Variabilités spatiales et temporelles de la mortalité
Le taux de mortalité est le nombre de décès annuels rapporté à la population sans double compte.

$$x_{(U_u),t}^{tx.MORT} = \frac{x_{(U_u),t}^{DECES}}{x_{(U_u),t}^{POP}}$$

Précision granulaire : Dénombrement : décès domiciliés à partir des bulletins statistiques de l'état civil (INSEE, 2012c).

Stratégie d'estimation : Variabilités spatiales et temporelles moyennes des accroissements naturels
Le taux annuel moyen d'accroissement naturel de la population - la différence entre le taux de natalité et le taux de mortalité, calculé entre deux dates de recensement, et rapporté à la quantité de population moyenne sans double compte (Vallin, 2001).

$$x_{(U_u),t}^{txAccNAT} = \frac{\left(x_{(U_u),t}^{NAISS} - x_{(U_u),t_{-1}}^{NAISS}\right) - \left(x_{(U_u),t}^{DECES} - x_{(U_u),t_{-1}}^{DECES}\right)}{\frac{1}{2} \cdot \left(x_{(U_u),t}^{POP} + x_{(U_u),t_{-1}}^{POP}\right) \cdot (t - t_{-1})} \times 100$$

Précision granulaire : Cet indicateur ne tient pas compte des mouvements migratoires *internes et externes* (Vallin, 2001).

Stratégie d'estimation : Variabilités spatiales et temporelles des accroissements démographiques
Les taux d'accroissement démographique évoluent de façon géométrique (Peguy, 1996), il convient de les estimer comme suit :

$$x_{(U_u),t}^{txAccPOP} = \exp\left(\frac{1}{(t - t_{-1})} \cdot \ln\left(\frac{x_{(U_u),t}^{POP}}{x_{(U_u),t_{-1}}^{POP}}\right)\right) - 1$$

Stratégie d'estimation : Variabilités spatiales et temporelles des niveaux culturels
Les niveaux culturels sont estimés en géographie de la santé par la proportion d'individus âgés de plus de 15 ans et ayant un diplôme au moins équivalent au baccalauréat (Rey, Jougla et al., 2009)

$$x_{(U_u),t}^{tx.BAC} = \frac{\sum_{niv=5}^{6} \sum_{sexe=1}^{2} \sum_{age=1}^{2} x_{(U_u),t}^{\{age,niv,sexe\}}}{\sum_{niv=1}^{6} \sum_{sexe=1}^{2} \sum_{age=1}^{2} x_{(U_u),t}^{\{age,niv,sexe\}}} \times 100$$

Spécificité granulaire : Les données INSEE comptabilisent le nombre d'individus, par tranche d'âge : {16 à 24 ans, de 25 ans et plus} et par sexe: {G, F}, en fonction de 6 catégories de diplômes: {Aucun, CEP, BEPC, CAP-BEP, Baccalauréat, Enseignement supérieur} (INSEE, 2012c).

Géographie des expositions potentielles à des substances toxiques (1) et aux pesticides (2) liées à la spécialisation socio-économique des territoires.
Spécification de l'hypothèse (1) :
La proportion d'individus exerçant un emploi dans un secteur d'activité où il existe des risques d'exposition à des substances toxiques est un moyen contextuel d'estimer les prédispositions géographiques morbides. Ces personnes sont formées à la prévention des risques, ce qui influence positivement leur consommation de soins mais l'accommodation aux risques engendre parallèlement des contextes de surexposition (Anses, 2012).

Stratégie d'estimation : Variabilités spatiales et temporelles des emplois exercés dans des secteurs d'activité à risques
Proportion d'individus exerçant un emploi à risques - le nombre d'individus exerçant un emploi dans un secteur d'activité où le risque d'exposition à des substances toxiques est suspecté, rapporté au nombre total d'emplois.

$$x_{(U_u),t}^{tx.EAR} = \frac{x_{(U_u),t}^{Emp.à.Ris}}{x_{(U_u),t}^{Emp.Tot}}$$

Précisions granulaires : L'INSEE comptabilise toutes les professions selon 15 fonctions transversales communes à tous les secteurs d'activité. Ont été supposées risquées les fonctions de : conception-recherche ; agriculture et pêche ; bâtiment et travaux publics ; fabrication ; transport et logistique ; entretien et réparation ; distribution (INSEE, 2012c).

Spécification de l'hypothèse (2) :
La proportion de surface allouée à l'agriculture et l'intensivité des activités agricoles sont dépendantes et liées à la spécialisation socio-économique des territoires, elle constitue un moyen contextuel, en conférant à la notion de *risque* une *dimension collective,* pour évaluer *les expositions géographiques potentielles* aux *pesticides* (Bailly et Beguin, 2005). Les pesticides comprennent les insecticides, fongicides, herbicides et antiparasitaires. Quelle que soit la spécialité des activités agricoles, elles utilisent toutes au moins l'une des quatre substances citées. Les agents toxiques contenus dans les pesticides sont vectorisés par les vents, les eaux de pluie et de surface vers tous les compartiments environnementaux (eau, air, sol, et milieu biologique) et contaminent les milieux de vie exposant les populations les plus proches à des risques morbides multiples (Anses, 2012).

Stratégie d'estimation : Variabilité spatiale et temporelle des ratios de Surface Agricole Utilisée
Le ratio entre la somme de toutes les classes de SAU recensées par l'INSEE et la surface communale de l'unité géographique communale :

$$x_{(U_u),t}^{tx.SAU} = \frac{1}{x_{(U_u)}^{SG}} \cdot \sum_{j=1}^{5} x_{(U_u),t}^{\{SAU.j\}}$$

Spécificité granulaire : La Superficie Agricole Utilisée (SAU) est une statistique européenne exprimée en hectares (ha) (INSEE, 2012a); les 5 types INSEE d'exploitations agricoles sont : terres labourables, céréales, fourragère principale, toujours en herbes et les fermages. Sur toutes ces surfaces des pesticides sont systématiquement utilisés (Anses, 2012).

Stratégie d'estimation : Variabilité spatiale et temporelle de l'Intensité des Activités Agricoles
Il s'agit du produit entre les Unités de Temps Annuel communal $x_{(U_u),t}^{UTA}$ et la somme des exploitations agricoles sur cette communes $x_{(U_u),t}^{\{Exp.AGRI.j\}}$ – *réduit* par les écarts-types de ces variables afin de ne pas engendrer un biais statistique (Saporta, 2006)

$$x_{(U_u),t}^{Int.AGRI} = \frac{x_{(U_u),t}^{UTA}}{\hat{\sigma}_{x_{(\cdot),t}^{UTA}} \cdot \hat{\sigma}_{x_{(\cdot),t}^{\cup(Exp.AGRI.j)}}} \cdot \sum_{j=1}^{6} x_{(U_u),t}^{\{Exp.AGRI.j\}}$$

Spécificité granulaire : $x_{(U_u),t}^{UTA}$ les Unités de Temps Annuel communal, une UTA équivaut au travail d'une personne à temps complet, dans une exploitation agricole, pendant un an ; Et $x_{(U_u),t}^{\{Exp.AGRI.j\}}$ est le nombre d'exploitations agricoles dont la définition correspond à l'une des 6 catégories Agrest (INSEE, 2012a)

Géographie des prédispositions contextuelles aux phénomènes morbides
Spécification des hypothèses :
Le concept de *défaveur* est défini comme une accumulation de désavantages socio-économiques ayant un impact néfaste sur la santé des populations (Townsend, 1987). A l'accoutumée, les niveaux géographiques de la *défaveur sociale* s'estiment par *des scores composites* calculés à partir de données socio-économiques étatiques, *généralement acquises lors de recensements* (Carstairs et Morris, 1989). Les indicateurs géographiques de *défaveur* ou de *carence sociale* permettent de caractériser les états de santé des populations mais leurs performances varient selon l'échelle, les catégories de population ciblées, les zones géographiques, et les maladies étudiées (Barnett, Wrigley et al., 2002). Parmi le large panel d'indicateurs *de défaveur sociale* proposé dans la littérature, *il convient d'en choisir un adapté à la*

*fois aux caractéristiques épidémiologiques et géographiques des PM** d'intérêt et *aux données communautaires disponibles* (Benach et Yasui, 1999).

Dans le cadre de cette recherche, la modélisation des niveaux géographiques de *défaveur sociale* se fonde sur l'indicateur *FDep.99*. Il a été retenu car il est peut être estimé avec des données INSEE et que sa significativité spatiale opère pour tous : *Les types d'occupation du sol* (rural, quasi-rural, quasi-urbain, urbain, Paris et sa banlieue), ce qui est d'une importance capitale en France où les gradients territoriaux ruraux/urbains sont très volatiles, les *niveaux géographiques d'analyse*, et *l'échelle des communes est celle où il affirme sa suprématie*, les individus quel que soit leur sexe ou leur âge, *les PM** ou presque, et il se montre *très pertinent pour les cancers*. En dépit de tous les attraits de l'index FDep99 énoncés, sa fiabilité temporelle n'a pas été démontrée en dehors de la période couverte par les données morbides INSERM utilisées, i.e. entre 1997 et 2001. Au-delà de ce laps de temps les auteurs ne *garantissent plus la même robustesse spatiale* (Rey,Jougla et al., 2009).

Définition de FDep99 : Cet index se construit à l'échelle des communes U_u à partir de quatre variables INSEE : *le revenu fiscal médian des ménages en 2001* : $x^{RevMed.m}_{(U_u),t=2001}$, *le pourcentage d'individus de plus de 15 ans ayant un diplôme au moins équivalent au baccalauréat en 1999* : $x^{tx.BAC}_{(U_u),t=1999}$, *la proportion d'ouvriers dans la population active exerçant un emploi en 1999* : $x^{tx.OUV}_{(U_u),t=1999}$ *et le taux de chômeurs en 1999 :* $x^{tx.CHÔM}_{(U_u),t=1999}$ - soit la matrice : $X^{INSEE.j}_{FDep.99}$. FDep.99 est défini comme la projection de $X^{INSEE}_{FDep.99}$ sur *la première composante d'une ACP* (Rey, Jougla et al., 2009).

Notions mathématiques : L'Analyse en Composante Principale (ACP) a pour dessein de maximiser l'hétérogénéité algébrique d'un jeu de données multidimensionnelles X – i.e. un nuage de points composé de n-individus à p-coordonnées vectorielles - par une projection vectorielle dans un sous-espace F_k de plus petite dimension $k \leq p$. Le premier axe d'une ACP $c_{.}^{j=1}$ renferme les coordonnées projetées des individus et il est engendré par *le vecteur propre* de la matrice *des facteurs principaux* $u_{.}^{j}$ associés à la plus *grande valeur propre* λ_k et estimée par la relation $MVu = \lambda u$. Avec : MV le produit matriciel d'une métrique de poids M et la matrice des variances-covariances *V* associées à X (Saporta, 2006) – (annexe.3).

Stratégie d'estimation : Variabilité spatiale et temporelle de la défaveur sociale de 1997 à 2001

L'estimation de FDep.99 a été calculée conditionnellement aux données préconisées :

$$X^{INSEE.j}_{FDep.99} = \left(x^{RevMed.m}_{(U_u),t=2001};\ x^{tx.BAC}_{(U_u),t=1999};\ x^{tx.OUV}_{(U_u),t=1999}; x^{tx.CHÔM}_{(U_u),t=1999} \right)$$

Cependant le principe d'estimation de FDep.99 n'est pas spécifié explicitement (Rey, Jougla et al., 2009). Par conséquent, FDep.99 a été calculé à partir des vecteurs propres de MV de sorte que l'estimation de V a été effectuée *en rendant inactives les communes* où les valeurs de $x^{RevMed.m}_{(U_u),t=2001}$ étaient soumises au secret statistique. La métrique ${}^{F}M$ choisie pour engendrer l'espace des individus est $D_{1/\hat{\sigma}^2}$ puisque *les unités* des variables de $X^{INSEE.j}_{FDep.99}$ *sont disparates*. Ensuite, comme les *données ne sont pas recueillies* aléatoirement, *mais dans une logique territoriale*, la métrique ${}^{E}M$ choisie pour engendrer l'espace des variables est $D = p_{(U_u),t} \cdot \mathbb{I}_n$. Cela revient à supposer que chaque U_u contribue à la construction des axes proportionnellement à sa quantité de population pour la tranche d'âge considérée :

$$p_{(U_u),t=1999} = \left(x^{POP\geq 15.ans}_{(U_u),t=1999} \Big/ \sum_{u=1}^{n(U_.)} x^{POP\geq 15.ans}_{(U_u),t=1999} \right)$$

Comme le nombre de communes actives est très important, les fonctions : prcomp(.)et princomp(·)du logiciel: R sont inopérantes (Institute for Statistics and Mathematics, 1997). *Un produit scalaire a dû être programmé* afin d'engendrer l'espace des variables et d'estimer les facteurs principaux permettant de définir les composantes principales de l'ACP - et enfin obtenir FDep99, tel que :

$$x^{FDep99}_{(U_u)} = \left\{ c^{1}_{(U_u),t=\{2001\ ;\ 1999\}} \right\}$$

Stratégie d'estimation : Variabilité spatiale et temporelle de la défaveur sociale de 2007 à 2011
Les niveaux géographiques de la variabilité de la *défaveur sociale* entre 2007 et 2011 sont modélisés par FDep.09. Le processus d'estimation est le même que celui utilisé pour FDep.99, à la différence près que les temporalités des données utilisées sont variables :

$$X_{FDep.09}^{INSEE.j} = \left(x_{(U_u),t=2010}^{RevMed.m};\ x_{(U_u),t=2009}^{tx.BAC};\ x_{(U_u),t=2009}^{tx.OUV};\ x_{(U_u),t=2009}^{tx.CHÔM} \right)$$

Quant aux poids de la métrique permettant d'engendrer l'espace des variables, ils ont été définis par $p_{(U_u),t=2009}$, ce qui a permis d'obtenir :

$$x_{(U_u)}^{FDep09} = \left\{ c_{(U_u),t=\{2009\,;\,2010\}}^{1} \right\}$$

Afin d'estimer la défaveur sociale sur une période plus longue donc plus cohérente avec celle recouverte par l'étude LEA, l'index FDep.XX est proposé.

Stratégie d'estimation : Variabilité spatiale de la défaveur sociale temporelle plausible entre 1997 et 2011
L'index FDep.XX est obtenu par la moyenne des valeurs réduites de $x_{(U_u)}^{FDep99}$ et de $x_{(U_u)}^{FDep09}$ - puisque les valeurs prises aux deux dates sont souvent radicalement différentes. De fait :

$$x_{(U_u)}^{FDepXX} = \frac{1}{2} \cdot \left(\frac{x_{(U_u)}^{FDep99}}{\widehat{\sigma}_{\left(x_{(U_\cdot)}^{FDep.99}\right)}} + \frac{x_{(U_u)}^{FDep09}}{\widehat{\sigma}_{\left(x_{(U_\cdot)}^{FDep.09}\right)}} \right)$$

APPLICATION AUX VARIABLES ONDRP

Géographie des niveaux de stress potentiellement perçu et induit par l'insécurité territoriale contextuelle

Spécification des hypothèses :
Les niveaux de stress lié au sentiment d'insécurité territoriale peuvent être interprétés comme une *carence politique* en matière de lutte contre la pauvreté, la délinquance, l'exclusion, l'insécurité… (Godin, 2007). Le stress chronique a des répercussions physiologiques qui causent la libération d'hormones en déréglant le système endocrinien, favorisant ainsi le développement de nombreuses maladies, dont les cancers font partie (Inserm - Expertise collective, 2011). Le stress chronique d'origine sociale est une *souffrance psychique* qui favorise les comportements à risques - individuels ou collectifs – et qui exerce, entre autres, une influence néfaste sur le recours aux soins (Furtos, 2007).
Dans le cadre de cette recherche le sentiment de stress est modélisé à partir des mesures spatiales d'infractions ONDRP disponibles sur l'intégralité de la France, à l'échelle des départements : D_u, sur une période qui s'étale de $t = \{1996, \dots, 2011\}$. Elles se déclinent sous la forme d'index spatiaux d'infractions, i.e. crimes et délits, thématiques : $isit_{(D_u),t}^{\{k|l\}}$. Les quatre types d'infractions thématiques sont les : *Atteintes aux biens* (index : vols à main armée, cambriolages, incendies volontaires, attentats…) ; *Atteintes volontaires à l'intégrité physique* (index : viols, harcèlements, homicides, violences, …) ; *Escroqueries et infractions économiques et financières* :(Faux en écritures publiques et authentiques ; Fausses monnaies ; Contrefaçons…) ; *Infractions relevées par l'action des services* (Recels ; Proxénétismes ; Trafics et revente de stupéfiants…). Les variables ONDRP sont obtenues par cumul des index d'une même thématique $x_{(D_u),t}^{l:ONDRP}$, donc des quantités d'infractions annuelles. Les variables peuvent être ramenées à l'échelle des populations locales, on parlera de ratios (ONDRP, 2012).
Toutefois, il convient de noter que les index des *Infractions relevées par l'action des services* sont redondants avec les trois autres, et que les ratios sont biaisés puisqu'ils surestiment la réalité géographique des infractions (ONDRP, 2011).
Dans le cadre de cette recherche la variabilité spatiale du sentiment de stress contextuel est modélisée à partir des $x_{(D_u),t}^{l:ONDRP}$ – qui permettent de confectionner des *i.st.e* Curieux* de test* car en dépit des effets manifestes du stress sur l'état de santé, *la Distance a-spatiale morbide*,* induite par l'imprécision d'échelle et l'éloignement avec les PM* d'intérêt, est manifestement grande.

Stratégie d'estimation : Variabilités spatiales annuelles du sentiment de stress potentiellement induit par les niveaux d'atteintes aux biens matériels ou d'atteintes à l'intégrité physique

Les cumuls des $\text{isit}_{(D_u),t}^{\{k|l\}}$ ont été préférés puisque les ratios sont biaisés. La thématique *Escroqueries et infractions économiques et financières* n'est pas pertinente et celle des *Infractions relevées par l'action des services* engendre des redondances. Les thématiques retenues sont : *Atteintes aux biens* et *Atteintes volontaires à l'intégrité physique,* et tous les index qui les constituent ont été jugés représentatifs du sentiment d'insécurité territoriale. De fait :

$$x_{(D_u),t}^{l:ONDRP} = \sum_{k=1}^{n_l^k} \text{isit}_{(D_u),t}^{\{k|l\}}$$

Avec : n_l^k le nombre total d'$\text{isit}_{(D_u),t}^{\{k|l\}}$ des index spatiaux d'infractions thématiques inclus dans la catégorie l: ONDRP ∈ {att. BIEN; att. PHY} ; Et $x_{(D_u),t}^{l:ONDRP}$ la somme annuelle des infractions.

Stratégie d'estimation : Variabilités spatiales annuelles du sentiment de stress lié à l'insécurité composite d'infractions protéiformes.

Il s'agit d'une proposition heuristique vouée à construire un indicateur spatial composite d'insécurité : $x_{(D_u),t}^{INSECU}$ permettant de fusionner les deux indicateurs précédents en les ramenant d'abord à une échelle de valeur commune puis en les normalisant (Saporta, 2006) :

$$x_{(D_u),t}^{INSECU} = \left\{ \sum_{l\in\{att.BIEN;att.PHY\}} \frac{\left(\dot{x}_{(D_u),t}^{l} - \hat{moy}\left(\dot{x}_{(\cdot),t}^{l}\right)\right)}{2\cdot\hat{\sigma}\left(\dot{x}_{(\cdot),t}^{l}\right)} \middle| \dot{x}_{(D_u),t}^{l} = \frac{x_{(D_u),t}^{l:ONDRP}}{\hat{max}\left(x_{(\cdot),t}^{l:ONDRP}\right) - \hat{min}\left(x_{(\cdot),t}^{l:ONDRP}\right)} \right\}$$

Remarque : $x_{(D_u),t}^{INSECU}$ est Sans Unité (SU). En revanche $x_{(D_u),t}^{l:att.BIENS}$ et $x_{(D_u),t}^{l:att.PHY}$ sont des cumuls d'infractions thématiques ce qui induit *a priori* un biais *géographique de masse*. Cependant l'optimisation de *l'effet information** sera finalisée lors du processus d'harmonisation des échelles – décliné dans la partie subséquente.

Les processus de fusion des informations géographiques disponibles se fondent sur des théories statistiques et sur celle des ensembles flous. Elle permet de disposer des variables spatiotemporelles dont la nature est adaptée à la modélisation des FE-SOCIO.ECO* pertinents* et Curieux* intégrables. Toutefois des conflits d'échelles demeurent

UNIFORMISATION A L'ECHELLE DES COMMUNES

Certaines variables INSEE sont déclinées à l'échelle des arrondissements Ar_u. Les variables ONDRP sont, quant à elles, mobilisables à l'échelle des départements De_u. Il convient donc de proposer des stratégies adaptées afin d'uniformiser les échelles.

PROPOSITION POUR LES DONNEES INSEE

Remarques liminaires :

Les FE-SOCIO.ECO* modélisés à partir des indicateurs INSEE doivent être agrégés à l'échelle des U_u, au moins pour les communes de 1ère espèce de Marseille, Lyon et Paris qui sont découpées en arrondissements. A l'instar des FE-SAN* le processus *mathématique d'agrégation spatiale* incorpore un pondérateur géographique.

Les poids utilisés pour les indicateurs spatiaux SAN permettaient d'effectuer une agrégation conditionnelle aux : $x_{(.),t}^{SG}$ – les surfaces territoriales - car les indicateurs spatiaux des DTA et des APL sont pondérés par la demande, i.e. la taille des populations locales. Or les sources INSEE sont des dénombrements, elles ne sont pas ramenées à l'échelle des populations spatialisées, et comme elles décrivent des conjonctures territoriales socio-économiques, le pondérateur géographique le plus

adapté pour procéder à une agrégation d'échelle est $x_{(.),t}^{\mathrm{POP}}$, i.e. les quantités de population situées dans les $\mathrm{Ar_u}$ des $\mathrm{U_u}$ (Lahousse et Piédanna, 1998).

Hypothèse :
Elle est identique à celle déclinée dans le processus diachronique d'appariement spatial des i.st.e* entrant dans la modélisation des FE-SAN*.

Proposition :
Elle est identique à celle déclinée dans le processus diachronique d'appariement spatial des i.st.e* voués à la modélisation des FE-SAN.

PROPOSITION POUR LES DONNEES ONDRP

Remarques liminaires :

Les variables ONDRP obtenues à la suite du processus de fusion sont des cumuls temporels d'index spatiaux - représentatifs des atteintes aux biens, des atteintes à l'intégrité physique et du niveau d'insécurité. Les $x_{(\mathrm{De_u}),t}^{\mathrm{l:ONDRP}}$ sont utilisés pour la géographie des FE-SOCIO.ECO* *Curieux** à cause de la *Distance a-spatiale morbide* probablement éloignée* avec les PM* d'intérêt et du fait qu'ils soient à l'échelle des départements $\mathrm{De_u}$. Afin d'être intégrés dans la procédure multidimensionnelle d'identification des DES*, les $x_{(\mathrm{De_u}),t}^{\mathrm{l:ONDRP}}$ doivent être ramenés à l'échelle des $\mathrm{U_u}$.

Hypothèses :

Le cumul des index ONDRP est biaisé et surestime légèrement la réalité géographique à cause de la méthode d'acquisition des données de criminalité. Les ratios ONDRP, i.e. ramenés à la taille des populations à la même date, sont doublement biaisés et sous-estiment la réalité géographique. Toutefois, la quantité d'infractions augmente avec celle de la population mais la relation n'est pas linéaire (ONDRP, 2011).

En France, la population croît de façon géométrique : 53,7 millions d'habitants en 1980 ; 56,6 millions en 1990 ; 58,9 millions en 2000 ; 62,7 millions en 2010 (Blanpain et Chardon, 2008) – donc la quantité d'infractions augmente et est doublement surestimée car les index ONDRP utilisés sont déclinés depuis 1996, or l'étude LEA commence en 1980.

En contrepartie, le sentiment d'insécurité perçu dépend de la taille des zones géographiques considérées et de la localisation du lieu de résidence. Par exemple, à Marseille le sentiment d'insécurité perçu est différent selon que les individus résident dans les quartiers Nord ou dans le VIIème arrondissement (Gérard, 2012).

Enfin, il existe dans la littérature de nombreuses méthodes de désagrégation d'échelle mais dans la pratique, celles fondées sur des stratégies simples sont généralement les plus représentatives de la réalité géographique (Powell, 1995), en particulier celles qui conservent les propriétés intrinsèques des sources (Tobler, 1979).

Principe d'estimation :

La stratégie de désagrégation des $x_{(\mathrm{De_u}),t}^{\mathrm{l:ONDRP}}$ à l'échelle ($\mathrm{U_u}$) proposée est donc à la fois fondée sur un processus mathématique simple de fusion de données géographiques environnementales conservatrice, i.e. que le nombre d'infractions avant et après les transformations est conservé sur la zone considérée ; Capable d'amoindrir la surestimation engendrée par l'utilisation des index spatiaux de la période temporelle disponible, en ramenant les $x_{(\mathrm{De_u}),t}^{\mathrm{l:ONDRP}}$ à la taille des populations communales à une date plus proche du début que de la fin de la période recouverte par LEA, et enfin apte à prendre en compte *l'effet tampon* de l'espace sur le sentiment d'insécurité perçu – autrement dit l'ampleur des surfaces géographiques communales. En d'autres termes :

$$x_{(\mathrm{U}_u),t}^{l} = \frac{x_{(\mathrm{U}_u)}^{\mathrm{POP}} \cdot x_{(\mathrm{De_u}),t}^{l}}{x_{(\mathrm{U}_u)}^{\mathrm{SG}} \cdot \sum_{u=1}^{n_{(\mathrm{U.})}} \left(x_{(\mathrm{U}_u)}^{\mathrm{POP}} \cdot \mathbb{1}_{\{\mathrm{U}_u \subseteq \mathrm{De_u}\}} \right)}$$

Avec : $\mathrm{x}_{(\mathrm{U_u})}^{\mathrm{POP}}$ et $\mathrm{x}_{(\mathrm{U_u})}^{\mathrm{SG}}$ respectivement la quantité de population (u) et la surface géographique (ha) de l'$\mathrm{U_u}$, en 2003 –de la BD-SIG utilisée. Cette valeur est estimée pour les 95 $\mathrm{De_u}$ situés en France métropolitaine.

L'effet information* est optimisé et a permis d'obtenir des variables spatiotemporelles plus précises, plus adaptées à la modélisation des FE-SOCIO.ECO. La granularité* d'échelle est à la fois uniforme et adaptée à l'échelle d'investigation.

MAXIMISATION DE L'EFFET DE SUPPORT

*L'effet information** est désormais optimisé, il convient maintenant de maximiser *l'effet de support** i.e. de construire des indicateurs spatiaux agrégés *verticalement* conditionnellement à un processus temporel probabiliste et permettant d'obtenir les i.st.e* $x^{l}_{(U_u)}$, qu'il conviendra ensuite d'apparier de façon cohérente aux U_k.

PROCESSUS D'HARMONISATION TEMPORELLE PROBABILISTE

Remarques liminaires :

Les variables INSEE et ONDRP se présentent sous la forme de chroniques temporelles localisées. Elles sont toutes de nature quantitative continue ou discrète à l'instar de toutes les séries temporelles localisées ou géo-localisées mobilisées pour la modélisation des FE-PHY.CHIM. La stratégie d'agrégation verticale qui est proposée ici est applicable à toutes les variables utilisées dans le cadre de cette thèse.

Les i.st.e* proposés sont uniques. Les *chroniques temporelles* dont l'effet information* a été optimisé doivent donc être agrégées *verticalement* pour satisfaire à cette condition. La statistique utilisée pour y parvenir doit être représentative de la stationnarité temporelle apparente des données disponibles.

Les séries temporelles localisées ONDRP $x^{l:ONDRP}_{(U_u),t}$ recouvrent une période Δ^{l}_{t} allant de : $[\![\{t_1 = 1996, \dots, t_{n^{l}_{t}} = 2011\}]\!]$ et à l'instar des données environnementales Météo-France, par exemple elles sont déclinées à un pas de temps constant, i.e. $(t_j - t_{j+1}) = (t_k - t_{k+1}), \forall\, j \neq k$.

En revanche ce n'est pas le cas des chroniques temporelles localisées INSEE qui recouvrent des périodes temporelles particulières et pour lesquelles le pas de temps varie – à l'instar des mesures temporelles localisées RNM* de la radioactivité dans l'environnement (chapitre.1).

Spécification des hypothèses :

Il s'agit de proposer, pour chaque unité géographique un i.st.e* unique $x^{l}_{(U_u),}$par un processus d'agrégation verticale des séries temporelles (Peguy, 1996). La dialectique est géographique, elle doit être conçue dans une logique territoriale et l'estimateur spatial doit être statistiquement représentatif de la variabilité temporelle du phénomène géographique décrit par les $x^{l}_{(U_u),t}$ (Charre, 1995). De plus, l'estimateur proposé doit être adapté à la *nature* de la variable et être *consistant* ou au moins supposé comme tel *en pratique* (Saporta, 2006). Par conséquent la volatilité des valeurs disponibles à différentes dates sera analysée à partir des caractéristiques statistiques des séries temporelles localisées (Lütkepohl, 1991), et le concept de *stationnarité - ou stabilité - temporelle apparente des accroissements temporels moyens à l'ordre un* - permettra de mettre au point un processus décisionnel probabiliste et de choisir un estimateur statistique adéquat (Hamilton, 1994). En sus, pour ce qui est de la modélisation des FE-SOCIO.ECO* intégrables, il convient de prendre en compte aussi le fait que la variabilité temporelle des contextes socio-économiques des territoires est cyclique car déterminée par les tendances locales des décideurs politiques en place (Bailly et Beguin, 2005)

Principe d'estimation :

L'estimateur statistique d'agréation verticale est unique et repose sur le concept de *stationnarité temporelle apparente*. Cette spécificité s'apprécie par l'analyse statistique des différenciations *temporelles* (Hamilton, 1994). En effet, les phénomènes géographiques ne sont pas stables, ils évoluent dans le temps et présentent, presque systématiquement, *des dérives temporelles*. Cependant, leurs *accroissements moyens à l'ordre un, sont généralement stationnaires*, peu ou prou (Peguy, 1996) et s'estiment de la façon suivante :

$$\Delta\left(x^{l}_{(U_u),t_i}\right) = \frac{1}{(t_{i+1} - t_i)} \cdot \left(x^{l}_{(U_u),t_{i+1}} - x^{l}_{(U_u),t_i}\right), \quad \forall\, i = \{1, \dots, n^{l}_{t}\}$$

On choisit l'estimateur de la médiane pondérée temporellement lorsque les accroissements temporels à l'ordre un suggèrent une stationnarité temporelle apparente. Celui-ci s'estime de la façon suivante

$$\hat{\mathrm{m}}\mathrm{ed}_{\mathcal{P}}^{t}\left(\mathrm{x}_{(\mathrm{U_u}),\mathrm{t}}^{\mathrm{l}}\right)=\hat{\mathrm{m}}\mathrm{ed}\left(\bigcup_{\mathrm{t_i}=1}^{\mathrm{n}_t^{l}}\left(\bigcup_{\mathrm{b}=1}^{[(\mathrm{t_{i+1}}-\mathrm{t_{i-1}})/2\cdot\mathrm{d_t^l}]}\left(\mathrm{x}_{(\mathrm{U_u}),\mathrm{t_i}}^{\mathrm{l}}\cdot\mathbb{1}_{\{b\}}\,\middle|\,\left\{\mathrm{t}_0\cup\mathrm{t}_{\mathrm{n_t^l}+1}\right\}\right)\right)\right)$$

Avec : $\mathrm{d_t^l}$ le pas de temps minimal de la série temporelle estimée considérée ; et n_t^{l} le nombre de temporalités mobilisées qui correspondent au mieux à la période recouverte par LEA.

Sinon, lorsque la série est *apparemment très volatile dans le temps*, ou que la médiane n'est pas un estimateur statistique consistant, i.e. lorsque en pratique $\mathrm{n_t^l}\leq 4$, c'est l'estimateur de la moyenne pondérée temporellement qui est utilisé – tel que :

$$\hat{\mathrm{m}}\mathrm{oy}_{\mathcal{P}}^{\mathrm{t}}\left(\mathrm{x}_{(\mathrm{U_u}),\mathrm{t}}^{\mathrm{l}}\right)=\frac{\left(2\cdot\mathrm{n_t^l}\right)^{-1}}{\left(\mathrm{t}_{\mathrm{n_t^l}}-\mathrm{t}_1\right)}\left(\sum_{\mathrm{i}=1}^{\mathrm{n_t^l}}(\mathrm{t_{i+1}}-\mathrm{t_{i-1}})\cdot\left(\mathrm{x}_{(\mathrm{U_u}),\mathrm{t_i}}^{\mathrm{l}}\middle|\left\{\mathrm{t}_0\cup\mathrm{t}_{\mathrm{n_t^l}+1}\right\}\right)\right)$$

Il convient de remarquer que l'estimation de ces statistiques nécessite, pour chaque variable, de connaitre les temporalités disponibles qui encadrent la période recouverte par l'étude LEA.

En d'autres termes, l'estimateur pondéré de la moyenne temporelle est utilisé pour caractériser les $\mathrm{U_u}$ où les populations sont soumises à des situations à risques induites par des conjonctures socio-économiques (ou des expositions environnementales chroniques à des substances physicochimiques toxiques) particulièrement changeantes dans le temps. La moyenne pondérée dans le temps est préconisée lorsque les séries temporelles sont *apparemment erratiques,* ce qui permet ainsi de grever l'estimateur agrégé *verticalement* de la présence *anormale* de minima temporels (ou à l'inverse de maxima temporels) - en utilisant les propriétés statistiques de cet estimateur, sensibles aux extrema, et ainsi obtenir une représentation plus réaliste des cycles décrits par les séries temporelles localisées. A l'inverse lorsque la série est *apparemment stationnaire dans le temps,* i.e. que ses accroissements temporels ne s'écartent pas d'un nombre de fois *anormalement* élevé $\mathrm{N_T^l}$ de l'accroissement moyen, il convient d'utiliser l'estimateur de la médiane temporelle pondérée. En effet, celui-ci n'est pas affecté par la présence occasionnelle d'extrema - à moins de constituer plus de 50% des effectifs (Saporta, 2006), donc d'éluder les variabilités temporelles anecdotiques, peu représentatives des caractéristiques environnementales habituelles auxquelles sont assujetties les populations localisées concernées. Ce faisant, la règle de décision proposée est la suivante :

$$x_{(\mathrm{U_u})}^{\mathrm{l}}=\begin{cases}\hat{\mathrm{m}}\mathrm{ed}_{\mathcal{P}}^{\mathrm{t}}\left(\mathrm{x}_{(\mathrm{U_u}),\mathrm{t}}^{\mathrm{l}}\right) & \text{lorsque: } \left\{\mathrm{card}\left(\bigcup_{\mathrm{t_i}=1}^{\mathrm{n_t^l}}\left(\Delta\left(x_{(\mathrm{U_u}),\mathrm{t_i}}^{l}\right)\right)\notin {}^{(\pm)}_{\mathbb{T}}\psi_{\Delta_{(\cdot)}}^{\mathrm{l}}\right)\leq \mathrm{N_T^l}\right\}\cap\left\{\mathrm{n_t^l}>4\right\}\\ \hat{\mathrm{m}}\mathrm{oy}_{\mathcal{P}}^{\mathrm{t}}\left(x_{(\mathrm{U_u}),\mathrm{t}}^{\mathrm{l}}\right) & \text{lorsque: } \left\{\mathrm{card}\left(\bigcup_{\mathrm{t_i}=1}^{\mathrm{n_t^l}}\left(\Delta\left(x_{(\mathrm{U_u}),\mathrm{t_i}}^{l}\right)\right)\notin {}^{(\pm)}_{\mathbb{T}}\psi_{\Delta_{(\cdot)}}^{\mathrm{l}}\right)> \mathrm{N_T^l}\right\}\cup\left\{\mathrm{n_t^l}\leq 4\right\}\end{cases}$$

Avec : ${}^{(\pm)}_{\mathbb{T}}\psi_{\Delta_{(\cdot)}}^{\mathrm{l}}$ un intervalle de confiance bilatère, de niveau α peu élevé, permettant de circonvenir les valeurs des accroissements temporels moyens représentatifs d'une stationnarité temporelle apparente, tel que

$${}^{(\pm)}_{\mathbb{T}}\psi_{\Delta_{(\cdot)}}^{\mathrm{l}}=\left[\!\left[\bar{\Delta}\left(x_{(\mathrm{U_u}),\cdot}^{\mathrm{l}}\right)+\mathbb{t}_{\frac{\alpha}{2}}\cdot\frac{\hat{\sigma}_{\Delta\left(x_{(\mathrm{U_u}),\cdot}^{\mathrm{l}}\right)}}{\sqrt{\mathrm{n_t^l}-1}};\ \bar{\Delta}\left(x_{(\mathrm{U_u}),\cdot}^{\mathrm{l}}\right)-\mathbb{t}_{\frac{\alpha}{2}}\cdot\frac{\hat{\sigma}_{\Delta\left(x_{(\mathrm{U_u}),\cdot}^{\mathrm{l}}\right)}}{\sqrt{\mathrm{n_t^l}-1}}\right]\!\right]$$

Avec: $\bar{\Delta}\left(x^{l}_{(U_u),.}\right)$ l'estimateur de la moyenne des $\Delta\left(x^{l}_{(U_u),t}\right)$; $\widehat{\sigma}_{\Delta\left(x^{l}_{(U_u),.}\right)}$ l'estimateur biaisé des écarts-types est $\Delta\left(x^{l}_{(U_u),t}\right)$; Et $\mathbb{t}_{\frac{\alpha}{2}}$ une variable gaussienne déterministe telle que :

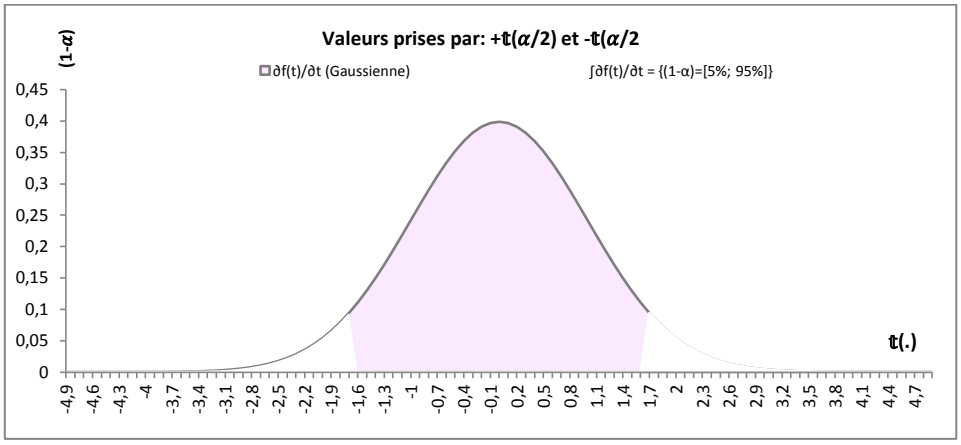


Figure 105 : Valeurs déterministes prises par $\mathbb{t}_{\left(\frac{\alpha}{2}\right)}$ lorsque $\alpha = 90\%$

Les paramètres associés à la caractérisation de la *stabilité temporelle apparente* sont:
$\mathbb{t}_{(\cdot)} \xrightarrow[n\to+\infty]{} Z \sim \mathcal{N}(0,1)$; $2 \cdot \mathbb{P}\left(\mathbb{t}_{\left(\frac{\alpha}{2}\right)} \leq Z \leq \mathbb{t}_{\left(1-\frac{\alpha}{2}\right)}\right) = \{1-\alpha = 90\%\}$; $N^{l}_{\mathbb{T}} = \left\lceil \mathbb{T} \cdot n^{l}_{t} \right\rceil$; $\{\mathbb{T} = 50\%\}$

<u>Application à des données socio-économiques :</u>
Afin d'illustrer la stratégie décisionnelle d'agrégation verticale proposée, elle est appliquée à $x^{tx.BAC}_{(U_u),t}$ - la proportion communale d'individus diplômés du baccalauréat ou plus - $U_u = "ABERGEMENT - CLEMENCIAT"$. La ligne de pointillés noirs représente les $\Delta\left(x^{l}_{(U_u),t}\right)$ sur la période $\{1975; 1982; 1990; 1999; 2009\}$, i.e. les seules temporalités disponibles et recouvrant l'étude LEA. Dans cette configuration il faut donc choisir entre la médiane et la moyenne, pondérée temporellement. Pour ce faire, on estime les bornes de l'intervalle probabiliste ${}^{(\pm)}_{\mathbb{T}}\psi^{l}_{\Delta_{(\cdot)}}$ représentatif de la stabilité temporelle apparente. Le nombre d'accroissements moyens outrepassant ces bandes de confiance est de deux alors que $\left\{N^{tx.BAC}_{\mathbb{T}} = 3\right\}$. Par conséquent l'estimateur choisi est celui de la médiane pondérée temporellement.

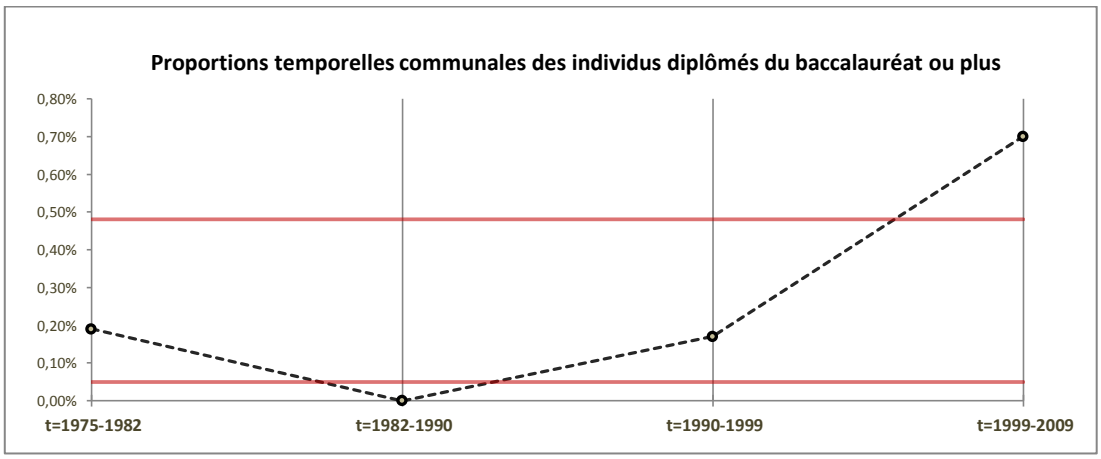


Figure 106 : Analyse de la variabilité temporelle du niveau de cultures dans la commune de Abergement-Clémenciat entre 1975 et 2010 ; Sources : INSEE

Cette procédure a été appliquée à tous les $x^{l:INSEE}_{(U_u),t}$ et $x^{l:ONDRP}_{(U_u),t}$ par le biais d'un algorithme programmé en VBA (Premium Consultants, 2008).

L'effet de support* temporel a été maximisé, par conséquent il convient de maximiser l'effet de support* spatial en appariant les $x_{(U_u)}^{l:SOCIO.ECO}$ aux Uk.

PROCESSUS D'HARMONISATION SPATIALE DIACHRONIQUE

Remarques liminaires :
A l'instar de la modélisation des FE-SAN* par les i.st. $x'^{l:SAN}_{(U_u)}$ à partir des variables DREES, les $x_{(U_u)}^{l:SOCIO.ECO}$ construits via les indicateurs géographiques INSEE et ONDR doivent être appariés aux Uk *de la BD-SIG géofla 2003* (IGN, 2004) – *dans une logique territoriale cohérente* - qui tient compte *des processus diachroniques de création/suppression ou de fusion/division de communes* (Bellin, Morin et al., 2011).
Hypothèse
Elle est identique à celle déclinée dans le processus diachronique d'appariement spatial des i.st.e* destinés la modélisation des FE-SAN* dans la section B.

Proposition :
Elle est identique à celle déclinée dans le processus diachronique d'appariement spatial des i.st.e* destinés à la modélisation des FE-SAN* dans la section B.

Les i.st.e* $x_{(U_k)}^{l:SOCIO.ECO}$ qui modélisent les FE-SOCIO.ECO* sont *a priori* robustes puisque leur biais conditionnel* a été minimisé. Les résultats obtenus sont présentés ci-après.

PRESENTATION DES RESULTATS ET REMARQUES

Les résultats cartographiques de la modélisation des FE-SOCIO.ECO* *pertinents* et Curieux**, par le biais des i.st.e* $x_{(U_k)}^{l:SOCIO.ECO}$ proposés, permettent de modéliser l'exposition géographique *à des situations à risques* ou *à des substances toxiques potentiellement présentes dans les milieux eau-air-sol et biologiques* à partir des caractéristiques spatiotemporelles des conjonctures territoriales.

Les FE-SOCIO.ECO* représentent la dimension *a spatiale de* l'accès aux soins et lorsqu'ils sont défavorables, ils peuvent induire des prédispositions géographiques morbides – et en l'occurrence ceux intégrés sont potentiellement liées aux PM* d'intérêt.

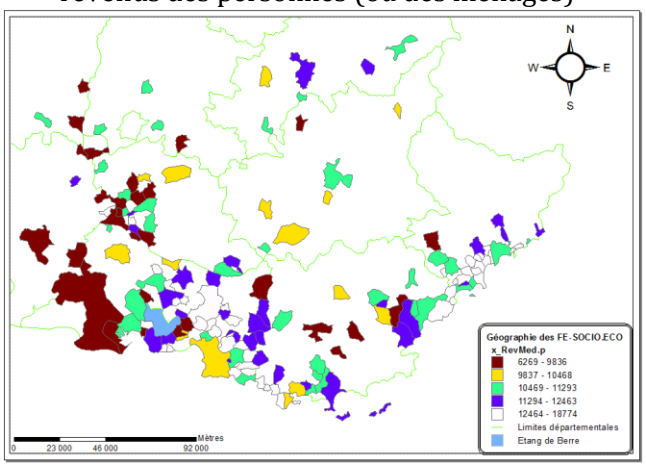
Pour les même raisons que pour les FE-SAN* les résultats cartographiques de la modélisation des FE-SOCIO.ECO* sont présentés uniquement pour les U_k de la région PACA et l'indicateur $SpaLea^2_{U_{k_o}}$ représentant l'incertitude associée à l'identification des U_k n'est pas affiché. Aussi, la documentation des cartes décline l'i.st.e* $x_{(U_k)}^{l:SOCIO.ECO}$, le type de variabilité spatiale modélisée et le tableau statistique joint renvoie aux principaux paramètres *de position et de dispersion*, estimés sur l'ensemble des U_k.

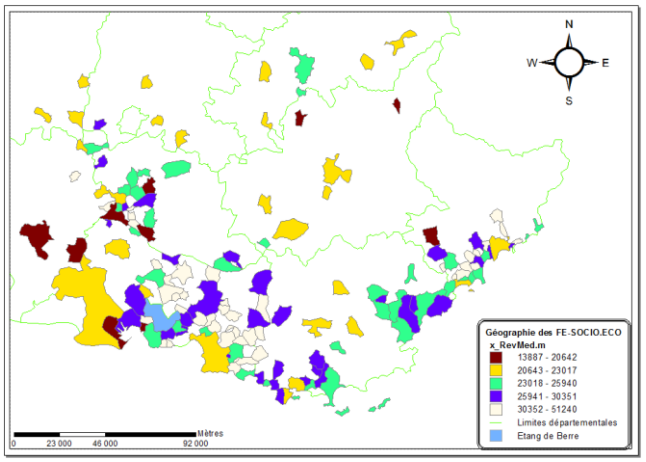
Des remarques sont déclinées uniquement lorsque des singularités spatiales sont observables.

CARTOGRAPHIES ET STATISTIQUES DE LA GEOGRAPHIE DES FE-SOCIO.ECO

Géographie des comportements vis-à-vis du recours à l'offre de soins territoriale.

Variabilité spatiotemporelle des niveaux de revenus des personnes (ou des ménages)

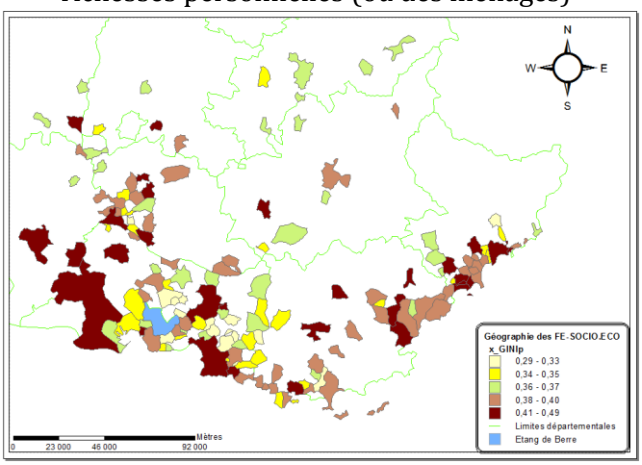


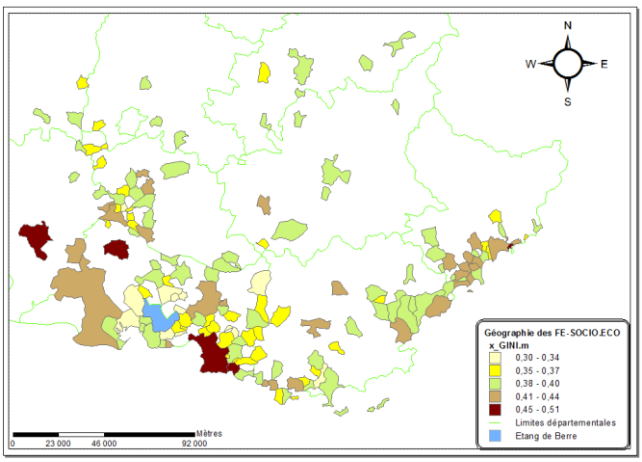


Statistique	$x_{(U_k)}^{RevMed.p}$ (€)	$x_{(U_k)}^{RevMed.m}$ (€)
$\hat{m}\hat{o}y(\cdot)$	11193	25603
$\hat{\sigma}(\cdot)$	1780	5773
$\hat{Q}_1(\cdot)$	10006	21235
$\hat{m}\hat{e}d(\cdot)$	10778	24313
$\hat{Q}_3(\cdot)$	12087	29120

Figure 107 : i.st.e* :$x_{(U_k)}^{RevMed.p}$ (au dessus), $x_{(U_k)}^{RevMed.m}$ (en-dessous) Les revenus nets médians spatiotemporels déclarés par personne, puis par ménage ou par personne entre 2001et 2010

Variabilité spatiotemporelle de la répartition des richesses personnelles (ou des ménages)

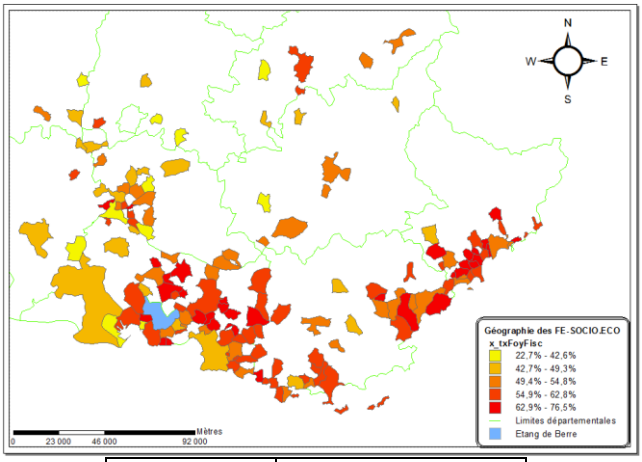




Statistique	$x_{(U_k)}^{GINI.p}$ (s.u.)	$x_{(U_k)}^{GINI.m}$ (s.u.)
$\hat{m}\hat{o}y(\cdot)$	0,38	0,36
$\hat{\sigma}(\cdot)$	0,036	0,037
$\hat{Q}_1(\cdot)$	0,36	0,33
$\hat{m}\hat{e}d(\cdot)$	0,37	0,36
$\hat{Q}_3(\cdot)$	0,40	0,39

Figure 108 : $x_{(U_k)}^{GINI.p}$ (au dessus), $x_{(U_k)}^{GINI.m}$ (en dessous) L'indice de GINI global spatiotemporel exprimé par individu, puis par ménage entre 2001et 2010

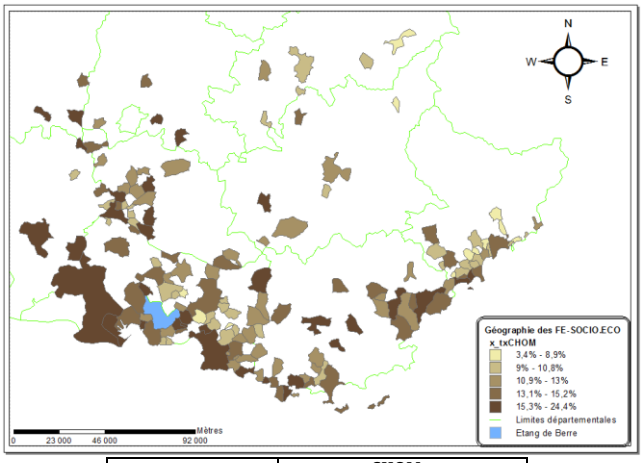
Variabilité spatiotemporelle des niveaux de vie



Statistique	$x_{(U_k)}^{tx.FoyFisc}$ (%)
môy(·)	52
$\hat{\sigma}(\cdot)$	10,6
$\hat{Q}_1(\cdot)$	44
mêd(·)	52
$\hat{Q}_3(\cdot)$	61

Figure 109 : i.st.e* : $x_{(U_k)}^{tx.FoyFisc}$ La proportion spatiotemporelle communale de foyers fiscaux imposables entre 1999et 2010

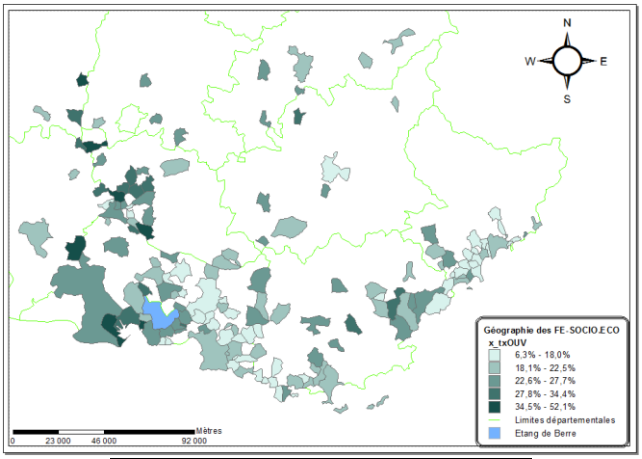
Variabilité spatiotemporelle des niveaux de précarité socioprofessionnelle



Statistique	$x_{(U_k)}^{tx.CHOM}$ (%)
môy(·)	**0,12**
$\hat{\sigma}(\cdot)$	**0,037**
$\hat{Q}_1(\cdot)$	**0,09**
mêd(·)	**0,12**
$\hat{Q}_3(\cdot)$	**0,14**

Figure 110 : i.st.e* $x_{(U_k)}^{tx.CHOM}$ La proportion spatiotemporelle de chômeurs dans la population active entre 1999 et 2010

Variabilité spatiotemporelle des niveaux des catégories socio-professionnelles désavantagées

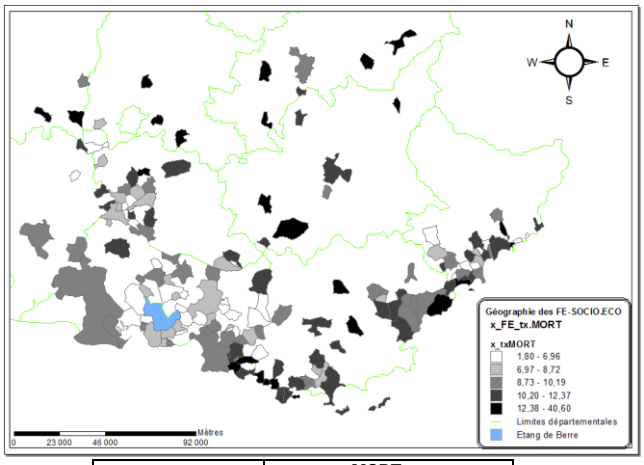


Statistique	$x_{(U_k)}^{tx.OUV}$ (%)
môy(·)	**0,26**
$\hat{\sigma}(\cdot)$	**0,091**
$\hat{Q}_1(\cdot)$	**0,19**
mêd(·)	**0,25**
$\hat{Q}_3(\cdot)$	**0,32**

Figure 111 : i.st.e* : $x_{(U_k)}^{tx.OUV}$La proportion spatiotemporelle d'ouvriers parmi tous les actifs et exerçant un emploi entre 1999 et 2010

Géographie de la qualité des politiques menées en matière de durabilité et de leurs répercussions sur les attraits sociaux, économiques et sanitaires des territoires

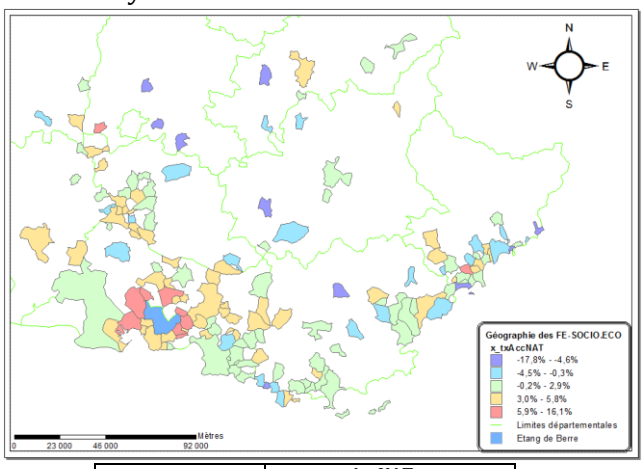
Variabilité spatiotemporelle des niveaux de mortalité



Statistique	$x_{(U_k)}^{tx.MORT}$ (‰)
$\hat{m}oy(\cdot)$	10,12
$\hat{\sigma}(\cdot)$	4,12
$\hat{Q}_1(\cdot)$	7,45
$\hat{m}ed(\cdot)$	9,33
$\hat{Q}_3(\cdot)$	11,62

Figure 112 : i.st.e* : $x_{(U_k)}^{tx.MORT}$: Le taux de mortalité spatiotemporelle entre 1999 et 2010

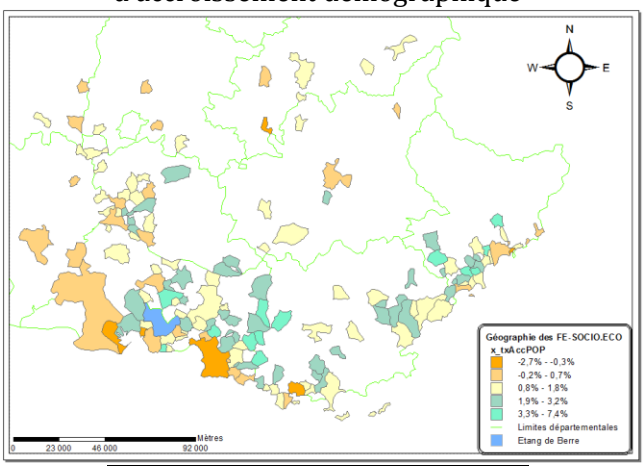
Variabilité spatiotemporelle des niveaux moyens de l'accroissement naturel



Statistique	$x_{(U_k)}^{txAccNAT}$ (%)
$\hat{m}oy(\cdot)$	0,02
$\hat{\sigma}(\cdot)$	0,041
$\hat{Q}_1(\cdot)$	0,00
$\hat{m}ed(\cdot)$	0,03
$\hat{Q}_3(\cdot)$	0,05

Figure 113 : i.st.e: $x_{(U_k)}^{txAccNAT}$ Le taux d'accroissement naturel spatiotemporel entre 1982 et 2010

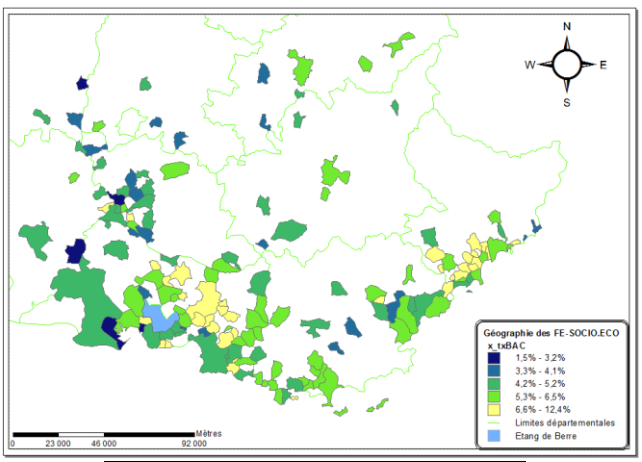
Variabilité spatiotemporelle des niveaux d'accroissement démographique



Statistique	$x_{(U_k)}^{txAccPOP}$ (%)
$\hat{m}oy(\cdot)$	0,01
$\hat{\sigma}(\cdot)$	0,013
$\hat{Q}_1(\cdot)$	0,00
$\hat{m}ed(\cdot)$	0,01
$\hat{Q}_3(\cdot)$	0,02

Figure 114 : i.st.e: $x_{(U_k)}^{txAccPOP}$ Le taux géométrique d'accroissement démographique spatiotemporel entre 1982 et 2010

Variabilité spatiotemporelle des niveaux culturels

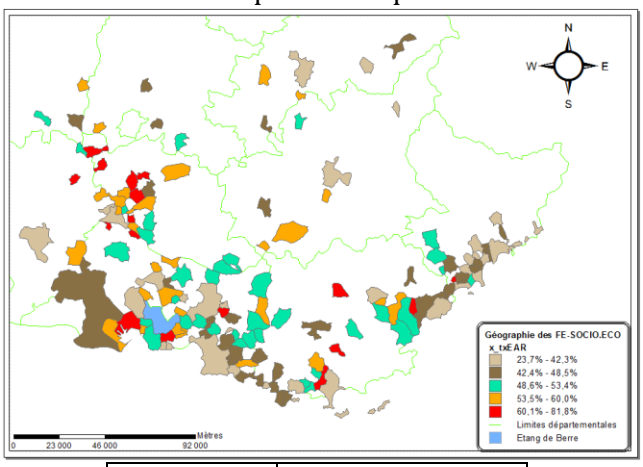


Statistique	$x_{(U_k)}^{tx.BAC}$ (%)
$\hat{m}oy(\cdot)$	5
$\hat{\sigma}(\cdot)$	2
$\hat{Q}_1(\cdot)$	3
$\hat{m}ed(\cdot)$	5
$\hat{Q}_3(\cdot)$	6

Figure 115 : i.st.e* $x_{(U_k)}^{tx.BAC}$: proportion spatiotemporelle d'individus diplômés au moins du baccalauréat entre 1982 et 2010

Géographie des expositions potentielles à des substances toxiques liées aux spécialisations socio-économiques et socio-professionnelles des territoires

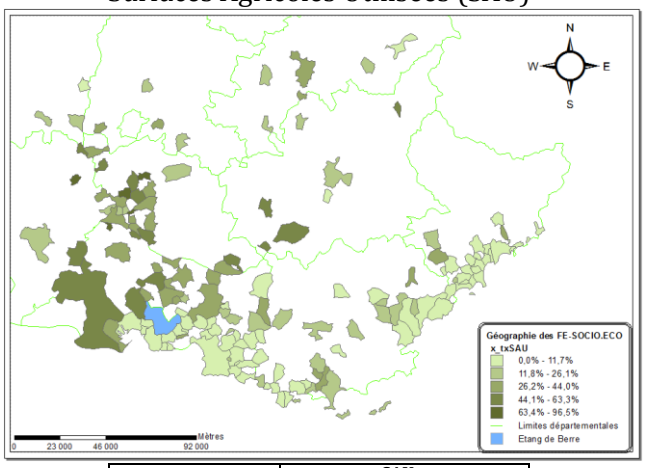
Variabilités spatiotemporelles des niveaux d'Emplois A Risques



Statistique	$x_{(U_k)}^{tx.EAR}$ (%)
$\hat{m}oy(\cdot)$	51
$\hat{\sigma}(\cdot)$	10,7
$\hat{Q}_1(\cdot)$	44
$\hat{m}ed(\cdot)$	51
$\hat{Q}_3(\cdot)$	58

Figure 116 : i.st.e* $x_{(U_k)}^{tx.EAR}$: proportion spatiotemporelle d'individus exerçant une fonction dans un secteur d'activité où l'exposition à des substances toxiques est possible, entre 1982 et 2010

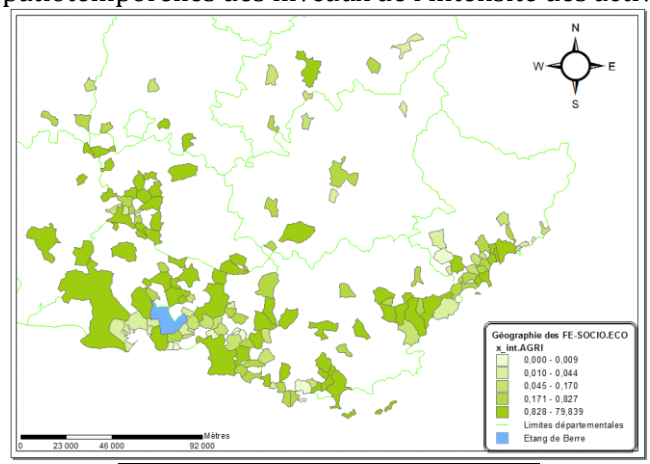
Variabilités spatiotemporelles des niveaux de Surfaces Agricoles Utilisées (SAU)



Statistique	$x_{(U_k)}^{tx.SAU}$ (%)
$\hat{m}oy(\cdot)$	28
$\hat{\sigma}(\cdot)$	22
$\hat{Q}_1(\cdot)$	9
$\hat{m}ed(\cdot)$	23
$\hat{Q}_3(\cdot)$	43

Figure 117 : i.st.e* $x_{(U_k)}^{tx.SAU}$: proportion spatiotemporelle des surfaces communales allouées à l'agriculture entre 1968 et 2000

Variabilités spatiotemporelles des niveaux de l'intensité des activités agricoles

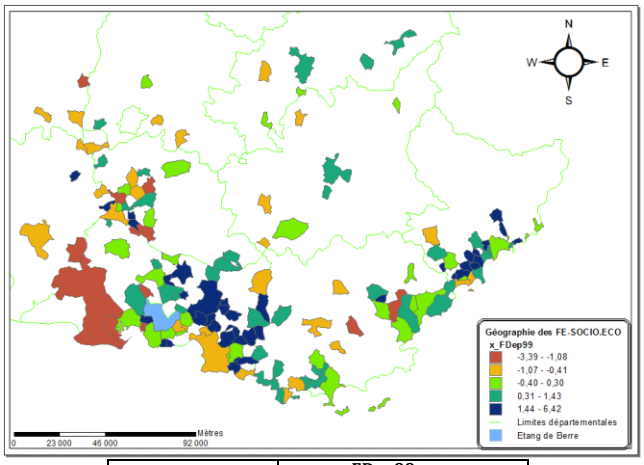


Statistique	$x_{(U_k)}^{int.AGRI}$ (SU)
$\hat{m}oy(\cdot)$	1,37
$\hat{\sigma}(\cdot)$	6,071
$\hat{Q}_1(\cdot)$	0,01
$\hat{m}ed(\cdot)$	0,08
$\hat{Q}_3(\cdot)$	0,47

Figure 118 :. i.st.e* $x_{(U_k)}^{int.AGRI}$: intensité spatiotemporelle des activités agricoles communales entre 1968 et 2000

Géographie des prédispositions contextuelles aux phénomènes morbides

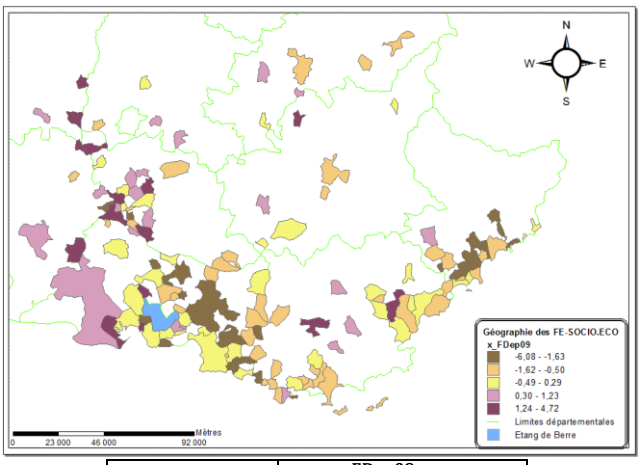
Variabilités spatiotemporelles des niveaux de la défaveur sociale 1999



Statistique	$x_{(U_k)}^{FDep99}$ (SU)
$\widehat{m}oy(\cdot)$	0,2
$\widehat{\sigma}(\cdot)$	1,053
$\widehat{Q}_1(\cdot)$	-0,87
$\widehat{m}\hat{e}d(\cdot)$	-0,06
$\widehat{Q}_3(\cdot)$	1,03

Figure 119 : i.st.e* $x_{(U_k)}^{FDep99}$: FDep.99 est la projection sur l'axe principal d'une ACP des revenus médians des ménages en 2001, du taux de chômage, de la proportion d'individus diplômés du bac, du taux d'emplois ouvriers - en 1999, pondéré par la population ciblée à la même date ; et caractérisant les états de santé entre 1997 et 2001

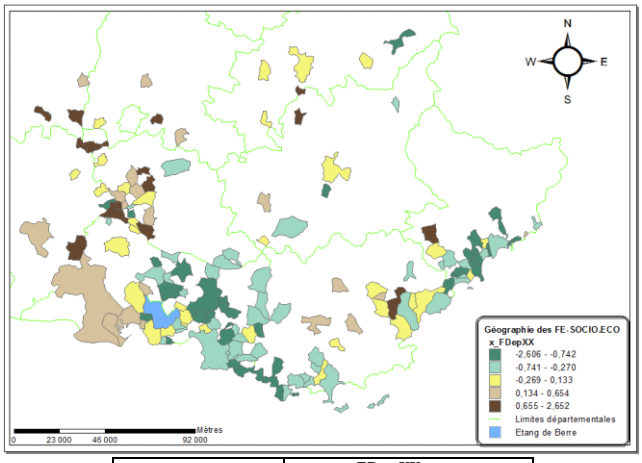
Variabilités spatiotemporelles des niveaux de la défaveur sociale 2009



Statistique	$x_{(U_k)}^{FDep09}$ (SU)
$\widehat{m}oy(\cdot)$	-0,19
$\widehat{\sigma}(\cdot)$	1,635
$\widehat{Q}_1(\cdot)$	-1,33
$\widehat{m}\hat{e}d(\cdot)$	-0,11
$\widehat{Q}_3(\cdot)$	1,10

Figure 120 : i.st.e* $x_{(U_k)}^{FDep09}$: FDep.09 est la projection sur l'axe principal d'une ACP : Des revenus médians des ménages en 2011 du taux de chômage, de la proportion d'individus diplômés du base, du taux d'emplois ouvriers - en 2009, pondéré par la population ciblée à la même date, et caractérisant, par hypothèse, les états de santé entre 2007 et 2010

Variabilités spatiotemporelles des niveaux probables de la défaveur sociale temporelle

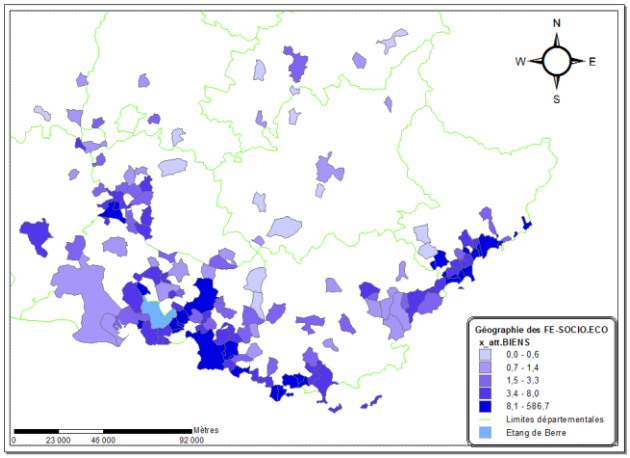


Statistique	$x_{(U_k)}^{FDepXX}$ (SU)
$\widehat{m}oy(\cdot)$	-0,05
$\widehat{\sigma}(\cdot)$	0,836
$\widehat{Q}_1(\cdot)$	-0,60
$\widehat{m}\hat{e}d(\cdot)$	-0,09
$\widehat{Q}_3(\cdot)$	0,54

Figure 121 : i.st.e* $x_{(U_k)}^{FDepXX}$: index FDep.XX est la moyenne réduite des index $x_{(U_k)}^{FDep99}$ et $x_{(U_k)}^{FDep09}$. Il caractérise probablement, par hypothèse, les états de santé des populations communales entre 1997 et 2010

Géographie du stress potentiellement perçu et induit par l'insécurité territoriale contextuelle

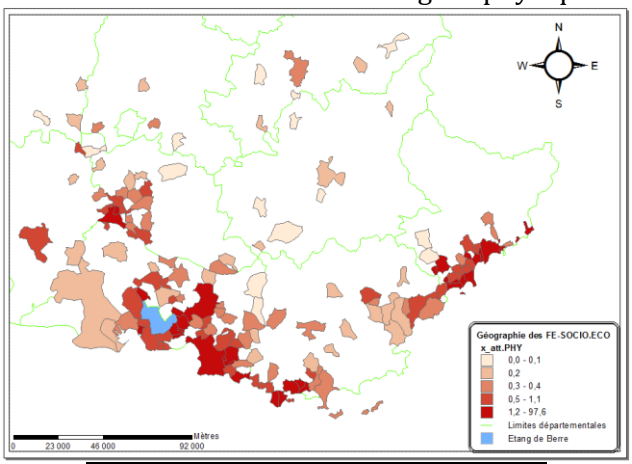
Variabilités spatiotemporelles des niveaux de stress lié à des atteintes aux biens matériels.



Statistique	$x^{att.BIENS}_{(U_k)}$(nb.infactions/ha)
$\hat{m}oy(\cdot)$	10,89
$\hat{\sigma}(\cdot)$	46,998
$\hat{Q}_1(\cdot)$	0,80
$\hat{m}ed(\cdot)$	2,17
$\hat{Q}_3(\cdot)$	6,68

Figure 119 : i.st.e* $x^{att.BIENS}_{(U_k)}$: Mesure spatiotemporelle du nombre global d'infractions annuelles, entre 1996 et 2010, perpétrées à l'encontre des biens matériels des populations in situ ,et rapporté à la taille des communes

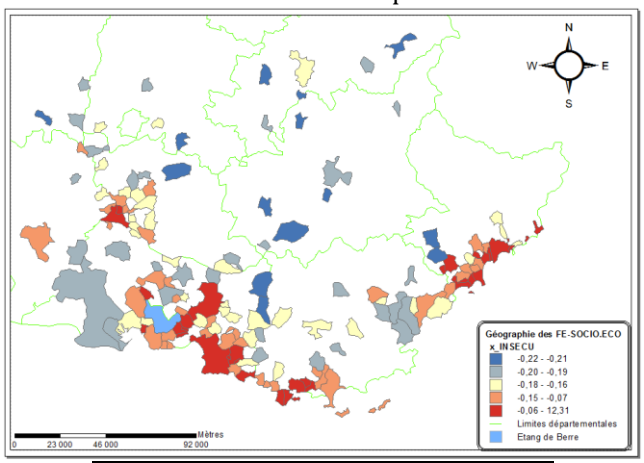
Variabilités spatiotemporelles des niveaux de stress lié à des atteintes à l'intégrité physique.



Statistique	$x^{att.PHY}_{(U_k)}$(nb.infactions/ha)
$\hat{m}oy(\cdot)$	1,6
$\hat{\sigma}(\cdot)$	7,765
$\hat{Q}_1(\cdot)$	0,11
$\hat{m}ed(\cdot)$	0,3
$\hat{Q}_3(\cdot)$	0,84

Figure 122 : i.st.e* $x^{att.PHY}_{(U_k)}$: Mesure spatiotemporelle du nombre global d'infractions annuelles, entre 1996 et 2010, portant atteinte à l'intégrité physique des populations in situ et rapporté à la taille des communes

Variabilités spatiotemporelles des niveaux de stress lié à un sentiment d'insécurité composite engendré par des infractions multiples.



Statistique	$x^{INSECU}_{(U_k)}$ (SU)
$\hat{m}oy(\cdot)$	0
$\hat{\sigma}(\cdot)$	1
$\hat{Q}_1(\cdot)$	-0,20
$\hat{m}ed(\cdot)$	-0,18
$\hat{Q}_3(\cdot)$	-0,10

Figure 123 : i.st.e* $x^{INSECU}_{(U_k)}$: indicateur spatiotemporel composite du sentiment d'insécurité globale annuel relatif à des infractions protéiformes perpétrées à l'encontre des populations communales, entre 1996 et 2010

REMARQUES

Les résultats cartographiques ainsi que les tableaux statistiques qui les complètent permettent de montrer que :

Globalement tous les $x_{(U_k)}^{l:FE-SOCIO.ECO}$ présentent de fortes variabilités spatiales, très hétéroclites, et intimement liées à la nature du FE-SOCIO.ECO* modélisé. En effet, certains écarts-types prennent des valeurs très élevées et parfois largement supérieures à la moyenne spatiale, qu'il s'agisse de ceux confectionnés à partir des indicateurs géographiques INSEE ou avec les index temporels spatialisés ONDRP. C'est le cas par exemple pour les i.st.e* $x_{(U_k)}^{INSECU}$, $x_{(U_k)}^{att.PHY}$; $x_{(U_k)}^{FDep09}$; $x_{(U_k)}^{RevMed.m}$ et $x_{(U_k)}^{tx.EAR}$. Il y a donc une quantité de chances non négligeable pour que certains d'entre eux soient identifiés comme des Déterminants Environnementaux de Santé* (DES).

Il convient cependant de remarquer que les i.st.e* $x_{(U_k)}^{RevMed.m}$ et $x_{(U_k)}^{RevMed.p}$, $x_{(U_k)}^{GINI.m}$ et $x_{(U_k)}^{GINI.p}$ ainsi que $x_{(U_k)}^{att.BIEN}$ et $x_{(U_k)}^{att.PHY}$ présentent des similarités visuelles sur les cartographie des U_k en région PACA et des caractéristiques statistiques relativement proches, parfois presque *fonctionnelles* (Saporta, 2006).

Par conséquent il est fort probable que ces i.st.e* SOCIO.ECO soient redondants. Il aurait peut-être été préférable de les fusionner à l'instar de : $x_{(U_k)}^{INSECU}$ à partir $x_{(U_k)}^{att.BIEN}$ et $x_{(U_k)}^{att.PHY}$, où de $x_{(U_k)}^{FDepXX}$ avec $x_{(U_k)}^{FDep99}$ et $x_{(U_k)}^{FDep09}$ et par la suite, à l'aune de ce qui a été fait, de supprimer les i.st.e* *intermédiaires* – i.e. ceux utilisés pour les confectionner – qui sont nécessairement auto-corrélés peu ou prou, avec les i.st.e* agrégés définitifs.

Toutefois, comme la stratégie de sélection VSURF – utilisée pour l'identification des DES* géographiques - est une procédure de dataminig fondée sur une méthode d'ensemble non paramétrique, randomisée, et qui plus est, qu'elle est éliminatrice, le fait d'introduire des variables redondantes n'altère en rien la puissance statistique de ce modèle. D'ailleurs seul VSURF sera à même d'évaluer le rôle explicatif ou contributif des i.st.e* proposés avec les i.st.m* construits à partir des données LEA et par la même occasion de mettre en évidence d'éventuelles redondances – puisqu'il s'agit aussi de l'un des objectifs intrinsèques de VSURF (Genuer, Poggi et al., 2013).

Les FE-SOCIO.ECO* pertinents* et Curieux* ont été modélisés par le biais des i.st.e* $x_{(U_k)}^{l:FE-SOCIO.ECO}$ *a priori* robustes puisque leur biais conditionnel* a été minimisé.

Il convient d'énoncer, à présent, les stratégies proposées pour modéliser les variabilités spatiotemporelles de la dernière composante environnementale, i.e. les FE-PHY.CHIM.

SECTION D) FACTEURS ENVIRONNEMENTAUX PHYSICOCHIMIQUES

Les Facteurs Environnementaux physicochimiques (FE-PHY.CHIM) caractérisent la qualité environnementale géographique des milieux vie qui est abordée dans toute sa plénitude (World Health Organization, 2009). Les FE-PHY.CHIM* modélisent la variabilité spatiale d'expositions chroniques *à de faibles doses* à des substances toxiques présentes dans les *milieux environnementaux*, i.e. l'eau, l'air, le sol et les matrices biologiques ou dans *les milieux de contact*, i.e. les éléments des milieux environnementaux susceptibles d'engendrer des contaminations humaines.

Les études cliniques manquent de recul pour évaluer les risques réels de ces expositions. Les recherches conduites en géographie de la santé sont controversées. Mais les liens de causalité sont prégnants. Avec la puissance des modèles mathématiques actuels (Cartier, Villani et al., 2012), la mise à disposition de BD environnementales (Zeitouni, 2006), et la mise en évidence de relations déterministes ou contributives significatives entre des expositions environnementales et des états de santé, *a fortiori* chez les sujets prédisposés, à partir de données cohortes (Gehring, Casas et al., 2013) un regain d'intérêt pour ce type d'expositions géographiques est observé en Europe (Afsset, 2009a), à tel point que l'intégration des expositions environnementales chroniques combinées à des substances toxiques liées aux caractéristiques industrielles ou naturelles des territoires est aujourd'hui inéluctable dans les études menées en géographie de la santé (Leux et Guénel, 2010).

Dans le cadre de cette recherche on distingue *les expositions environnementales potentielles* modélisées à partir de mesures environnementales ou d'indicateurs géographiques représentatifs *des milieux environnementaux* (Overmars, Verburg et al., 2008), *des expositions environnementales intrinsèques* qui sont estimées à partir de doses d'expositions ou d'indicateurs de risques caractérisant la contamination des organismes humains depuis les *milieux de contact* (Caudeville, Bonnard et al., 2012). Les i.st.e* estimées à partir de ces dernières ont une distance *a-spatiale morbide** plus proche des populations et peuvent sembler plus fiables pour caractériser les états de santé. Or, ils sont criblés d'incertitudes protéiformes qui posent un problème de fiabilité et les rendent parfois même *inopérants* à des échelles territoriales (Afsset, 2009a).

Le caractère *potentiel ou intrinsèque* de la variabilité des expositions dépend des données disponibles mais l'une comme l'autre de ces informations sont *pertinentes* dès lors que l'état des connaissances suggère une toxicité des doses environnementales d'*expositions,* ou au moins des effets biologiques délétères avérés au regard des PM* d'intérêt. Toutefois les stratégies d'intégration des FE-PHY.CHIM* se focalisent sur les méthodes et les caractéristiques granulaires des données mobilisées. Les BD retenues pour modéliser la géographie des FE-PHY.CHIM* *pertinents* et Curieux** contiennent des variables dont les caractéristiques granulaires d'échelle, de temporalité, de précision, et de lacune ont été décrites (chapitre.1).

En somme, la modélisation des FE-PHY.CHIM* pertinents* intégrés caractérise la variabilité spatiale des :

- Expositions environnementales *potentielles* à des paramètres géophysiques, et en l'occurrence au rayonnement global, aux températures et à la pluviométrie – à partir des données météorologiques (Météo-France, 2010) ainsi qu'aux différentiels des niveaux altimétriques - à partir des données géofla (IGN, 2004) ;
- Expositions à la radioactivité environnementale *intrinsèques* liées à la globalité des rayonnements γ dans l'air, et *potentielles* à l'ensemble des particules α et β ainsi que des isotopes radioactifs du Tritium dans l'eau, à la dégradation du Plutonium 238, du Césium 137 et de l'Antimoine 125 et les sols fins, ou à celle de l'iode 131, du Strontium 90 et du Césium 137 dans le lait - à partir des mesures du portail RNM* (ANS et IRSN, 2013) ;
- Expositions Géographiques à des Radionucléides Artificiels (EGRA) potentiellement diffusées dans les milieux environnementaux liés à la proximité spatiale des INB* en fonctionnement – à partir du répertoire de l'ANS (ASN, 2011)

- Expositions potentielles à la radioactivité tellurique liée à la présence du gaz radon dans les habitations – à partir de l'Atlas Radon (IRSN, 2001).
- Expositions *intrinsèques* des contaminations humaines par certains métalloïdes, en l'occurrence : le Chrome, le Cadmium et le Plomb, ou combinées, i.e. à l'ensemble des Eléments Traces Métalliques (ETM) présents ans l'air – à partir de variables issues de PLAINE (INERIS, 2012)
- Expositions *potentielles* à des substances physicochimiques multiples nocives inhérentes à l'usage anthropique des espaces et à ses répercussions sur l'occupation biophysique des sols - en l'occurrence par l'estimation de l'ampleur des zones géographiques agricoles ou industrialisées – à partir des données CORINE Land Cover (CGDD, SOeS, 2009).

Quant à la modélisation des FE-PHY.CHIM* Curieux* de test, elle a pour dessein d'évaluer la variabilité spatiale des expositions environnementales *potentielles liées* aux diffusions géographiques de substances toxiques suite à des feux de forêt de grand ampleur, ou à l'inverse, à l'effet protecteur des zones géographiques *a priori* préventives – toujours à partir des données CLC (CGDD, SOeS, 2009)

Les variabilités spatiales des FE-PHY.CHIM* *pertinents* et Curieux* de test* seront modélisées par des i.st.e* notés : $x_{(U_k)}^{l:PHY.CIM}$, et les stratégies d'intégration spatiotemporelles des données mobilisées sont déclinées subséquemment.

PROPOSITIONS HEURISTIQUES D'INTEGRATION DES FE-PHY.CHIM

Objectif :

Modéliser dans les communes de 1ère espèce : U_k la géographie des FE-PHY.CHIM* en proposant des stratégies robustes d'intégration spatiotemporelle permettant d'embrasser toutes les caractéristiques granulaires des données mobilisées. Les i.st.e* $x_{(U_k)}^{l:PHY.CHIM}$ doivent être représentatifs de la géographie des expositions *potentielles ou intrinsèques* aux substances toxiques intégrables, et aussi adaptés à la finalité recherchée, i.e. l'identification des DES* par une procédure de sélection de variables multidimensionnelles au regard des i.st.m* proposés pour modéliser la géographie des PM* d'intérêt.

Remarques liminaires :

Parmi la pléthore de variables retenues pour leur fiabilité, leur disponibilité sur l'intégralité de la France métropolitaine et leur capacité à modéliser les FE-PHY.CHIM* *pertinents* et Curieux**, la majorité n'est pas directement intégrable. Toutefois, aucune des variables utilisées ne contient de lacunes. En revanche, leurs caractéristiques granulaires sont très disparates et doivent être intégralement prises en compte dans les stratégies d'intégration spatiotemporelles proposées pour construire les i.st.e*

Les variables Météo-France $m_{(s_g,t)}^{l:MétéoFrance}$ sont des mesures environnementales climatologiques déclinées entre 1981 et 2011 avec un pas de temps constant et pour des stations géo-localisées : s_g. Il en est de même pour les données RNM* $m_{(s_g,t)}^{l:RNM}$ qui sont des mesures temporelles localisées des niveaux de radioactivité environnementale et disponibles entre 2008 et 2010, mais pour lesquelles le pas de temps est variable. Les mesures de radioactivité environnementale disponibles sont nombreuses. Celles retenues pour leurs effets suggérés ou avérés sur les PM* d'intérêt ont été spécifiées (chapitre.1).

Les variables du répertoire des autorisations d'exploitation des INB* de l'ASN : INB_i^{info} sont des données sémantiques grevées d'informations spatiales (commune de localisation), temporelles (date de service) et sémantiques (nature de l'installation) qui peuvent être intégrées dans un SIG afin d'estimer conjointement, pour chaque U_k, leur nombre et leur proximité géographique.

Les variables de l'Atlas Radon de l'IRSN : $AVR_{(U_u)}$ sont des indicateurs géographiques qualitatifs associés à des classes de valeurs continues, d'amplitudes inégales, représentant l'activité volumique moyenne du radon dans les habitations. Elles sont déclinées à l'échelle des communes U_u et de certains AR_u - une classe unique caractérise l'intégralité de la période temporelle qui s'étend de 1982 à 2000. L'information spatiotemporelle disponible est donc assez grossière. Cependant, les mesures de l'Activité Volumique du Radon (ARV) dans les habitations ont été collectées dans une logique

temporelle départementale et des statistiques attributaires quantitatives sont disponibles à cette échelle. Elles peuvent être fusionnées pour améliorer la qualité informationnelle des sources.
Les variables issues de PLAINE sont des données INERIS. Elles permettent de caractériser l'exposition spatiotemporelle intrinsèque de populations spécifiques ETM par des approches multi-milieux entre 1990 et 2009. La population ciblée se constitue d'enfants et d'adolescents.

En l'occurrence, $x_{(U_k)}^{DEJ(Cr)}$ représente la Dose Journalière des Expositions au Chrome par ingestion ou inhalation rapportée à la masse moyenne des populations ciblées (Caudeville J., Boudet C. et al., 2012); Ensuite, les trois *indicateurs de risque spatial* : $x_{(U_k)}^{irs(\cdot)}$ touchent à l'ingestion de particules de sol dans l'eau potable et de denrées alimentaires ainsi qu'à l'inhalation de particules présentes dans l'air et rapportées aux doses journalières maximales tolérées par les organismes humains pour le Plomb (Pb), le Cadmium (Cd) et le Nickel (Ni) (Caudeville, Bonnard et al., 2012); Enfin, un Proxy du Risque d'Exposition à des substances Chimiques $x_{(U_k)}^{PREC}$ par inhalation de toutes les substances chimiques présentes dans l'air ambiant émises depuis les installations industrielles recensées par l'IrEP (Inventaire des émissions polluantes) - ont été directement intégrés (Caudeville, 2011). Aucun traitement additionnel n'a été appliqué. Ces variables constituent les i.st.e. définitifs et leurs interactions plausibles avec les PM* étudiés ont été décrites dans l'état de l'art.

Les niveaux altimétriques moyens des U_k : $x_{(U_k)}^{TOPO}$ ont eux aussi été intégrés sans traitement spécifique. Il s'agit d'informations attributaires inhérentes à la BD SIG géofla 2003 (IGN, 2004). Ils ont initialement été utilisés comme variables auxiliaires de cokrigeage (voir intégration des variables Météo-France et RNM). L'interaction plausible des niveaux géographiques avec les PM* étudiés ont été décrits dans l'état de l'art.
Les variables de la BD géographique CORINE Land Cover (CLC) de l'AEE : $e.clc_t^{\varrho(CLC)}$ sont des entités géographiques vectorielles caractérisant l'occupation biophysique des sols avec une typologie constituée de 44 postes $\varrho(CLC)$ normalisés par une nomenclature européenne. La taille des objets géographiques $e.clc_t^{i \in \boldsymbol{\varrho(CLC)}}$ varie selon la continuité spatiale des occupations biophysiques homogènes. Les plus petites entités forment des surfaces de 25ha. Les $e.clc_t^{\boldsymbol{\varrho(CLC)}}$ sont pertinentes jusqu'à l'échelle des communes, en revanche, elles ne sont pas immédiatement adaptées à la logique territoriale - puisqu'elles chevauchent les U_u administratives des SIG. Par ailleurs, les variables CLC sont disponibles à trois dates spécifiques : 1990, 2000, 2006, mais dans la mesure où elles ont été confectionnées avec des technologies différentes, des conflits granulaires interviennent et l'utilisation d'une *base intermédiaire révisée* (CGDD, SOeS, 2009) est requise. Ce qui revient à considérer conjointement deux BD-CLC à $t = \{1990; 2000\}$ puis deux autres BD-CLC pour les temporalités : $t = \{2000.r;\ 2006\}$.

Hypothèse principale :
Les variables utilisées pour confectionner les i.st.e* $x_{(U_k)}^{l:PHY.CHIM}$ doivent être soumises à traitements spatiotemporels afin de minimiser le concept du *biais conditionnel** en embrassant les caractéristiques granulaires des données pour parvenir à des modélisations consistantes, i.e. représentatives de la réalité géographique des FE-PHY.CHIM* intégrables (Marcotte, 2008). Les stratégies d'intégration spatiotemporelles proposent des fonctions mathématiques permettant d'effectuer des fusions statistiques ou des transformations topologiques par incorporation de *données sémantiques attributaires* ou *de connaissances géographiques et médicales expertes*. Il s'agit d'une part d'optimiser *l'effet information**, i.e. amoindrir les incertitudes et les imprécisions entachant les sources utilisées, et d'autre part, de *maximiser l'effet de support**, i.e. améliorer leurs caractéristiques spatiotemporelles afin qu'elles soient adaptées à la finalité recherchée (Baillargeon, 2005).

Le principe d'intégration repose sur la même hypothèse que pour les autres FE, puisqu'il s'agit de proposer des processus d'estimation d'i.st.e* $x_{(U_k)}^{l}$ *minimisant le biais conditionnel**. Mais dans la mesure où les caractéristiques granulaires des sources utilisées pour modéliser les FE-PHY.CHIM* sont

complètement disparates, les méthodes ainsi que la chronologie des processus d'optimisation de *l'effet information** et de maximisation *de l'effet de support** diffèrent radicalement d'une variable input à l'autre.

Proposition méthodologique principale :
Les stratégies d'estimation des i.st.e* destinées à modéliser la géographie des FE-PHY.CHIM* en minimisant le *biais conditionnel** se fondent sur des méthodes pluridisciplinaires.
Les SIG sont les instruments qui ont été retenus car ils sont *parfaitement adaptés* à l'intégration spatiotemporelle de *données physicochimiques diverses* permettant de caractériser *la qualité environnementale des espaces géographiques* (Overmars, Verburg et al., 2008). Les transformations topologiques des caractéristiques granulaires visent à améliorer les dimensions spatiales et temporelles des sources afin de concevoir des *espaces géographiques numériques de simulation* adaptés à l'analyse des interactions *homme-milieu*, dont la thématique santé-environnement fait partie (Tissot et Cuq, 2004). La technologie SIG –en l'occurrence celle d'ArcGis.10 (ESRI, 2013) – propose des outils et des méthodes robustes qui permettent d'acquérir et d'intégrer tous les types de données d'expositions à des substances physicochimiques, pour tous les milieux *environnementaux ou de contact,* et par suite, de procéder à des modélisations géographiques parfaitement représentatives de leurs variabilités spatiotemporelles (Jacquez, Goovaerts et al., 2005).

L'optimisation de *l'effet information** dépend donc de la variable considérée et peut intervenir à différentes phases :

- spécification bibliographique des *substances physicochimiques les plus pertinentes lorsque celles-ci sont douteuses, redondantes ou trop nombreuses pour être toutes prises en compte* (Chasles et Fervers, 2011) - variables : ASN et CLC ;

- *fusions d'informations spatiotemporelles attributaires par des processus de randomisation probabilistes* (Saporta, 2006) ou par injection de connaissances fondées sur la *théorie des ensembles flous* (Dubois Didier, Prade Henri, 2004) et adaptées à la logique *des statistiques territoriales* (Charre, 1995) - outil SIG : Spatial Statistics (ESRI, 2013) ; Variable : Atlas Radon ; ASN ;

- construction de couches de surface à partir *de variables régionalisées** par le biais de processus d'interpolation spatiale (Myers Donald, 1994) et en particulier de méthodes stochastiques issues des géostatiques *uni-variables* (Matheron, 1965) et *multi-variables* (Wackernagel, 2003) – outil SIG : Geostatistical Analyst (ESRI, 2013) ; variables Météo-France et RNM* ;

- capture d'informations géographiques par des processus dynamiques de dilatation/rétraction – ou de désagrégation raster d'entités vectorielles – combinés à des opérateurs ensemblistes de statistiques spatiales (Voiron-Canicio, 1995) (Processus SIG : ModelBuilder (ESRI, 2013) - outils SIG : Analysis (ESRI, 2013), Spatial Analyst (ESRI, 2013) ; variables : ASN ou CLC ;

- uniformisation des échelles à partir de processus *d'agrégation spatiale pondérés par des variables géographiques* (Pumain et Saint-Julien, 1997) - outil SIG : Analysis (ESRI, 2013) ; variable : Atlas Radon).

La maximisation *de l'effet de support** peut intervenir à deux niveaux :

- *l'agrégation verticale* de chroniques spatiotemporelles (Peguy, 1996) à partir d'un processus décisionnel probabiliste, fondé sur le concept de *stabilité temporelle apparente* (Lütkepohl, 1991), permettant de choisir une statistique consistante - sans biais (Saporta, 2006) *, adaptée à la logique territoriale* (Charre, 1995)

- *l'injection de connaissances – ou aux degrés de croyance – expertes,* grâce à la *théorie des ensembles flous* (Dubois Didier, Prade Henri, 2004) - outil VBA : macro-commande (Microsoft, 2013) ; variables : Météo-France et RNM, ou CLC) ; Harmonisation diachronique spatiale - aux U_k - par des procédures d'appariement numériques conçues dans une logique diachronique et territoriale cohérente (Lahousse et Piédanna, 1998) – Outil VBA : macro-commande (Microsoft, 2013) ; Variable Atlas Radon.
Les stratégies d'intégration des variables sont donc spécifiées indépendamment pour chacune des BD mobilisées. A l'exception des $m_{(s_g,t)}^{l:MétéoFrance}$ et des $m_{(s_g,t)}^{l:RNM}$ qui sont des mesures environnementales aux granularités* analogues et pour lesquelles la stratégie d'intégration proposée est identique.

STRATEGIE D'INTEGRATION DES VARIABLES METEO-FRANCE ET RNM

Afin de *minimiser le concept de biais conditionnel** des i.st.e* proposés, la stratégie d'intégration s'opère : d'abord par une maximisation de l'effet de support*, par un processus d'agrégation des chroniques temporelles afin d'obtenir des indicateurs probabilistes verticaux spatialisés : $m^{l}_{(s_g)}$ et ensuite par une *optimisation de l'effet information** en deux étapes : La première est une reconstitution spatiale géostatistique des valeurs inconnues de $m^{l}_{(s_g)}$, notées $m^{l}_{(s_o)}$, le biais d'indicateurs probabilistes verticaux interpolés horizontalement : $\hat{m}^{l}_{(s_o)}$ permettant d'obtenir une estimation en tous points de l'espace géographique*. La seconde est un processus d'estimation de *statistiques zonales* par un *opérateur spatial ensembliste* permettant de capturer les valeurs de $\hat{m}^{l}_{(s_o)}$ contenues dans chaque U_k – afin de supputer les i.st.e. $x^{l}_{(U_k)}$ définitifs. La synoptique de la stratégie du processus d'estimation des $x^{l:MétéoFrance}_{(U_k)}$ et des $x^{l:RNM}_{(U_k)}$ est donnée dans le schéma suivant :

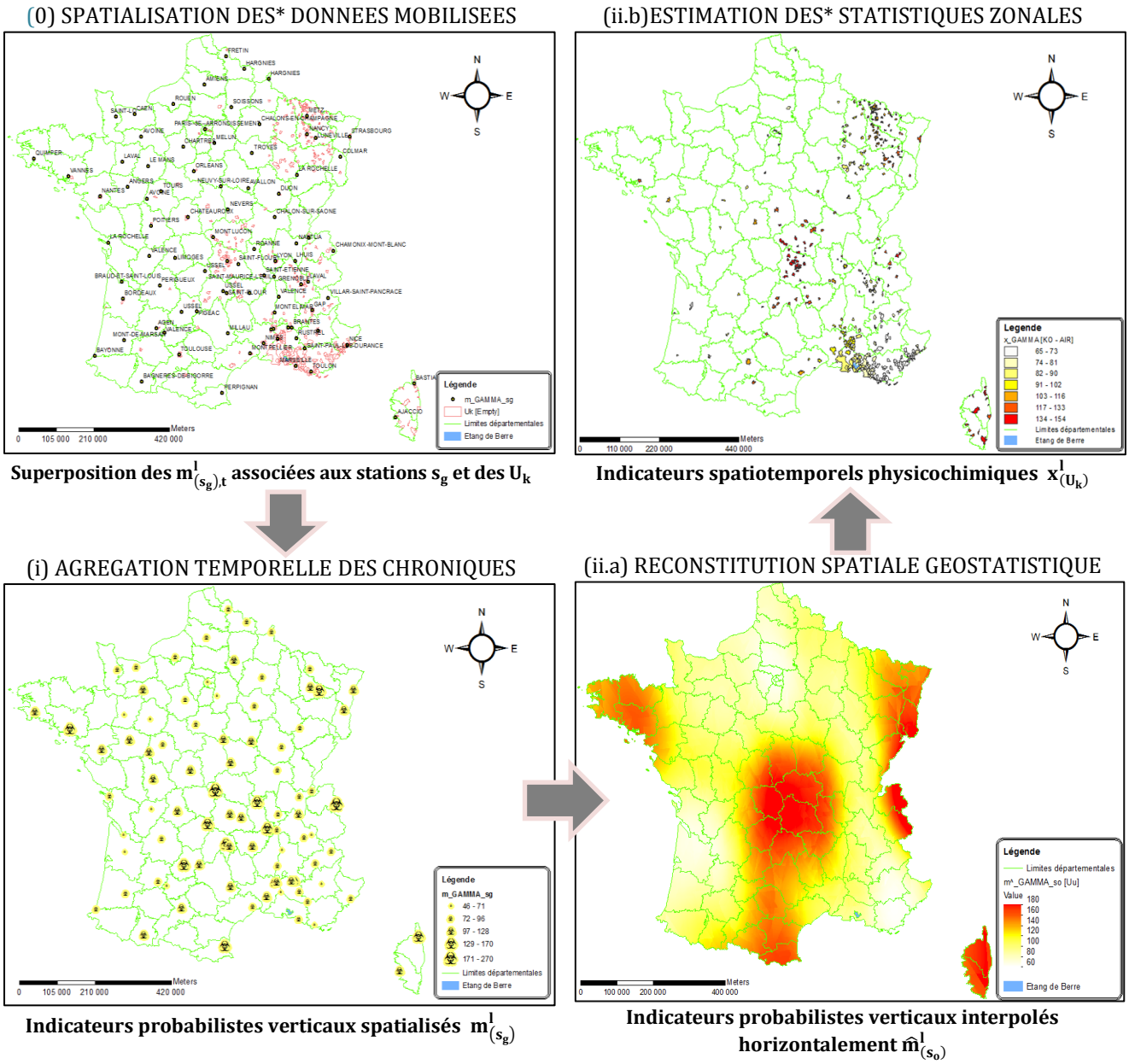


Figure 124 : Synoptique d'intégration spatiotemporelle des FE-PHY.CHIM* à partir des données Météo-France et RNM

La granularité* des chroniques temporelles Météo-France et RNM* et les distances *a-spatiales morbides* avec les PM* d'intérêt ont été déclinées (chapitre.1)

MAXIMISATION DE L'EFFET DE SUPPORT

L'intégration des mesures RNM* et Météo-France nécessite de procéder d'abord à la *maximisation de l'effet de support** avant *d'optimiser l'effet information**, l'objectif étant d'obtenir un estimateur statistique temporel unique pour chaque s_g.

PROCESSUS D'HARMONISATION TEMPORELLE PROBABILISTE

Le processus d'harmonisation temporelle a pour objet l'estimation des indicateurs probabilistes verticaux spatialisés, notés : $m^{l}_{(s_g)}$

Remarques liminaires :

Les variables Météo-France sont des chroniques temporelles géo-localisées $m^{l:Météo-France}_{(s_g,t)}$ au niveau des stations de mesures : s_g et déclinées avec un pas de temps : $\Delta_t^{l:Météo-France}$ constant.

Les variables RNM* sont des chroniques temporelles qui ont été localisées au niveau du centroïde des communes d'appartenance, la notation de site s_g est conservée : $m^{l:RNM}_{(s_g,t)}$, elles sont déclinées avec un pas de temps : $\Delta_t^{l:RNM}$variable.

L'objectif étant d'obtenir pour chaque site s_g une mesure temporelle géo-localisée unique conformément au processus *vertical d'agrégation probabiliste* (Peguy, 1996) fondé sur le concept de *stationnarité temporelle apparente* et par l'analyse statistique des différenciations *temporelles* (Hamilton, 1994), si ce n'est qu'il est appliqué à des mesures géo-localisées, non plus à des indicateurs géographiques territoriaux. Il s'agit donc d'obtenir les indicateurs probabilistes verticaux spatialisés $m^{l}_{(s_g)}$.

Hypothèse

Elle est identique à celle déclinée dans le processus diachronique *vertical d'agrégation probabiliste* destiné à la modélisation des FE-SOCIO.ECO.

Proposition :

Elle est identique à celle déclinée dans le processus diachronique *vertical d'agrégation probabiliste* destiné à la modélisation des FE-SOCIO.ECO.

Application aux données Météo-France :

Pour illustrer la méthode, la stratégie *d'agrégation probabiliste verticale* est appliquée aux paramètres météorologiques : $m^{RAY}_{(s_g,t)}$ i.e. : des moyennes mensuelles des cumuls journaliers du rayonnement global exprimés en j/cm^2. Les données présentées sont celles de la station météo située sur la commune U_u de Marignane et ses coordonnées géographiques $s_g = \left(x.géo_g = 833\,400.\,m;\ y.géo_g = 1830\,400.\,m\right)$. La période temporelle disponible s'étend de 1981 à 2012. Or, celle recouverte par l'étude LEA est $\Delta_t^{LEA} = \{1980, \ldots, 2010\}$. Par conséquent les temporalités utilisées commencent à $t_{1^{RAY}} = jan.\,1981$ et se terminent à $t_{n_t^{RAY}} = jan.\,2011$. Dans la mesure où l'on travaille sur des accroissements, il convient d'utiliser une temporalité supplémentaire.

Sur le graphique, la courbe jaune représente les valeurs des accroissements temporels mensuels : $\Delta\left(m^{RAY}_{(s_g,t)}\right)$.

Pour améliorer la lisibilité, seuls les mois des trois dernières années sont représentés. Les valeurs des bandes de confiance : $\bar{}_{\mathbb{T}}\psi^{RAY}_{\Delta_{(\cdot)}}$ et $^{+}_{\mathbb{T}}\psi^{RAY}_{\Delta_{(\cdot)}}$ permettant de circonvenir les accroissements suggérant une

stabilité temporelle apparente sont représentées par des lignes continues rouges. La valeur moyenne de l'accroissement temporel : $\bar{\Delta}\left(m^{RAY}_{(s_g,t)}\right)$ est représentée par une ligne discontinue noire.

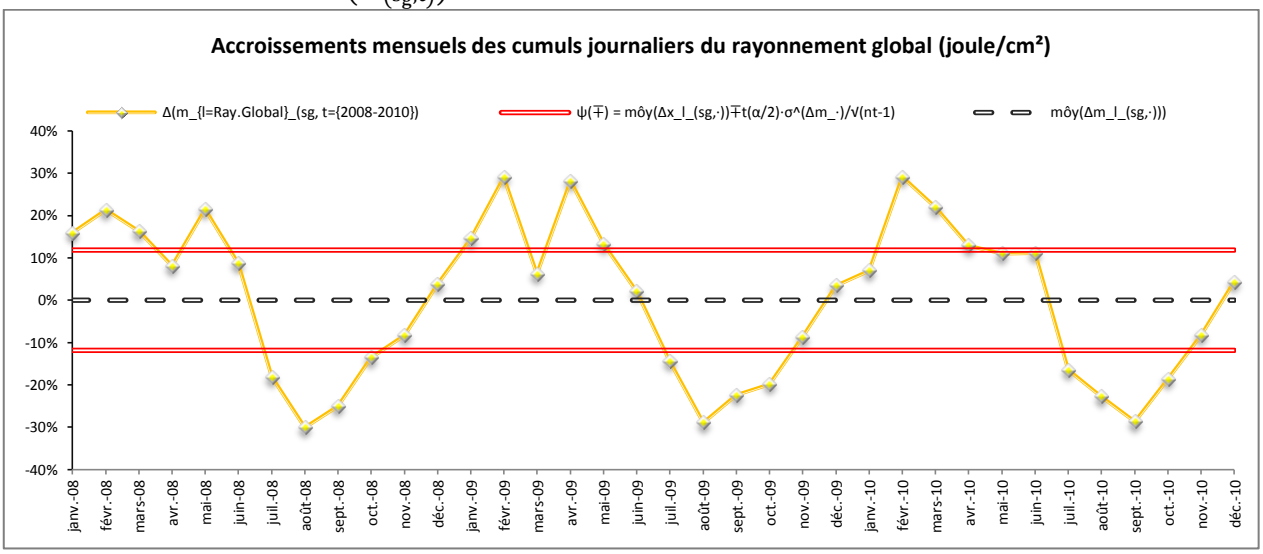


Figure 125 : Diagramme des accroissements mensuels des cumuls journaliers du rayonnement global (joule/cm²) entre 2008 et 2010

Pour que cette chronique soit caractérisée *d'apparemment stationnaire* le nombre de $\Delta\left(m^{RAY}_{(s_g,t)}\right)$ outrepassant les bornes : ${}^{-}_{\mathbb{T}}\psi^{RAY}_{\Delta_{(\cdot)}}$ et ${}^{+}_{\mathbb{T}}\psi^{RAY}_{\Delta_{(\cdot)}}$ doit être strictement inférieur à $N^{l}_{\mathbb{T}} = 277$. Or, ce nombre s'élève à 328. Par conséquent, la chronique temporelle $m^{RAY}_{(s_g,t)}$ de Marignane est qualifiée *d'apparemment volatile*. Il convient de choisir un estimateur représentatif de la variabilité temporelle du rayonnement global auquel sont assujetties les populations localisées. Par conséquent, c'est l'indicateur temporel pondéré de la moyenne spatialisée qui est choisi – qui est identique à celui de la moyenne empirique puisque le pas de temps est constant :

$$m^{RAY}_{(s_g=\text{Marignane})} = môy^{t}_{\mathcal{P}}\left(m^{RAY}_{(s_g,t)}\right) = môy\left(m^{RAY}_{(s_g,t)}\right) = 568\ 122\ \text{joules/cm}^2$$

Cette procédure a été appliquée à chaque site s_g correspondant aux stations Météo-France à la localisation des chroniques RNM. Le processus a été automatisé grâce à une macro-commande (Microsoft, 2013). De fait, tous les *indicateurs verticaux spatialisés* $m^{l:RNM}_{(s_g)}$ et $m^{l:MétéoFrance}_{(s_g)}$ ont pu être estimés. Mais la maximisation de *l'effet de support** ne suffit pas pour obtenir les i.st. $x^{l}_{(U_k)}$. Pour cela il faut reconstituer les valeurs inconnues $m^{l}_{(s_o)}$, sur l'intégralité du territoire français, afin de pouvoir estimer, dans les U_k, la variabilité spatiale des FE-PHY.CHIM.

OPTIMISATION DE L'EFFET INFORMATION

L'optimisation de l'effet information* va permettre d'obtenir les i.st.e* représentatifs de la géographie des expositions à la radioactivité environnementale et aux paramètres météo - $x^{l}_{(U_k)}$ - par un processus en deux étapes, en commençant par estimer les valeurs inconnues des phénomènes d'intérêt sur l'intégralité du territoire $\widehat{m}^{l}_{(s_o)}$, et par la suite, en agrégeant ces valeurs dans les U_k par le biais des statistiques zonales.

TRANSFORMATION TOPOLOGIQUE : INTERPOLATION SPATIALE STOCHASTIQUE DE L'INFORMATION GEOGRAPHIQUE

Le processus d'harmonisation temporel a permis d'obtenir les indicateurs probabilistes verticaux spatialisés : $m^{l}_{(s_g)}$. Il est désormais nécessaire d'estimer leurs valeurs sur l'intégralité du territoire français métropolitain par le biais des indicateurs probabilistes verticaux interpolés horizontalement - notés : $\widehat{m}^{l}_{(s_o)}$.

Remarques liminaires :

Il s'agit de proposer une méthode permettant d'estimer, sur l'intégralité du territoire français métropolitain, les variabilités spatiotemporelles des phénomènes physicochimiques d'intérêt, ces derniers n'étant connus qu'à travers les réalisations fragmentaires des indicateurs probabilistes verticaux spatialisés $m^{l}_{(s_g)}$ estimés auparavant.

Le nombre de sites localisés où les $m^{l}_{(s_g,)}$ sont disponibles est identique à celle des chroniques temporelles inputs $m^{l}_{(s_g,t)}$ spécifiées (chapitre.1)

Spécification de l'hypothèse principale :

Mettre […] l'espace en lumière, [puisque] le temps [a été sorti de] l'ombre (Brunet, 1968).

L'interpolation spatiale permet justement, à partir d'un jeu de données spatialisées - $m^{l}_{(s_g,)}$, d'estimer dans l'espace géographique* toutes les valeurs inconnues $m^{l}_{(s_o)}$. La qualité d'une interpolation spatiale dépend du choix de la méthode utilisée. Elle doit être adaptée à la nature et à la structure spatiale des données d'échantillonnage (Myers Donald, 1994).

La géostatistique s'impose, et concurrence toutes les méthodes déterministes d'interpolation spatiale en inscrivant l'idée du hasard dans la dialectique. C'est une méthode stochastique d'interpolation et les techniques de krigeage sont parfaitement adaptées aux jeux de données géographiques (Baillargeon, 2005).

En considérant les $m^{l}_{(s_g)}$ comme les mesures géo-localisées d'une variable régionalisée (v.r.)., les $m^{l}_{(s_o)}$ peuvent être estimés, en tous points du champ : $\mathcal{D}(\omega_{.})$, par le biais *d'indicateurs probabilistes verticaux interpolés horizontalement :* $\widehat{m}^{l}_{(s_o)}$. En supposant que les accroissements spatiaux des $m^{l}_{(s_g,)}$ sont stationnaires on peut interpréter la v.r. comme les réalisations fragmentaires d'une Fonction Aléatoire Intrinsèque* (FAI). La plausibilité de cette hypothèse peut être appréciée grâce à l'analyse des variogrammes. De fait, un cadre de travail est fixé et les techniques géostatistiques de krigeage sont utilisables (Matheron, 1962).

Lorsque les données d'échantillonnage, les $m^{l}_{(s_g)}$, sont à la fois denses et uniformément réparties dans $\mathcal{D}(\omega_{.})$, le krigeage uni-variable optimise automatiquement *l'effet information** (Marcotte, 2008). Ce faisant les variabilités spatiales de $m^{l}_{(s_o)}$ peuvent être estimées en tous points du territoire français métropolitain grâce au-Krigeage Ordinaire (KO). Cet opérateur géostatistique est qualifié de BLUP : *best linear unbiased predictor* (Matheron, 1963). Et quand la densité spatiale d'échantillonnage est *défaillante,* l'introduction de variables auxiliaires $m^{k}_{(s_g)}$ grevées de l'hypothèse de *stationnarité d'ordre deux conjointe* permet aussi d'estimer en tous points les valeurs prises par $m^{l}_{(s_o)}$ - grâce au Co-Krigeage Ordinaire (CKO), estimateur géostatistique qui possède les mêmes propriétés que le KO (Wackernagel, 2003).

Stratégie d'estimation :

Le vocabulaire, les notions, les spécificités des opérateurs géostatistiques ainsi que la mise en œuvre du KO et du CKO sont spécifiés dans l'annexe 4 : complément théorique sur l'interpolation spatiale : Géostatistiques Uni-Variables et Multi-Variables. L'essentiel de ce complément théorique inhérent à la

compréhension de l'application illustrative de la stratégie d'optimisation de l'effet information* - sur le rayonnement solaire global (RAY) - est décliné ici :

L'estimation des valeurs inconnues de $m^{l}_{(s_o)}$ par les opérateurs du Krigeage Ordinaire (KO) ou du Co-Krigeage Ordinaire (CKO) commence par l'analyse la distribution spatiale des $m^{l}_{(s_{g,})}$ en les assimilant aux mesures d'une v.r.

Le variogramme est un outil géostatistique adapté au krigeage et au cokrigeage. Afin d'apprécier l'hypothèse de stationnarité intrinsèque, le variogramme : $\gamma^{l}(h)$ est estimé par $\hat{\gamma}^{l}(h)$, conditionnellement aux données d'échantillonnage $m^{l}_{(s_{g,})}$.

Il est généralement évalué pour des valeurs de h inférieures à la moitié de la distance maximale d'échantillonnage. Lorsque $\hat{\gamma}^{l}(h)$ dénote une stationnarité intrinsèque le variogramme peut être ajusté à un modèle théorique : $\tilde{\gamma}^{l}(h)$.

Le choix du modèle d'ajustement $\tilde{\gamma}^{l}(h)$ se fait *a priori* selon les caractéristiques *de forme* et *d'échelle inhérentes* décrites par $\hat{\gamma}^{l}(h)$ – dont les principales à estimer sont : L'effet pépite $\hat{\sigma}^{2}_{o,l}$; Le palier $\hat{\sigma}^{2}_{a,l}$ et la portée $\hat{a}_l$.

L'analyse des variogrammes directionnels $\tilde{\gamma}^{l}(h;\theta_k)$ permet de déceler certaines anisotropies spatiales. Leurs causes sont généralement inexplicables. L'anisotropie géométrique est très courante et présente l'avantage d'être facilement décelable par l'allure *en rose des sables* que présente la superposition des $\tilde{\gamma}^{l}(h;\theta_k)$, et corrigeable par une ellipse d'iso-valeur. Toutefois la correction n'a de sens que si le facteur d'anisotropie : F^{l}_{a} est supérieur ou égal à 1.5.

Le variogramme empirique sert de support à l'ajustement $\tilde{\gamma}^{l}(h)$ – et le modèle Gaussien additionné d'une *pépite* est souvent utilisé.

Conséquemment l'opérateur du KO peut être construit. La résolution du système de krigeage s'effectue conditionnellement à $\tilde{\gamma}^{l}(h)$ et aux données disponibles $m^{l}_{(s)}$. Le KO permet de s'affranchir de l'estimation globale de la moyenne qui est constamment ré-estimée dans le *voisinage de krigeage* défini par une fenêtre glissante de sélection : $V^{l}(s_o)$. Ses caractéristiques de forme et de segmentation sont spécifiées à la discrétion du modélisateur, conditionnellement au variogramme, à la densité et à la structure spatiale des $m^{l}_{(s_g)}$. La sélection du voisinage doit être la plus uniforme possible, le nombre de $m^{l}_{(s_g)}$ inclus pour l'estimation ne doit pas excéder 20 points, mais être au moins égal à 10 – quitte à utiliser parfois des valeurs situées à l'extérieur de $V^{l}(s_o)$.

L'estimateur du KO est donné par :

$$\hat{m}^{l}_{(s_o)} = \sum_{i=[1]}^{\left[n^{l}_{(V)}\right]} \hat{\lambda}^{l}_{i} \cdot m^{l}_{(s_i)}$$

L'estimateur de la variance de KO est :

$$\hat{\sigma}^{2}_{\hat{m}^{l}_{(s_o)}} \quad \sum_{i=[1]}^{\left[n^{k}_{(V)}\right]} \hat{\lambda}_{i} \cdot \tilde{\gamma}^{l}(h_{io}) + \hat{\mu}$$

Le CKO est l'extension géostatistique multi-variables du KO. Il fournit aussi un estimateur spatial linéaire, sans biais et de variance minimale. Le CKO permet de modéliser – par le biais de FAI - une v.r. $x^{l}_{(s_g)}$ particulière et conditionnellement à d'autres v.r. auxiliaires $x^{k}_{(s_g)}$ en tenant compte de leurs propriétés d'autocorrélation spatiales intrinsèques et croisées.

A l'instar du KO, le CKO utilise les variogrammes : $\gamma^{l}(h), \gamma^{k}(h)$ pour modéliser les autocorrélations intrinsèques.

Les corrélations spatiales croisées sont modélisées par des covariances-croisées $\mathbb{C}^{l,k}(h)$ car les variogrammes-croisés : $\gamma^{l,k}(h)$ ne sont pas adaptés à *l'hétérotopie totale ou partielle*, i.e. à des v.r. dont

les contreparties mesurables : $m^{l}_{(s_g)}$ et $m^{k}_{(s_g)}$ sont localisées sur des sites disjoints – ce qui est le cas de toutes les $m^{l}_{(s_g)}$ interpolées par un CKO dans le cadre de cette thèse.

L'analyse empirique des auto-corrélations spatiales simples et croisées est nécessaire à la mise en œuvre du CKO. En effet, le CKO n'a de sens que si les $m^{l}_{(s_g)}$ et les $m^{k}_{(s_g)}$ présentent des stationnarités intrinsèques et qu'elles entretiennent des corrélations simples et croisées significatives. C'est-à-dire que les coefficients de corrélation : simples $\left|\hat{\rho}\left(m^{k}_{(s)}\ ; m^{l}_{(s)}\ \right)\right| \geq 0{,}7$ et croisées : $\left|\left\{\hat{\rho}^{lk}(h=0)\right\}\right| \geq 0{,}5$.

Ensuite, à l'instar du KO, le CKO nécessite l'ajustement : des semi-variogrammes intrinsèques : $\hat{\gamma}^{l}$ (h), $\hat{\gamma}^{k}$ (h), à des modèles théoriques : $\tilde{\gamma}^{l}$ (h), $\tilde{\gamma}^{k}$ (h) et des covariances croisées $\hat{\mathbb{C}}^{l,k}(h)$ à des modèles mathématiques admissibles $\tilde{\mathbb{C}}^{l,k}(h)$. Lorsque les fonctions choisies sont différentes, les modèles peuvent être uniformisés à une fonction unique – appelée Modèle Linéaire de Corégionalisation (MLC) - afin de simplifier la résolution du système d'équations du CKO.

Les caractéristiques de la fenêtre de voisinage $V^{l}(s_o)$ et $V^{k}(s_o)$ sont à spécifier conditionnellement : aux : $\tilde{\gamma}^{l}$ (h), $\tilde{\gamma}^{k}$ (h) et aux structures spatiales des $m^{l}_{(s_g)}$ et $m^{k}_{(s_g)}$.

L'estimateur du CKO s'écrit :

$$\hat{m}^{l}_{(s_o)} = \sum_{k=1}^{N} \sum_{i=[1]}^{\left[n^{k}_{(V)}\right]} \hat{\lambda}^{k}_{i} \cdot m^{k}_{(s_i)}$$

L'estimateur de la variance de CKO est :

$$\hat{\sigma}^{2}_{\left(\hat{m}^{l}_{(s_o)}\right)} = \tilde{\mathbb{C}}^{l,l}(\{h_o = \|s_o - s_o\|\}) - \sum_{k=1}^{N} \sum_{i=[1]}^{\left[n^{k}_{(V)}\right]} \hat{\lambda}^{k}_{i} \cdot \tilde{\mathbb{C}}^{l,k}(\{h_{io} = \|s_i - s_o\|\}) - \hat{\mu}_{l}$$

Les opérateurs du KO et du CKO sont *exacts* et la qualité des modèles spécifiés est appréciée par le biais d'une procédure de validation croisée (c.v.). Elle a pour but de valider le choix du modèle de variogramme et de(s) covariance(s) (dans le cadre d'un CKO), la calibration des paramètres ainsi que les critères de voisinage spécifiés.

La qualité de l'estimateur krigeage est donnée par des indicateurs calculés à partir des résidus de c.v. On dira qu'un modèle est *adéquat* lorsque : Residial Mean ≈ 0, Root. MSE = petit, AKSE. cv $\approx$ {Root. MSE}, Residual Mean Standardized ≈ 0 et Root. MSE Standardized ≈ 1 , que les valeurs d'échantillonnage et les ré-estimées de c.v. présentent une bonne corrélation graphique, et que les résidus *sont approximativement gaussiens*.

Remarques :

Lorsque la procédure de c.v. permet de qualifier un modèle de KO ou de CKO *d'adéquat*, il n'y a presque aucune chance cependant qu'il soit *optimal* mais il est tout de même admissible (Marcotte, 2008).

Les traitements ont été effectués avec l'extension : Geostatistical Analyst (ESRI, 2013).

Le KO et le CKO sont des estimateurs ponctuels. Or, les logiciels SIG renvoient les résultats sous forme de couches de surface de type : image raster ou TIN. Il convient d'agréger ces informations numériques, à l'échelle des U_k, par des opérateurs statistiques spatiaux afin d'obtenir les $x^{l}_{(U_k)}$.

UNIFORMISATION A L'ECHELLE DES COMMUNES

Le processus de transformation topologique et statistique permet d'obtenir en tous points de l'espace français métropolitain des indicateurs probabilistes verticaux interpolés horizontalement : $\hat{m}^{l}_{(s_o)}$. Afin d'obtenir les i.st.e* finaux $x^{l}_{(U_k)}$ il s'agit d'agréger les valeurs de $\hat{m}^{l}_{(s_o)}$ contenues dans chaque U_k par des opérateurs ensemblistes spatiaux afin d'estimer des statistiques zonales.

Remarques liminaires :

Les résultats des interpolations spatiales renvoyés par les SIG sont des couches de surface raster continues dans l'espace. Leur rugosité dépend de leur résolution : $res_{(\cdot)}$. Elles peuvent donc être interprétées comme des grilles de points spatialisés - portant les valeurs : $\hat{m}^{l}_{(s_o)}$.

Les rasters générés avec la résolution par défaut : $res_{(default)}$ ne sont pas adaptés à l'estimation des $x^{l}_{(U_k)}$. Elle est trop grossière, inadapté au processus d'estimation. En effet, comme le montre la Figure 123, les opérateurs statistiques d'agrégation spatiale ensemblistes ne parviennent pas à capturer des $\hat{m}^{l}_{(s_o)}$ dans les *petites* U_k ou dans celles ayant une morphologie complexe.

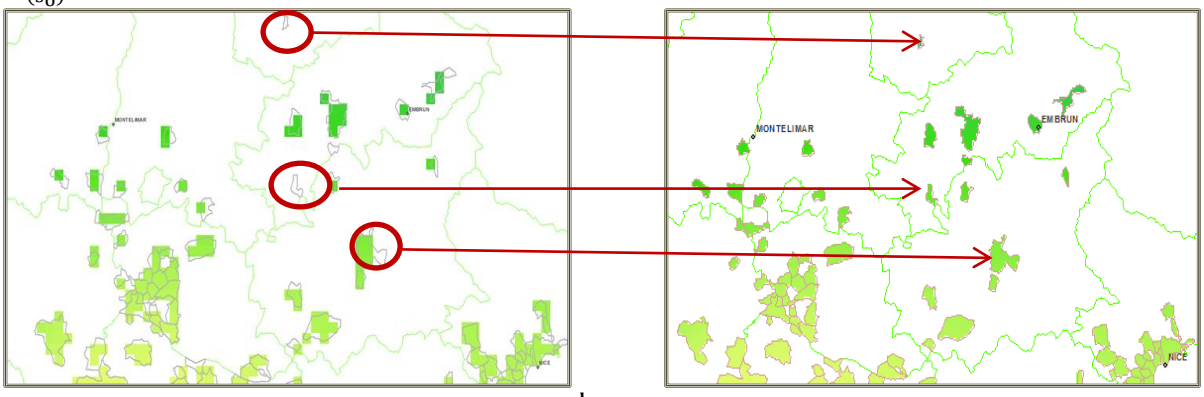


Figure 126 : Schéma cartographique de la capture des $\hat{m}^{l}_{(s_o)}$ dans les U_k par des opérateurs spatiaux ensemblistes. - à gauche, avec la résolution par défaut - et à droite, avec une résolution adaptée

Les opérateurs KO et du COKO sont des estimateurs ponctuels. Les $\hat{m}^{l}_{(s_o)}$ sont donc calculables en tous points de l'espace. Par conséquent la résolution du support raster des $\hat{m}^{l}_{(\boldsymbol{s_o})}$ est sans limite.

La résolution du fichier raster est à la discrétion du modélisateur. Elle peut être spécifiée pour des rugosités raster infinitésimales. Cependant, la taille du fichier, et par extension, le temps d'estimation des statistiques zonales seront considérablement augmentés (ESRI, 2012).

Et surtout, plus la maille de la grille spatialisée des estimés est fine et plus la variance de krigeage ou de cokrigeage : $\hat{\sigma}^2\left(\hat{m}^{l}_{(s_o)}\right)$ augmente, i.e. plus l'erreur commise sur les $\hat{m}^{l}_{(s_o)}$ est grande. En particulier pour les sites s_o situés dans des zones sous-échantillonnées (Matheron, 1963).

Spécification de l'hypothèse principale :

Le processus d'harmonisation d'échelle permet de finaliser la phase d'optimisation de *l'effet information** et donc de minimiser le *biais conditionnel** des i.st.e* $x^{l}_{(U_k)}$ pour garantir leur robustesse spatiotemporelle (Marcotte, 2008).

Il s'agit de proposer un processus permettant de spécifier un paramètre de résolution $res_{(\cdot)}$ pour les matrices spatiales des $\hat{m}^{l}_{(\boldsymbol{s_o})}$ de sorte que sa rugosité soit assez fine pour capter, dans chaque U_k, au moins une valeur de $\hat{m}^{l}_{(\boldsymbol{s_o})}$, et en contrepartie suffisamment grossière pour ne pas augmenter l'erreur en terme de variance ni rallonger inutilement les temps de calcul. Il s'agit donc de spécifier un *critère permettant d'optimiser la capture spatiale* des opérateurs ensemblistes supputant *les statistiques zonales* (Voiron-Canicio, 1995)

Stratégie d'estimation :

Les $x^{l}_{(U_k)}$ sont des *statistiques zonales d'analyse spatiale*. Plus exactement, il s'agit de moyennes des $\hat{m}^{l}_{(s_o)}$ estimées contenues dans chacune des U_k :

$$x^{l}_{(\mathbf{U}_{\boldsymbol{k}})} = \frac{1}{n^{l}_{(\boldsymbol{s_o}\subseteq U_k)}} \cdot \sum_{\forall \boldsymbol{s_o}\subseteq \boldsymbol{\mathcal{D}}(\boldsymbol{\omega}.)} \left(\hat{m}^{l}_{(\boldsymbol{s_o})} \cdot \mathbb{1}_{\{\boldsymbol{s_o}\subseteq U_k\}}\right)$$

Avec : $n^{l}_{(s_o\subseteq U_k)}$ le nombre de valeurs interpolées contenues dans U_k.

Afin que des $x^{l}_{(U_k)}$ puissent être estimés il convient d'optimiser un critère de capture spatiale permettant de spécifier une résolution adaptée à l'estimation de $x^{l}_{(U_k)}$robustes. Il doit permettre de définir une valeur de résolution optimisée : $res_{(opt)}$, i.e. suffisamment fine pour satisfaire la condition de capture *d'au moins une* $\hat{m}^{l}_{(s_o)}$ *dans chaque* U_k mais assez grossière pour ne pas altérer la précision des estimés : $\hat{m}^{l}_{(s_o)}$ et engendrer des temps de calcul interminables.

Le critère de capture spatiale a été spécifié de la façon suivante :

$$res_{(opt)} = \underset{res_{(\cdot)}\in\{\mathbb{N}^{+}_{*}|\Delta_{res}\}}{\operatorname{argmin}} \left\{ \min\left(\bigcup_{k=1}^{q_1} \left(\operatorname{card}\left(\bigcup_{\forall s_o\subseteq \mathcal{D}(\omega.)} \left\{ \hat{m}^{l}_{(\boldsymbol{s_o}|res_{(\cdot)})} \subseteq U_k \right\} \right) \right) \right) = 1 \right\}$$

A l'initialisation, $res_{(\cdot)}$ est réglée à $res_{(defaut)}$; à chaque itération, on soustrait Δ_{res} qui a été fixée à 500m ; la résolution optimisée permettant de satisfaire aux conditions énoncées est :

$$res_{(opt)} = \{(cell.size.h.opt \times cell.size.v.opt) = (1000m \times 1000m)\}$$

Remarque :

L'optimisation itérative de la contrainte de capture spatiale, i.e. la $res_{(opt)}$ a été évaluée grâce à l'outil de géo-traitement ModelBuilder (ESRI, 2013) et le calcul des statistiques zonales a été effectué avec l'extension Spatial Analyst (ESRI, 2013).

L'ensemble des procédures d'estimation des i.st.e* : $x^{l:Météo-France}_{(U_k)}$ et $x^{l:RNM}_{(U_k)}$ obtenus par la stratégie déclinée est *a priori* robuste car leur le biais conditionnel* a été minimisé. La proposition est illustrée dans l'application subséquente.

APPLICATION

Cette application décline les principales étapes de la stratégie d'intégration des chroniques temporelles Météo-France et RNM. Dans un premier temps les $m^{l}_{(s_g),t}$ sont spatialisées afin de se faire une idée de la densité et de la répartition spatiale des stations s_o mesurant la substance physicochimique concernée. Ensuite, l'effet de support* est maximisé conformément au processus d'harmonisation temporel, ce qui permet d'obtenir les indicateurs probabilistes verticaux spatialisés : $m^{l}_{(s_g)}$ - une illustration de sa mise en œuvre a déjà été déclinée, en début de section sur les cumuls du rayonnement global, elle ne sera pas rappelée.

Ici, il est question de maximiser l'effet information*. Pour ce faire, il s'agit d'abord d'estimer les indicateurs probabilistes verticaux interpolés horizontalement $\hat{m}^{l}_{(s_o)}$. La majorité des $\hat{m}^{l}_{(s_o)}$ est obtenue par krigeage ordinaire (KO). Cependant la densité d'échantillonnage et la répartition spatiale de trois $m^{l}_{(s_g)}$ *sont défaillantes*. Les substances physicochimiques concernent Le rayonnement global (RAY) et l'activité volumique du Plutonium 238 et de l'Antimoine 125 dans les sols. Pour ces variables le nombre

de $m^{l}_{(s_g)}$ disponibles sur l'intégralité de la France est respectivement de $\{n^{RAY}_{s.} = 27\}$, $\{n^{238Pu}_{s.} = 25\}$ et $\{n^{125Sb}_{s.} = 26\}$. De fait, les $\widehat{m}^{l}_{(s_o)}$ ont été obtenus par le biais d'un cokrigeage ordinaire (CKO) en introduisant des v.r. auxiliaires $m^{k}_{(s_g)}$ correctement échantillonnées et pourvues de corrélations simples et croisées. Comme la mise en œuvre d'un CKO reprend toutes les étapes d'un KO en rajoutant celles inhérentes à la modélisation des corrélations croisées, alors la synoptique d'estimation des $x^{l}_{(U_k)}$ est illustrée pour $x^{RAY}_{(U_k)}$. Le calcul des $\widehat{m}^{RAY}_{(s_o)}$ fait intervenir deux variables auxiliaires qui sont spécifiées subséquemment.

Le plutonium.238 est un émetteur α presque pur, les $\widehat{m}^{238Pu}_{(s_o)}$ sont obtenus en introduisant les $m^{ALPHA}_{(s_g)}$. Quant à l'Antimoine 125, sa présence dans l'environnement est essentiellement artificielle et liée à des activités anthropiques, à l'instar du Tritium radioactif. Les $\widehat{m}^{125Sb}_{(s_o)}$ sont obtenus à partir des $m^{3H}_{(s_g)}$.
La résolution de la couche de surface raster portant les $\widehat{m}^{l}_{(s_o)}$ est spécifiée : $res_{(opt)}$, conformément aux préconisations du processus d'harmonisation d'échelle. Enfin les i.st.e* $x^{l}_{(U_k)}$ sont estimés par l'opérateur de la moyenne spatiale des $\widehat{m}^{l}_{(s_o)}$ incluses partiellement ou intégralement dans les U_k

Les inputs sont les chroniques temporelles $m^{\mathbf{RAY}}_{(s_g,t)}$ transmises par Météo-France. Afin d'assurer une cohérence temporelle dans les $m^{\mathbf{RAY}}_{(s_g,t)}$ utilisées, ce paramètre météo n'a pu être intégré que pour $\{n^{\mathbf{RAY}}_{(s)} = 27\}$ stations s_g. En dépit du fait que les $m^{\mathbf{RAY}}_{(s_g,t)}$ sont uniformément réparties dans le champ, la densité d'échantillonnage est trop faible pour qu'un KO permette de modéliser précisément les disparités géographiques des $m^{\mathbf{RAY}}_{(s_o)}$ et par suite supputer des $x^{\mathbf{RAY}}_{(U_k)}$ robustes. L'estimation des $m^{\mathbf{RAY}}_{(s_o)}$ est effectuée par CKO en introduisant deux variables auxiliaires.

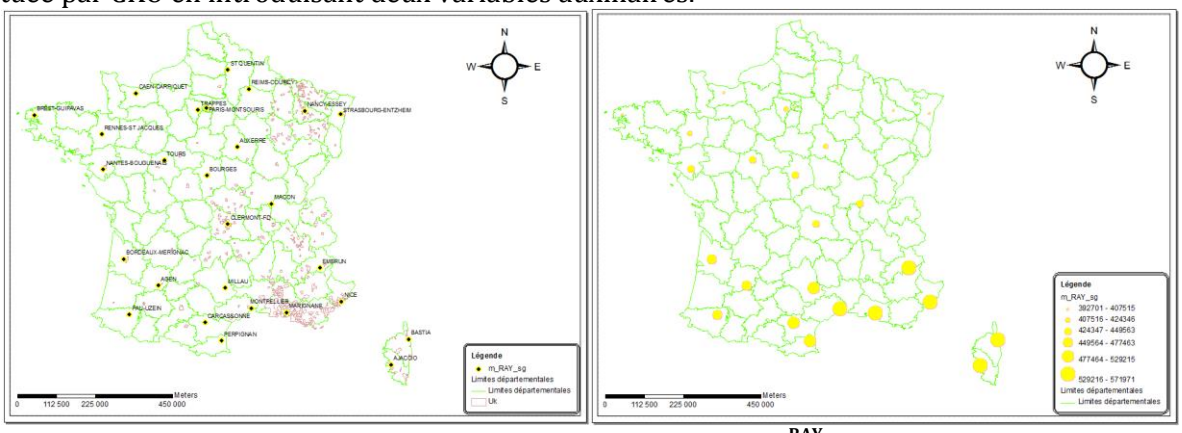


Figure 127 : Localisation des stations mesurant le rayonnement global : $m^{\mathbf{RAY}}_{(s_g,t)}$ (à gauche) ; Représentation de la variabilité géographique des indicateurs probabilistes verticaux : $m^{\mathbf{RAY}}_{(s_g)}$ (à droite)

La cartographie de gauche localise les stations météo : s_g pour lesquelles des chroniques temporelles spatialisées $m^{RAY}_{(s_g,t)}$ ont été mobilisées. Elles ont été mises en perspective des $\mathcal{U}^{1ère}_{.}$ matérialisées en rose.

La cartographie de droite présente les valeurs des indicateurs probabilistes verticaux spatialisés $m^{RAY}_{(s_g)}$ calculés conformément à la procédure d'agrégation temporelle. La cartographie des $m^{RAY}_{(s_g)}$ dénote de fortes disparités spatiales.

Les $m^{RAY}_{(s_g)}$ peuvent être interprétés comme les réalisations fragmentaires d'une v.r. – dont la faiblesse de la densité d'échantillonnage les rend indigentes pour modéliser, de façon précise à l'échelle des communes, les variabilités géographiques du rayonnement global.
L'interpolation d'une v.r. dont la distribution spatiale est défaillante peut néanmoins être effectuée par un CKO lorsqu'une ou plusieurs v.r. auxiliaires sont mesurées de façon uniforme et dense sur $\mathcal{D}(\omega)$. Typiquement les variables secondaires utilisées pour reconstituer le rayonnement global sont les températures et l'altimétrie. En l'occurrence les $m^{TEMP}_{(s_g)}$ ont été estimées à partir des $m^{TEMP}_{(s_g,t)}$ fournies par Météo-France. Et les niveaux altimétriques moyens communaux ont été récupérés dans la BD SIG géofla $x^{TOPO}_{(U_u)}$ pour les 36 600 communes françaises U_u. Un échantillonnage stochastique uniforme a été appliqué aux données altimétriques afin de réduire leur nombre à 500 : $m^{TOPO}_{(s_g)}$, et éviter ainsi de construire une matrice de CKO quasi-singulière (Marcotte, 2008), ce qui biaiserait l'estimation des $m^{RAY}_{(s_o)}$.

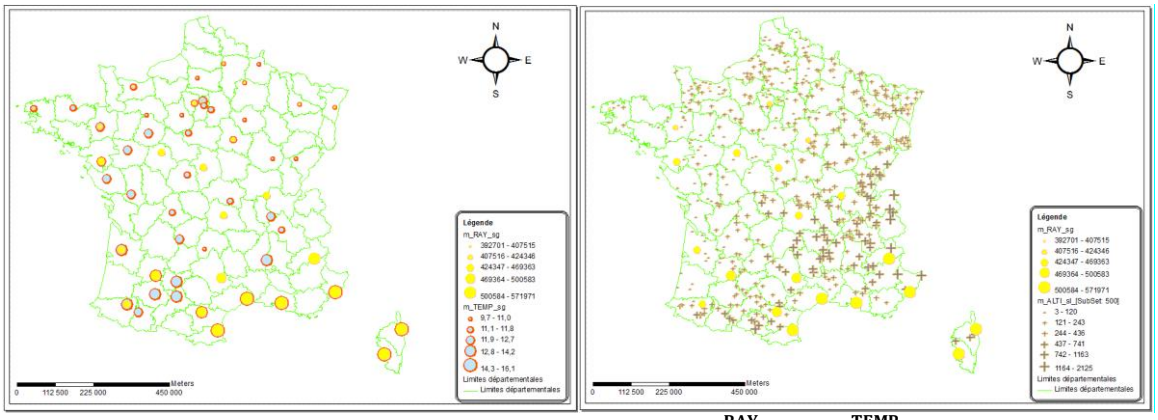


Figure 128 : Superposition des indicateurs probabilistes verticaux : $m^{RAY}_{(s_g)}$ et des $m^{TEMP}_{(s_g)}$ (à gauche) ; des indicateurs probabilistes verticaux : $m^{RAY}_{(s_g)}$ et des $m^{TOPO}_{(s_g)}$ (à droite)

La superposition des indicateurs verticaux spatialisés $m^{RAY}_{(s_g)}$ et des $m^{TEMP}_{(s_g)}$ montre que l'on se positionne dans un contexte d'hétérotopie partielle. Quant à la superposition des $m^{RAY}_{(s_g)}$ et des $m^{TOPO}_{(s_g)}$ - localisés au niveau des centroïdes communaux – elle engendre un contexte d'hétérotopie totale.
Avant de procéder à une reconstitution spatiale des $m^{RAY}_{(s_o)}$ par CKO, les interactions spatiales entre les différentes variables ont été étudiées à partir des variogrammes et covariogrammes expérimentaux. Si les hypothèses de stationnarité et de stationnarité conjointe des accroissements sont vérifiées, les autocorrélations simples et croisées sont ajustées à des modèles de variogrammes et de covariances croisées.

Au préalable une analyse de la distribution statistique des $m^{RAY}_{(s_g)}$ a été effectuée. Elle a permis de mettre en évidence une dérive qui a été corrigée par une fonction de Kernel de type exponentiel. L'intégrité des extrema de $m^{RAY}_{(s_g)}$ a été vérifiée. Des explorations analogues ont été effectuées sur les $m^{TOPO}_{(s_g)}$ et $m^{TEM}_{(s_g)}$afin de faciliter la mise en œuvre du CKO. De plus, une analyse préliminaire des corrélations simples et croisées a permis de valider l'intérêt d'intégrer simultanément ces deux variables supplétives. Par conséquent les valeurs de $m^{RAY}_{(s_o)}$ sont estimées par l'opérateur du CKO à partir d'un Modèle de Corégionalisation Linéaire (MCL).

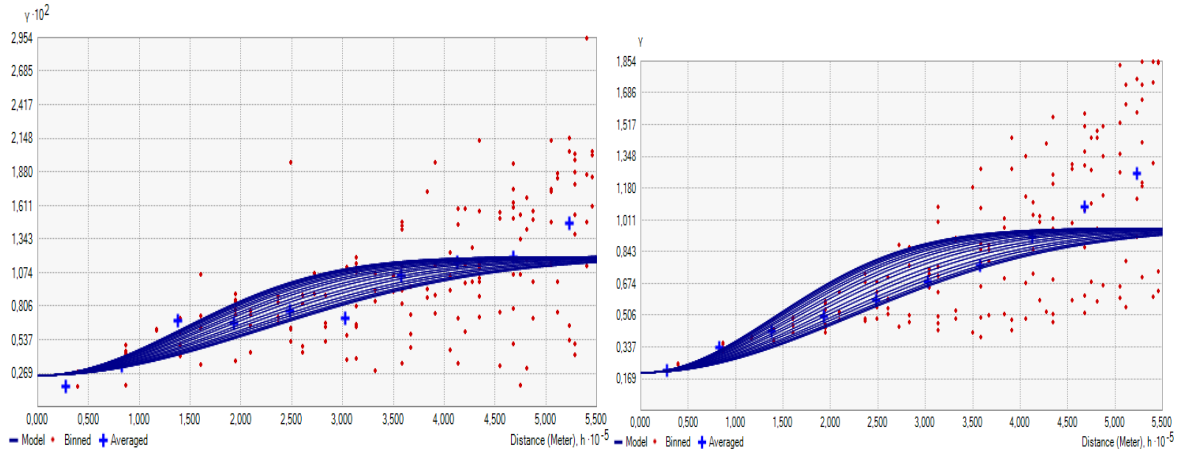


Figure 129 : Superposition des variogrammes empiriques et des modèles directionnels de : $m_{(s_g)}^{RAY}$ (à gauche) ; Et de $m_{(s_g)}^{TOPO}$ à droite

L'analyse variographique permet de vérifier l'hypothèse d'accroissements stationnaires sur la variable d'intérêt $m_{(s_g)}^{RAY}$ (variogramme de gauche) et les variables auxiliaires : $m_{(s_g)}^{TEMP}$ (variogramme non présenté) et $m_{(s_g)}^{TOPO}$ (variogramme de droite). En considérant les trois $m_{(s_g)}^{k}$ comme des FAI on fixe un cadre de travail. Tous les variogrammes $\hat{\gamma}^k(h), \forall\, k = \{RAY, TEMP, TOPO\}$ présentent de forts niveaux d'autocorrélations à petite, moyenne et grande échelles. De fait le MCL est adapté à la modélisation. La portée critique spécifiée est fixée à $\left\{h_{crit} \approx \frac{1}{2} \cdot h_{max}\right\}$. L'allure sigmoïdale de $\hat{\gamma}^k(h)$ suggère l'ajustement de $\tilde{\gamma}^k(h)$ par la somme d'une *pépite* et d'une *gaussienne*. Des anisotropies géométriques ont systématiquement été décelées, hypothèse d'autant plus prégnante au vu de l'allure en « rose des sables » dessinée par la superposition des $\tilde{\gamma}^k(h, \theta_j)$ et qui a fait l'objet d'une correction du fait que $F_a^k \in [\![1{,}5\,;\,2]\!]$; Les paramètres des variogrammes $\hat{\vartheta}^k$ sont spécifiés par le critère des mcp avec les poids de Cressie.

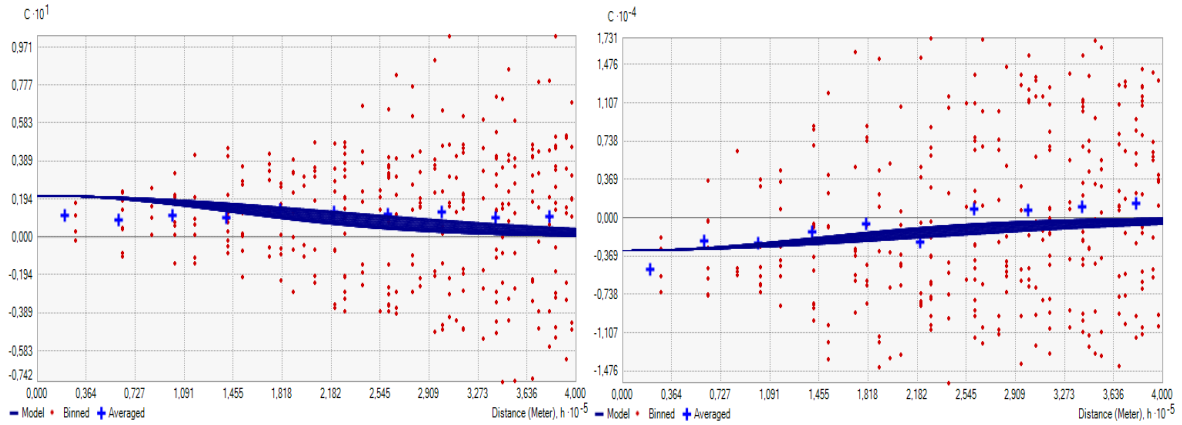


Figure 130 : Superposition des covariogrammes croisés et des modèles directionnels ajustés pour le rayonnement global avec la température (à gauche) et avec la topographie (à droite)

S'agissant des autocorrélations entre les indicateurs temporels agrégés du rayonnement global et ceux auxiliaires (températures et topographie), l'estimation des coefficients de corrélation simple est conforme : $\left|\hat{\rho}^{\{l,k\}}\right| \in [\![0{,}65\,;\,0{,}85]\!]$. L'estimation des autocorrélations croisées maximales est telle que : $\max_{h\to 0^+}\left|\hat{\rho}^{\{l,k\}}(h)\right| \in [\![0{,}45\,;\,0{,}70]\!]$.

L'analyse des diagrammes des covariances croisées centrées réduites, avec à gauche $\hat{\mathbb{C}}^{\{RAY,TEMP\}}(h)$ et à droite $\hat{\mathbb{C}}^{\{RAY,TOPO\}}(h)$ révèle respectivement des corrélations croisées positives et négatives. Bien que les portées des autocorrélations croisées soient plus faibles que celles des autocorrélations intrinsèques, l'ajustement des modèles est effectué sur au moins 7 covariances agrégées par classe de

distance. La qualité de l'ajustement de $\tilde{\mathbb{C}}^{\{RAY,TEMP\}}(h)$ et de $\tilde{\mathbb{C}}^{\{RAY,TOPO\}}(h)$ à leur contrepartie empirique, par le critère des mcp., à un MCL, est appréciée visuellement.
En somme, les variogrammes ont permis de valider l'hypothèse intrinsèque. Les covariogrammes ont permis de vérifier l'hypothèse de stationnarité des accroissements conjoints et donc de justifier l'utilisation de l'opérateur du CKO et du MCL ainsi que l'introduction des réalisations fragmentaires de $m_{(s_g)}^{TEMP}$ et de $m_{(s_g)}^{TOPO}$. S'agissant de la spécification des critères de voisinage de CKO, elle est effectuée indépendamment sur chacune des variables puisque les distributions spatiales sont hétérotopiques.

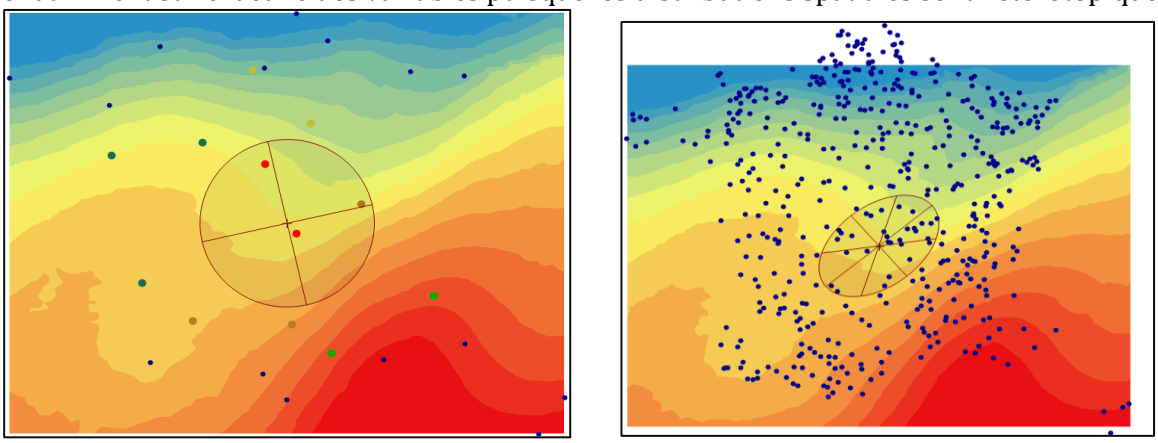


Figure 131 : Spécification des paramètres de la fenêtre de voisinage de krigeage pour le rayonnement global et la topographie

Les tailles du voisinage : $V_{s_o}^{RAY}$ à gauche, de $V_{s_o}^{TEMP}$ (non présenté) et de $V_{s_o}^{TOPO}$ à droite sont spécifiées au moyen d'un compromis entre les distributions spatiales des données d'échantillonnage et les critères du MCL. La répartition des données et le nombre de réalisations fragmentaires disponibles ou utilisées n_s^k ainsi que les valeurs du facteur d'anisotropie $\{F_a^k \in [\![1{,}65\,;2]\!]\}$ ont permis de partitionner, d'orienter et de dimensionner $\{\hat{a}_g^k, \hat{a}_p^k, \hat{\theta}_g^k\}$ les fenêtres de sélection du voisinage – dans le respect des règles empiriques énoncées dans la partie théorique.
En particulier pour $V_{s_o}^{TOPO}$ une attention particulière a été portée à la règle du moins de 20 points afin de ne pas créer de singularité dans la matrice de CKO.
Enfin, l'ajustement définitif des paramètres du MLC et des critères de voisinage – tels qu'ils sont présentés – a été affiné manuellement grâce à la procédure de validation croisée.

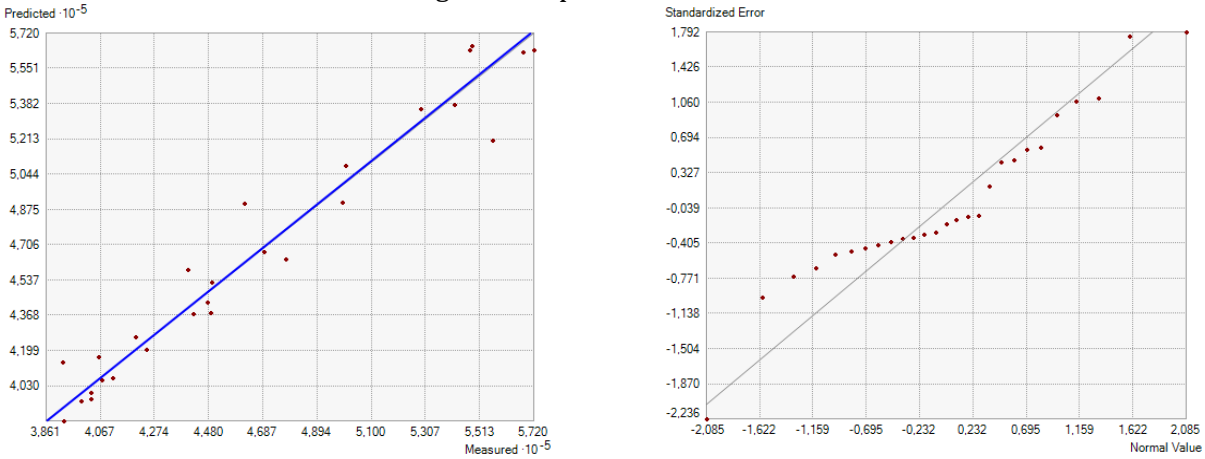


Figure 132 : Diagramme du nuage de points des valeurs observées et prédites (à gauche) et diagramme QQ.plot Normal (à droite), et obtenus dans le cadre de la procédure de validation croisée

La mise en concurrence expérimentale de nombreux modèles de CKO a permis d'obtenir un modèle pour lequel les scores de validation croisée sont adéquats. En l'occurrence {Residial Mean = 19} ; {Root. MSE = 12839} ; {Residual Mean Standardized = −0,002} ; {Root. MSE Standardized = 0,87} ; {AKSE. cv = 15662} ; A cela s'ajoutent une bonne corrélation entre les valeurs d'échantillonnage et les

estimés de c.v. ainsi que le caractère très approximativement gaussien des résidus de c.v.. Ces caractéristiques s'apprécient visuellement sur les graphiques présentés. Désormais une estimation *a priori* robuste des $m_{(s_o)}^{RAY}$ est disponible sur l'intégralité de la France.

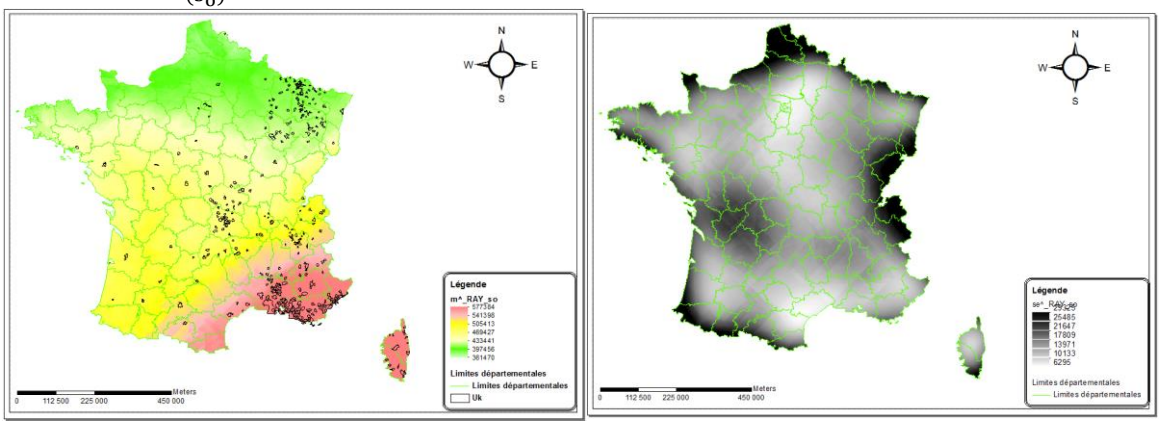


Figure 133 : Résultats du CKO, valeurs prédites (à gauche) ; Estimateurs des écarts-types (à droite)

Les cartographies présentent les résultats du CKO avec à gauche les $\widehat{m}_{(s_o)}^{RAY}$ auxquelles ont été superposées les $\mathcal{U}_{.}^{1ère}$, et à droite, les $\widehat{\sigma}_{(s_o)}^{RAY}$ de CKO pour chaque prédiction effectuée. Les indicateurs probabilistes verticaux interpolés horizontalement $\widehat{m}_{(s_o)}^{I}$ sont désormais évalués et de fait, les i.s.pc : $x_{(U_k)}^{RAY}$ sont estimables de façon robuste.

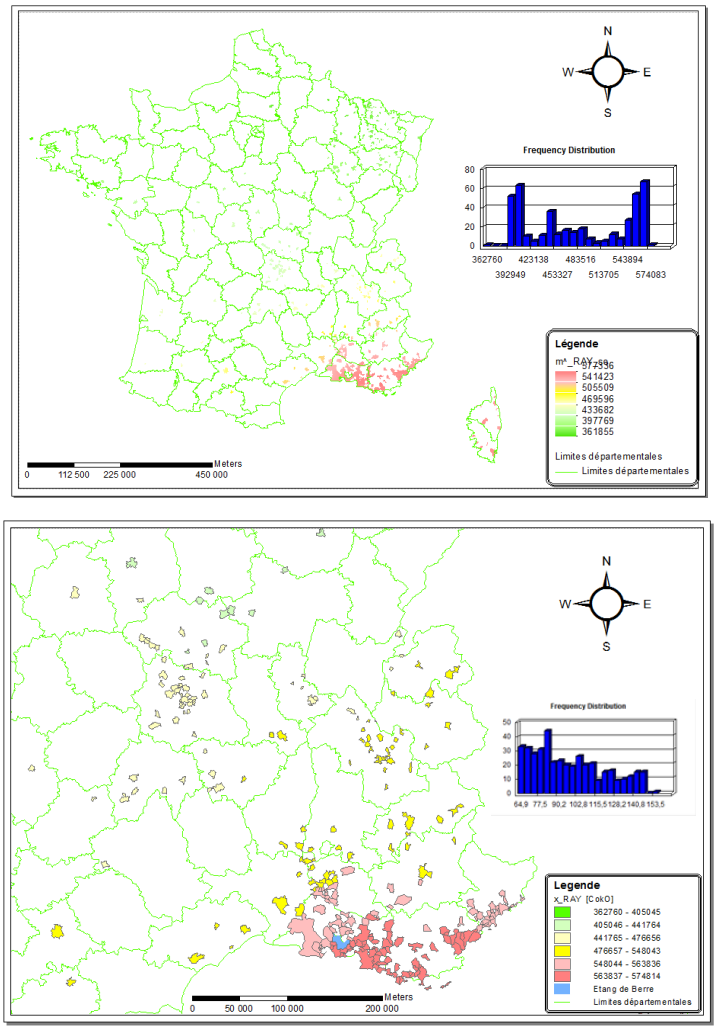


Figure 134 : valeurs des i.st.e* modélisant les niveaux géographiques du rayonnement global spatiotemporel pour l'intégralité des U_k (à gauche) ; pour celles situées en PACA et aux alentours (à droite)

Les **i.st.e**: $x_{(U_k)}^{RAY}$ sont des estimateurs statistiques zonaux évalués conformément à la procédure décrite dans la phase d'harmonisation d'échelle. La cartographie de gauche présente les résultats obtenus pour l'intégralité des $\mathcal{U}_{.}^{1ère}$ et l'histogramme des $x_{(U_k)}^{RAY}$ dénote la présence de disparités géographiques.

La cartographie de droite est un zoom in de la première. Une discrétisation des couleurs plus adaptée – par exemple en isolant les communes de PACA - permettrait de mieux percevoir visuellement l'ampleur des variabilités spatiales qui sont atténuées lorsque l'histogramme des couleurs de la carte est construit sur l'ensemble du territoire français métropolitain. Les $x_{(U_k)}^{RAY}$modélisent l'exposition environnementale potentielle des patients aux doses globales mensuelles de rayonnement solaire et sont exprimées en joules/cm^2.

STRATEGIE D'INTEGRATION DES VARIABLES DE L'ATLAS RADON

Les variables issues de l'Atlas Radon : $AVR_{(U_u)}$ sont uniques sur l'intégralité de la période couverte par la campagne de mesures de 1982 à 2000. Elles sont disponibles à l'échelle des communes – à l'exception de Paris - et correspondent aux U_u de la BD-SIG géofla 2003. Par conséquent, il n'y a pas lieu de maximiser *l'effet de support**.
Les $AVR_{(U_u)}$ sont des indicateurs géographiques qualitatifs, relatifs à des classes modales *grossières* et d'amplitudes inégales. Des informations géographiques auxiliaires à l'échelle des départements sont mises à disposition par l'IRN. Elles vont permettre de construire deux i.st.e* $x_{(U_k)}^{I:RADON}$ et $x''^{I:RADON}_{(U_k)}$ pour lesquels le biais conditionnel* est minimisé.

OPTIMISATION DE L'EFFET INFORMATION

La stratégie proposée pour l'optimisation de l'effet information* s'opère en deux étapes :

- D'abord, par une stratégie de transformation topologique par le biais de fusions statistiques randomisées des informations disponibles à l'échelle des De_u ;
- Ensuite, par une stratégie d'harmonisation, géographiquement pondérée, à l'échelle des U_k.

TRANSFORMATION TOPOLOGIQUE : FUSION STATISTIQUE RANDOMISEE D'INFORMATIONS ATTRIBUTAIRES

Remarques liminaires :
$AVR_{(U_u)}$ est une variable représentative *des expositions domestiques au Radon*. Elle est de nature qualitative et comprend 5 classes C_j^{AVR} discrétisées par un code couleur (IRSN, 2001).

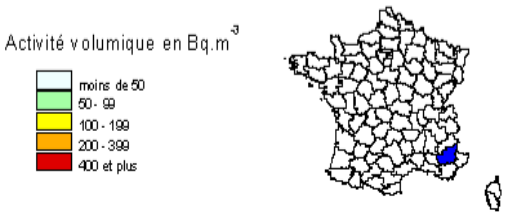


Figure 135 : Amplitude des valeurs associées aux classes modales de la variable $AVR_{(U_u)}$ de l'Atlas Radon

Les valeurs numériques moyennes de l'Activité Volumique du Radon $\overline{m}^{RADON}_{(\{s_g \subset U_u\})}$ enregistrées par les stations IRN s_g, dans l'ensemble des habitations d'une même commune U_u, ne sont pas disponibles.
Les $AVR_{(U_u)}$ sont associées à des intervalles continus, bornés par des valeurs exprimées en $Bq \cdot m^{-3}$ et d'amplitudes inégales censées circonvenir $m^{RADON}_{(\{s_g \subset U_u\})}$, tel que :

$$\overline{m}^{RADON}_{(\{s_g \subset U_u\})} \in AVR_{(U_u)} = \left[\!\left[AVR^{-}_{\left(U_u \middle| C_j^{AVR}\right)} ; AVR^{+}_{\left(U_u \middle| C_j^{AVR}\right)} \right]\!\right]$$

Or, de fortes incertitudes spatiales entachent ces variables. En l'occurrence, quelle est la signification géographique de : $AVR_{(U_u)} = \{B = Blanc\}$, i.e. "moins de 50 $Bq.m^{-3}$? Ou de $AVR_{(U_u)} = \{R = Rouge\}$, i.e. plus de 400 $Bq.m^{-3}$?
Cependant, la collecte des mesures a été effectuée dans une logique temporelle départementale et l'IRSN a rendu publiques les statistiques à cette échelle. En l'occurrence :

Le nombre de mesures moyennes communales disponibles : $\text{card}\left(\overline{m}^{RADON}_{(s_g \subset U_u \subset De_u)}\right)$;

La valeur communale moyenne minimale : $\text{mîn}\left(\overline{m}^{RADON}_{(s_g \subset U_u \subset De_u)}\right)$; sur l'ensemble des départements français métropolitains le minimum a été estimé à1 $Bq.m^{-3}$;

La valeur communale moyenne maximale : $\hat{\max}\left(\bar{m}^{RADON}_{(s_g \subset U_u \subset De_u)}\right)$; sur l'ensemble des départements français métropolitains le minimum a été estimé à 4 964 Bq. m^{-3} ;

La valeur communale moyenne départementale : $\bar{\bar{m}}\left(m^{RADON}_{(s_g \subset U_u \subset De_u)}\right)$; sur l'ensemble des départements français métropolitains la moyenne a été estimée à 87,4 Bq. m^{-3} ;

La valeur communale de l'écart-type départemental des moyennes communales : $\hat{\sigma}\left(\bar{m}^{RADON}_{(s_g \subset U_u \subset De_u)}\right)$; sur l'ensemble des départements français métropolitains l'écart-type a été estimé à 113 Bq. m^{-3} ;

L'analyse des statistiques départementales met en évidence de fortes disparités géographiques tant au niveau des extrema - par le biais des $\hat{\min}(\cdot)$ et des $\hat{\max}(\cdot)$, que de la variabilité de la distribution spatiale des $\bar{m}^{RADON}_{(\cdot)}$ – avec $\bar{\bar{m}}(\cdot)$ très différentes d'un De_u à l'autre et des écarts-types $\hat{\sigma}\left(\bar{m}^{RADON}_{(\cdot)}\right)$ toujours très élevés – généralement supérieurs aux $\bar{\bar{m}}(\cdot)$ (IRSN, 2001)

Spécification des hypothèses :

Il convient de prendre en compte ces informations spatiales attributaires dans la construction des i.st.e* des expositions géographiques au radon.

Pour ce faire, un processus de *fusion* par injection *des statistiques départementales extrêmes* fondé sur la *théorie des ensembles flous* (Dubois Didier, Prade Henri, 2004) *; et un processus stochastique intégrant le différentiel des niveaux géographiques des variabilités spatiales observées* (Saporta, 2006) sont couplés dans une logique adaptée à la statistique *territoriale* (Charre, 1995).

Spécification de la stratégie d'estimation :

Le premier i.st.e* proposé consiste simplement à apparier les modalités des $AVR_{(U_u)}$ aux U_k :

$$x^{RADON}_{(U_k)} = \bigcup_{u=1}^{n_{(U_u)}} \left(AVR_{(U_u)} \middle| U_u = U_k\right)$$

Ensuite cet i.st.e* est transformé en un indicateur spatial quantitatif intermédiaire prenant la valeur moyenne de la classe modale associée tout en tenant compte des statistiques extrêmes départementales disponibles, tel que :

$$x'^{RADON}_{(U_k|U_k \subset De_u)} = \begin{cases} \left[\frac{1}{2} \cdot \left(AVR^{-}_{\left(U_u \middle| c_j^{AVR}\right)} + AVR^{+}_{\left(U_u \middle| c_j^{AVR}\right)}\right)\right] & \text{lorsque: } x^{RADON}_{(U_k=U_u)} \in \{V; J; O\} \\ \left[\frac{1}{2} \cdot \left(\hat{\min}\left(\bar{m}^{RADON}_{(U_u \subset De_u)}\right) + 49\right)\right] & \text{lorsque: } \left\{x^{RADON}_{(U_k=U_u)} = B\right\} \\ \left[\frac{1}{2} \cdot \left(400 + \hat{\max}\left(\bar{m}^{RADON}_{(U_u \subset De_u)}\right)\right)\right] & \text{lorsque: } \left\{x^{RADON}_{(U_k=U_u)} = R\right\} \end{cases}$$

L'i.st.e* terminal est obtenu par un processus stochastique contrôlé, i.e. permettant de randomiser les valeurs de $x'^{RADON}_{\cdot}$ en intégrant, de façon aléatoire, l'ampleur de la volatilité spatiale des valeurs communales moyennes de l'activité du radon dans les habitations, tout en veillant à ce que le bruit de fond induit n'outrepasse pas l'amplitude des valeurs communales possibles, autrement dit, que :

$$x''^{RADON}_{(U_k)} = \left\{ x'^{RADON}_{(U_k|U_k \subset De_u)} + \frac{\mathbb{t}^{\mp}_{\left(\frac{\alpha}{2}\right)} \cdot \hat{\sigma}\left(\bar{m}^{RADON}_{(U_u \subset De_u)}\right)}{\sqrt{card\left(\bar{m}^{RADON}_{(U_u \subset De_u)}\right)}} \right\}$$

$$\in \left[\!\left[AVR^{-}_{\left(U_k=U_u \middle| \hat{\min}\left(\bar{m}^{RADON}_{(U_u \subset De_u)}\right)\right)} ; AVR^{+}_{\left(U_k=U_u \middle| \hat{\max}\left(\bar{m}^{RADON}_{(U_u \subset De_u)}\right)\right)} \right]\!\right]$$

Avec comme paramètres du processus de randomisation contrôlé définis tel que :

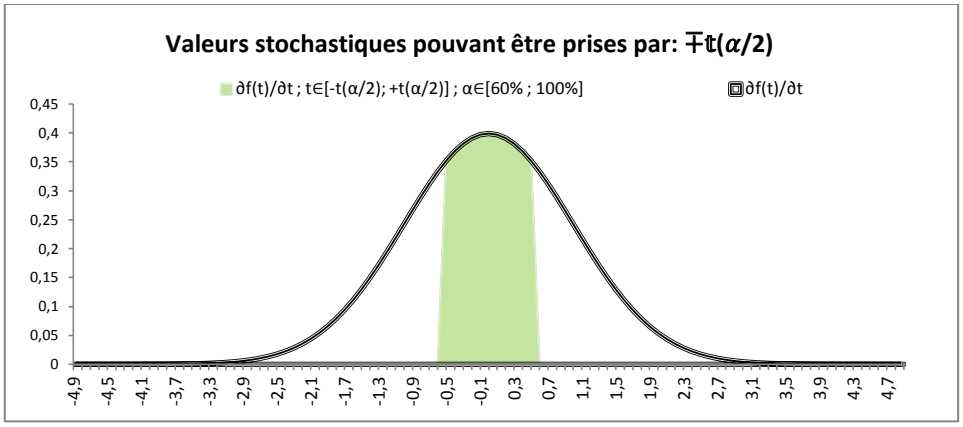


Figure 136 : Amplitude des valeurs prises par $\mathbb{t}^{\mp}_{\left(\frac{\alpha}{2}\right)}$ en fonction de la valeur stochastique de $(1-\alpha)$

$$\mathbb{t}^{\mp}_{\left(\frac{\alpha}{2}\right)} \xrightarrow[n\to+\infty]{} U^{RADON}_{\mathbb{t}} \sim \mathcal{N}(0,1),\ \ \mathbb{P}\left(U^{RADON}_{\mathbb{t}} \in \left[\!\left[t^{-}_{\frac{\alpha}{2}}; t^{+}_{\frac{\alpha}{2}}\right]\!\right]\right) = \{(1-\alpha) \sim \mathbb{U}[\![60\%; 100\%]\!]\}$$

UNIFORMISATION A L'ECHELLE DES COMMUNES

Remarques liminaires:

Les variables de l'atlas Radon sont déclinées à l'échelle des U_u, à l'exception de Paris qui est découpé en arrondissements Ar_u. Or la commune de Paris fait partie des U_k. Pour celle-ci les i.st. ont été estimés à l'échelle des Ar_u. Il convient donc de les fusionner par un processus *mathématique d'agrégation spatiale* qui incorpore un pondérateur géographique.

Afin de proposer un $x^{\wp.radon}_{(Ar_u)}$ adapté, il convient de remarquer que le nombre de mesures IRSN effectuées dans chaque unité géographique a été déterminé en fonction de leur Surface Géographique:$x^{SG}_{(.)}$, et de la quantité de population résidente $x^{POP}_{(.)}$ (IRSN, 2001)

Spécification de l'hypothèse :

Elle est identique à celle déclinée dans le processus diachronique d'appariement spatial des i.st.e* voués à la modélisation des FE-SAN.

Spécification de la proposition :

Elle est identique à celle déclinée dans le processus diachronique d'appariement spatial des i.st.e* voués à la modélisation des FE-SAN.

Les $x^{\wp.radon}_{(Ar_u)}$ sont obtenus par combinaison des deux variables fragmentaires citées. Elles ont été *normalisée*s de façon à conserver leurs variabilités géographiques et en même temps, que leur système unitaire n'altère pas les propriétés statistiques des poids (Saporta, 2006).

$$x^{\wp.radon}_{(Ar_u)} = \left\{ môy\left(\dot{x}^{SG}_{(Ar_u),t}\ ; \dot{x}^{POP}_{(Ar_u),t}\right) \middle| \dot{x}^{l}_{(Ar_u),t} = \frac{x^{l}_{(Ar_u),t} - \bar{x}^{l}_{(Ar_u),t}}{\hat{\sigma}\left(x^{l}_{(Ar_u),t}\right)} \right\}$$

Remarques :

La variabilité spatiale des expositions géographiques au gaz radon dans les habitations est décrite par deux i.st.e* singulièrement différents. Le premier :r $x^{RADON}_{(U_k)}$ est de nature qualitative multi-classes. Et le second : $x''^{RADON}_{(U_k)}$ est de nature quantitative continue.

Les traitements ont été effectués grâce à des macro-commandes programmées en VBA (Microsoft, 2013).

STRATEGIE D'INTEGRATION DES VARIABLES ASN

Les variables du répertoire de l'Agence de Sureté Nucléaire (ASN) sont des données sémantiques grevées d'informations spatiotemporelles relatives aux Installations Nucléaires de Base (INB) ayant fait l'objet d'une demande d'autorisation pour une activité effective entre 1980 et 2011. Certaines variables peuvent être intégrées dans un SIG et par conséquent seul l'effet information* peut faire l'objet d'une maximisation.

L'i.st.e* $x_{(U_k)}^{EGRA}$ est voué à la modélisation des Expositions Géographiques potentielles à des Radionucléides Artificiels, diffusés légalement ou accidentellement dans les milieux : eau, air, sol et biologiques du fait de la présence d'INB* en fonctionnement. L'i.st.e* somme toutes les INB* présentes dans des zones tampons dynamiques, pondérées par la valeur des rayons géographiques de dilatation des U_k.

OPTIMISATION DE L'EFFET INFORMATION

La stratégie de maximisation de *l'effet information** est d'abord exprimée de façon théorique, puis un applicatif permet d'illustrer la synoptique d'estimation et les outils de géo-traitement utilisés.
Le processus d'estimation de $x_{(U_k)}^{EGRA}$ s'opère en trois étapes : (o) Une phase d'initialisation ayant pour objet l'intégration des informations spatiotemporelles ASN, la spatialisation des INB* et la spécification des paramètres du processus d'estimation ; (i) Une procédure de dénombrement dynamique, par un processus de dilatation en tâche d'huile d'entités géographiques, et de pondération par un paramètre spatial de capture ; (ii) Une phase de finalisation permettant d'agréger, par un opérateur spatial ensembliste, les dénombrements fragmentaires pondérés des Expositions Géographiques à des Radionucléides Artificiels estimées dans l'étape précédente.

TRANSFORMATION TOPOLOGIQUE : FUSION STATISTIQUE PAR DES OPERATEURS LOGIQUES ET SPATIAUX DE DILATATION ET DE CAPTURE

Remarques liminaires :
D'une manière générale : *all types of ionizing radiation are carcinogenic to humains* (IARC, 2012), i.e. sur THYR et TUM2, et les expositions environnementales sont très suspectées d'avoir un effet sur l'incidence des CATA (Jacob, Bertrand et al., 2010).

Les informations extraites du répertoire de l'ASN pour chaque INB* sont : la commune d'implantation CI_i, le nom de l'exploitant $EXPL_i$, le type d'installation$TYPE_i$, les dates de début et de fin de mise en service DMS_i ; DFS_i, avec : $INB_i^{info} = (CI_i;\ EXPL_i;\ DMS_i; TYPE_i)$.

Toutes les INB* manipulent ou transforment des produits radioactifs toxiques et rejettent de façon réglementée ou accidentelle des radionucléides dans l'environnement. La nature et les quantités diffusées varient selon le type d'INB* et sont difficilement estimables (Afsset, 2009a).

Les coordonnées géographiques des INB* ne sont pas publiques, seul le nom de la commune d'implantation l'est. Le nombre communal d'INB* ayant opéré un fonctionnement effectif sur la période recouverte par LEA varie 0 à 20 - pour la commune de Saint-Paul-Lès-Durance (ASN, 2011).

La granularité* informationnelle correspond à l'échelle d'investigation. Mais les INB* sont des entités géographiques ponctuelles et doivent être considérées comme telles, au moins à l'initialisation. La maille la plus fine garantissant l'intégrité des données géofla est de $(\{\varepsilon_x = 100\text{m}\} \times \{\varepsilon_y = 100\text{m}\})$, soit un hectare (IGN, 2004).

Spécification des hypothèses :

Le processus de *fusion* dynamique doit prendre en compte les remarques liminaires et les informations spatiotemporelles *relatives aux INB** en se fondant sur la *théorie des ensembles flous* (Dubois Didier, Prade Henri, 2004). Pour ce faire, il s'agit de *coupler un processus ensembliste itératif d'estimation de statistiques spatiales* permettant de dénombrer les INB* (Saporta, 2006) *à des méthodes d'analyse spatiale par des* dilations *en tâche d'huile* des U_k à partir de rayons de capture r_j (Voiron-Canicio, 1995) et un système de pondération géographique élaboré dans une logique adaptée aux *statistiques territoriales* (Charre, 1995).

Afin de spécifier les paramètres du modèle, le risque d'exposition est supposé proportionnellement croissant au nombre d'INB* sises dans les communes voisines, mais inversement proportionnel à la distance géographique qui les séparent (Afsset, 2009a), en tenant compte aussi des connaissances théoriques sur *les interactions spatiales multi-échelles* (Dauphiné et Voiron-Canicio, 1988) et sur *la propagation de la radioactivité artificielle dans l'environnement et l'exposition des populations vivant à proximité d'installations nucléaires* (AECB, 1991).

Spécification de la stratégie d'estimation :

La synoptique du processus d'estimation dynamique des $x_{(U_k)}^{EGRA}$ s'opère en trois étapes.

PHASE 0 : Initialisation des paramètres

Les données sémantiques ASN mobilisées INB_i^{info} sont intégrées dans le SIG pour les INB* situées sur $\mathcal{D}(\omega_.)$, i.e. en France métropolitaine, et dont la Date de Mise en Service chevauche celle de l'étude LEA.

$$INB_i^{SIG} = \bigcup_{i=1}^{n_{INB}} \left(INB_i^{info} \middle| \{CI_i \equiv U_u \subseteq \mathcal{D}(\omega_.)\} \cap \{DMS_i \in [\![1980; 2010]\!]\}\right)$$

Parmi les 126 INB, 125 ont été retenues. Elles sont spatialisées au niveau des centroïdes communaux de l'U_u concernée et par un processus stochastique afin d'éviter qu'elles se chevauchent :

$$s_{(INB_i)} = \left(\left(x_{(U_u \equiv CI_i)} + u_{(\varepsilon_x)}\right); \left(y_{(U_u \equiv CI_i)} + u_{(\varepsilon_y)}\right)\right)$$

Avec : $\{u(\varepsilon_x) \overset{def}{=} u(\varepsilon_y)\} \xrightarrow[n\to+\infty]{} u \sim \mathbb{U}\left[\!\left[\left\{a_u = -\frac{\varepsilon_x}{2}\right\}; \left\{b_u = \frac{\varepsilon_x}{2}\right\}\right]\!\right]$. Puis pour chaque commune située dans $\mathcal{D}(\omega_.)$ le nombre d'INB* contenues dans son enceinte est dénombré par $x_{(U_u)}^{EGRA.\Re_0}$.

Pour finir, la valeur des rayons de capture : r_j qui déterminent la morphologie des dilatations en tâche d'huile $\mathfrak{Z}_{(U_k)}^{r_.}$ des U_k est spécifiée dans une table attributaire. Les valeurs de ces paramètres sont fixées conditionnellement aux connaissances expertes disponibles en géographie de la santé et en analyse spatiale. Les valeurs spécifiées pour les $x_{(U_k)}^{EGRA}$ sont :

$$r_. = (\{r_0 = -1\ m\};\ \{r_1 = 2500\ m\};\ \{r_2 = 5000\ km\};\ \{r_3 = 10\ 000\ km\};\ \{r_4 = 20\ 000\ km\})$$

PHASE 1 : Statistiques d'opérateurs spatiaux pondérées géographiquement

Il s'agit d'un processus dynamique de dilatation d'entités géographiques et de dénombrement itératif de statistiques pondérées géographiquement par la valeur du rayon de capture. Des statistiques fragmentaires sont estimées pour chaque rayon de capture $r_j \in \{r_1; r_2; r_3; r_4\}$ par le processus dynamique qui s'opère en trois temps :

(i) Construction des buffers de capture : $\mathfrak{Z}_{(U_k)}^{r_.}$ par des dilatations, en tâche d'huile, des U_k. conditionnellement à la valeur du rayon de capture r_j, tel que :

$$\mathfrak{Z}_{(U_k)}^{r_j} = (U_u \oplus r_j), \forall$$

(ii) Somme des valeurs des $x_{(U_u)}^{EGRA.\Re_0}$ complètement ou partiellement contenues dans les zones de capture : $\mathfrak{Z}_{(U_k)}^{r.}$, tel que :

$$\Upsilon_{(U_k)}^{EGAR.r_j} = \sum_{u=1}^{n_u} x_{(U_u)}^{EGRA.\Re_0} \cdot \mathbb{1}_{\left\{U_u \subseteq \mathfrak{Z}_{(U_k)}^{r_j}\right\}}, \quad \forall\, k \in \{1, \dots, q_1\}$$

(iii) Estimation des i.st.e* fragmentaires $x_{(U_k)}^{EGAR.r_j}$ qui correspondent à la somme des INB* capturées par les $\mathfrak{Z}_{(U_k)}^{r_j}$ et réduite par la valeur absolue du rayon de capture r_j, exprimée en mètre, soit :

$$x_{(U_k)}^{EGAR.r_j} = \frac{\Upsilon_{(U_k)}^{EGAR.r_j}}{|r_j|} \times \left(1 + 999 \cdot \mathbb{1}_{\{r_j \neq -1\}}\right)^{-1}$$

Les $x_{(U_k)}^{EGAR.r_j}$ représentent des Expositions Géographiques à des Radionucléides Artificiels potentiellement présents dans les U_k – pondérées géographiquement par la proximité spatiale avec les communes où des INB* sont localisées.

PHASE 2 : Agrégation ensembliste de statistiques spatiales fragmentaires

Les $x_{(U_k)}^{EGRA.r_j}$ embrassent pleinement les spécificités spatiales, et de manière partielle les spécificités temporelles des Expositions Géographiques à des Radionucléides Artificiels liées à la présence d'INB. L'i.st.e* final est défini comme la somme des statistiques spatiales fragmentaires. Il est arrondi à l'entier le plus proche, de façon à conserver les propriétés statistiques des variables initiales:

$$x_{(U_k)}^{EGRA} = \left[\sum_{r_j=r_0}^{r_4} x_{(U_k)}^{EGAR.r_j}\right]$$

Afin d'illustrer cette stratégie d'estimation, les résultats cartographiques pour les communes situées en région PACA et aux alentours, la synoptique du processus de géo-traitement et les outils SIG utilisés sont déclinés dans le paragraphe suivant.

SYNOPTIQUE DU PROCESSUS DE GEOTRAITEMENT D'ESTIMATION

Workflow SIG du processus d'estimation :

La chaine synoptique du processus de géo-traitement ainsi que les outils SIG utilisés sont déclinés dans le schéma ci-dessous.

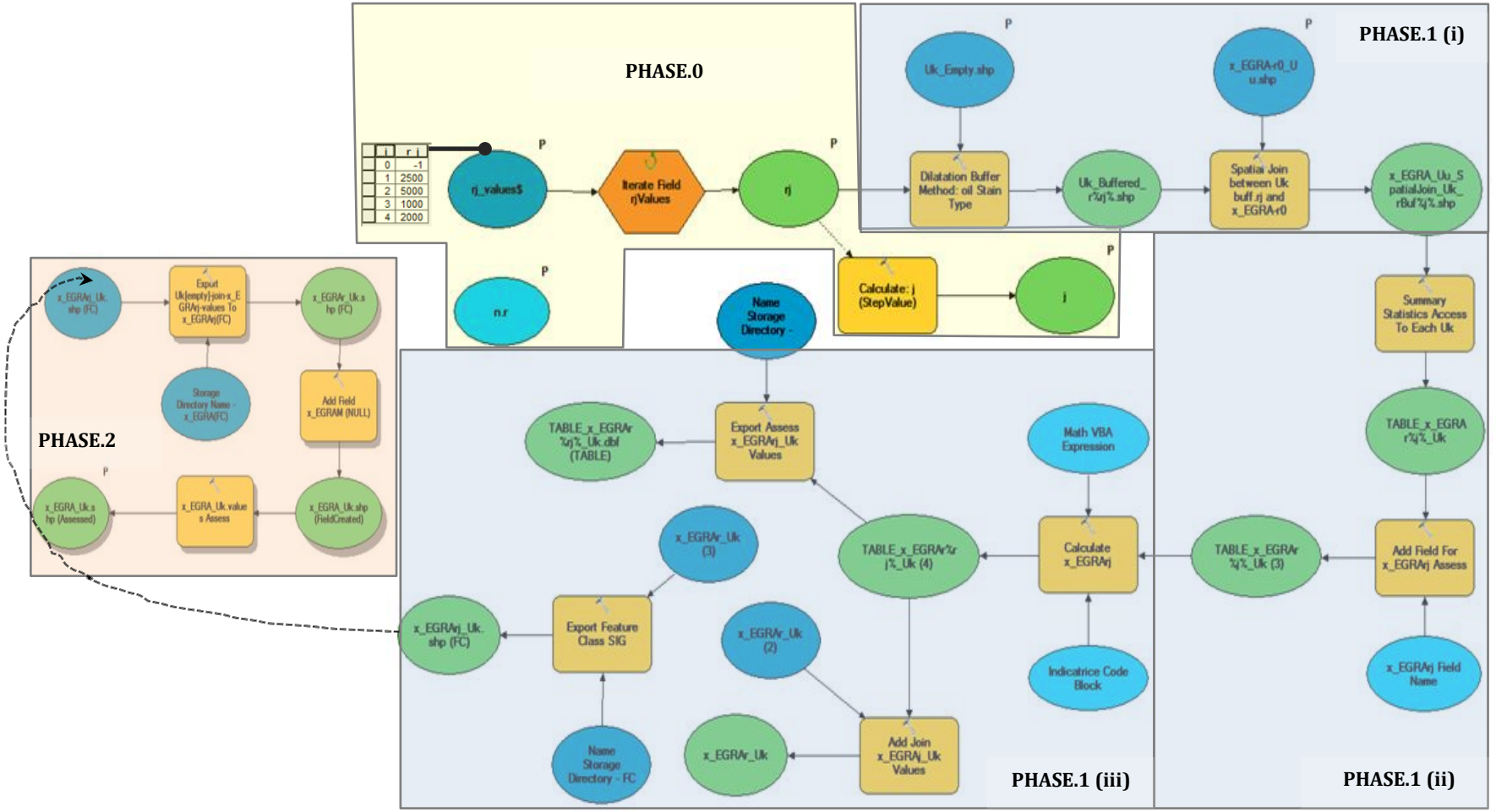


Figure 137 : Workflow ModelBuilder des outils SIG associés aux différentes phases du processus d'estimation

Un workflow est une sorte de programme visuel permettant de concaténer l'exécution d'outils SIG de géo-traitement et de construire un processus adapté à l'estimation $x_{(U_k)}^{EGRA}$. Il se décompose en trois étapes :

PHASE 0 : initialisation des paramètres :

Sélection des informations ASN pertinentes ; spatialisation des INB* et dénombrements communaux $x_{(U_u)}^{EGRA.\Re_0}$; implémentation des paramètres du modèle : r_j

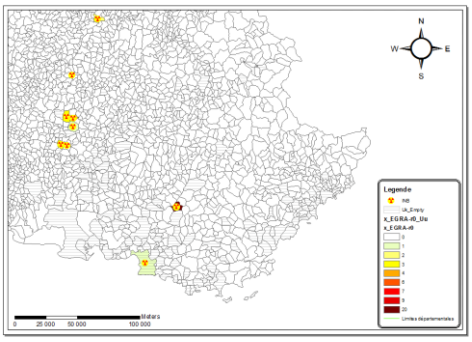


Figure 138 : Superposition des INB* spatialisées et des U_k

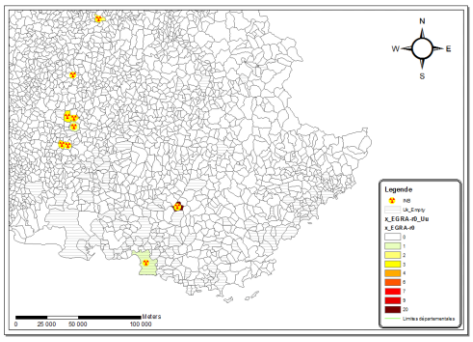


Figure 139 : Superposition des INB* spatialisées et des $x_{(U_u)}^{EGRA.\Re_0}$

PHASE 1 : Statistiques itératives d'opérateurs spatiaux pondérées géographiquement

(i) Dilatation en tâche d'huile des $U_k$et construction des zones de capture $\mathfrak{Z}_{(U_k)}^{r_j}$;

(ii) Dénombrement des INB_i^* sises dans les U_u capturées : $\Upsilon_{(U_k)}^{EGAR.r_j}$;

(iii) Estimation des i.st.e* EGRA pondérés intermédiaires : $x_{(U_k)}^{EGRA.r_j}$

Les schémas cartographiques situés dans la colonne de gauche représentent les $x_{(U_u)}^{EGRA.\Re_0}$ capturées par les $\mathfrak{Z}_{(U_k)}^{r_j}$ pour certaines valeurs de r_j. Et, ceux dans la colonne de droite sont les résultats obtenus pour certains $x_{(U_k)}^{EGAR.r_j}$.

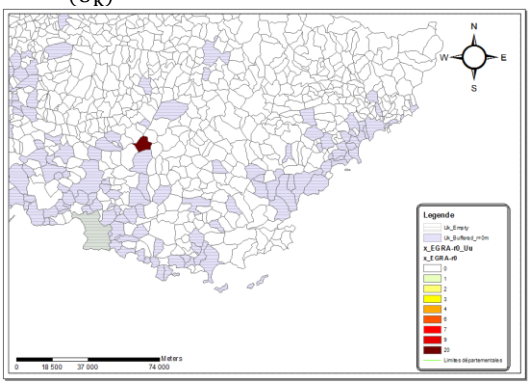


Figure 140 : superposition des $x_{(U_u)}^{EGRA.\Re_0}$ des U_k et des; $\mathfrak{Z}_{(U_k)}^{\{r_j=0m\}}$

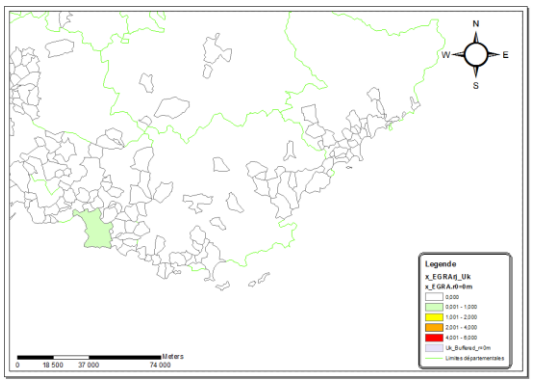


Figure 141 : Superposition des $\mathfrak{Z}_{(U_k)}^{\{r_j=0m\}}$ et des $x_{(U_k)}^{EGRA.\{r_j=0m\}}$

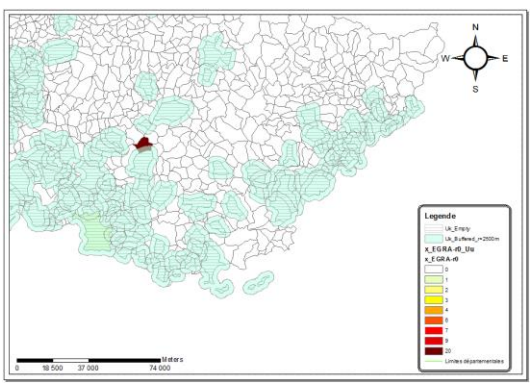


Figure 142 : Superposition des $\mathfrak{Z}_{(U_k)}^{\{r_j=2\ 500m\}}$ et $x_{(U_u)}^{EGRA.\{r_j=2\ 500m\}}$

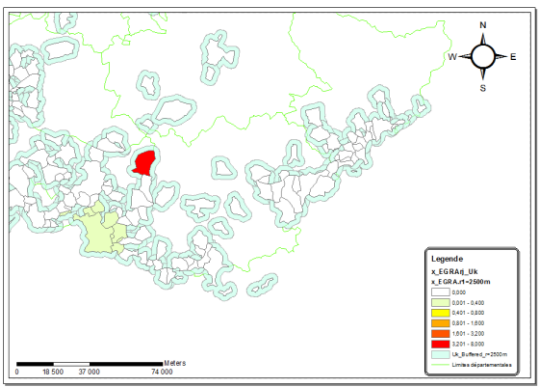


Figure 143: Superposition des $x_{(U_k)}^{EGRA.r_j}$; des U_k ; des $\mathfrak{Z}_{(U_k)}^{\{r_j=2\ 500m\}}$

. (:) (:)

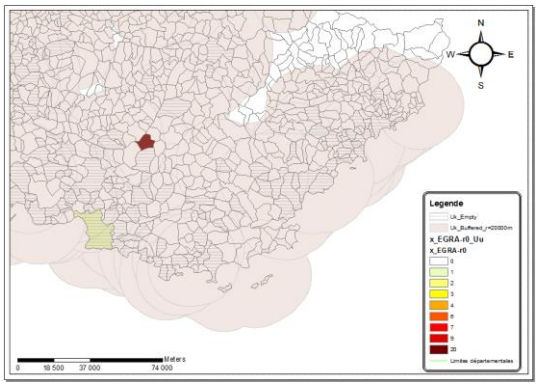


Figure 144 : Superposition des $\mathfrak{Z}_{(U_k)}^{\{r_j=20\ 000m\}}$ et des $x_{(U_u)}^{EGRA.\{r_j=20\ 000m\}}$

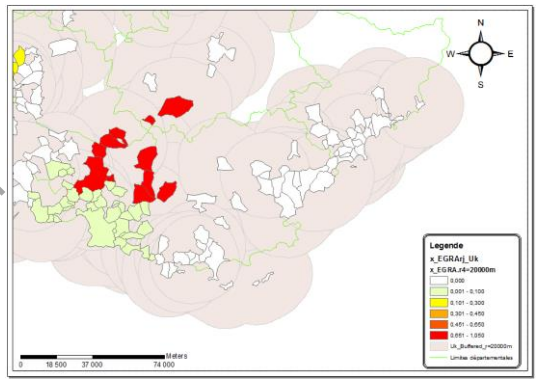


Figure 145 : Superposition des $x_{(U_k)}^{EGRA.r_j}$; des U_k ; des $\mathfrak{Z}_{(U_k)}^{r_j=20\ 000m}$

PHASE 2 : Agrégation ensembliste de statistiques spatiales fragmentaires

Sommation pour chaque U_k des i.st.e* EGRA intermédiaires : $x_{(U_k)}^{EGRA.r_j}$ arrondie à l'entier : $x_{(U_k)}^{EGRA}$

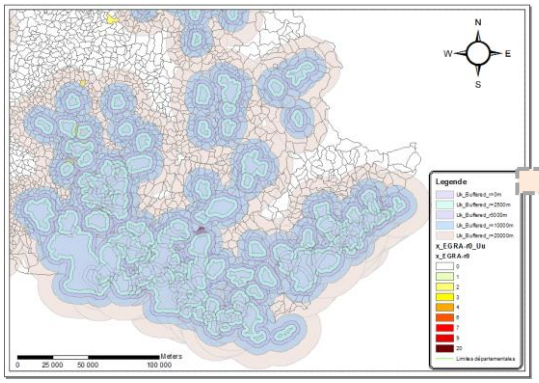


Figure 146 : Superposition des $x_{(U_u)}^{EGRA.\Re_0}$ et des $\mathfrak{Z}_{(U_k)}^{r_j}$

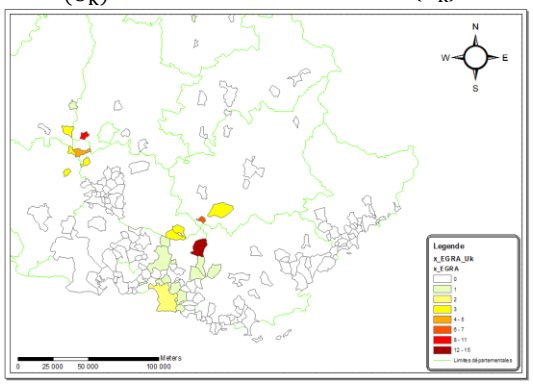


Figure 147 : Valeurs des $x_{(U_k)}^{EGRA.r_j}$ en PACA et aux alentours

Outils SIG utilisés :

- ModelBuilder, application de création et de gestion d'outils personnalisés de géo-traitements - (ArcGIS: ModelBuilder, 2013) ;
- Construction de zones Tampons ou Buffers $\mathfrak{Z}_{(U_k)}^{r_j}$ - qui s'exécute si et seulement si $\{r_j \in \mathbb{R}_*\}$ - (ArcGis: Analysis toolbox, 2013).
- Outils : *Jointure spatiale* et *résumé statistique* qui permettent l'appariement spatial des U_u avec les $\mathfrak{Z}_{(U_k)}^{r_j}$ et par suite d'estimer : $x_{(U_u)}^{EGRA.\Re_0}$ et les $x_{(U_k)}^{EGRA.r_j}$ (ArcGis: Analysis toolbox, 2013).
- La procédure est répétée par le biais d'un *itérator SIG* en fonction du nombre de r_j spécifiés (ArcGIS: Iterator, 2013).

Conclusion :

En somme $x_{(U_k)}^{EGAR}$ est un estimateur spatiotemporel qui a pour dessein de modéliser, dans les communes de 1ère espèce : U_k, l'Exposition Géographique (interne et/ou externe) potentielle à des Radionucléides Artificiels, émis de façon chronique et/ou accidentelle, dans les différents milieux : eau, air, sol et biologique en fonction du nombre d'INB* ayant opéré une activité effective depuis 1980 et pondéré par leur proximité spatiale euclidienne avec chacune des U_k.

STRATEGIE D'INTEGRATION DES VARIABLES CLC

Les données spatiales CORINE Land Cover (CLC) sont utilisées dans le but de modéliser l'exposition géographique *potentielle ou curieuse* à des agents physicochimiques toxiques protéiformes présents dans l'environnement des milieux de vie à partir de l'occupation biophysique des sols.
Afin de *minimiser le biais conditionnel** des $x_{(U_k)}^{l:CLC}$ la stratégie propose d'abord d'estimer des indicateurs spatiaux à différentes dates $x_{(U_k),t}^{l:CLC}$ en *optimisant l'effet information** par *des fusions d'entités vectorielles* suivies d'une méthode *d'acquisition d'informations spatiales adaptée à l'échelle* des U_k, et ensuite, de proposer un processus d'agrégation temporelle adapté aux discordances diachroniques des bases CLC.

OPTIMISATION DE L'EFFET INFORMATION

L'optimisation de *l'effet information** permet d'obtenir, à différentes dates, des indicateurs spatiaux aux différentes dates $x_{(U_k),t}^{l:CLC}$ par un processus en deux étapes. En commençant par fusionner les informations spatiales $e.clc_t^l$ en tenant compte des informations déclinées dans l'état de l'art (chapitre.1), puis par un processus de capture permettant d'estimer des statistiques zonales à l'échelle des U_k.

TRANSFORMATION TOPOLOGIQUE : FUSION STATISTIQUE D'INFORMATIONS GEOGRAPHIQUES VECTORIELLES

Remarques liminaires :
Les variables CLC sont des entités géographiques vectorielles : $e.clc_t^l$ décrivant l'occupation biophysique des sols à différentes dates $t = \{1990;\ 2000; 2000.r;\ 2006\}$ (CGDD, SOeS, 2009)

Les $e.clc \in \{1, \dots, n_t^{e.clc}\}$ caractérisent l'intégralité du territoire français métropolitain : $\mathcal{D}(\omega_.)$. Le niveau d'analyse le plus fin est celui qui est utilisé, i.e. CLC n°3 qui discrétise l'espace en 44 postes - notés : $l \in \varrho(CLC) = \{111, \dots, 523\}$.

La nomenclature des $e.clc_t^l$ garantit que les postes de niveau 3 sont adaptés à la dialectique spatiale jusqu'à l'échelle des communes (CGDD, SOeS, 2009).

Spécification de l'hypothèse principale :
Elaborer un processus de *fusion* des : $e.clc_t^l$, fondé sur la *théorie des ensembles flous,* à partir des remarques et des connaissances en santé environnementale* énoncées dans l'état de l'art (Dubois Didier, Prade Henri, 2004). Il s'agit d'une agrégation horizontale basée sur des connaissances expertes. C'est une stratégie très courante en géographie de la santé où l'exposition à des substances toxiques est supposée proportionnelle à une occupation des sols qui la suggère (Peguy, 1996).

Stratégie d'estimation :
Les zones géographiques d'expositions potentielles à des agents physicochimiques protéiformes nocifs sont obtenues par l'agrégation spatiale de $e.clc_t^l$ particuliers, tel que

$$e.clc_t^{NOM.EXPO} = \bigcup_{\forall\, l \in \varrho(NOM.EXPO)} \left(e.clc_t^l \subseteq \mathcal{D}(\omega_.)\right), \qquad \forall \varrho(NOM.EXPO) \subseteq \varrho(CLC)$$

Zones potentielles d'exposition aux pesticides définies par les postes :

$$\varrho(PEST) = \{141,142, 211, 212, 213,221, 222, 223, 241, 242, 243, 244\}$$

Ces postes caractérisent les : terres arables, i.e. des espaces où l'on pratique la culture de céréales, de fleurs, de légumes et de riz… ; cultures permanentes composées de : vignobles, vergers, oliveraies… et

les espaces verts urbains, i.e. les équipements de sport et de loisir comme les hippodromes, les golfs, les campings et les terrains de sport (CGDD, SOeS, 2009).

Zones potentielles d'exposition à des nuisances urbaines et industrielles néfastes multiples :

$$\varrho(\text{URIN}) = \{111\ 112, 121, 122, 123, 124, 131, 132, 133\}$$

Les deux premiers postes représentent des zones fortement urbanisées, les quatre suivants des zones industrielles, commerciales, portuaires, aéroportuaires ainsi que des infrastructures de transport de grande ampleur. Quant aux trois derniers, il s'agit de chantiers, de carrières et de décharges (CGDD, SOeS, 2009).

Zones curieuses d'exposition à des substances toxiques diffusées dans l'environnement suite à des feux de forêt de grande ampleur :

$$\varrho(\text{FEFO}) = \{334\}$$

Ce poste cartographie les matériaux carbonisés encore présents au moment où les ortho-photographies ont été réalisées (CGDD, SOeS, 2009). Il convient de rappeler que les feux de forêts diffusent d'importantes quantités d'agents chimiques toxiques comme le formaldéhyde, le méthane, des MES et des dioxines : tétrachlorodibenzopara… (chapitre.1).

Zones curieuses préventives de non exposition à des substances toxiques

$$\varrho(\text{PREV}) = \{311, 312, 313, 321, 322, 323\}$$

Les trois premiers postes représentent des forêts de feuillus, de conifères ou composées d'arbres mixtes ; Les trois autres décrivent des surfaces arbusives telles que : des pâturages naturels, des landes, des broussailles, des végétations sclérophylles (i.e. : composées : d'arbustes persistants, de maquis et de garrigues (CGDD, SOeS, 2009)).

Remarque :
Cette phase de fusion d'entités géographiques vectorielles a été répétée par des requêtes SQL grâce au module ModelBuilder (ArcGIS: ModelBuilder, 2013).

Afin d'estimer les indicateurs spatiaux diachroniques $x^{\text{l:CLC}}_{(U_k),t}$ les surfaces d'occupation biophysique des sols caractérisant des expositions potentielles ou curieuses $\varrho(\text{EXPO})$ doivent être évaluées à l'échelle des U_k, grâce à un processus de *capture spatiale par superposition* des $\text{e.clc}_t^{\text{PREV}}$; $\text{e.clc}_t^{\text{FEFO}}$; $\text{e.clc}_t^{\text{URIN}}$; $\text{e.clc}_t^{\text{PREV}}$.

UNIFORMISATION A L'ECHELLE DES COMMUNES

Remarques liminaires :
Il s'agit de proposer une stratégie de défragmentation des $\text{e.clc}_t^{\varrho(\cdot)}$, sur une maille – ou surface raster $\text{clc}^{\varrho(\cdot)}_{(\text{cell}),t}$, permettant d'estimer précisément, dans chaque U_k, la proportion de surfaces diachroniques associées, i.e. à partir de :

$$\text{e.clc}_t^{\varrho(\cdot)} :\rightarrow \text{clc}^{\varrho(\cdot)}_{(\text{cell}),t}$$

Selon la documentation technique la précision des données CLC est assurée pour des surfaces supérieures à 25 ha (CGDD, SOeS, 2009), i.e. que :

$$\left\{\text{res}_{(\text{CLC})} \stackrel{\text{def}}{=} (\text{cell. size. h. aquis} \times \text{cell. size. v. aquis}) \leq (500\text{m} \times 500\text{m})\right\}$$

La transformation : $\text{e.clc}_t^{\varrho(\cdot)} :\rightarrow \text{clc}^{\varrho(\cdot)}_{(\text{cell}),t}$ doit être effectuée sur une maille dont la résolution : $\text{res}_{(.)}$ est adaptée à l'estimation d'i.s.e diachroniques $x^{\text{l:CLC}}_{(U_k),t}$ robustes. Or $\text{res}_{(\text{CLC})}$ ne l'est pas.
Tous les jeux d'entités vectorielles peuvent être transformés en raster $\text{clc}^{\varrho(\cdot)}_{(\text{cell}),t}$. Ils peuvent être ré-échantillonnés. Le gain de précision au niveau de la granularité* des informations spatiales est limité

par la résolution de l'input, i.e. $\text{res}_{(\text{CLC})}$. Cependant dans le cadre d'une problématique de capture de surface le ré-échantillonnage peut aller au-delà de la résolution des données. Il ne s'agit pas de mieux caractériser l'espace mais de mieux épouser la forme des entités vectorielles initiales, i.e. les $\text{e.clc}_{t}^{\varrho(\cdot)}$ (ESRI, 2013).

Afin d'illustrer le processus de capture par superposition spatiale de cellules rasters $\text{e.clc}_{t=1990}^{\varrho(\text{PEST})}$ obtenues par défragmentation - avec une résolution adaptée à l'estimation des $x_{(U_k),t}^{\text{PEST}}$ – à partir des postes CLC fusionnés : $\text{e.clc}_{t=1990}^{\varrho(\text{PEST})}$ qui caractérisent des zones potentielles d'exposition aux pesticides en 1990 ont été superposés aux U_k – est présentée par la figure suivante :

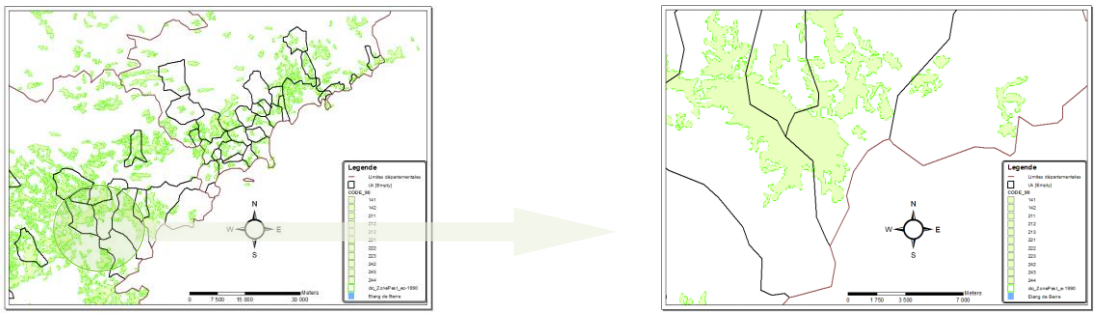


Figure 148 : Superposition spatiale des $\text{clc}_{(\text{cell}),t=1990}^{\varrho(\text{PEST})}$ avec les postes fusionnés $\text{e.clc}_{t=1990}^{\varrho(\text{PEST})}$ obtenus par défragmentation spatiale avec une $\text{res}_{(\cdot)}$ optimisée

Dans le cadre d'une problématique de capture de surface une résolution grossière engendre de fortes imprécisions. En revanche, comme le montre le schéma ci-dessous, une résolution trop fine peut aussi s'avérer spécieuse. De plus, la taille du fichier output et les temps d'estimation des statistiques zonales – permettant de calculer les $x_{(U_k),t}^{l}$ – sont inversement proportionnels à la résolution (ESRI, 2013).

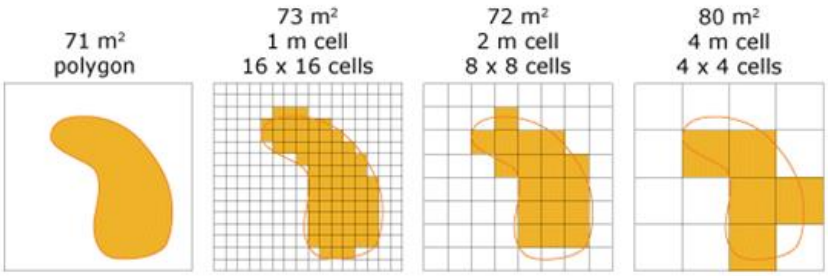


Figure 149 : Schéma de principe d'estimation de surfaces par capture spatiale de cellules raster et des résultats en fonction de différentes résolutions : $\text{res}_{(\cdot)}$

Il convient de prendre en compte ces spécificités dans l'estimation des indicateurs spatiaux diachroniques.

<u>Spécification de l'hypothèse :</u>

*Maximiser l'effet information** en déterminant une résolution raster en adéquation avec l'échelle d'investigation (Marcotte, 2008), permettant de *supputer des statistiques zonales consistantes* circonscises par des *limites administratives territoriales* (Groupe CHADULE, 1997), à partir d'un processus de *capture spatiale par superposition raster* - calibré par un algorithme itératif décisionnel adapté à la logique de l'analyse spatiale (Voiron-Canicio, 1995)

<u>Principe stratégique d'estimation :</u>

Il s'agit de spécifier la résolution la plus grossière pour laquelle le rapport entre les surfaces géographiques géofla $x_{(U_k)}^{SG}$ et celles estimées pour une résolution testée : $\text{res}_{(\text{test})}$, à partir de l'ensemble

des cellules rasters représentatives de tous les postes CLC : ϱ(CLC), pour chaque U_k, i.e. $x^{\varrho(CLC)}_{(U_k|res_{(test)})}$ - est inférieur à un *critère d'optimisation spatiale :* $\aleph_{opt}$ dont la valeur est *petite*, soit :

$$res_{(CLC.opt)} = \underset{res_{(test)} \in \{res_{(CLC)} \mp ite \cdot \Delta_{res}\}}{argmax} \left\{ mâx \left(\bigcup_{k=1}^{n_{(U_k)}} \left(\left\{ 1 - \frac{mîn\left(x^{\varrho(CLC)}_{(U_k|res_{(test)})};\ x^{SG}_{(U_k)} \right)}{mâx\left(x^{\varrho(CLC)}_{(U_k|res_{(test)})};\ x^{SG}_{(U_k)} \right)} \leq \aleph_{opt} \right\} \right) \right) \right\}$$

Avec comme paramètres itératifs spécifiés : $\{\Delta_{res} = 50m\}$, de sorte que l'intervalle testé couvre : $res_{(test)} \in [\![\{50, \dots, 750\}]\!]$ et une contrainte d'optimisation fixée à : $\{\aleph_{opt} = 2\%\}$;
La résolution raster *optimale* a été estimée à :

$$\{res_{(CLC.opt)} \overset{def}{=} (cell.size.h.clc.opt \times cell.size.v.clc.opt) = (100m \times 100m)\}$$

Les indicateurs spatiaux diachroniques d'exposition, ou de non exposition, potentielle à des agents physicochimiques délétères à partir des données CLC sont données par un rapport de surface, tel que :

$$x^{l:\varrho(\cdot)}_{(U_k),t} = \frac{res_{(CLC.opt)}}{10\,000 \cdot x^{SG}_{(U_k)}} \cdot \sum_{\forall cell \subseteq \mathcal{D}(\boldsymbol{\omega}.)} \left(\mathbb{1}_{\left\{ clc^{l=\varrho(\cdot)}_{(cell),t} \subseteq U_k \middle| clc^{l=\varrho(\cdot)}_{(cell),t} \neq \phi \right\}} \right), \quad \forall\, k \in \{1, \dots, q_1\}$$

Avec : $res_{(CLC.opt)}$ exprimée en en m^2 ; $x^{SG}_{(U_k)}$ en ha – des données attributaire géofla (IGN, 2004)

Remarque :
La recherche opérationnelle itérative de $res_{(CLC.opt)}$ et le calcul des $x^{l:\varrho(\cdot)}_{(U_k),t}$ ont été effectués grâce à : *ModelBuilder* et à l'outil d'analyse spatiale : *Zonal Statistics as Table* (ESRI, 2013).

MAXIMISATION DE L'EFFET DE SUPPORT

L'optimisation de *l'effet information** permet d'obtenir des $x^{l:CLC}_{(U_k),t}$. *horizontaux* consistants. La maximisation *verticale* de *l'effet de support** permet d'obtenir les i.st.e* $x^{l:CLC}_{(U_k)}$ définitifs.

PROCESSUS D'HARMONISATION TEMPORELLE PROBABILISTE

Remarques liminaires :
Les données CLC ont été acquises à partir de méthodes différentes qui ont suivi l'évolution des processus technologiques et des innovations techniques. Par conséquent les données CLC sont disponibles à trois dates distinctes mais une des bases a dû être révisée et par conséquent quatre temporalités sont disponibles

$$t = \{1990;\ 2000;\ 2000.r;\ 2006\}$$

Il convient de prendre en compte *le degré de croyance* quant à la *qualité temporelle associée* au processus d'innovation inhérent aux méthodes d'acquisition et de traitement du signal.

Hypothèses :
Il s'agit de proposer des densités spatiales uniques, i.e. en agrégeant dans le temps les $x^{l:CLC}_{(U_k),t}$ à partir d'un système de pondération subjectif mais néanmoins adapté à la dialectique géographique (Pumain et Saint-Julien, 1997) – dans l'idée d'une *fusion temporelle floue associée, à partir d'un degré de croyance quant au gain de qualité apporté par l'innovation technologique* (Dubois Didier, Prade Henri, 2004). En veillant cependant à supputer un i.st.e* consistant, i.e. sans que les poids n'induisent de biais statistique (Saporta, 2006).

Principe d'estimation :
Les i.st.e* proposés sont des moyennes temporelles pondérées par des facteurs de croyance que le géographe accorde au gain de qualité induit par les innovations technologiques des méthodes d'acquisition de données spatiales :

$$x^{l}_{(U_k)} = \frac{1}{n_t^{CLC}} \sum_{t=1}^{n_t^{CLC}} \left(x^{l=\varrho(\cdot)}_{(U_k),t} \cdot \rho_t \right) \stackrel{def}{=} \frac{1}{n_t^{CLC}} \cdot \rho_{CLC} \left(x^{l=\varrho(\cdot)}_{(U_k),t} \right)^{t}$$

Avec : n_t^{CLC} le nombre de bases CLC temporelles; ρ_{CLC} est le vecteur des facteurs de croyance en l'innovation – garantissant la contrainte de non biais $\sum_{t=1}^{n_t^{CLC}} \left(\frac{\rho_t}{n_t^{CLC}} \right) = 1$ – définis subjectivement pour chaque temporalité – tel que :

$$\rho_{CLC} = (\rho_{t=1990}\ ;\ \rho_{t=2000};\ \rho_{t=200.r};\ \rho_{t=2006}\ ) = (0{,}8;\ 0{,}8;\ 1{,}2;\ 1{,}2)$$

Remarques:
ρ_{CLC} est introduit dans l'optique de parvenir à une représentation plus robuste et plus juste des expositions environnementales potentielles et curieuses - entre 1990 et 2006 et dans les $\mathcal{U}^{1ère}$ – à partir des données CLC agrégées par fusion des postes retenus $\varrho(\cdot)$.
Les i.st.e. $x^{l}_{(U_k)}$ ont été estimés par le biais d'une macro programmée en VBA (Microsoft, 2013).

PRESENTATION DES RESULTATS ET REMARQUES

Les résultats cartographiques de la modélisation des FE-PHY.CHIM* *pertinents* et Curieux**, par le biais les i.st.e* $x^{l:PHY.CHIM}_{(U_k)}$ proposés, représentent la variabilité spatiotemporelle territoriale de l'exposition *à des substances toxiques - physiques ou chimiques - présentes dans les milieux environnementaux, que sont : l'eau, l'air le sol et des éléments biologiques*. Ils sont estimés par des stratégies d'intégration spatiotemporelle à partir de mesures environnementales, d'indices de risques, de doses d'exposition ou d'indicateurs géographiques caractérisant la biophysique des sols.

Les FE-PHY.CHIM* modélisent les disparités géographiques des *expositions environnementales* (i.e. à des doses faibles mais chroniques) à des substances toxiques dont les effets déterministes ou contributifs ont un impact avéré ou fortement suspecté vis-à-vis des PM* d'intérêt : les cataractes, les tumeurs thyroïdiennes et les cancers en général. La *pertinence* des FE-PHY.CHIM* repose sur des connaissances bibliographiques établies dans l'état de l'art (chapitre.1).

Pour les mêmes raisons que pour les autres FE* modélisés, les résultats cartographiques de la modélisation des FE-PHY.CHIM* sont présentés uniquement pour les U_k sises en région PACA, et l'indicateur $SpaLea^{2}_{U_{k_o}}$ représentant l'incertitude associée à l'identification des U_k n'est pas affiché.

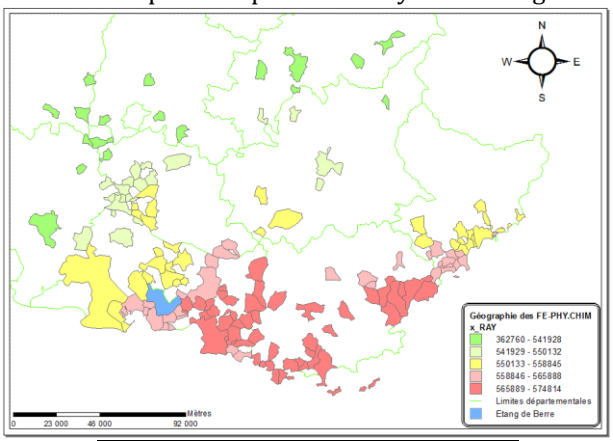
La documentation des cartes décline l'i.st.e* $x^{l:PHY.CHIM}_{(U_k)}$, le type de variabilité spatiale modélisée et le tableau qui l'agrémente renvoie pour tous les i.st.e* de nature : Quantitative, les principaux *paramètres* statistiques *de position et de dispersion* estimés sur l'ensemble des U_k. ; Qualitative, en l'occurrence $x^{RADON}_{(U_k)}$, la $\mathbb{P}_{F_n}\left(x^{l}_{(U_k)} = c^{l}_{j}\right)$ i.e. la proportion observée de chacune des modalités c^{l}_{j} prises par l'i.st.e* considéré, estimée sur l'ensemble des U_k.

Des remarques sont déclinées uniquement lorsque des singularités spatiales sont observables.

CARTOGRAPHIES ET STATISTIQUES DE LA GEOGRAPHIE DES FE-PHY.CHIM

Géographie des expositions potentielles à des paramètres géophysiques.

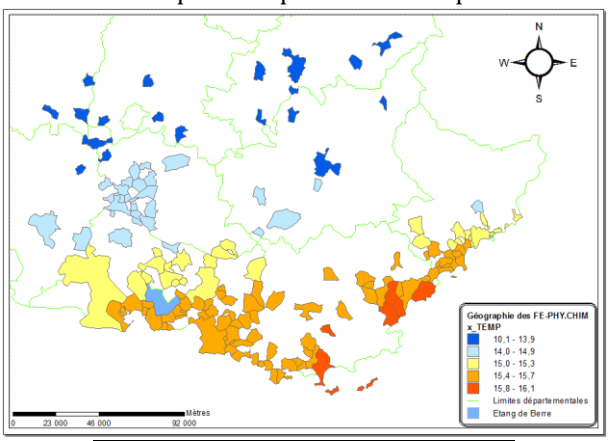
Variabilité spatiotemporelle du rayonnement global



Statistique	$x_{(U_k)}^{RAY}$ (joule/cm²)
môy(·)	484921
$\hat{\sigma}(\cdot)$	66314
$\hat{Q}_1(\cdot)$	410210
mêd(·)	476656
$\hat{Q}_3(\cdot)$	557769

Figure 150 : i.st.e* : $x_{(U_k)}^{RAY}$: doses spatiotemporelles des rayonnements solaires cumulés annuels entre 1981et 2010

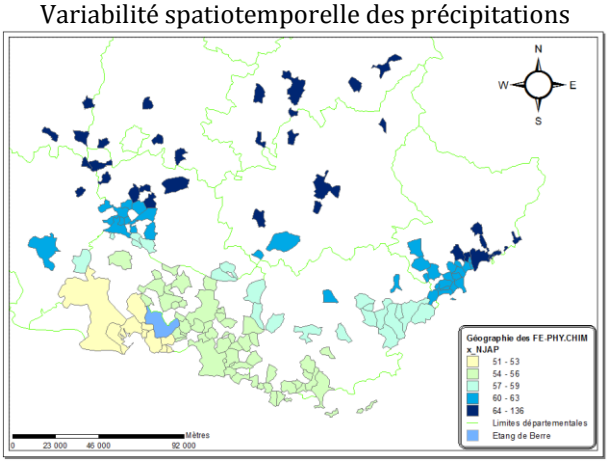
Variabilité spatiotemporelle des températures



Statistique	$x_{(U_k)}^{TEMP}$ (°Celsius)
môy(·)	12,72
$\hat{\sigma}(\cdot)$	2,07
$\hat{Q}_1(\cdot)$	10,82
mêd(·)	11,62
$\hat{Q}_3(\cdot)$	15,23

Figure 151 : i.st.e* : $x_{(U_k)}^{TEMP}$: températures spatiotemporelles annuelles moyennes 1981et 2010

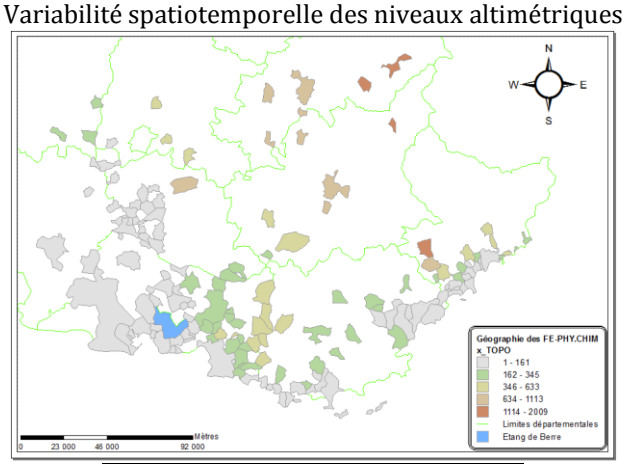
Variabilité spatiotemporelle des précipitations



Statistique	$x_{(U_k)}^{NJAP}$ (Jours)
môy(·)	**93,7**
$\hat{\sigma}(\cdot)$	**29,2**
$\hat{Q}_1(\cdot)$	**60,7**
mêd(·)	**105,0**
$\hat{Q}_3(\cdot)$	**122,2**

Figure 152 : i.st.e* : $x_{(U_k)}^{NJAP}$: valeurs spatiotemporelles du Nombre de Jours Annuels Pluvieux 1981et 2010

Variabilité spatiotemporelle des niveaux altimétriques

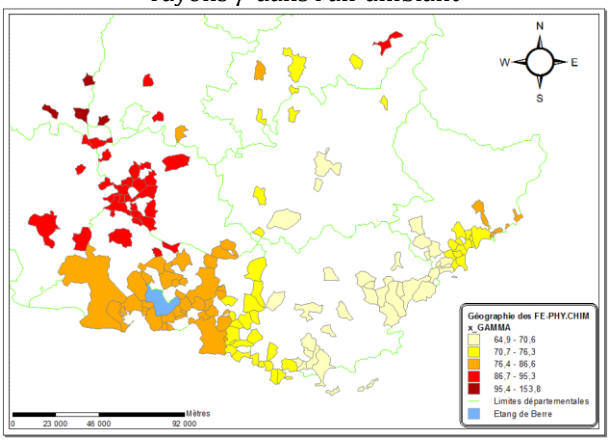


Statistique	$x_{(U_k)}^{TOPO}$ (m)
môy(·)	**307,7**
$\hat{\sigma}(\cdot)$	**275,2**
$\hat{Q}_1(\cdot)$	**135,0**
mêd(·)	**243,0**
$\hat{Q}_3(\cdot)$	**385,0**

Figure 153 : i.st.e* : $x_{(U_k)}^{TOPO}$: Niveaux altimétriques moyens communaux estimés par l'IGN en 2003

Géographie des expositions intrinsèques ou potentielles à la radioactivité environnementale.

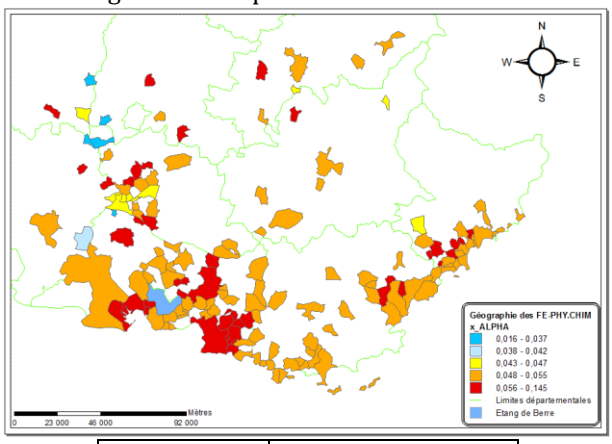
Variabilité spatiotemporelle des émissions globales des rayons γ dans l'air ambiant



Statistique	$x_{(U_k)}^{GAMMA}$(NanoSv/h)
môy(·)	98,7
$\hat{\sigma}(\cdot)$	23,7
$\hat{Q}_1(\cdot)$	79,9
mêd(·)	92,8
$\hat{Q}_3(\cdot)$	115,2

Figure 154 : i.st.e* : $x_{(U_k)}^{GAMMA}$Doses efficaces des rayonnements γ spatiotemporels émis par tous les radionucléides entre 2008 et 2010

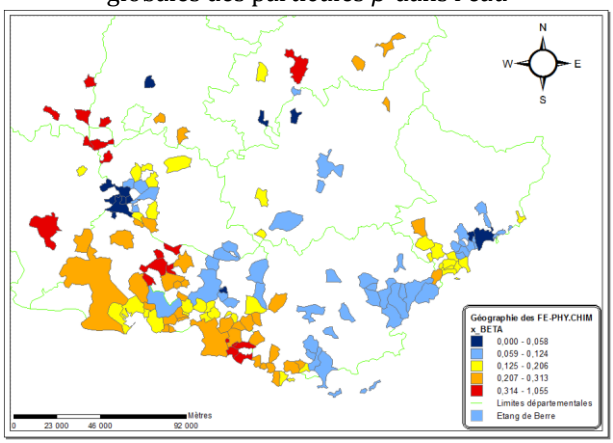
Variabilité spatiotemporelle des émissions radioactives globales des particules α dans l'eau



Statistique	$x_{(U_k)}^{ALPHA}$(Bq/litre)
môy(·)	0,05
$\hat{\sigma}(\cdot)$	0,013
$\hat{Q}_1(\cdot)$	0,05
mêd(·)	0,05
$\hat{Q}_3(\cdot)$	0,06

Figure 155 : i.st.e* : $x_{(U_k)}^{ALPHA}$: activités volumiques radioactives spatiotemporelles des émetteurs de particules α entre 2008 et 2010

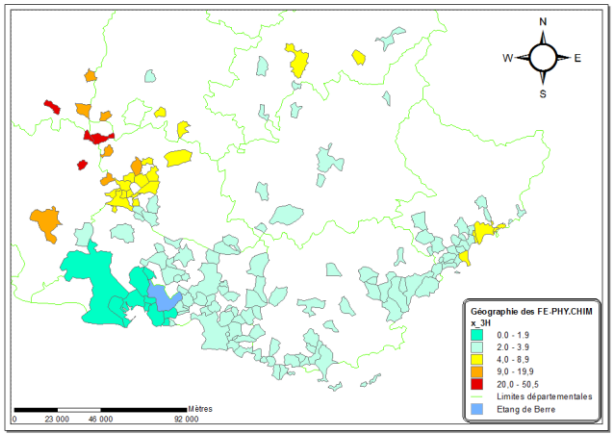
Variabilité spatiotemporelle des émissions radioactives globales des particules β dans l'eau



Statistique	$x_{(U_k)}^{BETA}$(Bq/litre)
môy(·)	0,25
$\hat{\sigma}(\cdot)$	0,155
$\hat{Q}_1(\cdot)$	0,14
mêd(·)	0,21
$\hat{Q}_3(\cdot)$	0,33

Figure 156 : i.st.e* : $x_{(U_k)}^{BÊTA}$: activités volumiques radioactives spatiotemporelles des émetteurs de particules β entre 2008 et 2010

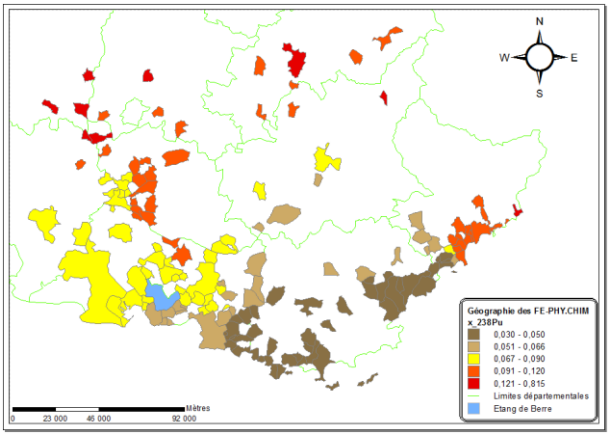
Variabilité spatiotemporelle des émissions radioactives totales du ^3H dans les eaux douces



Statistique	$x_{(U_k)}^{3H}$(Bq/litre)
môy(·)	12,82
$\hat{\sigma}(\cdot)$	17,90
$\hat{Q}_1(\cdot)$	2,68
mêd(·)	7,83
$\hat{Q}_3(\cdot)$	21,31

Figure 157 : i.st.e* : $x_{(U_k)}^{3H}$: activités volumiques radioactives spatiotemporelles des isotopes radioactifs du Tritium entre 2008 et 2010

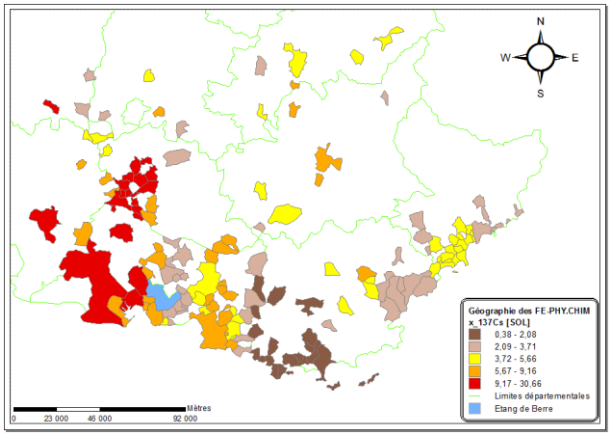
Variabilité spatiotemporelle des émissions radioactives du ^{238}Pu dans les sols fins



Statistique	$x_{(U_k)}^{238Pu}$(Bq/kg.mat.sec)
môy(·)	0,19
$\hat{\sigma}(\cdot)$	0,139
$\hat{Q}_1(\cdot)$	0,09
mêd(·)	0,15
$\hat{Q}_3(\cdot)$	0,24

Figure 158 : i.st.e* : $x_{(U_k)}^{238Pu}$: activités volumiques spatiotemporelles liées à la dégradation du Plutonium 238 entre 2008 et 2010

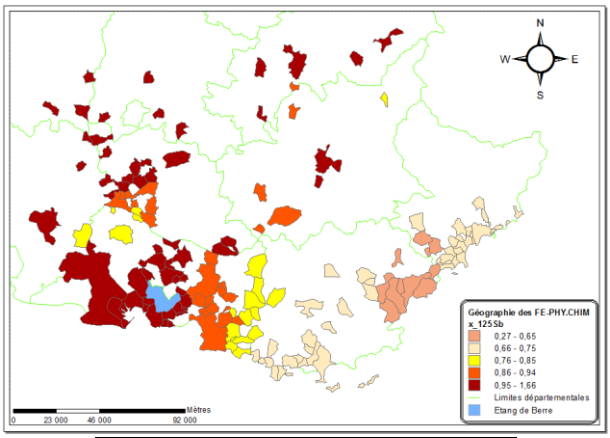
Variabilité spatiotemporelle des émissions radioactives du ^{137}Cs dans les sols fins



Statistique	$x_{(U_k)}^{137Cs-sol}$(Bq/kg.mat.sec)
môy(·)	8,09
$\hat{\sigma}(\cdot)$	4,65
$\hat{Q}_1(\cdot)$	4,10
mêd(·)	7,66
$\hat{Q}_3(\cdot)$	11,23

Figure 159 : i.st.e* : $x_{(U_k)}^{137Cs-sol}$: activités volumiques spatiotemporelles liées à la dégradation du Césium 137 entre 2008 et 2010

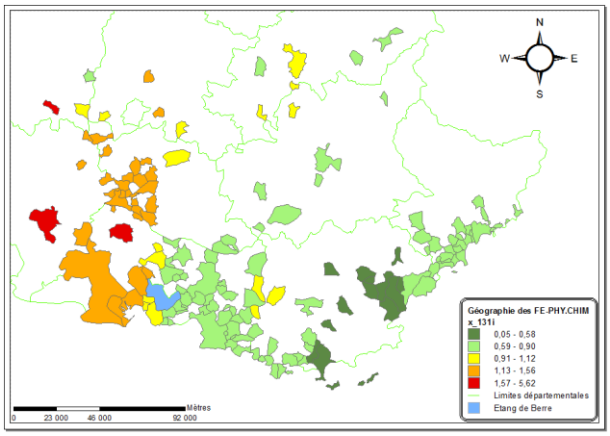
Variabilité spatiotemporelle des émissions radioactives du ^{125}Sb dans les sols fins



Statistique	$x_{(U_k)}^{125Sb}$(Bq/kg.mat.sec)
môy(·)	0,89
$\hat{\sigma}(\cdot)$	0,222
$\hat{Q}_1(\cdot)$	0,70
mêd(·)	0,88
$\hat{Q}_3(\cdot)$	1,00

Figure 160 : i.st.e* : $x_{(U_k)}^{125Sb}$: activités volumiques spatiotemporelles liées à la dégradation de l'Antimoine 125 entre 2008 et 2010

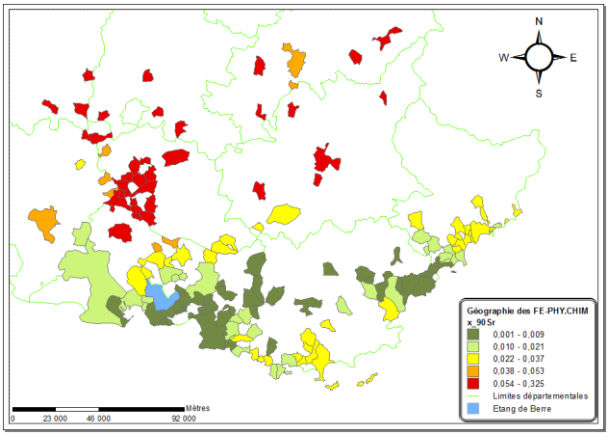
Variabilité spatiotemporelle des émissions radioactives du ^{131}i dans le lait



Statistique	$x_{(U_k)}^{131i}$(Bq/litre)
môy(·)	1,71
$\hat{\sigma}(\cdot)$	1,361
$\hat{Q}_1(\cdot)$	0,78
mêd(·)	1,20
$\hat{Q}_3(\cdot)$	2,02

Figure 161 : i.st.e* : $x_{(U_k)}^{131i}$: activités volumiques spatiotemporelles liées à la dégradation de l'Iode 131 entre 2008 et 2010

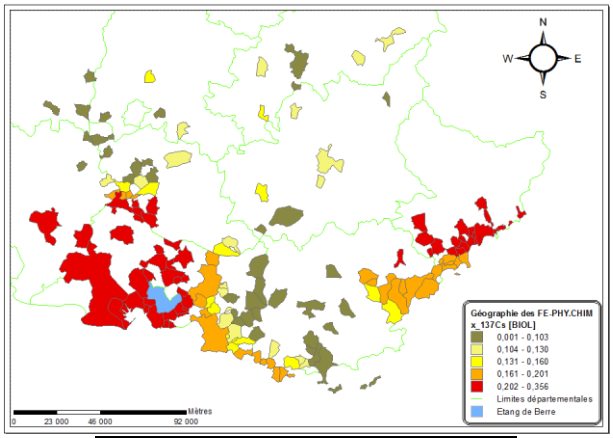
Variabilité spatiotemporelle des émissions radioactives du ^{90}Sr dans le lait



Statistique	$x_{(U_k)}^{90Sr}$(Bq/litre)
$\hat{m}oy(\cdot)$	0,10
$\hat{\sigma}(\cdot)$	0,077
$\hat{Q}_1(\cdot)$	0,03
$\hat{m}ed(\cdot)$	0,08
$\hat{Q}_3(\cdot)$	0,16

Figure 162 : i.st.e* : $x_{(U_k)}^{90Sr}$: activités volumiques spatiotemporelles liées à la dégradation du Strontium 90 entre 2008 et 2010

Variabilité spatiotemporelle des émissions radioactives du ^{137}Cs dans le lait

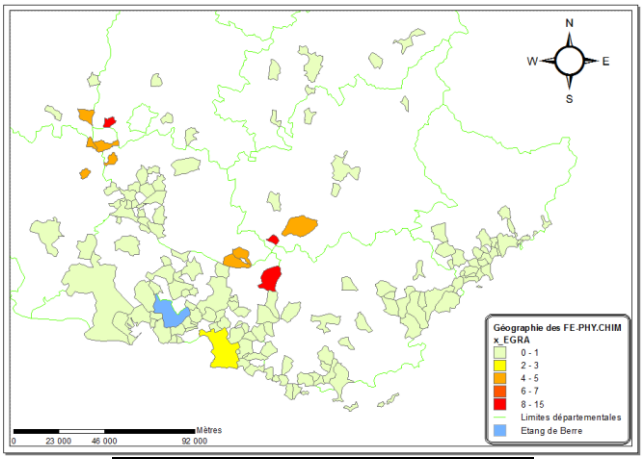


Statistique	$x_{(U_k)}^{137Cs-biol}$(Bq/litre)
$\hat{m}oy(\cdot)$	0,14
$\hat{\sigma}(\cdot)$	0,061
$\hat{Q}_1(\cdot)$	0,10
$\hat{m}ed(\cdot)$	0,13
$\hat{Q}_3(\cdot)$	0,17

Figure 163 : i.st.e* : $x_{(U_k)}^{90Sr}$: activités volumiques spatiotemporelles liées à la dégradation du Césium 137 entre 2008 et 2010

Géographie des expositions potentielles à la radioactivité liée à la présence d'INB.

Variabilité spatiotemporelle des Expositions Géographiques des Radionucléides Artificiels (EGRA)

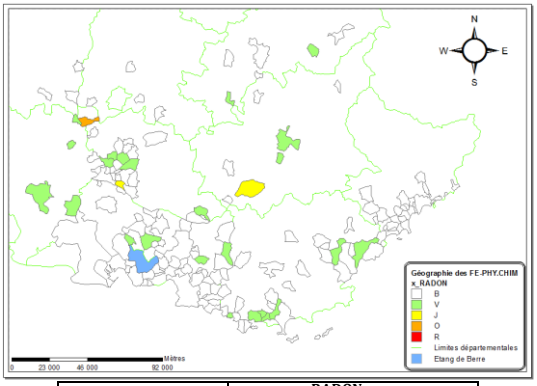


Statistique	$x_{(U_k)}^{EGRA}$(U/m)
$\hat{m}oy(\cdot)$	0,35
$\hat{\sigma}(\cdot)$	1,333
$\hat{Q}_1(\cdot)$	0,00
$\hat{m}ed(\cdot)$	0,00
$\hat{Q}_3(\cdot)$	0,00

Figure 164 : i.st.e* : $x_{(U_k)}^{EGRA}$: nombre spatiotemporel d'INB* ayant opéré une activité effective entre 1980 et 2010, captée par des zones dynamiques de capture en tâche d'huile, et pondérée par la valeur des rayons de dilatation itérée.

Géographie des expositions potentielles à la radioactivité tellurique.

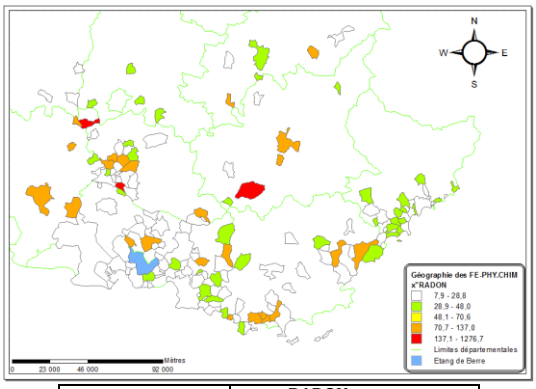
Variabilité de l'Activité Volumique du Radon dans les habitations



Statistique	$x^{RADON}_{(U_k)}$ (%)
$\mathbb{P}_{F_n}(x^I_{(.)} = B)$	66,7
$\mathbb{P}_{F_n}(x^I_{(.)} = V)$	20,0
$\mathbb{P}_{F_n}(x^I_{(.)} = J)$	8,8
$\mathbb{P}_{F_n}(x^I_{(.)} = O)$	3,6
$\mathbb{P}_{F_n}(x^I_{(.)} = R)$	1,0

Figure 165 : i.st.e* : $x^{RADON}_{(U_k)}$: Activité Volumique du Radon dans les habitations telle qu'elle est décrite par l'Atlas Radon.

Variabilité spatiotemporelle de l'Activité Volumique du Radon dans les habitations

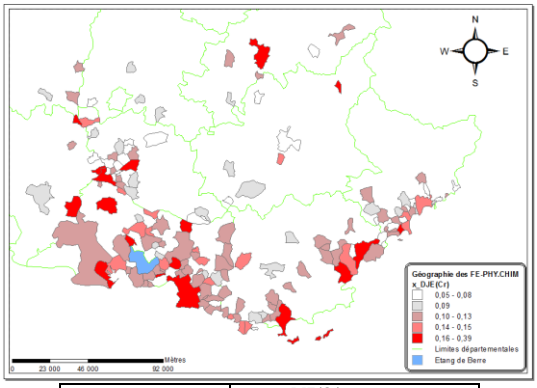


Statistique	$x''^{RADON}_{(U_k)}$ (Bq/m³)
môy(·)	66,5
$\widehat{\sigma}(\cdot)$	111,4
$\widehat{Q}_1(\cdot)$	27,00
mêd(·)	30,20
$\widehat{Q}_3(\cdot)$	73,90

Figure 166 : i.st.e* : $x''^{RADON}_{(U_k)}$ niveaux spatiotemporels de l'Activité Volumique du Radon dans les habitations entre 1982 et 2000.

Géographie des expositions intrinsèques aux Eléments Métalliques Traces.

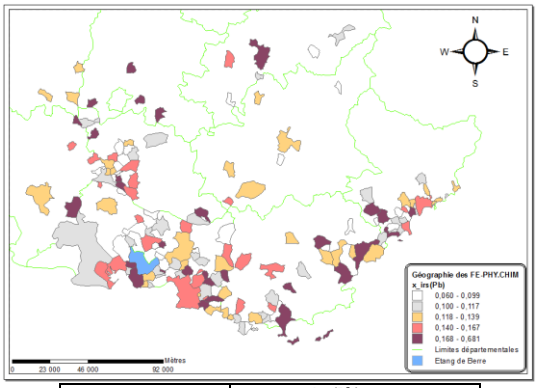
Variabilité spatiotemporelle des doses journalières de Chrome



Statistique	$x^{DJE(Cr)}_{(U_k)}$ (mg/kg)
môy(·)	0,11
$\widehat{\sigma}(\cdot)$	0,034
$\widehat{Q}_1(\cdot)$	0,09
mêd(·)	0,10
$\widehat{Q}_3(\cdot)$	0,12

Figure 167 : i.st.e* : $x^{DEJ(Cr)}_{(U_k)}$: Doses Journalières des Expositions spatiotemporelles liées à l'ingestion ou à l'inhalation de particules de Chrome (Cr), rapportées à la masse moyenne des mineurs, entre 1990 et 2009

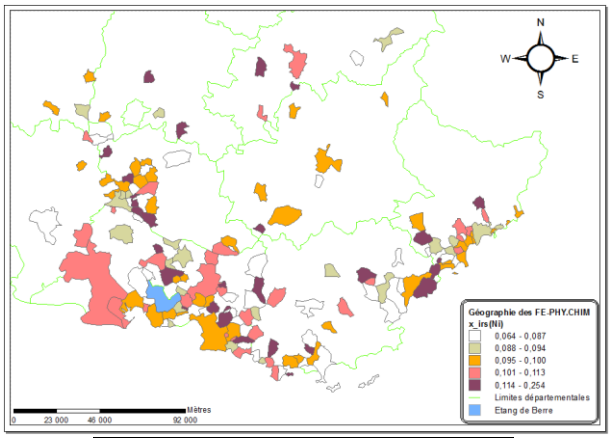
Variabilité spatiotemporelle des risques d'exposition au Plomb



Statistique	$x^{irs(Pb)}_{(U_k)}$ (SU)
môy(·)	0,14
$\widehat{\sigma}(\cdot)$	0,064
$\widehat{Q}_1(\cdot)$	0,10
mêd(·)	0,13
$\widehat{Q}_3(\cdot)$	0,16

Figure 168 i.st.e* : $x^{irs(Pb)}_{(U_k)}$: indicateur spatiotemporel du risque de contamination au plomb par ingestion ou par inhalation rapporté aux doses journalières maximales tolérées par des mineurs, entre 1990 et 2009

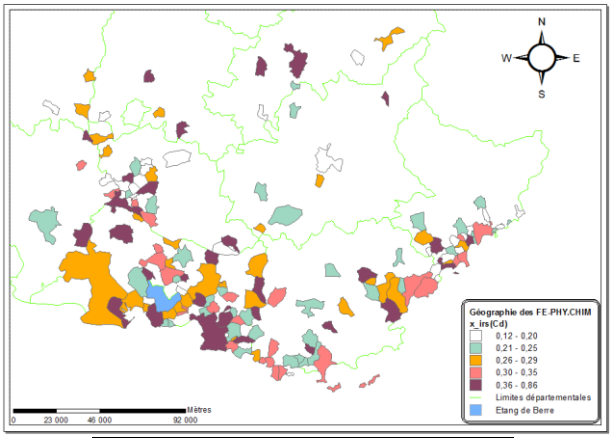
Variabilité spatiotemporelle des risques d'exposition au Nickel



Statistique	$x_{(U_k)}^{isr(Ni)}$ (SU)
$\hat{m}oy(\cdot)$	0,1
$\hat{\sigma}(\cdot)$	0,028
$\hat{Q}_1(\cdot)$	0,09
$\hat{m}ed(\cdot)$	0,10
$\hat{Q}_3(\cdot)$	0,11

Figure 169 : i.st.e* : $x_{(U_k)}^{irs(Ni)}$: indicateur spatiotemporel du risque de contamination au Nickel par ingestion ou par inhalation rapporté aux doses journalières maximales tolérées par des mineurs, entre 1990 et 2009

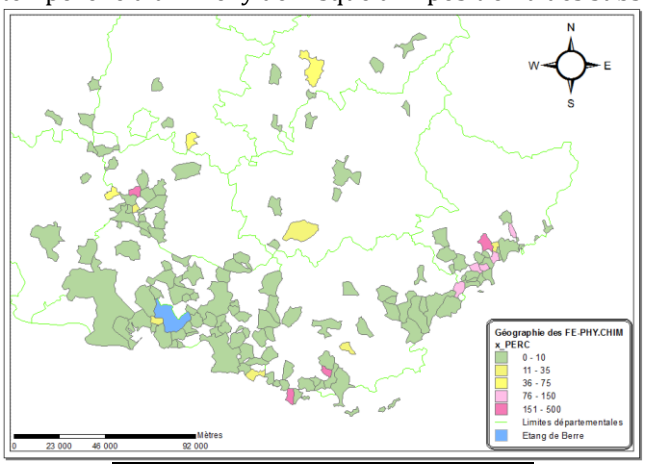
Variabilité spatiotemporelle des risques d'exposition au Cadmium



Statistique	$x_{(U_k)}^{isr(Cd)}$ (SU)
$\hat{m}oy(\cdot)$	0,28
$\hat{\sigma}(\cdot)$	0,105
$\hat{Q}_1(\cdot)$	0,21
$\hat{m}ed(\cdot)$	0,27
$\hat{Q}_3(\cdot)$	0,33

Figure 170 : i.st.e* : $x_{(U_k)}^{irs(Cd)}$: indicateur spatiotemporel du risque de contamination au Cadmium par ingestion ou par inhalation rapporté aux doses journalières maximales tolérées par des mineurs, entre 1990 et 2009

Variabilité spatiotemporelle d'un Proxy de Risque d'Exposition à des substances chimiques

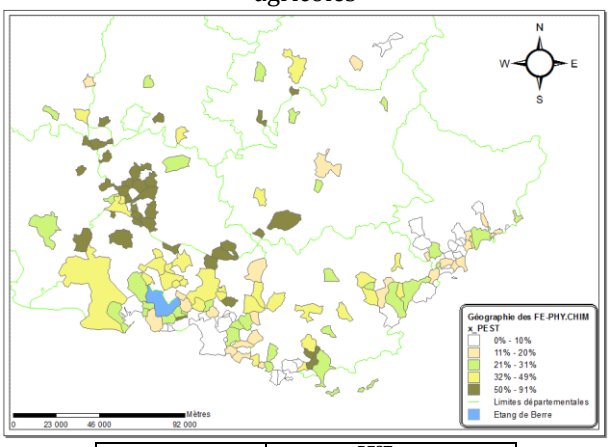


Statistique	$x_{(U_k)}^{PREC}$ (SU)
$\hat{m}oy(\cdot)$	357,33
$\hat{\sigma}(\cdot)$	2617,9
$\hat{Q}_1(\cdot)$	0,00
$\hat{m}ed(\cdot)$	0,00
$\hat{Q}_3(\cdot)$	0,02

Figure 171 : i.st.e* : $x_{(U_k)}^{PREC}$: Proxy spatiotemporel du Risque d'Exposition à des substances chimiques multiples liées à des activités industrielles polluantes, pour des mineurs, entre 1990 et 2009

Géographie des expositions potentielles à des substances nocives combinées.

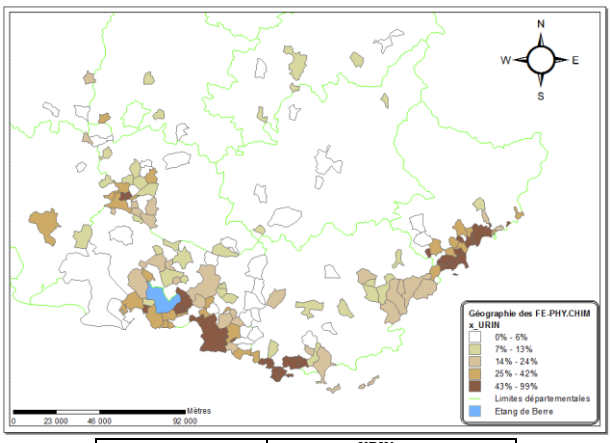
Variabilité spatiotemporelle de l'ampleur des zones agricoles



Statistique	$x_{(U_k)}^{PEST}$ (%)
$\hat{m}oy(\cdot)$	29,0
$\hat{\sigma}(\cdot)$	21,4
$\hat{Q}_1(\cdot)$	12,0
$\hat{m}ed(\cdot)$	25,0
$\hat{Q}_3(\cdot)$	44,0

Figure 172 : i.st.e* : $x_{(U_k)}^{PEST}$: proportions spatiotemporelles des surfaces communales où l'occupation biophysique des sols suggère l'utilisation récurrente de pesticides, entre 1990 et 2006.

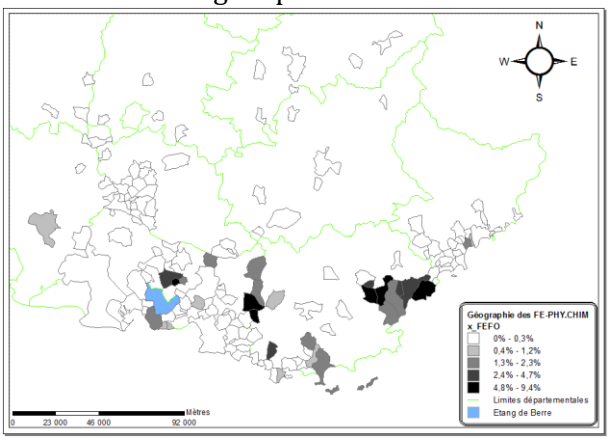
Variabilité spatiotemporelle de l'ampleur des zones urbanisées ou industrialisées



Statistique	$x_{(U_k)}^{URIN}$ (%)
$\hat{m}oy(\cdot)$	25,0
$\hat{\sigma}(\cdot)$	23,6
$\hat{Q}_1(\cdot)$	8,0
$\hat{m}ed(\cdot)$	17,0
$\hat{Q}_3(\cdot)$	36,0

Figure 173 : i.st.e* : $x_{(U_k)}^{URIN}$: proportions spatiotemporelles des surfaces communales où l'occupation biophysique des sols suggère des activités anthropiques polluantes intensives par leur degré d'urbanisation ou d'industrialisation, entre 1990 et 2006

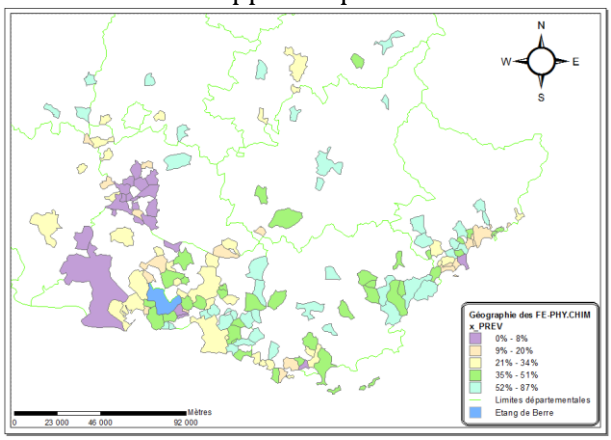
Variabilité spatiotemporelle de l'ampleur des zones ravagées par des feux de forêt



Statistique	$x_{(U_k)}^{FEFO}$ (%)
$\hat{m}oy(\cdot)$	0,0
$\hat{\sigma}(\cdot)$	0,9
$\hat{Q}_1(\cdot)$	0,0
$\hat{m}ed(\cdot)$	0,0
$\hat{Q}_3(\cdot)$	0,0

Figure 174 : i.st.e* : $x_{(U_k)}^{FEFO}$: proportions spatiotemporelles des surfaces communales où l'occupation biophysique des sols suggère des diffusions toxiques liées à des feux de forêt de grande ampleur, entre 1990 et 2006

Variabilité spatiotemporelle de l'ampleur des zones supposées préventives



Statistique	$x_{(U_k)}^{PREV}$ (%)
$\hat{m}oy(\cdot)$	30,0
$\hat{\sigma}(\cdot)$	21,40
$\hat{Q}_1(\cdot)$	10,0
$\hat{m}ed(\cdot)$	28,0
$\hat{Q}_3(\cdot)$	47,0

Figure 175 : i.st.e* : $x_{(U_k)}^{PREV}$: proportions spatiotemporelles des surfaces communales où l'occupation biophysique des sols ne suggère pas explicitement l'exposition à des substances toxiques, entre 1990 et 2006

REMARQUES

Les résultats cartographiques ainsi que les tableaux statistiques qui les complètent permettent de montrer que globalement, les $x_{(U_k)}^{l:PHY.CHIM}$ présentent de fortes variabilités spatiotemporelles. En effet, certains écarts-types prennent des valeurs très élevées et parfois largement supérieures à la moyenne spatiale, e.g. $x_{(U_k)}^{FEFO}$; $x''_{(U_k)}^{RADON}$; $x_{(U_k)}^{EGRA}$; $x_{(U_k)}^{3H}$.
La géographie des FE-PHY.CHIM* est modélisée par des variabilités spatiotemporelles parfaitement hétéroclites tant au niveau :

- Des distorsions *distances a-spatiales morbides*, notamment entre : Les *expositions intrinsèques* estimées à partir d'indices spatiaux de contamination organique par des ETM: $x_{(U_k)}^{PREC}$ ou par des doses efficaces de rayons γ : $x_{(U_k)}^{GAMMA}$; Et *les expositions potentielles* toxiques estimées à partir de l'occupation biophysique des sols et liées à de fortes densités urbaines et industrielles : $x_{(U_k)}^{URIN}$ ou à l'ampleur des zones agricoles $x_{(U_k)}^{PEST}$
- Que de la nature statistique des i.st.e* physicochimiques : Qualitative multi-classes - $x_{(U_k)}^{RADON}$; Quantitative discrète -$x_{(U_k)}^{EGAR}$; Quantitative continue et à valeur dans zéro et un - $x_{(U_k)}^{PREV}$; Ou quantitative continue pouvant prendre toutes sortes de valeurs - $x_{(U_k)}^{90Sr}$.

Les $x_{(U_k)}^{l}$ modélisent des variabilités spatiotemporelles d'expositions géographiques *potentielles* ou *curieuses* à des substances physicochimiques toxiques. Cependant il convient de remarquer que s'ils sont tous connotés *négativement*, ce n'est pas le cas de $x_{(U_k)}^{PREV}$. Cet i.st.e* est connoté positivement – ce qui rend son caractère de FE* *Curieux* de test* discutable.

Les cartographies des $x_{(U_k)}^{Météo-France}$ et $x_{(U_k)}^{RNM}$ affichent une variabilité spatiale de contiguïté qui peut sembler atténuée à l'aune de celle observée pour les autres FE-PHY.CIM. En effet, les reconstitutions géostatistiques par les techniques de KO et de CKO inscrivent l'hypothèse de stationnarité intrinsèque au cœur des processus de modélisation, ce qui engendre un lissage – particulièrement accentué dans les zones géographiques sous échantillonnées. Cette remarque concerne surtout les $x_{(U_k)}^{RNM}$. Quant aux contiguïtés spatiales apparentes des $x_{(U_k)}^{Météo-France}$ observables en PACA, il s'agit d'un biais d'affichage. Les variabilités spatiales communales de contiguïté sont masquées par des gradients géographiques Nord/Sud particulièrement forts. Ces remarques ne remettent nullement en question la capacité des $x_{(U_k)}^{Météo-France}$ et des $x_{(U_k)}^{RNM}$ à être éligibles au statut de DES.
Toutefois, la modélisation des variabilités spatiotemporelles à micro-échelle peut être améliorée pour les paramètres météo en augmentant le nombre de points de mesures – par une demande auprès de Météo-France. En revanche, ce n'est pas le cas pour la radioactivité environnementale, toutes les données RNM* ont été extraites.

L'activité volumique de certains radionucléides n'était pas mesurée en Corse. Sur l'île de beauté les variabilités spatiotemporelles des expositions potentielles au : Césium.137 ; Antimoine.125 et à l'ensemble des particules α et β dans l'eau, ne sont pas significatives.

Les structures spatiales d'échantillonnage : des doses efficaces de rayonnement global – RAY – (Météo-France), de l'activité volumique du Plutonium 238, et de l'Antimoine 125 (RNM), étaient défaillantes, les i.st.e* modélisant la géographie des FE-PHY.CHIM* correspondants ont été obtenus par un Co-Krigeage Ordinaire (CKO). Il est donc probable qu'ils soient auto-corrélés, peu ou prou, avec les variables auxiliaires utilisées – les vecteurs concernés sont : $\left(x_{(U_k)}^{RAY}; x_{(U_k)}^{TOPO}; x_{(U_k)}^{TEMP}\right)$; $\left(x_{(U_k)}^{238Pu}; x_{(U_k)}^{ALPHA}\right)$; $\left(x_{(U_k)}^{125Sb}; x_{(U_k)}^{3H}\right)$. D'ailleurs, certains de ces i.st.e* présentent, en PACA, des similarités visuelles parfois marquées (Saporta, 2006).

L'idée d'une redondance peut être étendue au couple $\left(x_{(U_k)}^{RADON}; x''_{(U_k)}^{RADON}\right)$. Mais la plausibilité de cette hypothèse reste très faible, d'une part parce qu'il n'y a pas de similitude spatiale prégnante, et d'autre part parce que le processus d'estimation de $x''_{(U_k)}^{RADON}$ est fondé sur la variabilité des extrema et des écarts-types départementaux qui est particulièrement forte. L'i.st.e* $x_{(U_k)}^{RADON}$ est exactement la variable déclinée par l'Atlas Radon. Elle est *imprécise* et peu consistante d'un point de vue spatiotemporel - son *biais conditionnel** n'a pas été minimisé.

La géographie des FE-PHY.CHIM* *Curieux* de test* est modélisée par les i.st.e* $x_{(U_k)}^{PREV}$ et $x_{(U_k)}^{FEFO}$, ils ne devraient jamais *a priori* être identifiés comme des DES. Ces i.st.e* ont pour objet de vérifier la capacité de VSURF à éliminer *les variables environnementales de bruit*.

L'i.st.e* $x_{(U_k)}^{RADON}$ est qualitatif multi-classes (4), il a été intégré pour vérifier les suspicions qui pèsent sur le fait que la procédure multidimensionnelle de caractérisation des FE* pourrait être biaisée en faveur de ce type de variable… *mais ceci n'est qu'une supposition* (Genuer, 2010).

Enfin, comme la stratégie VSURF - vouée à l'identification des DES* géographiques - est une procédure de dataminig : non paramétrique, randomisée et surtout éliminatrice - l'introduction des variables redondantes n'altère en rien sa puissance statistique. Au contraire, cela permet de tester sa capacité à disjoindre, dans le paquet de variables explicatives, celles qui sont justement redondantes dans le but de créer *un paquet qui s'adapte à la parcimonie prédictive* (Genuer, Poggi et al., 2013).

Les FE-PHY.CHIM* *pertinents* et Curieux** sont modélisés par des i.st.e* $x_{(U_k)}^{l:PHY.CIM}$. Ils caractérisent les espaces géographiques par la variabilité des expositions à *des contextes environnementaux à risques* ou à *des substances physicochimiques environnementales ayant un lien avéré ou suspecté avec les pathologies étudiées. Les i.st.e* sont supposés* robustes puisque leur *biais conditionnel** a été minimisé.

Il en est de même pour les *FE* pertinents* et Curieux** des autres composantes environnementales intégrées - dont les variabilités spatiotemporelles sont caractérisées par les i.st.e* $x_{(U_k)}^{l:SAN}$; $x_{(U_k)}^{l:SOCIO.ECO}$ et $x_{(U_k)}^{l:FIM}$. Il s'agit désormais d'identifier les DES* pour les PM* d'intérêt – *séquelles : CATA, THYR, TUM2* – ce qui est l'objet du chapitre 4.

SYNTHESE DU CHAPITRE 3

L' objectif de ce chapitre était de proposer des i.st.e* robustes $x^{l}_{(U_k)}$ i.e. capables de modéliser de façon précise et fiable, à l'échelle des U_k, la géographie de tous les FE* pertinents* ou de *Curieux** de sorte qu'ils soient adaptés à finalité recherchée, i.e à l'identification des DES* à partir des i.st.m* pondérés EpiGéoStat et intégrant le concept *du temps d'exposition à l'environnement* proposés : $z'^{j}_{(U_k),.}$.
La modélisation géographique des PM* étudiés permet aussi de caractériser les communes en fonction d'une typologie des Risques d'Expositions Géographiques (REG) - ce qui a permis de mettre en évidence des *clusters*, et par extension, de conjecturer un effet environnement évident, peut-être déterministe ou au moins contributif.

FE* pertinents* caractérisent les expositions environnementales à des situations contextuelles à risques ou à des substances physicochimiques toxiques qui ont des effets avérés ou très fortement suspectés sur les états de santé étudiés. Ils sont décrits comme tels chez les individus sains, et supposés, *a fortiori, dangereux* chez ceux prédisposés, puisqu'il n'existe aucune littérature sur les séquelles. Les *FE* Curieux* de test* sont construits à partir de données dont la *Distance a-spatiale morbide** est éloignée des pathologies généralement à cause d'une défaillance granulaire. Cependant ils restent pertinents au regard des connaissances bibliographiques. Aussi ils sont parfois intégrés dans l'optique de valider les résultats de la procédure de sélection des DES* – et devraient être identifiés comme du bruit environnemental.

Le Distance a-spatiale morbide* : ce concept caractérise à la fois la plausibilité du lien de causalité entre les FE* retenus et les PM* étudiés au vu des connaissances actuelles, et la qualité des données disponibles puisqu'elle conditionne fortement la modélisation géographiques.

Les variabilités des expositions environnementales modélisées sont soit qualifiées de potentielles, lorsqu'elles sont modélisées à partir de mesures environnementales ou d'indicateurs communautaires, soit d'intrinsèques, lorsqu'elles sont modélisées à partir de doses d'exposition ou de variables individuelles. Cette distinction permet de différencier les inputs et *leur Distance a-spatiale morbide** avec la population ciblée.
Le gain de précision apparent des expositions intrinsèques est maculé de biais et de bruits de fond qui, en pratique, ne les rendent pas plus robustes pour caractériser les états de santé.

La modélisation des FE* pertinents* et *Curieux** a pour objet d'intégrer horizontalement l'espace et verticalement la temporalité au cœur du processus de modélisation géographique. Les stratégies d'intégration sont élaborées pour **minimiser le concept de biais conditionnel*** des données mobilisées - afin d'obtenir des i.st.e* $x^{l}_{(U_k)}$ représentatifs des variabilités spatiotemporelles de FE/FIM* intégrables et de les adapter à l'identification des DES. Pour ce faire l'idée est :

D'optimiser l'effet information* par des processus de :

- Comblement des lacunes avec : d*es statistiques consistantes* estimées dans une logique : *individus-centrée** (LEA) ou *territoriale* - pour les lacunes induites par *le secret statistique* (INSEE)
- Transformations topologiques telles que des fusions d'informations spatiales, temporelles ou sémantiques - *dans l'idée de la théorie des ensembles flous* – et injections de connaissances expertes : épidémiologiques (LEA), sanitaires (DREES), géopolitiques (INSEE, ONDRP), ou par des : R*andomisation*s spatiales avec des données géographiques environnementales connexes (IRSN), méthodes d'analyse en statistiques descriptives multidimensionnelles (INSEE), reconstitutions spatiales d'informations géographiques discontinues par le biais de techniques géostatistiques :

Krigeage Ordinaire - KO ou Co-Krigeage Ordinaire - CKO (Météo-France ; RNM),_*captures d'entités géographiques* par des *dynamiques SIG de dilatation/rétraction (ASN)*, de *désagrégation raster d'entités vectorielles* (CLC), combinées à des *opérateurs ensemblistes de statistiques spatiales*.

Uniformisation de l'échelle par le biais de : *techniques d'analyse spatiale d'agrégation / désagrégation pondérées* dans une *logique territoriale* à partir des variables : administratives (DREES), géo-démographiques (IRSN) ou *géo-composites* et *conservatrices des propriétés granulaires* des sources (ONDRP ; INSEE), *métriques floues géographiques* à partir de *facteurs de certitude spatiotemporelle* υ_i^l permettant d'injecter *des données et des connaissances expertes* : Epidémiologiques, Géographiques ; Statistique (*EpiGéoStat*) dans le processus d'agrégation ensembliste des Caractéristiques Individuelles et Médicales*(CIM) des patients, dans lesU_k (LEA).

Maximiser l'effet de support* grâce à des processus d' :

- Harmonisation temporelle avec des *agrégations verticales* à partir de stratégies décisionnelles probabilistes fondées sur le concept de *stabilité temporelle apparente* et permettant de choisir une statistique consistante (*Météo-France ; ONDRP ; INSEE ; IRSN) ; Ou des estimateurs statistiques sans biais et pondérés par l'injection de croyances expertes (CLC)*, fusion d'informations spatiales, de données sémantiques connexes (ASN) ; de données géographiques randomisées pour pallier *l'inconsistance** de variables déclinées à une temporalité unique (DREES).
- Harmonisation spatiale grâce à une stratégie d'appariement adaptée à la logique de la variabilité diachronique des limites territoriales (INSEE ; DREES).

La mise en œuvre des propositions stratégiques heuristiques, des processus mathématiques d'intégration et des méthodes de modélisation utilisées a été illustrée par des applications. Elle nécessite trois logiciels : R, ArcGis.10 et VBA. Les FE/FIM* intégrés, les $x^l_{(U_k)}$ qui les caractérisent, et les traitements appliqués sont déclinés pour chaque composante environnementale.

Les Facteurs Individuels et Médicaux (FIM) :

Ils caractérisent la variabilité spatiotemporelle de la population ciblée, i.e. les patients spatialisés de la Cohorte LEA – par la géographie de leurs Caractéristiques Individuelles et Médicales*(CIM). L'omission des FIM* serait un biais à l'analyse géographique. Les i.st.e* sont construits à partir de données *individus-centrées** dont la *Distance a-spatiale morbide** avec les patients est plus proche que celle des variables utilisées pour modéliser les FE* - qui sont empreintes de bruits de fond environnementaux. Les facteurs de certitude EpiGéoStat ne pondèrent pas de manière aberrante les CIM. Ils permettent donc de modéliser de façon fiable la réalité géographique des FIM. Et en contrepartie, ils induisent *une distorsion a-spatiale morbide* afin de les adapter à la phase d'identification des DES.

FIM* PERTINENTS ET CURIEUX*			**PROCESSUS DE TRAITEMENT**	
Géographie des expositions	Inputs	Variabilité spatiotemporelle & i.st.e	Optimisation de l'effet information*	Maximisation de l'effet de support*
Intrinsèques liées aux Caractéristiques Individuelles et Médicales*(CIM) des patients	*LEA*	Genre : $x'^{SEXE}_{(U_k)}$	Comblement des lacunes Fusion topologique Uniformisation d'échelle	Aucun
		Leucémies traitées : $x'^{TYPLEUC}_{(U_k)}$		
		Age au diagnostic : $x'^{AGE\_DIAG}_{(U_k)}$		
		Durées de suivi : $x'^{DSUIVI}_{(U_k)}$		
		Rechutes : $x'^{RECHUT}_{(U_k)}$		
		Greffes : $x'^{GREF}_{(U_k)}$		
		Irradiations corporelles totales : $x'^{IRCAT}_{(U_k)}$		
		Protocoles de traitement : $x'^{PROTOC}_{(U_k)}$		
	LEA	Activités physiques : $x'^{ACPHY}_{(U_k)}$		

Tableau 33 : Synthèse des FIM* pertinents et Curieux* modélisés par des i.st.e* : $x^{l:FIM}_{(U_k)}$:

Remarques :
Le système de pondération EpiGéoStat reste subjectif. Par conséquent l'identification des DES* des FREC et des FREPA, sera abordée d'abord par le prisme de la géographie et ensuite, par celui de l'épidémiologie – i.e. directement sur l'état de santé des patients.
La géographie des FIM* pertinents et Curieux* est modélisée essentiellement par des i.st.m* qualitatifs qui mettent en exergue, par nature, des disparités géographiques.
La géographie des patients greffés présente des analogies avec celle de ceux ayant subi des irradiations corporelles totales.
Enfin il existe une composante géographique manifeste dans le type de protocole reçu. Mais ce n'est pas un problème d'accès au traitement de référence – la nature et le dosage des principes actifs sont à peu près identiques. De fait, $x'^{PROTOC}_{(U_k)}$ ne devrait pas avoir d'influence...

Les Facteurs Sanitaires (FE-SAN) :
Ils caractérisent les espaces géographiques par la qualité des tissus sanitaires* territoriaux. Les FE-SAN* constituent la *dimension spatiale* de l'accès géographique aux soins. Les i.st.e* SAN proposés modélisent la variabilité spatiotemporelle des expositions potentielles à des situations à risque liées à l'accès aux items sanitaires*, i.e. aux praticiens de santé libéraux et aux plateaux techniques des établissements de santé : services médicaux et Equipements Matériels Lourds (EML*). Les effets des FE-SAN* sur l'incidence et la gravité des séquelles étudiées sont très suspectés par le Professeur Auquier et le Professeur Michel, les responsables de l'étude LEA. Les modélisations géographiques sont effectuées à partir des Distances Temporelles d'Accès (DTA) en 2007 aux items sanitaires* et des Accès Potentiels Localisés (APL) en 2010 aux praticiens libéraux.

FE-SAN* PERTINENTS ET CURIEUX*			**PROCESSUS DE TRAITEMENT**	
Géographie des expositions	Inputs	Variabilité spatiotemporelle & i.st.e	Optimisation de l'effet information*	Maximisation de l'effet de support*
Potentielles liées à l'accessibilité aux items sanitaires* territoriaux	*DREES*	Temps d'accès aux généralistes : $x'^{GENE}_{(U_k)}$	Transformation topologique Uniformisation de l'échelle	Harmonisation spatiale
		Temps d'accès aux ophtalmologues : $x'^{OPHT}_{(U_k)}$		
		Temps d'accès aux ORL : $x'^{ORL}_{(U_k)}$		
		Temps d'accès aux pédiatres : $x'^{PEDIA}_{(U_k)}$		
		Temps d'accès aux radiologues : $x'^{RADIO}_{(U_k)}$		
		Temps d'accès : service neurologie : $x'^{NEUROs}_{(U_k)}$		
		Temps d'accès : service ORL : $x'^{ORLs}_{(U_k)}$		
		Temps d'accès : service endocrinologie : $x'^{ENDOs}_{(U_k)}$		
		Temps d'accès : service hématologie : $x'^{HEMAs}_{(U_k)}$		
		Temps d'accès : service ophtalmologie : $x'^{OPTHs}_{(U_k)}$		
		Temps d'accès à un TEP : $x'^{TEP}_{(U_k)}$		
		Temps d'accès à un Scanner* : $x'^{SCAN}_{(U_k)}$		
		Temps d'accès à une caméra à scintillation $x'^{CAME}_{(U_k)}$		
		Temps d'accès à IRM : $x'^{IRM}_{(U_k)}$		
		Accès Potentiel Localisé* - généralistes $x^{APL\_GEN}_{(U_k)}$	Aucun	Harmonisation spatiale
		l'Accès Potentiel Localisé* - ophtalmologues $x^{APL\_OPHT}_{(U_k)}$		

Tableau 34 : Synthèse des FE-SAN* pertinents et Curieux* modélisés par des i.st.e* : $x^{I:SAN}_{(U_k)}$:

Remarques :
Très peu d'indicateurs géographiques sanitaires sont disponibles en France. Par conséquent aucun FE-SAN* *Curieux* de test* n'a été intégré. Cependant, des i.st.e* qui n'ont *a priori* aucun lien avec les PM* d'intérêt seront injectés à ce titre dans la procédure d'identification des DES* des FREC et des FREPA ; Par exemple les $x'^{I:ORL}_{(U_k)}$ pour la séquelle cataractes (CATA).
Les $x'^{I:SAN}_{(U_k)}$ estimés à partir des DTA 2007 aux plateaux techniques des établissements de santé mettent en exergue de fortes disparités géographiques. Ils sont donc *a priori* robustes. Cependant, ils se

montrent défaillants pour caractériser les variabilités spatiotemporelles d'accès à des praticiens libéraux - excepté pour les radiologues.
En revanche, les $x'^{l:SAN}_{(U_k)}$ construits à partir des indicateurs APL 2010 modélisent parfaitement les disparités géographiques d'accès aux généralistes et aux ophtalmologues. Il est regrettable que ces indicateurs soient l'apanage de ces professions libérales, et que les DTA robustes ainsi que les APL ne soient disponibles qu'à une date unique.

Les Facteurs Socio-Economiques (FE-SOCIO.ECO)

Ces facteurs modélisent la géographie des contextes sociaux, économiques et démographiques qui constituent la dimension *a-spatiale* de l'accès aux soins. Ils ont un impact sur les milieux de vie communautaire, donc des répercussions sur *les capacités collectives* et *les conduites individuelles* vis-à-vis du recours aux soins. Les i.st.e* proposés caractérisent les variabilités spatiotemporelles des conjonctures socio-économiques qui, lorsqu'elles sont défavorables, induisent des tendances contextuelles morbides. En épidémiologie et en géographie de la santé, *l'effet de contexte* se modélise, à l'échelle des territoires, grâce à des statistiques géographiques étatiques.

FE-SOCIO.ECO* PERTINENTS ET CURIEUX*			**PROCESSUS DE TRAITEMENT**	
Géographie des expositions	Inputs	Variabilité spatiotemporelle des niveaux de ; i.st.e	Optimisation de l'effet information*	Maximisation de l'effet de support*
Potentielles des comportements à risques vis-à-vis du recours à l'offre de soins	INSEE	Niveaux de Vie : $x^{tx.FoyFisc}_{(U_k)}$	Comblement des lacunes Transformation topologique Uniformisation de l'échelle	Harmonisation temporelle Harmonisation spatiale
		Revenus fiscaux $x^{RevMed.p/m}_{(U_k)}$		
		Répartition des richesses $x^{GINI.p/m}_{(U_k)}$		
		Précarité socioprofessionnelle : $x^{tx.CHOM}_{(U_k)}$		
		Catégories socio-professionnelles $x^{tx.OUV}_{(U_k)}$		
Potentielles de la qualité des politiques de durabilité et de leurs répercussions sur les attraits des territoires	INSEE	Mortalité : $x^{tx.MORT}_{(U_k)}$		
		Accroissements naturels : $x^{txAccNAT}_{(U_k)}$		
		Accroissements démographiques : $x^{txAccPOP}_{(U_k)}$		
		Niveaux culturels : $x^{tx.BAC}_{(U_k)}$		
Potentielles à des substances toxiques socioprofessionnelles liées à la spécialisation socio-économique des territoires	INSEE	Emplois A Risque : $x^{tx.EAR}_{(U_k)}$		
		Surfaces Agricoles Utilisées : $x^{tx.SAU}_{(U_k)}$		
		Intensité des activités agricoles : $x^{int.AGRI}_{(U_k)}$		
Potentielles liées à des contextes conjoncturels induisant des prédispositions morbides	INSEE	Défaveur sociale 99 : $x^{FDep99}_{(U_k)}$		
		Défaveur sociale 09 : $x^{FDep09}_{(U_k)}$		
		Probables de la défaveur sociale temporelle $x^{FDepXX}_{(U_k)}$		
Potentielle au stress perçu - induit par l'insécurité territoriale contextuelle	ONDRP	atteintes aux biens matériels : $x^{att.BIENS}_{(U_k)}$	Transformations topologiques Uniformisation de l'échelle	Harmonisation temporelle
		atteintes à l'intégrité physique : $x^{att.PHY}_{(U_k)}$		
		insécurité composite territoriale d'infractions multiples : $x^{INSECU}_{(U_k)}$		

Tableau 35 : Synthèse des FE-SOCIO.ECO* pertinents et Curieux* modélisés par des i.st.e* : $x^{l:SOCIO.ECO}_{(U_k)}$

Remarques :

Globalement les i.st.e* proposés modélisent des phénomènes conjoncturels de nature très hétéroclite et qui présentent de fortes disparités géographiques. En effet, ce constat est particulièrement bien illustré par les variabilités spatiotemporelles mises en exergue par : $x^{INSECU}_{(U_k)}$, $x^{att.PHY}_{(U_k)}$; $x^{FDep09}_{(U_k)}$; $x^{RevMed.p/m}_{(U_k)}$ et $x^{tx.EAR}_{(U_k)}$.

Certains i.st.e* SOCIO.ECO sont redondants, en particulier ceux qui modélisent la géographie de la défaveur sociale et des niveaux de stress liés à l'insécurité territoriale. Des stratégies de fusion auraient pu être opportunes. Cependant, l'intégration d'i.st.e* redondants présente aussi l'avantage d'offrir un

moyen de tester la capacité de la procédure de sélection des variables à disjoindre, parmi les variables explicatives, celles qui ont une efficience forte de celles qui ont un impact plus faible sur les PM. En effet, la méthode proposée isole un paquet de variables adaptées à la parcimonie prédictive qui s'opère au regard des redondances (chapitre.4).

Les Facteurs Physicochimiques (FE-PHY.CHIM) :

Ils modélisent la géographie des expositions *potentielles ou intrinsèques* mais *chroniques* à des substances physiques ou chimiques toxiques et présentes en quantité variable, souvent faible, dans les *milieux environnementaux* ou dans *les milieux de contact*. L'effet des FE-PHY.CHIM* en Santé Environnementale est encore controversé. Cependant, grâce à la puissance des modèles statistiques et l'émergence de BD environnementales protéiformes, un regain d'intérêt est observé à ce sujet. La plausibilité des effets combinés déterministes, ou au moins contributifs, sur l'état de santé des populations ne saurait être négligée en géographie de la santé.

FE-PHY.CHIM* PERTINENTS ET CURIEUX*			**PROCESSUS DE TRAITEMENT**	
Géographie des expositions	Inputs	Variabilité spatiotemporelle & i.st.e	Optimisation de l'effet information*	Maximisation de l'effet de support*
Potentielles ou intrinsèques aux paramètres géophysiques	Météo-France	Rayonnement global : $x_{(U_k)}^{RAY}$	Transformation topologique	Harmonisation temporelle
		Températures : $x_{(U_k)}^{TEMP}$		
		Précipitations : $x_{(U_k)}^{NJAP}$		
	Géofla	Altimétriques : $x_{(U_k)}^{TOPO}$	Aucun	Aucun
Potentielles ou intrinsèques à la radioactivité environnementale	RNM	Rayons γ (air) : $x_{(U_k)}^{GAMMA}$	Transformation topologique	Harmonisation temporelle
		Particules α (eau) : $x_{(U_k)}^{ALPHA}$)		
		Isotopes radioactifs du Tritium (eau) : $x_{(U_k)}^{3H}$		
		Plutonium 238 (sol) : $x_{(U_k)}^{238Pu}$		
		Césium 137 (sol) : $x_{(U_k)}^{137Cs-sol}$		
		Antimoine 125 (sol) : $x_{(U_k)}^{125Sb}$		
		Strontium 90 (biol-lait):$x_{(U_k)}^{90Sr}$		
		Cesium137 (biol-lait)$x_{(U_k)}^{137Cs-biol}$		
		Iode 131 (biol: lait) : $x_{(U_k)}^{131i}$		
Potentielles à la Radioactivité liée à la proximité d'INB	ASN	Expositions Géographiques à des Radionucléides Artificiels : $x_{(U_k)}^{EGRA}$	Transformation topologique	Harmonisation temporelle
Potentielles et domestiques à la Radioactivité tellurique	IRSN	Au Radon $x_{(U_k)}^{RADON}$	Aucun	Aucun
		Spatiotemporelle du Radon $x''^{RADON}_{(U_k)}$	Transformation topologique Uniformisation d'échelle	Aucun
Intrinsèques aux Eléments Métalliques Traces (ETM)	INERIS	Doses journalières de Chrome $x_{(U_k)}^{DEJ(Cr)}$	Aucun	Aucun
		Risques d'exposition au Plomb $x_{(U_k)}^{irs(Pb)}$		
		Risques d'exposition au Nickel $x_{(U_k)}^{irs(Ni)}$		
		Risques d'exposition au Cadmium $x_{(U_k)}^{irs(Cd)}$		
		Proxy de Risque d'Exposition à des substances chimiques $x_{(U_k)}^{PREC}$		
Potentielles à des substances nocives combinées	CLC	zones agricoles $x_{(U_k)}^{PEST}$	Transformation topologique	Harmonisation temporelle
		zones urbanisées ou industrialisées $x_{(U_k)}^{URIN}$		
	CLC	zones incendiées par des feux de forêt $x_{(U_k)}^{FEFO}$		
		zones supposées préventives $x_{(U_k)}^{PREV}$		

Tableau 36 : Synthèse des FE-PHY.CHIM* pertinents et Curieux* modélisés par des i.st.e* : $x_{(U_k)}^{i:PHY.CHIM}$

Remarques :
La géographie des FE-PHY.CHIM* est modélisée par des i.st.e* hétéroclites tant au niveau des *distances a-spatiales morbides* entre *les faits de santé* intégrés et les *états de santé* étudiés, qu'à celui de la nature statistique des $x_{(U_k)}^{l:PHY.CHIM}$ qualitative multi-classes, quantitative discrète, continue bornée ou pas.
Le *biais conditionnel** de l'i.st.e* : $x_{(U_k)}^{RADON}$ n'a pas été minimisé. De plus, sa nature qualitative multi-classes lui confère la capacité de tester la robustesse de la procédure d'identification des DES. Il pourrait donc caractériser un FE* *Curieux*.*
Tous les $x_{(U_k)}^{l}$, quelle que soit la composante environnementale, modélisent des expositions géographiques potentiellement dangereuses à l'exception de $x_{(U_k)}^{PREV}$, qui a une connotation positive.
La majorité $x_{(U_k)}^{l:PHY.CHIM}$ proposés met en évidence de fortes variabilités spatiotemporelles, en particulier : $x_{(U_k)}^{FEFO}$; $x''_{(U_k)}^{RADON}$; $x_{(U_k)}^{EGRA}$; $x_{(U_k)}^{3H}$; $x_{(U_k)}^{238Pu}$; $x_{(U_k)}^{isr(Pb)}$; $x_{(U_k)}^{URIN}$.
En revanche, les variabilités spatiales de contiguïté sont parfois atténuées pour les i.st.e* obtenus à partir de reconstitutions géostatistiques : $x_{(U_k)}^{Météo-France}$ et $x_{(U_k)}^{RNM}$. Les techniques de KO et de CKO supposent une stationnarité intrinsèque et engendrent un lissage exagéré dans les zones géographiques sous échantillonnées. Des modélisations plus fines sont envisageables pour les paramètres météo en augmentant la densité d'échantillonnage, mais ce n'est pas le cas pour les $x_{(U_k)}^{RNM}$ - toutes les variables RNM* disponibles ont été intégrées.
La structure spatiale d'échantillonnage de défaillance de certaines mesures RNM* rend inopérants les i.st.e* : $x_{(U_k)}^{137Cs[sol]}$; $x_{(U_k)}^{125Sb}$; $x_{(U_k)}^{ALPHA}$; $x_{(U_k)}^{BETA}$, en Corse uniquement.
Enfin, les i.st.e* $\left(x_{(U_k)}^{RAY}; x_{(U_k)}^{TOPO}; x_{(U_k)}^{TEMP}\right)$; $\left(x_{(U_k)}^{238Pu}; x_{(U_k)}^{ALPHA}\right)$; $\left(x_{(U_k)}^{125Sb}; x_{(U_k)}^{3H}\right)$ sont très probablement auto-corrélés puisque le premier, des trois vecteurs colonnes cités, est construit par CKO à partir du ou des autres variables auxiliaires. Cependant, l'intégration d'i.st.e* corrélés n'altère pas la puissance de la procédure d'identification des DES, des FREC et des FREPA (chapitre 4). Au contraire, elle permet de tester sa capacité à disjoindre les FE* de bruit des FE* déterminants. Et conséquemment, de hiérarchiser l'efficience des FE* justement au regard des redondances.

CONCLUSION DU CHAPITRE 3

La géographie des FE/FIM* pertinents et Curieux* est désormais modélisée par des i.st.e* supposés *robust*es – puisqu'ils sont fondés sur le concept de minimisation du biais conditionnel*. Cependant les i.st.e* proposés peuvent parfois modéliser des phénomènes géographiques radicalement différents de ceux qu'ils sont sensés caractériser. Ce constat est fréquent, il est décrit en géographie de la santé et en épidémiologie spatiale.
La géographie des FE/FIM* ne donne pas lieu à des commentaires approfondis. L'analyse des interactions statistiques n'est envisageable qu'en mettant en perspective simultanément tous les i.st.e. avec les i.st.m* proposés. Le caractère multidimensionnel ainsi que les phénomènes aléatoires qui maculent ces interactions induisent une complexité top grande pour que les instruments d'analyse spatiale classiques puissent caractériser leurs interactions. Pour répondre à la problématique posée, une méthode de *sélection de variables adaptée aux jeux de données* géographiques *de grande dimension* est requise. Ceci est justement l'objet du chapitre 4.
L'identification des DES, des FREC et des FREPA est fondamentale sur le plan humain. Elle permet de caractériser l'état de santé des populations par des indicateurs spatiaux représentatifs de la qualité environnementale des territoires. Et, par suite de concevoir, mettre en place et justifier socialement les mesures financières politiques prises pour réduire les inégalités géographiques d'accès à une bonne santé environnementale*.

CHAPITRE 4 : IDENTIFICATION DE FACTEURS ENVIRONNEMENTAUX GEOGRAPHIQUES EXPLICATIFS DES ETATS DE SANTE

La géographie des PM* étudiés et des FE/FIM* pertinents* et curieux* intégrés est modélisée respectivement par : des i.st.m* quantitatifs $z'^{j}_{(Uk),c}$ ou qualitatifs $z'^{j}_{(Uk),q}$, et des i.st.e* multidimensionnels $x^{l}_{(Uk)}$. Ils sont estimés à partir de stratégies et de méthodes *a priori* consistantes d'un point de vue spatiotemporel (chapitre.2), (chapitre.3).

Le chapitre 4 est dévolu à l'application d'une méthode de *sélection de variables* permettant de caractériser les interactions statistiques entre les i.st.m* et les i.st.e* afin d'identifier des Déterminants Environnementaux de Santé* (DES). On distinguera les DES* des Facteurs Environnementaux* (FE) de Risques Contributifs (FREC) et des Facteurs Environnementaux* (FE) de Risques Probablement Aggravant (FREPA). La méthode sur laquelle se fondent les propositions heuristiques se nomme : *Variable Selection Using Random Forest* (VSURF). Elle permet de disjoindre, dans des jeux de données multidimensionnelles, les variables de bruit de celles qui permettent d'expliquer, puis de prédire, un phénomène observé.

La section A dresse un état des connaissances, du langage, des méthodes, des outils de datamining, des stratégies dédiées *à la sélection de variables*, et spécifie les caractéristiques de l'algorithme choisi - *randomForest** (Breiman, 2001) - ainsi que celles de VSURF dont la version bêta a récemment été mise à disposition (Genuer, Poggi et al., 2013) – *mais qui ne l'était pas au moment où les traitements ont été effectués*. De plus, le package R VSURF ne permet pas de calibrer l'algorithme et les items renvoyés ne sont pas adaptés à l'analyse géographique des états de santé étudiés. La stratégie a été adaptée à la dialectique dans un algorithme nommé MyVsurfGéo* (MVG) -testé et validé en collaboration avec l'inventeur de VSURF (Genuer, 2013). Enfin, une application à des jeux de données jouées permet d'illustrer la dialectique proposée et les résultats renvoyés par MVG. Les parties : théories et application, sont nécessaires à la compréhension des autres sections.

La section B se compose d'un rappel théorique, des notations utilisées, et d'une stratégie d'interprétation des résultats renvoyés par MVG. Ce dernier est ensuite appliqué aux jeux de données géographiques confectionnés afin d'identifier les i.st.e* représentatifs de la géographie des FE/FIM* intégrés qui permettent d'expliquer et de prédire la géographie des séquelles : CATA, THYR, TUM2. Ensuite, une évaluation de la robustesse des prédictions et une analyse des DES*, des FREC et des FREPA identifiés sont proposées.

La section C se décompose en deux parties. La première a pour objectif de valider les DES, les FREC et les FREPA identifiés pour chaque séquelle, par une approche *individus-centrée**. Une évolution heuristique de MVG est proposée à cet effet : BoostMyVsurfGéo* (BVMG). Ce nouvel algorithme intègre un processus de boosting* randomisé géographiquement, permettant d'augmenter la puissance de MVG et de l'adapter à complexité des jeux de validation utilisés, dans le cadre de cette approche à mi-chemin entre l'épidémiologie et la géographie. La seconde partie est le corollaire et la conséquence de la première. Elle a pour dessein de caractériser les espaces par des niveaux de Risques d'Expositions Géographiques (REG) morbides à partir des connaissances spatiales acquises sur les DES, FREC et FREPA. La qualité des REG prédits est évaluée par comparaison aux REG estimés sur les états de santé observés (chapitre 2).

SECTION A) LA SELECTION DE VARIABLES APPLIQUEE A LA GEOGRAPHIE DE LA SANTE

Cadre théorique

Les cancers sont des maladies au long sillage. Les effets secondaires des traitements ont des conséquences sur l'état de santé des patients mais ne suffisent pas à les expliquer. Et l'environnement exerce sans doute une influence. La caractérisation des interactions spatiales entre les états de santé et l'environnement est complexe mais le géographe a pour devoir de s'y intéresser (Peguy, 1996). En géographie, *la complexité* définit les phénomènes spatiotemporels qui présentent des formes *a priori* aléatoires. Cependant, *la caractéristique fondamentale des systèmes complexes est de présenter des similitudes spatiales* dès lors qu'ils sont considérés à des échelles macroscopiques (Pumain, 2004).

Les i.st.m* $z^{j}_{(U_k)}$ proposés pour modéliser la géographie des séquelles, comme les soixante-dix i.st.e* $x^{l}_{(U_k)}$ modélisant celle des FE/FIM, ont été estimés dans cette logique. Ils modélisent des caractéristiques morbides et environnementales dans un *espace géographique* numérique multidimensionnel,* constitué d'*objets - ou agrégats spatiotemporels* (Levy J., Lussault M., 2003).

Les outils statistiques d'analyse spatiale, implémentés dans les SIG (ESRI, 2013) sont indigents pour lever le voile de la complexité multidimensionnelle des jeux caractérisant ces espaces. Pourtant, les modèles statistiques et probabilistes restent *a priori* les meilleurs outils *d'exploration de la complexité spatiale* (Charre, 1995). Le cadre mathématique est celui de la caractérisation des variables dans *des jeux en grandes dimensions* (Tuleau-Malot, 2005). A l'heure actuelle seules certaines méthodes de datamining* basées sur *du scoring* permettent, à leur façon, de démêler le problème (Ghattas et Ben Ishak, 2008).

Objectif

Proposer un algorithme de datamining* et une méthode de sélection de variables - *quitte à en modifier les contours* – afin d'identifier parmi tous les FE/FIM* intégrés ceux qui déterminent la géographie des PM* étudiés. Il s'agit de sélectionner un ensemble d'i.st.e* $x^{l}_{(U_k)}$ contenant une quantité d'informations suffisante pour expliquer la variabilité des $z^{j}_{(U_k)}$ - en dépit de la complexité de ce type de jeux de données, qui tendent à se généraliser en géographie de la santé.

ETAT DE L'ART SUR L'APPRENTISSAGE STATISTIQUE ET CHOIX D'UNE METHODE ADAPTEE AUX DONNEES GEOGRAPHIQUES

Le datamining* ou *Machine Learning** est constitué d'un ensemble de méthodes informatiques différentielles fondées sur des algorithmes de modélisation statistique et probabiliste, adaptés à des *jeux de données d'apprentissage caractérisant des phénomènes complexes* (Han et Kamber, 2006).

ETAT DES CONNAISSANCES ET NOTATIONS VERNACULAIRES

Etat des connaissances sur le datamining

Autrefois l'analyse statistique s'appliquait à des jeux de données constitués de quelques centaines d'individus dont les caractéristiques étaient recueillies avec des protocoles spécifiques et normés. Les modèles utilisés se fondaient sur des hypothèses probabilistes fortes et restrictives. De fait, la compression du phénomène et les prédictions effectuées étaient limitées à la validité des hypothèses et ne pouvaient que porter sur un nombre restreint de données. Avec l'avènement des ordinateurs les premières analyses numériques ont permis de travailler sur des jeux de données contenant quelques centaines d'individus contenus dans des tableaux figés de type : *individus x variables*. La capacité prédictive des modèles était améliorée mais la compréhension des phénomènes étudiés restait

discutable. Les progrès scientifiques et techniques en informatique, en statistique et en probabilité ont engendré le datamining*. Désormais, il est possible de traiter des jeux de données contenant plusieurs millions d'individus décrits par des milliers de variables de nature quantitative, qualitative et parfois même textuelle. Les données peuvent être recueillies *ex-post* ou être intégrées *ex-ante*. Les protocoles de collecte sont simplifiés. Certains modèles offrent l'opportunité de mélanger des données hétéroclites - parfois initialement destinées à d'autres fins. D'autres algorithmes traitent les données imparfaites, contenant des erreurs de saisie, des lacunes, des valeurs aberrantes… Certaines permettent de travailler sur des populations en perpétuelle évolution, par des calculs rapides - parfois même en temps réel. Les procédures de calibration et de validation sont automatisées. La puissance des outils informatiques a permis d'introduire des processus mobilisant beaucoup de ressources comme des randomisations itératives et des modèles *non paramétriques*, i.e. s'affranchissant de toute hypothèse sur les lois de probabilité. En somme, se regroupe sous le terme de datamining*, ou de Machine Learning*, des algorithmes qui couplent statistiques, probabilités, intelligence artificielle et théories de l'apprentissage. Les algorithmes actuels permettent de prédire et aussi de comprendre de nombreux phénomènes complexes à partir d'observations diverses et variées (Han et Kamber, 2006).
Curieusement, *prédire* du point de vue statistique s'avère moins compliqué que *comprendre*. *Prédire* c'est anticiper, à peu près, l'évolution future d'un phénomène. Alors que le *comprendre* signifie identifier toutes les variables - corrélées ou pas - qui interagissent avec le phénomène observé. *Peut-on prédire sans comprendre? La question peut choquer mais au-delà du débat philosophique les outils de calcul statistiques semblent monter que oui.* (Saporta, 2006).
Désormais, l'objectif est moins de prédire que de comprendre. Afin de nourrir ce dessein, des *procédures de sélection de variables* – généralement fondées sur du *scoring* – ont été mises au point à partir de *Machine Learning**. Elles permettent d'identifier et de distinguer les facteurs explicatifs de ceux qui n'interagissent pas avec le phénomène d'intérêt. Et corollairement, elles décèlent les redondances et les corrélations qui entachent les capacités prédictives des modèles (Han et Kamber, 2006). *La sélection de variables* est une problématique de recherche qui fait couler beaucoup d'encre en mathématiques. Plus récemment le datamining* *s'ouvre à la sélection de modèles*. Le niveau d'inférence statistique va au-delà de celui de la sélection de variables. Mais cette perspective est l'apanage des recherches actuellement menées en mathématiques (Baraud, Giraud et al., 2009).

Notations utilisées en apprentissage statistique
La compréhension des méthodes de *sélection de variables* requiert l'utilisation d'un vocabulaire et de notations vernaculaires. Les jeux de données d'apprentissage sont notés $\mathcal{L}^{j}_{\mathrm{n}_i}$. Ils se constituent de n_i-individus et $p = (p_l + 1)$ variables. Par convention, le phénomène à expliquer est défini par une variable aléatoire *cible* notée $\mathrm{Y}^{\mathrm{j}}_{\cdot}$ ou $\mathrm{Z}^{\mathrm{j}}_{\cdot}$. La première notation est utilisée dans le cadre de l'approche *individus-centrée**, i.e. pour définir les variables LEA $\mathrm{y}^{\mathrm{j}}_{\mathrm{i}}$ (section.C). La seconde, dans l'approche géographique avec les i.st.m* – quantitatifs $\mathrm{z'}^{\mathrm{j}}_{(\mathrm{U}_k),\mathrm{c}}$ et qualitatifs $\mathrm{z'}^{\mathrm{j}}_{(\mathrm{U}_k),\mathrm{c}}$.
Les données d'apprentissage forment une matrice contenant $p_l \in \mathbb{N}$ variables potentiellement explicatives définies par $\mathcal{X}^j$ et une j-variable à expliquer définie par $\mathcal{Y}^j$.

$$\mathcal{L}^{j}_{\mathrm{n}_i} = \{\mathcal{X}^j \cup \mathcal{Y}^j\} = \left\{\left\{\mathrm{X}^{\{l|j\}}_{\cdot} = \left(\mathrm{X}^1_i, \dots, \mathrm{X}^{p_l}_i\right) \in \mathcal{X}^j, \forall \mathrm{i} = \{1, \dots, \mathrm{n}_i\}\right\} \cup \left\{\mathrm{Y}^j_{\cdot} = \left(\mathrm{Y}^j_i, \dots, \mathrm{Y}^j_{\mathrm{n}_i}\right) \in \mathcal{Y}^j\right\}\right\}$$

La variable aléatoire *cible* ou *réponse* est décrite par les observations mesurables du phénomène d'intérêt $y^j = \left(\mathrm{y}^j_1, \dots, \mathrm{y}^j_{\mathrm{n}_i}\right)$. Plusieurs variables *réponses* peuvent représenter, à leur façon, le même phénomène d'intérêt, de fait $j \in \{1, \dots, \mathrm{p}_j\}$. La nature quantitative ou qualitative de la réponse définit le Contexte statistique*, soit respectivement celui de la *régression* ou de la *classification*.

Les variables *potentiellement explicatives ou prédictives* appartiennent à l'ensemble $\mathcal{X}^j$, une matrice de dimension $(\mathrm{n}_i \times p_l)$. Les *coordonnées* des individus : e_i sont les valeurs mesurées et mesurables de

l'échantillon d'apprentissage $\mathcal{L}^{j}_{\mathrm{n}_i}$. Elles sont notées : $x^{l}_{\cdot} = \left(x^{1}_{i}, \ldots, x^{p_l}_{i}\right), \forall\, i = \{1, \ldots, \mathrm{n}_i\}$ et stockées dans des matrices, dont une représentation archétypique est donnée par la figure suivante :

L.j.ni	Variable								Cible
Individus	X.1	X.2	X.3		X.l		X.pl-1	X.pl	Y.j
e1	x11								y1j
e2									
e3									
ei					xil				ylj
e(ni-1)									
e(ni)								xni,pl	ynij

Figure 176 : Archétype d'un tableau statistique et notations conventionnelles d'un échantillon d'apprentissage $\mathcal{L}^{j}_{\mathrm{n_i}}$

L'idée de *la sélection de variables* est de réduire la dimension de $\mathcal{X}^{j}$, sans pour autant transformer les coordonnées statistiques comme dans les méthodes factorielles classiques. Le but est d'identifier les *variables de bruit, les variables explicatives et celles qui s'adaptent à la parcimonie prédictive*.
Les méthodes de *sélection de variables* les plus puissantes utilisent des algorithmes de datamining* et permettent d'atteindre l'objectif évoqué sur des jeux de données dits de *grandes dimensions, i.e. lorsque* $p_l \gg n_i$ (Tuleau-Malot, 2005).

CHOIX D'UN ALGORITHME ET D'UNE METHODE DE SELECTION

Etat de l'art sur la *sélection de variables*

L'efficience *des stratégies de sélection* est liée aux méthodes statistiques et probabilistes implémentées dans les algorithmes de datamining*. Leur puissance s'apprécie par leur capacité à répondre aux objectifs universels, à savoir: *Explorer, Modéliser, Comprendre, et Prédire*. Il n'existe pas de méthode prééminente et chacune opère sur des jeux de données particuliers. Les variables contenues dans $\mathcal{L}^{j}_{\mathrm{n}_i}$ sont-elles quantitatives (continues ou discrètes), qualitatives (booléennes ou multimodales), incomplètes, incertaines, lacunaires, en grandes dimensions ? Généralement les méthodes de datamining* utilisées par les stratégies de sélection permettent d'estimer l'importance des variables en attribuant des scores à chaque $x^{l}_{\cdot}$ (Han et Kamber, 2006).

Les stratégies *de sélection de variables* les plus citées dans la littérature et les plus connues pour leurs performances dans de nombreux domaines scientifiques sont les : *Support Vector Machine*, *Generalized Model Linear, Arbres de décisions*. Une étude menée dans un contexte de classification binaire et appliquée successivement à un jeu de *données contrôlées* où : $p_l =$ "grand" puis à des jeux de grandes dimensions, a permis de décrire les caractéristiques de chacune d'elles et d'identifier les jeux de données sur lesquelles elles performent (Ghattas et Ben Ishak, 2008).

Les Support Vector Machine (SVM) : sont des techniques d'apprentissage robustes. Ces Machine Learning* construisent un hyperplan qui maximise un critère de *marge* et qui optimise un *terme quadratique* (Vapnik, 1995). Plusieurs méthodes de sélection des variables utilisent les SVM et manifestement la plus puissante s'appelle : SVM-RFE* (Ben Ishak et Ghattas, 2005). Les SVM présentent l'avantage d'être parfaitement fiables dans un contexte de classification binaire. Aussi, les SVM offrent la possibilité de travailler sur des variables potentiellement explicatives de nature quantitative ou qualitative. Cependant il faut, au préalable, leur appliquer des transformations non-linéaires. Autre inconvénient, les scores d'importance des variables sont attribués de *façon indirecte* ce qui complexifie la mise en œuvre de cette technique. Et ils sont instables *lorsque les variables ne sont pas linéairement séparables*. En sus, les scores SVM ne sont pas pertinents* lorsque *la variable cible* est de nature qualitative multi-classes (Ghattas et Ben Ishak, 2008).

Les Generalized Model Linear (GML) sous contrainte de type L_1 : les GML constituent la jonction entre l'exploration topologique de données et les modèles généralisés de régression. La stratégie de sélection de variables la plus courante : *GLMpath*, est fondée sur l'approche *stepwise* (McCullagh et Nelder, 1989). Elle présente l'avantage d'identifier les variables importantes avec une précision chirurgicale si tant est que les données d'apprentissage forment un vecteur Gaussien. De nombreuses démonstrations mathématiques existent et ont permis de montrer qu'il s'agit de la seule méthode permettant d'obtenir des *estimateurs oracles*.
Elles peuvent être appliquées dans tous les contextes statistiques - sans que les $y_{\cdot}^{j}$ soient *linéairement séparables*. Cependant, les GML restent des *modèles paramétriques*. Par conséquence, ils ne sont valides que si les hypothèses faites sur les lois de probabilités correspondent à celles qui régissent $\mathcal{L}_{n_i}^{j}$, ce qui n'est jamais garanti. En outre, la stratégie *GLMpath* ne permet pas d'attribuer directement des scores d'importance aux variables. En sus, les GML gèrent très difficilement les $\mathcal{L}_{n_i}^{j}$ constitués simultanément de variables quantitatives et qualitatives (Ghattas et Ben Ishak, 2008). Enfin, des études montrent une forte dépendance à l'ordre de variable (Somol, Pudil et al., 1999) - ce qui en dit long sur leur fiabilité...

Les arbres de décision : font partie de la famille des *méthodes d'ensemble* (annexe.5). Il existe deux algorithmes permettant d'attribuer des scores d'importance aux variables : *CART* – acronyme qui signifie Classification And Regression Tree*s - (Breiman, Stone et al., 1984) et son successeur : *randomForest** (Breiman, 2001).
Les Forêts Aléatoires* (FA) sont générées à partir de nombreux arbres CART randomisés et constituent l'aboutissement des derniers travaux de Léo Breiman. En 1984, Breiman, assisté par l'informaticienne Cluter, créent l'algorithme Classification And Regression Tree* - CART (Breiman, Stone et al., 1984). En 1996 un nouvel algorithme voit le jour, il est connu sous la dénomination *Bagging de Breiman*. Il s'agit *d'une méthode d'ensemble* qui consiste à randomiser et partitionner l'espace des individus par une procédure *d'échantillonnage Boostrap,* puis à appliquer la règle CART, et enfin à agréger les prédicteurs en un seul - en fonction du Contexte statistique* (Breiman. Leo, 1996). En 2001, de nouvelles perturbations aléatoires sont implémentées pour randomiser *l'espace des variables*. Elles interviennent au niveau des nœuds des arbres CART, créant ainsi des arbres doublement randomisés. Le couplage du Bagging et du *Random-Subspace* constituent la pierre angulaire des FA, donnant ainsi naissance à l'algorithme *randomForest** (Breiman, 2001). Par la suite, Léo Breiman améliorera *randomForest** avec des processus de randomisation des outputs, de stabilisation des scores d'importance des variables, de comblement des lacunes… La cristallisation de *randomForest*.V4* signe la retraite de son créateur (Breiman, 2004).

Les performances explicatives et prédictives de *randomForest** dépassent largement celles de CART. De plus, les scores CART sont biaisés en faveur des variables qualitatives multi-classes *raison pour laquelle ils ne sont plus utilisés* (Genuer, 2010). Les FA présentent l'avantage d'être *des modèles non paramétriques*. De fait, aucune hypothèse n'est à faire sur les lois de probabilité. L'algorithme est accessible, l'attribution des scores *est directe* et ils sont particulièrement stables. L'atout majeur des FA est qu'elles sont capables de s'adapter à tous les contextes statistiques, quelle que soit la nature des variables contenues dans $\mathcal{L}_{n_i}^{j}$ et sans qu'il importe que les données soient linéairement séparables, que leurs unités soient hétéroclites et qu'il y ait des lacunes.
Les FA comptent deux inconvénients majeurs : d'abord il n'existe aucune preuve théorique permettant de légitimer leur puissance, et ensuite leurs performances sont intimement liées à la qualité de l'estimateur. Or, les FA sont très difficiles à paramétrer. Dans l'étude comparative dont il est question, les FA semblent les moins pertinentes pour détecter les variables explicatives dans les jeux de données en grandes dimensions. Elles sont concurrentielles lorsque le nombre de variables : $p_l =$ "grand. En contrepartie, les FA constituent le modèle fournissant les scores les plus stables (Ghattas et Ben Ishak, 2008).

Choix de l'algorithme et de la méthode de sélection

Dans l'étude comparative des méthodes de datamining*, les FA montrent la plus grande flexibilité aux jeux de données, i.e. qu'ils peuvent contenir des variables aux granularités* complètement hétéroclites. Cependant, les scores *randomForest** semblent être moins pertinents* pour caractériser l'importance des variables - par comparaison aux scores GML et SMV (Ghattas et Ben Ishak, 2008).

Remarques liminaires

Dans l'étude comparative, *randomForest** est utilisé avec les paramètres par défaut, implémentés par Breiman (Breiman, 2001). Or, les modèles de FA par défaut ne sont pas fiables, ils sont mêmes spécieux en classification et *a fortiori* en régression. Nonobstant cette faiblesse heuristique, depuis, des stratégies de calibration numérique ont récemment été proposées, testées, et leurs performances ont été validées (Genuer, 2010).

Au regard des jeux de données géographiques d'apprentissage utilisés dans cette thèse $\mathcal{L}_{n.}^{j}$:

Les variables réponses $\mathcal{Y}^{j}$ *sont,* dans le cadre de l'approche *individus-centrée*,* de nature qualitative booléenne – les variables séquelles LEA y_i^j (chapitre.1), dans le cadre de l'approche géographique de nature quantitative continue – les prévalences spatiales pondérées patients-années $z'^{j}_{(U_k),c}$, ou qualitative multi-classes – les propensions spatiales pondérées $z'^{j}_{(U_k),q}$ (chapitre.2).

Les variables potentiellement explicatives $\mathcal{X}^{j}$ ont des systèmes unitaires hétéroclites, elles prennent des gammes de valeurs disparates et peuvent être de nature : quantitative continue - i.e. $x'^{\mathrm{RADON}}_{(U_k)}, x^{137\mathrm{Cs}}_{(U_k)}, x^{\mathrm{tx.CHOM}}_{(U_k)}, x^{\mathrm{TEMP}}_{(U_k)}$ - ou quantitative discrète - i.e. $x^{\mathrm{EGRA}}_{(U_k)}$; $x^{\mathrm{OPHT}}_{(U_k)}$- qualitative booléenne - i.e. $x^{\mathrm{SEXE}}_{(U_k)}$; $x^{\mathrm{LEUC}}_{(U_k)}$, ou multi-classe - i.e. $x^{PROTO}_{(U_k)}$; $x^{RADON}_{(U_k)}$ (chapitre.3).

Propositions

Puisque les modèles de FA peuvent être calibrés et compte tenu du caractère complètement hétéroclites des variables constituant les jeux de données géographiques, l'outil de datamining* retenu est l'algorithme randomForest* (Breiman, 2001).

En dépit de la robustesse des scores *randomforest*, ces derniers ne présument pas de la façon dont les variables interagissent entre elles. Il existe peu de méthodes de sélection de variables avec cet algorithme. La plus adaptée à la finalité, i.e. à la caractérisation des interactions des FE/FIM* géographiques intégrés avec les états de santé étudiés est :Variable Selection Using RandomForest* (VSURF).

Il s'agit d'une récente stratégie de *sélection par seuillage* qui permet de disjoindre, dans un premier temps, *les variables de bruit des variables explicatives. Et, dans un second temps,* de construire un *paquet de variables qui s'adapte à la parcimonie prédictive* (Genuer, 2010).

Avant d'appliquer VSURF pour répondre à la problématique, il convient de décliner ses fondements théoriques, ceux de *randomForest** et aussi, les modestes compléments heuristiques destinés à adapter VSURF à la dialectique géographique

THEORIE SUR LES FORETS ALEATOIRES, SUR LA METHODE VSURF ET PROPOSITIONS HEURISTIQUES

La stratégie de sélection de variables qui a été retenue est Variable Selection Using RandomForest* (VSURF). Elle se fonde la théorie des FA et permet de construire trois paquets.
Le premier $\mathcal{X}^{j}_{\text{bruit}}$ contient toutes les variables de bruit, i.e. celles qui n'interagissent pas, au sens de VSURF, avec la réponse notée $y^{j}_{.}$ ou $z^{j}_{.}$. Le second $\mathcal{X}^{j}_{\text{explic}}$ contient toutes les variables $x^{l}_{.}$, corrélées ou pas, qui ont une influence significative sur la cible – i.e. qui permettent de l'expliquer. Enfin, le troisième paquet $\mathcal{X}^{j}_{\text{pred}}$ forme un sous-ensemble du second. Il contient uniquement les $x^{l}_{.}$ explicatives qui s'adaptent à la parcimonie prédictive (Genuer, 2010).

Spécification sur l'algorithme de sélection utilisé
La procédure VSURF se fonde sur l'algorithme *randomForest** qui est disponible dans un package.R (Liaw, 2013). En revanche, la procédure VSURF n'était implémentée dans aucun package.R au moment où les traitements ont été effectués. Par conséquent, *elle a été programmée dans un algorithme nommée MyVsurfGéo**. Une version béta de VSURF a cependant été mise à disposition entre temps sur le site CRAN (Genuer, Poggi et al., 2013).
Or VSURF n'est pas adaptée à la dialectique géographique. D'abord parce que ce package.R ne permet pas de calibrer les FA, et ensuite parce qu'aucun item graphique n'est renvoyé. Quant aux items statistiques, ils ne permettent pas de mettre en œuvre les *modestes propositions heuristiques* de cette thèse - qui touchent à la sélection d'un panel plus large de variables explicatives, et aux prédictions géographiques. Les fondements théoriques de MyVsurfGéo* (MVG) et les résultats obtenus sur des jeux de données jouées sont identiques à ceux de la thèse du concepteur de VSURF (Genuer, 2010). Enfin, les résultats obtenus par la version bêta de VSURF et MVG ont été comparés sur les données « iris » (Fisher, 1936) et se sont avérés identiques à l'instabilité près de l'algorithme *randomForest** – i.e. à cause des effets de bord* induits par les processus stochastiques implémentés (Genuer, 2013).

Les capacités de *randomForest** ont été testées et validées empiriquement sur des jeux de données jouées et des jeux de données contrôlées. Les performances prédictives et explicatives de l'algorithme sont *stupéfiantes* à tel point qu'elles se montrent parfois *compétitives avec les méthodes fondées sur du Boosting* - réputées pour être la technique d'apprentissage la plus efficace* (Bernard, Heutte et al., 2008). Depuis des modules complémentaires de calibration et de représentation graphique ont été introduits par différents mathématiciens. La version des FA utilisée dans le cadre cette thèse est celle du package.R *randomForest*.V4.6-7* (Liaw, 2013).

Avant de procéder à l'identification des FE/FIM* géographiques qui interagissent avec les PM* d'intérêt et les *modestes propositions heuristiques d'adaptation des concepts à la dialectique géographique*, il convient de décliner les éléments théoriques des FA, de *randomForest** et de VSURF nécessaires à la compréhension de la section.B et de la section.C.

Les deux sous-parties subséquentes constituent un résumé sommaire de l'annexe 7 et de l'annexe 8.

PRINCIPE DE L'ALGORITHME RANDOMFOREST

Les Forêts Aléatoires* (FA) se composent d'arbres CART randomisés. Avant d'introduire les principes des FA il convient de décliner sommairement ceux de Classification And Regression Tree* (CART).

Principe de l'algorithme CART

CART est un modèle d'inférence statistique *non paramétrique*. Les éléments théoriques subséquents sont identiques à ceux implémentés de la fonction *rpart* - la* seule version de l'algorithme conforme aux travaux de Breiman (Therneau, Atkinson et al., 2013). Cette partie s'appuie essentiellement sur la thèse de Christine Malot (Tuleau-Malot, 2005).

Le prédicteur CART $f^{j}_{CART}(X^{l}_{\cdot})$ est constant par morceaux. Sa forme diffère en fonction du Contexte statistique*. En régression, il prend la forme d'un modèle linéaire de type arbre de décision. En classification il s'agit de l'estimateur de Bayes.

Le principe de construction des arbres CART repose sur une logique dyadique*. A chaque nœud : t_k est associée une question binaire qui permet, à partir d'une division optimale $\delta^{*}_{l,k}$, de partitionner l'espace des individus au regard de leurs coordonnées et en maximisant une fonction d'hétérogénéité.

La division des nœuds t_k s'opère récursivement jusqu'à obtenir un arbre maximal : T^{j}_{max} dont toutes les feuilles $\tilde{t}_k$ sont pures. Autrement dit, chaque prédiction : $\tilde{t}_k(e_i) \rightarrow \hat{y}^{j}_{i}$ correspond à la valeur observée : y^{j}_{i}. Le prédicteur associé à $T^{j}_{max}(x^{l}_{\cdot})$ est sans biais, i.e. que : $R\left(T^{j}_{max}\right) = 0$. Où : $R\left(T^{j}_{k}\right)$ est l'estimateur de la variance en régression et celui de l'indice de Gini en classification.

Les prédictions In-Sample (IS) de T^{j}_{max} sont exactes. En revanche, elles sont médiocres Out-Of-Sample (OOS).

Pour améliorer l'erreur en terme de variance l'algorithme CART procède à une phase d'élagage dans l'optique d'identifier un arbre optimal : T^{j}_{opt}. Il est construit à partir $T^{j}_{max}(x^{l}_{\cdot})$ par une suite d'opérations récursives permettant d'obtenir une collection d'arbres optimisés et uniques connue sous le nom de *sous-suite de Breiman* : $\mathcal{T}^{K}_{Breiman}$. L'arbre optimal appartient nécessairement à la sous-suite $T^{j}_{opt} \in \mathcal{T}^{K}_{Breiman}$. Il est généralement spécifié, par une technique de cross-validation (c.v.) qui permet, par le biais d'un critère pénalisé : $Crit_{\alpha}\left(T^{j}_{k}\right)$, de maximiser un compromis biais/variance.

L'estimateur de T^{j}_{opt}est celui dont l'erreur de c.v. : $R^{cv}\left(T^{j}_{ko}\right)$ est inférieure au seuil d'élimination de c.v. : φ^{cv}_{ko} – qui fait intervenir un coefficient de complexité $\dot{\alpha}^{cv}_{k}$ – et qui est fondé sur la règle du *min + 1SD*. Dans la pratique, l'arbre optimal est obtenu par l'élagage de T^{j}_{max} en spécifiant une complexité α'_{opt} à partir de la complexité optimale estimée par c.v. : $\dot{\alpha}^{cv}_{opt}$, en privilégiant soit le biais, soit la variance afin de prendre en compte *l'effet de bord** lié à l'instabilité de l'algorithme.

L'estimateur CART optimal est noté $\hat{y}^{j}_{\cdot} = \hat{f}^{j}_{CART}(x^{l}_{\cdot}) \overset{def}{=} \hat{T}^{j}_{opt}(x^{l}_{\cdot})$. Les éléments essentiels de la méthode CART nécessaire à la compression des Forêts Aléatoires* et surtout de l'estimation du seuil d'élimination des variables de bruit de VSURF sont désormais énoncés – pour plus de précision se rapporter à l'annexe.5.

Principes de l'algorithme FA

Les Forêts Aléatoires* (FA) font partie *des méthodes d'ensemble* et sont fondées sur un modèle d'inférence statistique *non paramétrique* : CART et à laquelle *des processus stochastiques* ont été rajoutés. Cette partie s'appuie essentiellement sur la thèse de Robin Génuer (Genuer, 2010). Les principes énoncés sont ceux implémentés dans le package R : *randomForest*-V4.6-7* - la version officielle des Forêts Aléatoires* (Liaw, 2013).

Définition et notations : les FA, au sens où l'entend Breiman, se constituent d'une collection d'arbres CART doublement perturbés $\mathrm{T}_{\mathrm{cart}}^{j}\left(\mathcal{L}_{n_i}^{j}\left\{\Theta_{\mathrm{k}} \cap \left(\Xi_{\mathrm{t_v}}\right)_{1\leq \mathrm{v}\leq \mathrm{V_k}}\right\}\right)$, et sont construites conditionnellement à un échantillon d'apprentissage $\mathcal{L}_{n_i}^{j}$. La forme de l'opérateur FA : $\mathrm{Y}_{\cdot}^{\mathrm{j}} = f_{FA}^{j}\left(\mathrm{X}_{\cdot}^{\mathrm{l}}\right)$ dépend du Contexte statistique*, i.e. *la régression* lorsque : $y_{\cdot}^{j} \in \mathbb{R}^{n_i}$ ou *la classification quand :* $y_{\cdot}^{j} \in \mathcal{C}^{j} \subseteq \mathbb{N}^{|\mathcal{C}^{j}|}$. A l'initialisation, l'algorithme *randomForest** confectionne des *échantillons Bootstrap classiques* $\mathcal{L}_{n_i}^{\Theta_{\mathrm{k}},j}$ où Θ_{k} sont les indices des individus qui les constituent les graines qui vont donner naissance aux K-arbres CART de la FA. Puis, ces derniers sont développés jusqu'à *atteindre une taille maximale perturbée* par du Random-Subspace, i.e. que la division optimale de chaque nœud $\delta_{\mathrm{t},l,k}^{*}$ n'est pas estimée sur l'ensemble des coordonnées $x_{\cdot}^{l}$ - comme avec *rpart** - mais sur un sous-ensemble de m-variables choisies récursivement et de façon aléatoire $\Xi_{\mathrm{t_v}}$ dans une multinomiale uniforme discrète, afin de les *décorréler* (Breiman et Cutler, 2005).

Modèle de FA par défaut : le nombre d'arbres implémentés par défaut dans *randomForest** est fixé à : $\{\mathrm{ntree} = 500\}$. D'une manière générale plus ntree est grand et plus les FA sont robustes, mais plus les temps de calcul sont longs (Liaw et Wiener, 2006).
Le nombre de coordonnées par défaut regardées pour la construction des nœuds perturbés, suggéré par Léo Breiman dépend du Contexte statistique* :

$$\mathrm{mtry} = \begin{cases} \left[{}^{p_l}/_{3}\right] & \text{régression} \\ \left[\sqrt{p_l}\right] & \text{classification} \end{cases}$$

La valeur de mtry a une influence forte sur les performances des FA. A l'aune de ntree les performances des forêts ne sont pas proportionnelles à la valeur de mtry. Lorsque $\mathrm{mtry} = 1$ les prédicteurs individuels sont bien décorrélés mais complétement dégradés. Lorsque $\mathrm{mtry} = p_l$ on retombe sur le Bagging de Breiman avec des arbres maxima et trop corrélés. Ce paramètre doit être impérativement être calibré car les valeurs par défaut sont spécieuses (Genuer, 2010).

Un paramètre permet de juguler la taille des arbres, i.e de contrôler la transformation des nœuds en feuille : nodesize. Par défaut les valeurs implémentées sont : $\mathrm{nodesize} = 5$ en régression et $\mathrm{nodesize} = 1$ en classification. Ce paramètre a une influence importante sur la qualité des prédicteurs individuels (Tuleau-Malot, 2005).
Lorsque les paramètres des prédicteurs individuels sont spécifiés, la phase d'agrégation ensembliste construit l'estimateur FA. Il permet de prédire Out-Of-Bag (OOB) les valeurs de la *cible*, i.e. soit par un *vote majoritaire* soit par l'estimation d'une *moyenne empirique* des prédictions fragmentaires de chaque arbre mais uniquement avec les e_i qui n'ont pas été utilisés pour leur construction :

$$\hat{y}_{\cdot}^{OOB,j} = \hat{f}_{FA}^{j}\left(x_{\cdot}^{\mathrm{l}} \middle| \mathrm{ntree}; \mathrm{mtry}; \mathrm{nodesize}\right)$$

Qualité du modèle FA : elle s'estime sur ses capacités prédictives OOB par le biais d'une erreur de généralisation notée $\mathrm{R}^{\mathrm{OOB}}(\mathrm{FA})$. Cette erreur de généralisation s'évalue au regard des $y_{\cdot}^{j}$ et $\hat{y}_{\cdot}^{OOB,j}$ et en fonction du contexte : en classification par l'err.OOB, une proportion OOB d'individus mal classés, en régression par la mse.OOB, l'estimateur de la variance OOB. Cette dernière étant difficilement interprétable, elle est généralement convertie en *pourcentage OOB de variance expliquée*, noté : $\mathbb{var}.\mathbb{explain}_{\mathrm{FA}}^{\mathrm{OOB,j}}$.

Calibration des FA : l'erreur de généralisation $\mathrm{R}^{\mathrm{OOB}}(\mathrm{FA})$ est intimement liée à la qualité des données d'apprentissage et aux paramètres spécifiés pour la construction de l'estimateur FA. Aussi, les scores d'importance des variables montrent une forte dépendance à la qualité du modèle qu'il convient d'optimiser. Pour ce faire, il s'agit d'identifier les valeurs du vecteur de paramètres : $\Lambda_j = (\mathrm{ntree}\ ;\ \mathrm{mtry}\ ;\ \mathrm{node.size})$ qui minimisent $\mathrm{R}^{\mathrm{OOB}}(\mathrm{FA})$.

A ce jour, il n'existe aucune procédure permettant d'obtenir la FA optimale. Les paramètres cités semblent être les plus importants et sont très débattus dans la littérature.
S'agissant de node. size les valeurs par défaut sont *a priori* robustes. Cependant, en régression, il peut être intéressant de le *diminuer*, i.e. en le portant à {nodesize = 3} afin de ne pas exagérer la dégradation sur les prédicteurs individuels - déjà altérée par mtry (Tuleau-Malot, 2011).

La difficulté touche à la calibration de {ntree ; mtry }. La valeur calibrée de ntree est celle qui n'améliore plus de façon significative les performances de la FA, au regard des temps de calcul et en supposant que {ntree ≤ 1 000}. Quant au paramètre mtry, la technique proposée par Robin Génuer semble être la plus robuste car elle permet de tester toute l'étendue des valeurs pouvant être prises :

$$\Omega.\,\text{genuer} = \bigcup_{\forall p_l \in \mathbb{N}^+} \left(\left\{ \left[1; \frac{\sqrt{p_l}}{2}; \sqrt{p_l}; 2\sqrt{p_l}; 4\sqrt{p_l}; \frac{p_l}{4}; \frac{p_l}{3}; \frac{p_l}{2}; \frac{3}{4}p_l; p_l \right] \right\} \middle| \text{ntree} \in \{250 ; 500 ; 1\,000\} \right)$$

Dans la mesure où l'algorithme *randomForest* est instable,* l'optimisation de la FA est spécifiée par minimisation de l'erreur de généralisation stabilisée sur 10 forêts - ou *runs* – notée : $\overline{R}_{\cdot}^{OOB}(FA)$. Le vecteur $\widehat{\Lambda}_j$ est spécifié conditionnellement au processus d'optimisation suivant :

$$\widehat{\Lambda}_j = \underset{\forall\, \Lambda \in \{\Omega.\text{genuer} | \text{nodesize}\}}{\operatorname{argmin}} \left\{ \overline{R}_{\cdot}^{OOB}(FA) = \frac{1}{10} \sum_{r=1}^{10} R_r^{OOB}(FA) \right\}$$

Il convient de noter que pour calibrer mtry il est possible aussi d'utiliser le package.R tuneRF (Ligges, 2013) . Mais cet algorithme d'optimisation présente les désavantages classiques, à savoir une forte dépendance à *la condition initiale* et au *pas de calcul* choisis – deux nouveaux paramètres impossibles à spécifier *a priori* sauf pour des jeux de données de petite taille (Genuer, 2010).

Importance des variables : afin de mesurer l'efficience des $x_{\cdot}^{l}$, sur les observations de la réponse d'intérêt $y_{\cdot}^{j}$, quelles que soient leur nature et les corrélations qu'elles peuvent entretenir - Léo Breiman suggéra des scores d'importance des variables : $VI(X^l)$. Ces scores sont supputés en utilisant de la façon la plus topique qui soit, les propriétés des échantillons Bootstrap et du Random-Subspace. Ils sont estimés à partir de l'erreur de généralisation OOB $R^{OOB}(T_k|\widehat{\Lambda}_j)$ et d'une erreur de généralisation OOB permutée $R^{O\tilde{O}B^l}(T_k|\widehat{\Lambda}_j)$. Pour affecter des scores à la coordonnée $x_{\cdot}^{l}$, elle est systématiquement substituée, pour chaque e_i, par une autre – dont l'indice est tiré aléatoirement dans une loi uniforme discrète – ce qui permet d'évaluer $R^{O\tilde{O}B^l}(T_k|\widehat{\Lambda}_j)$. Puisque l'algorithme est instable les scores d'importance des variables sont stabilisés, en pratique, sur $n_r = 50$ forets: $\overline{VI}_j(x^l)$.
L'interprétation des $\overline{VI}_j(x^l)$ est très intuitive, plus les permutations engendrent une forte augmentation de $R^{O\tilde{O}B^l}(T_k|\widehat{\Lambda}_j)$, plus le score d'importance des variables tend à prendre des valeurs fortes, donc plus la variable substituée est importante au sens de randomForest*.
En théorie les $VI(X^l) \geq 0$. Dans la pratique il arrive parfois que $VI(X^l) < 0$.

Les estimateurs biaisés de l'écart-type des scores : $\hat{\sigma}\left(VI_j(x^l)\right)$ sont renvoyés par l'algorithme *randomForest*.* Ils permettent de *réduire* les scores : $\overline{VI}_j^{\mathbb{r}}(x^l)$ – en les ramenant à une échelle unitaire – et de mieux percevoir leurs variabilités. L'algorithme *renvoie* aussi les scores historiques : $\overline{VI}_j^{CART}(x^l)$ – mais comme ils sont biaisés, ils ne présentent plus guère d'intérêt (Genuer, 2010).
Les principes fondamentaux permettant de construire les estimateurs FA et CART tels qu'ils sont renvoyés par les algorithmes *randomForest** et *rpart** ont été décrits pour faciliter la compréhension de la stratégie de sélection de variables : VSURF.

PRINCIPE DE LA STRATEGIE VSURF ET DE MYVSURFGEO

La méthode Variable Selection Using Random Forest (VSURF) est une stratégie statistique qui permet de disjoindre par seuillage les variables *de bruit*, i.e. celles qui n'interagissent vraiment pas avec la cible $y_{\cdot}^{j}$, des *variables explicatives* i.e., les variables $x_{\cdot}^{l}$, redondantes ou corrélées, qui entretiennent des liens avec $y_{\cdot}^{j}$. Et par suite, d'identifier celles qui s'adaptent à la *parcimonie prédictive* i.e. un ensemble frustre de $x_{\cdot}^{l}$ explicatives qui interagissent très fortement avec $y_{\cdot}^{j}$ et qui permettent d'anticiper correctement son évolution par $\hat{y}_{\cdot}^{j}$

Les principes énoncés correspondent exactement à la théorie VSURF (Genuer, 2010). Les compléments récemment ajoutés à la méthode et implémentés dans la version bêta de VSURF sont aussi déclinés (Genuer, Poggi et al., 2013). Toutefois comme le package.R : VSURF n'est pas parfaitement adapté à la dialectique géographique, deux modestes propositions heuristiques ont été formulées et intégrées dans l'algorithme MyVsurfGéo* (MVG) – pour plus de détails se référer à l'annexe.6.

Phase 0 – Initialisation & Calibration des paramètres

A l'initialisation l'algorithme considère un jeu de données d'apprentissage multidimensionnel $\mathcal{L}_{\mathrm{n_i}}^{\mathrm{j}}$ contenant une *cible* $y_{\cdot}^{j}$ et une matrice de variables auxiliaires $x_{\cdot}^{l} = (x_{\cdot}^{1}, \dots, x_{\cdot}^{p_l})$.
Ensuite, les paramètres {ntree; mtry; nodesize} sont calibrés ; La FA optimisée est celle dont le vecteur $\widehat{\Lambda}_j$ permet de minimiser l'erreur de généralisation : $\overline{\mathrm{R}}_{\cdot}^{\mathrm{OOB}}\left(\mathrm{FA}\middle|\Lambda_j\right)$ - stabilisée sur 10 run.
Enfin, la stratégie de VSURF est appliquée. Elle s'opère en trois phases :

Phase 1 - Hiérarchisation descendante & Elimination des variables de bruit

L'objectif est de disjoindre les variables de *bruit avéré* de celles qui sont *potentiellement explicatives* – en se basant sur les écarts-types des scores d'importance des variables.
Il s'agit d'estimer, à partir d'un modèle de FA optimisé, la moyenne $\overline{\mathrm{VI}}_j\left(x^{\mathrm{l}}\right)$ et les écarts-types $\widehat{\sigma}\left(\mathrm{VI_{j,r}}\left(\mathrm{x^{l}}\right)\right)$ des scores d'importance des variables - stabilisés sur $\{\mathrm{n_r} = 50\}$ forêts.
Puis de hiérarchiser les variables par ordre décroissant de score et constituer ainsi, en fonction de leur rang, une matrice *de vecteurs colonnes* : $x_{\cdot}^{(l)} = \left(x_{\cdot}^{(1)}, \dots, x_{\cdot}^{(l)}, \dots, x_{\cdot}^{(p_l)}\right)$.

> Proposition heuristique vouée à l'interprétation et à l'analyse géographique (i)
> Présenter la valeur moyenne $\overline{\mathrm{VI}}_j\left(x^{\mathrm{l}}\right)$ et la dispersion des $\mathrm{VI_j}\left(\mathrm{x^{l}}\right)$, conjointement pour toutes les variables x^{l}, grâce à un *diagramme de Tukey* – plus connu sous la dénomination *de boite à moustaches* - afin de pouvoir les comparer plus facilement (Saporta, 2006)

Ensuite la stratégie VSURF consiste à disjoindre les variables *potentiellement explicatives* des variables *de bruit avéré*. Elle est dite *conservatrice*. Les variables conservées sont incluses dans le paquet $\mathcal{X}_{\mathrm{conserv}}^{\mathrm{j}}$ constitué de toutes celles pouvant potentiellement interagir, même de façon faiblement efficiente, avec $y_{\cdot}^{j}$.

Les variables éliminées constituent le paquet : $\mathcal{X}_{\mathrm{bruit}}^{\mathrm{j}}$. Il contient des *variables de bruit avéré*, i.e. celles dont le $\overline{\mathrm{VI}}_j\left(x^{\mathrm{l}}\right)$ est inférieur à un seuil d'élimination de bruit, noté : $\psi_{\mathrm{bruit}}^{\mathrm{j}}$.

Ce seuil représente le *saut moyen* qu'une variable même très faiblement explicative peut effectuer mais qu'une variable de bruit est incapable de faire. Le seuil $\psi_{\mathrm{bruit}}^{\mathrm{j}}$ correspond à la valeur minimale estimée de $\widehat{\sigma}\left(\mathrm{VI_{j,r}}\left(x_{\cdot}^{(l)}\right)\right)$, en fonction de son rang, par le prédicteur CART de la sous-suite de Breiman dont l'erreur de validation croisée est la plus petite.

Spécificité de l'algorithme MyVsurfGéo* (MVG)

L'arbre CART permettant d'estimer ψ^{j}_{bruit} est plus profond que celui de VSURF dans la mesure où *le critère de complexité* utilisé : $\alpha^{MyVsurfGeo}_{min}$, visant à réduire *l'effet de bord** de l'algorithme *rpart**, est spécifié pour maximiser la contrainte *de non biais* (annexe.5). Par conséquence MyVsurfGéo* est encore plus *conservatrice* que VSURF.

Phase 2 - Sélection des variables explicatives

Presque sûrement $\mathcal{X}^{j}_{conserv}$ contient l'intégralité des variables explicatives mais aussi systématiquement une quantité, parfois importante, de variables de bruit avéré. L'objectif de cette seconde phase est d'identifier les variables qui interagissent de façon évidente avec les réalisations du phénomène d'intérêt : y^{j} – en vue de constituer le paquet : $\mathcal{X}^{j}_{explic}$.

L'idée consiste à constituer k-modèles de FA imbriqués et paramétrés par défaut $FA\left(x^{(l)}_{imbr.k}\middle|\dot{\Lambda}\right)$ – contenant chacun les k-premières-variables les plus importantes du paquet $\mathcal{X}^{j}_{conserv}$. Puis de ne retenir que l'ensemble le plus frustre de variables $x^{(l)}_{imbr.k}$ qui améliore significativement, au sens statistique du terme, l'erreur OOB stabilisée, i.e. le modèle dont $\overline{R}^{OOB}_{\left(x^{j}_{imbr}\right)}$ est inférieure ou égale à un seuil explicatif, noté : ψ^{j}_{explic}

Le seuil ψ^{j}_{explic} est estimé à partir d'une *astuce statistique classique* permettant de prendre en compte l'instabilité de *randomForest*, à savoir le minimum des erreurs OOB stabilisées – parmi tous les modèles imbriqués – augmenté ensuite par son SD*

Remarque : généralement $\mathcal{X}^{j}_{explic}$ ne contient que des variables réellement explicatives mais pas toutes, certaines d'entre elles sont systématiquement éliminées.

Proposition heuristique vouée l'interprétation et à l'analyse géographique (ii)

Introduire une règle supplétive afin de retenir un paquet de variables plus exhaustif pour pallier le risque d'élimination de variables explicatives peu efficientes. Les jeux de données géographiques sont entachés par des bruits de fond environnementaux ce qui complexifie l'identification des FE/FIM* qui influencent les états de santé étudiés. Le paquet « Bourrelly » : proposé est noté : $\mathcal{X}^{j}_{0.explic}$, il se construit identiquement à celui de « Génuer » : $\mathcal{X}^{j}_{explic}$, i.e. qu'il correspond à celui du modèle imbriqué dont l'erreur OOB est inférieure ou égale à ψ^{j}_{explic} et dont le nombre de variables qui le composent est inférieur ou égal au maximum entre le sous-ensemble le plus frustre et 60% des variables contenues dans $\mathcal{X}^{j}_{conserv}$

Remarque : $\mathcal{X}^{j}_{0.explic}$ contient généralement plus de variables réellement explicatives que $\mathcal{X}^{j}_{explic}$ mais aussi plus de variables de bruit (annexe.6).

Intérêt : Obtenir un ensemble de variables explicatives plus grand dont la pertinence sera évaluée *a posteriori* à partir des connaissances expertes. Il est utilisé dans l'analyse des résultats MVG (section.B). Et aussi, dans le cadre de l'approche *individus-centrée**, par la méthode BoostMuVsurfGéo (section.C).

Phase 3 - Identification des variables prédictives

Cette dernière phase a pour dessein d'identifier parmi le paquet de *variables explicatives* celles qui *s'adaptent à la parcimonie prédictive* - en les concaténant dans le paquet : $\mathcal{X}^{j}_{\text{pred}}$. Le but étant d'éluder le *sur-apprentissage* qui dégrade les performances prédictives des FA.

L'idée consiste à *injecter séquentiellement* les variables explicatives de $\mathcal{X}^{j}_{\text{explic}}$ par ordre décroissant de score et d'éliminer toutes celles qui n'améliorent pas significativement la qualité du modèle.
Le gain significatif d'erreur OOB est modélisé par *un seuil d'élimination prédictif :* ψ^{j}_{pred} . Il correspond à la moyenne des différences d'ordre un, en valeur absolue, des erreurs OOB de FA par défaut construites avec toutes les variables explicatives : $\mathcal{X}^{j}_{\text{explic}}$ et ce même modèle dans lequel sont introduites les variables éliminées dans la phase 2, contenues dans $\mathcal{X}^{j}_{\text{elim}}$.

A l'initialisation $x_{\cdot}^{(1)}$ est incluse dans $\mathcal{X}^{j}_{\text{pred}}$ ensuite toutes les variables de $\mathcal{X}^{j}_{\text{explic}}$ sont injectées *séquentiellement* en fonction de leur rang $x_{\cdot}^{(l+1)}$ et un gain d'erreur OOB, stabilisé sur 50 forêts, est estimé: $\overline{\Delta}\left(R_l^{OOB}\right)$. Il correspond au rapport d'erreur entre un modèle de FA par défaut contenant les $x_{\cdot}^{(l)}$ retenues dans $\mathcal{X}^{j}_{\text{pred}}$ *au pas de calcul : l* et celui contenant les mêmes variables ainsi que celle testée : $x_{\cdot}^{(l+1)} \in \mathcal{X}^{j}_{\text{explic}}$. La procédure concatène itérativement dans $\mathcal{X}^{j}_{\text{pred}}$ toutes les variables testées $x_{\cdot}^{(l+1)}$ lorsque $\overline{\Delta}\left(R_l^{OOB}\right) > \psi^{j}_{\text{pred}}$, sinon elles sont éliminées.

Remarques

Les principes fondamentaux de VSURF et les spécificités implémentées dans MVG ont été résumés sommairement dans cette sous-partie (annexe.6).

Récapitulatif des propositions heuristiques :

Les différences entre MVG et VSURF sont les suivantes : les paramètres de la FA sont optimisés numériquement $\widehat{\Lambda}_{j}$, l'arbre CART permettant d'estimer ψ^{j}_{bruit} est plus profond que celui de VSURF, deux paquets explicatifs sont spécifiés - celui de Genuer $\mathcal{X}^{j}_{\text{explic}}$ et celui de Bourrelly $\mathcal{X}^{j}_{0.\text{explic}}$ qui est plus exhaustif – et, les prédictions géographiques *In-Sample* (IS) sont renvoyées $\hat{y}^{j}_{\cdot}$ à partir des $x^{l}_{\cdot}$ explicatives.

Aussi, il est important de notifier qu'entre la proposition méthodologique théorique et la programmation de la version béta de VSURF une évolution heuristique a été apportée. Elle a aussi été programmée dans MVG. Elle permet de pondérer les différents seuils : ψ^{j}_{bruit} ; ψ^{j}_{pred}; ψ^{j}_{explic} par des coefficient spécifiés subjectivement à partir de connaissances expertes : $c^{j}_{\psi.\text{bruit}}$; $c^{j}_{\psi.\text{explic}}$; $c^{j}_{\psi.\text{pred}}$, (Genuer, Poggi et al., 2013).

Avant d'appliquer MyVsurfGéo* (MVG) aux jeux de données géographiques - en vue de répondre à la problématique, i.e. d'identifier les Facteurs Environnementaux* (FE) ayant une influence sur la géographie des différents phénomènes morbides étudiés - l'algorithme est appliqué à des jeux de données jouées afin de se faire une idée de sa robustesse, de se familiariser avec les sorties graphiques et les items statistiques renvoyés et d'apprendre à les interpréter. Tous les traitements ont été effectués avec le logiciel R (Institute for Statistics and Mathematics, 1997).

APPLICATION DIDACTIQUE DE MVG ET ESTIMATION DE SON POTENTIEL

Dans cette sous-section MyVsurfGéo* (MVG) est appliquée à des jeux de données jouées, i.e. données dont les variables $x_{.}^{l} = (x_{.}^{1}, \dots, x_{.}^{p_l})$ interagissent avec la cible et dont l'efficience sur y^j est connue *a priori*. Le cheminement dialectique et les notations utilisées sont conformes à ce qui est spécifié dans la sous partie théorique (annexe.5) et (annexe.6).

Remarques liminaires

Au moment où cette application fictive a été effectuée les coefficients experts de pondération : $c_{\psi.}^{j}$ applicables aux différents seuils $\psi_{.}^{j}$ n'avaient pas encore été proposés. D'ailleurs, la méthode n'était pas baptisée VSURF (Genuer, Poggi et al., 2013). Cette spécificité n'était pas programmée dans MVG, par conséquent ils ne seront pas utilisés.

De plus, les $c_{\psi.}^{j}$ ne présentent aucun intérêt dans les applications fictives. En revanche, ils sont utilisés lors de l'application de MVG aux jeux de données géographiques. En l'occurrence $c_{\psi.\text{bruit}}^{j}$ qui permet d'augmenter la valeur de ψ_{bruit}^{j}, dans la mesure où MVG est encore plus conservatrice que ne l'est VSURF (annexe.6).

Objectifs

Objectif n°1 : proposer une application didactique afin se familiariser avec les résultats renvoyés par MVG et apprendre à les interpréter et à les analyser sur les jeux de données jouées : $\mathcal{L}_{n_i}^{\text{Friedman.1}}$ et $\mathcal{L}_{n_i}^{\text{ToyData}}$ aux caractéristiques statistiques analogues à ceux des jeux géographiques : $\mathcal{L}_{U_k}^{\text{CATA}}$; $\mathcal{L}_{U_k}^{\text{TUM2}}$; $\mathcal{L}_{U_k}^{\text{THYR}}$ – puisqu'ils touchent au contexte de la classification et de la régression.

Objectif n°2 : présenter la dialectique qui sera utilisée lors de l'application de MVG aux jeux de données géographiques. La procédure MVG, à l'instar de VSURF, n'est pas automatique : La phase 0 de calibration nécessite de spécifier ntree sur un compromis entre qualité du modèle et temps de calcul. La calibration de mtry se fait au regard de sa capacité à minimiser $\overline{R}^{OOB}(FA)$ mais plusieurs valeurs peuvent aboutir à des modèles aux performances identiques. Chacune des phases est associée à un seuil $\psi_{.}^{j}$, qui peut être pondéré par $c_{\psi.}^{j}$ et qui conditionne la phase suivante. En sus, l'identification des variables explicatives requiert de choisir entre le paquet Génuer : $\mathcal{X}_{\text{explic}}^{j}$ – conforme à VSURF - et le paquet Bourrelly $\mathcal{X}_{0.\text{explic}}^{j}$ proposé dans MVG en vue de s'adapter à la complexité géographique (annexe.6) et utilisé aussi dans la proposition heuristique de BoostMyVsurfGéo* (section.C).

Objectif n°3 : il s'agit d'évaluer la puissance de MVG sur des jeux de données où les règles sont connues afin de se faire une idée des perspectives prometteuses espérées lors de son application aux jeux de données géographiques, en testant : sa capacité à disjoindre les variables de bruit avéré des variables explicatives, l'utilité des propositions heuristiques effectuées et notamment celle du paquet Bourrelly $\mathcal{X}_{0.\text{explic}}^{j}$, la qualité des prédictions d'abord In-Sample (IS) - i.e. conformément à ce qui pourra être fait avec les jeux de données géographiques - et ensuite Out-Of-Sample (OOS) - i.e. à partir de nouveaux jeux de données jouées avec des règles stochastiques identiques afin d'évaluer les performances prédictives – ce qui ne sera pas faisable sur les données réelles.

Dans un premier temps les caractéristiques des jeux de données jouées sont présentées, puis chacune des phases d'application de MVG est décrite par : un rappel synoptique théorique, un aperçu des résultats renvoyés par MVG et une analyse des résultats intermédiaires permettant de comprendre l'intérêt de chaque phase et de passer à la suivante.

PRESENTATION DES JEUX DE DONNEES UTILISES

Jeux de données : ToyData

$\mathcal{L}_{\mathrm{n}_i=50}^{\mathrm{DataToy}(100)} = \{\mathcal{Y}^{\mathrm{DataToy}}\} \cup \{\mathcal{X}^{\mathrm{DataToy}}\}$ a trait au Contexte statistique* de la classification binaire *i.e.* avec : $y_{\cdot}^{\mathrm{DataToy}} \in \{-1, 1\}$. Les paramètres du générateur stochastique sont réglés de façon à générer un jeu de données de *grandes dimensions,* i.e. $\{\mathrm{n}_i = 50\}$ individus et $\{p_l = 100\}$ variables auxiliaires, et de sorte que :

La variable réponse résulte d'un tirage aléatoire dans une loi de Bernoulli, tel que :

$$\mathbb{P}\left(\{y_{\cdot}^{\mathrm{DataToy}} = 1\}\right) \sim \mathcal{B}(p = 0{,}3)$$

Les trois premières sont corrélées et interagissent fortement avec la *réponse*, tel que :

$$\mathbb{P}\left(x_{\cdot}^{l} = (x_{\cdot}^{1}\,;\; x_{\cdot}^{2};\; x_{\cdot}^{3})\right) \sim \mathcal{B}(p = 0{,}7) = \begin{cases} p & \text{lorsque: } x_i^l \sim \mathcal{N}\left(\left(y_i^j \cdot l\right);\; 1\right) \\ (1-p) & \text{lorsque: } x_i^l \sim \mathcal{N}(0;\; 1) \end{cases}$$

Les trois suivantes sont corrélées mais interagissent plus faiblement avec $y_{\cdot}^{j}$:

$$\mathbb{P}\left(x_{\cdot}^{l} = (x_{\cdot}^{4}\,;\; x_{\cdot}^{5};\; x_{\cdot}^{6})\right) \sim \mathcal{B}(p = 0{,}7) = \begin{cases} p & \text{lorsque: } x_i^l \sim \mathcal{N}(0;\; 1) \\ (1-p) & \text{lorsque: } x_i^l \sim \mathcal{N}\left(y_i^j \cdot (l-3);\; 1\right) \end{cases}$$

Toutes les autres variables représentent un Bruit Blanc Fort (BBF), i.e. :

$$(x_{\cdot}^{7}\,;\; \ldots\;;\; x_{\cdot}^{p_l}) \sim \mathcal{N}(0; 1)$$

Jeux de données : Friedman.1

$\mathcal{L}_{\mathrm{n}_i=50}^{\mathrm{Friedman1}(50)} = \{\mathcal{X}^{\mathrm{Friedman1}}\} \cup \{\mathcal{Y}^{\mathrm{Friedman1}}\}$ a trait au Contexte statistique* de la régression, *i.e.* que : $y_{\cdot}^{\mathrm{Friedman.1}} \in \mathbb{R}^{\mathrm{n}_i}$. Les paramètres du générateur stochastique sont réglés de façon à générer un jeu de données à la frontière entre des *jeux statistiques classiques* et ceux *de grandes dimensions,* i.e. $\{\mathrm{n}_i = 50\}$ individus et $\{p_l = 50\}$ variables auxiliaires, et de sorte que :

Cinq variables interagissent avec la cible y_i^j. Elles sont tirées indépendamment dans des lois uniformes continues paramétrées de la façon suivante :

$$(x_{\cdot}^{1}\,;\; x_{\cdot}^{2};\; x_{\cdot}^{3}; x_{\cdot}^{4};\; x_{\cdot}^{5}) \sim \mathcal{U}([\![0; 1]\!])$$

La somme standardisée des autres variables auxiliaires de sorte que celle-ci puisse être assimilée à un Bruit Blanc Fort :

$$\frac{1}{\sqrt{p_l - 5}} \cdot \sum_{l=6}^{p_l} \left(x_{\cdot}^{l}\right) = \varepsilon_{\dot{x}} \sim \mathcal{N}(0; 1)$$

L'union de ces deux ensembles disjoints permet de créer la variable cible par une combinaison linéaire et d'attribuer plus ou moins de poids aux cinq variables explicatives, tel que :

$$y_{\cdot}^{\mathrm{Friedman.1}} = 10 \cdot \sin\left(\pi \cdot x_i^1 \cdot x_i^2\right) + 20 \cdot \left(x_i^3 - 0{,}5\right)^2 + 10 \cdot x_i^4 + 5 \cdot x_i^5 + \varepsilon_{\dot{x}}$$

Résultats renvoyés par MVG, analyse et remarques

Classification -ToyData : histogramme des fréquences, fonction de répartition empirique.

Régression - Friedman.1 : histogramme empirique, boite à moustache, graphique des valeurs intrinsèques.

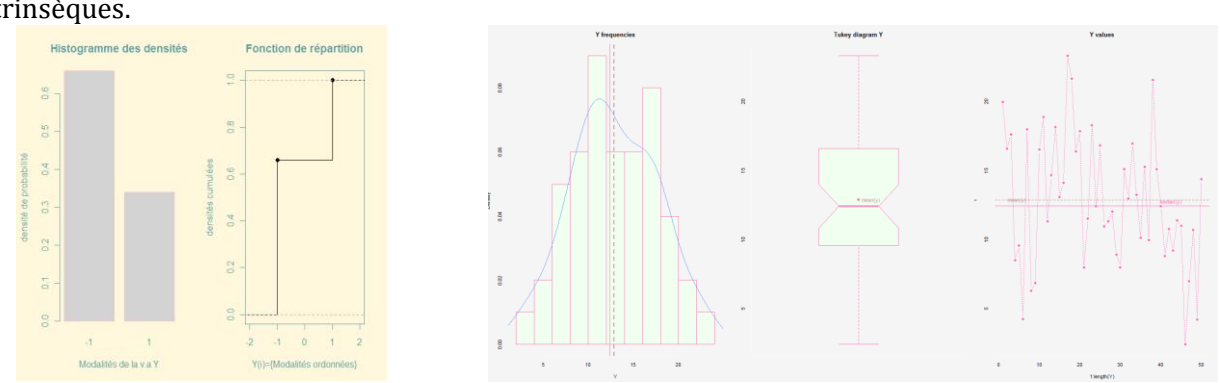


Figure 177 : Graphiques renvoyés par MVG en classification à gauche et en régression à droite

Analyse

En classification : les graphiques permettent de se faire une idée de l'incidence moyenne des pathologies.

En régression : les diagrammes permettent de juger de la variabilité de la cible, de l'amplitude des valeurs prises, de la présence d'éventuelles données aberrantes ou extrêmes.

Les graphiques sont agrémentés d'un item statistique renvoyant l'estimateur de la moyenne et l'écart-type en régression : $\widehat{\text{moy}}\left(y_i^j\right)$ et $\widehat{\sigma}\left(y_i^j\right)$. Et, de la proportion de modalité ainsi que son cardinal, en classification : $\hat{p}$ et $(n \cdot \hat{p})$

Remarques : Le cadre statistique des jeux de données géographiques est plus proche de $\mathcal{L}_{n_i=50}^{\text{Friedman1}(50)}$ que de $\mathcal{L}_{n_i=50}^{\text{DataToy}(100)}$. Bien que le nombre de dimensions soit encore inférieur à celle de $\mathcal{L}_{n_i=50}^{\text{Friedman1}(50)}$ la complexité qui macule les $\mathcal{L}_{U_k}^{j}$ est plus grande. Dans la mesure où MVG performe en classification, afin de se faire une meilleure idée de la qualité des résultats espérés, la complexité de $\mathcal{L}_{n_i=50}^{\text{DataToy}(100)}$ le positionne dans le cadre des jeux *de grandes dimensions*. La procédure MVG s'initialise par une phase de calibration.

PHASE 0 : CALIBRATION DES FA

Objectif

Optimiser les modèles de FA en calibrant les valeurs du vecteur $\Lambda_j = (\text{ntree} ; \text{mtry} ; \text{node.size})$

Synoptique du processus

Le paramètre nodesize est spécifié à partir de connaissances expertes en fonction du Contexte statistique*. Les paramètres ntree et mtry sont évalués numériquement de sorte qu'ils minimisent l'erreur OOB stabilisée $\overline{R}_{\cdot}^{OOB}(FA)$ - sur 10 forêts. L'étendue des valeurs possibles est d'abord balayée par l'ensemble Ω. genuer, puis affinée avec tuneRF.

Résultats renvoyés par MVG, analyse et remarques

L'axe des ordonnées représente la variation de $\overline{R}_{\cdot}^{OOB}(FA)$ en fonction du Log. des mtry qui est porté en abscisse. Les valeurs de ntree ne sont pas supérieures à 500 car elles n'amélioraient pas les résultats mais augmentaient les temps de calcul. Les lignes verticales en pointillés correspondantes à mtry.class et mtry.reg représentent les paramètres par défaut de Breiman.

Classification : ToyData

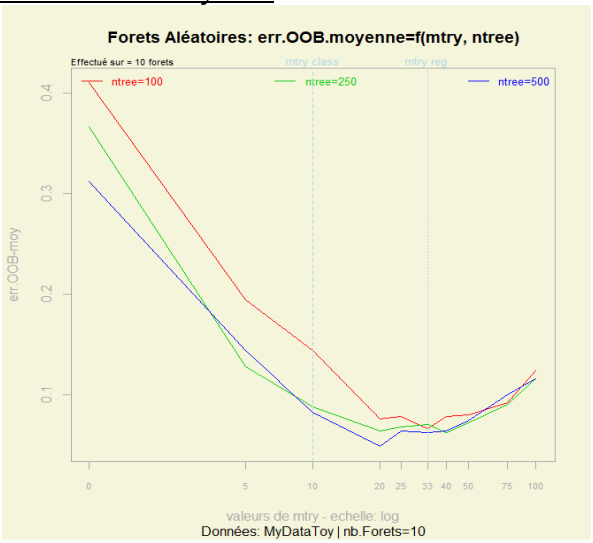


Figure 178 : Optimisation des valeurs de Λ_{ToyData}

Régression : Friedman.1

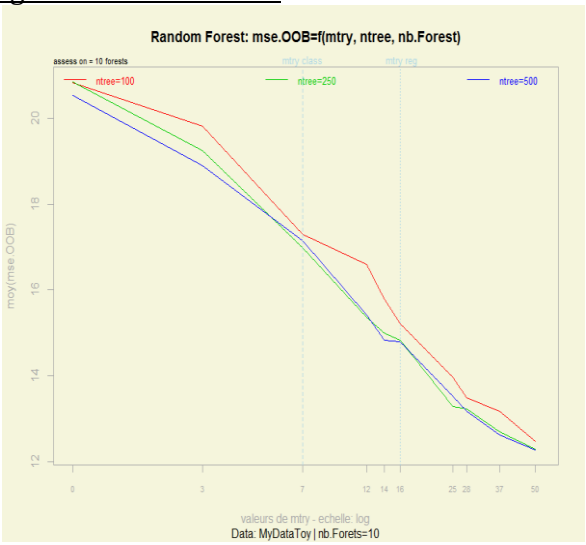


Figure 179 : Optimisation des valeurs $\Lambda_{\text{Friedman1}}$

Analyse et remarques
Les vecteurs permettant d'optimiser les modèles de FA sont :

Classification -ToyData : $\widehat{\Lambda}_{\text{DataToy}} = (\text{ntree} = 500\ ; \text{mtry} = 20; \text{nodesize} = 1)$
Classification -Régression : $\widehat{\Lambda}_{\text{Friedman1}} = (\text{ntree} = 500\ ; \text{mtry} = 50\ ; \text{nodesize} = 3)$

En classification comme en régression, les valeurs de mtry implémentées par défaut ne sont pas adaptées. Il convient de les calibrer.
La fonction tuneRF n'a pas permis d'affiner les résultats, seulement de les valider. Sa mise en œuvre est longue, elle ne sera plus utilisée par la suite.
Les $\overline{\text{R}}_{\cdot}^{\text{OOB}}$(FA) sont substantiellement diminuées - par rapport à des FA-naïves - les modèles de FA.opt peuvent être utilisés pour estimer des scores d'importance robustes.

PHASE 1 : ELIMINATION DES VARIABLES DE BRUIT

Objectifs
La première phase de MVG vise à valider deux objectifs : estimer les scores *randomForest** et procéder à une hiérarchisation descendante à partir des $\text{VI}_{\text{j,r}}(\text{x}^{\text{l}})$ stabilisés sur 50 forêts : $\overline{\text{VI}}_{\text{j}}(\text{x}^{\text{l}})$ afin d'obtenir les vecteurs $x_{\cdot}^{(l)}$, le second est de disjoindre $x_{\cdot}^{(l)}$ en deux paquets – i.e. celui des variables potentiellement explicatives $\mathcal{X}^{\text{j}}_{\text{conserv}}$ et celui des variables de bruit avéré $\mathcal{X}^{\text{j}}_{\text{bruit}}$

Synoptique du processus – 2 étapes
i. Estimation des $\overline{\text{VI}}_{\text{j}}(\text{x}^{\text{l}})$ et hiérarchisation descendante des variables : $x_{\cdot}^{(l)}$
ii. Estimation du seuil d'élimination de bruit $\psi^{\text{j}}_{\text{bruit}}$ et disjonction de $\mathcal{X}^{\text{j}}_{\text{conserv}}$ et de $\mathcal{X}^{\text{j}}_{\text{bruit}}$

Etape : 1 Résultats renvoyés par MVG, analyse et remarques
Remarques liminaires
Les valeurs des $\overline{\text{VI}}_{\text{j}}(\text{x}^{\text{l}})$ et $\text{VI}_{\text{j,r}}(\text{x}^{\text{l}})$ sont présentées sur des diagrammes de Tuckey, conformément à la proposition MVG.
Les x^{l} réellement explicatives sont les six premières variables situées à gauche des graphiques

Classification -ToyData :

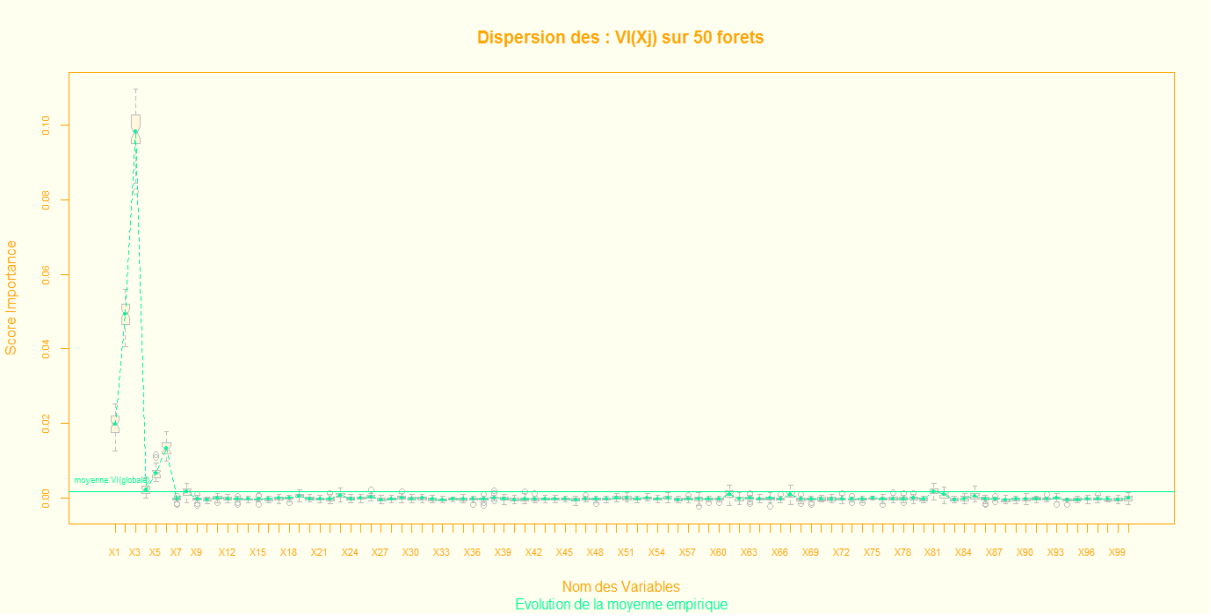


Figure 180 : Valeurs des $\overline{\text{VI}}_{\text{DataToy}}(\text{x}^{\text{l}})$ et variabilité des $\text{VI}_{\text{DataToy,r}}(\text{x}^{\text{l}})$, exprimées en fonction des x^{l}.

Régression - Friedman.1

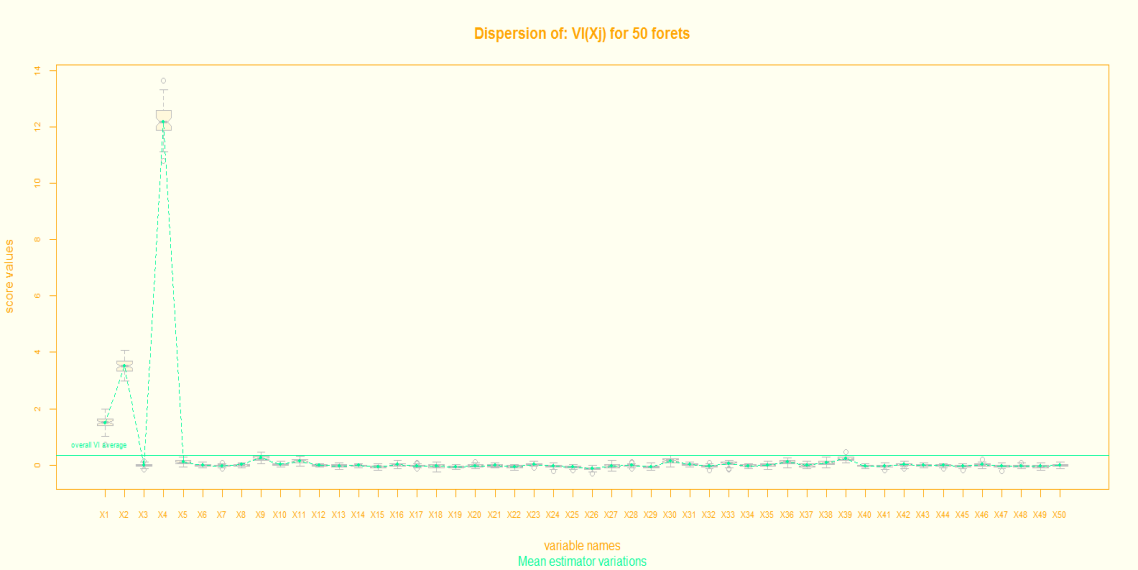


Figure 181 : Valeurs des $\overline{\mathrm{VI}}_{\mathrm{Friedman1}}(\mathrm{x}^{\mathrm{l}})$ et variabilité des $\mathrm{VI}_{\mathrm{Friedman1,r}}(\mathrm{x}^{\mathrm{l}})$ exprimées en fonction des x^{l}

Analyse et remarques

Classification -ToyData : les variables ($x_{.}^{1}$; $x_{.}^{2}$; $x_{.}^{3}$) entretiennent des interactions fortes avec $y_{.}^{\mathrm{ToyData}}$. Les variables ($x_{.}^{6}$; $x_{.}^{7}$; $x_{.}^{7}$) interagissant plus faiblement. Les interactions sont proportionnelles au rang des variables et sont détectées comme telles.

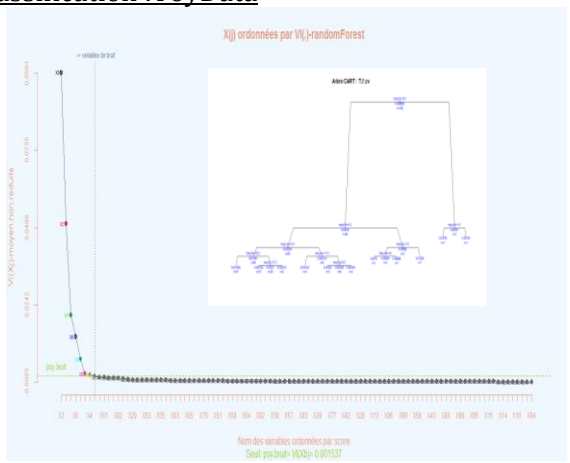
Régression – Friedman.1 : la variable ($x_{.}^{4}$) a une influence mathématique forte sur $y_{.}^{\mathrm{Friedman1}}$. Il en est de même, dans une mesure plus modérée, pour ($\mathrm{x}_{.}^{1}$; $\mathrm{x}_{.}^{2}$) qui sont corrélées entre elles. Les $\overline{\mathrm{VI}}_{\mathrm{Friedman1}}(\mathrm{x}^{\mathrm{l}})$ permettent de les détecter comme telles. Par contre, les variables explicatives ($\mathrm{x}_{.}^{3}$; $\mathrm{x}_{.}^{5}$) ont des scores proches de zéro alors qu'elles ont un pouvoir faiblement explicatif sur la cible.

Etape : 2 Résultats renvoyés par MVG, analyse et remarques

Remarques liminaires

Les $c_{\psi.}^{\mathrm{j}}$ ne sont pas utilisés. Le seuil $\psi_{\mathrm{bruit}}^{\mathrm{j}}$ est estimé par un arbre CART $\widehat{\mathrm{T}}_{\mathrm{MVG}}^{\mathrm{j}}((l))$ et sa structure est présentée. Les variables conservées sont mises en emphase et se situent à gauche de la ligne grise verticale pointillée. Les variables contenues dans $\mathcal{X}_{\mathrm{conserv}}^{\mathrm{j}}$ se trouvent à gauche de cette coupure et celles du paquet de bruit $\mathcal{X}_{\mathrm{bruit}}^{\mathrm{j}}$ se situent à droite.

Classification :ToyData



Régression : Friedman.1

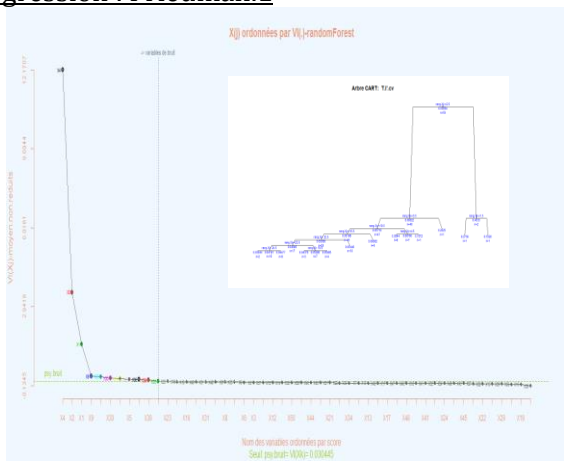


Figure 182 : Représentation graphique des valeurs : des $\overline{\mathrm{VI}}_{\mathrm{j}}(\mathrm{x}^{(\mathrm{l})})$ associées aux variables ordonnées $\mathrm{x}^{(\mathrm{l})}$, du seuil $\psi_{\mathrm{bruit}}^{\mathrm{j}}$ – ligne verte horizontale, disjonction de $\mathcal{X}_{\mathrm{conserv}}^{\mathrm{j}}$ $\mathcal{X}_{\mathrm{bruit}}^{\mathrm{j}}$ par la ligne grise verticale, et affichage de $\widehat{\mathrm{T}}_{\mathrm{MVG}}^{\mathrm{j}}((l))$

Analyse et remarque:

Classification -ToyData : le paquet $\mathcal{X}_{\text{conserv}}^{\text{ToyData}}$ se compose de six variables : ($x_{.}^{3}$; $x_{.}^{2}$; $x_{.}^{1}$; $x_{.}^{6}$; $x_{.}^{5}x_{.}^{4}$). Il s'agit des six variables explicatives et elles sont ordonnées exactement en fonction de leur influence sur $y_{.}^{\text{ToyData}}$. $\mathcal{X}_{\text{bruit}}^{\text{ToyData}}$ se compose des 94 variables de bruit.

Régression – Friedman.1 : onze variables sont conservées dans : $\mathcal{X}_{\text{conserv}}^{\text{Friendman.1}}$ et ordonnées de la façon suivante ($x_{.}^{4}$; $x_{.}^{2}$; $x_{.}^{1}$; $x_{.}^{9}$; $x_{.}^{39}x_{.}^{30}$; $x_{.}^{11}$; $x_{.}^{5}$; $x_{.}^{36}$; $x_{.}^{38}$; $x_{.}^{33}$) ; 4/5 sont des variables réellement explicatives, les 7 autres sont des variables de bruit $\mathcal{X}_{\text{bruit}}^{\text{Friedman.1}}$ se compose des 43 variables de bruit dont une est explicative.

Les graphiques des $\overline{\text{VI}}_{\text{j}}(\text{x}^{\text{l}})$ permettent de mettre en évidence des variables ayant une efficience statistique forte sur x^{j}. La phase d'élimination semble être plus performante en classification qu'en régression – surtout que ToyData est un jeu de données en grandes dimensions.

PHASE 2 : IDENTIFICATION DES VARIABLES EXPLICATIVES

Objectif

Scinder $\mathcal{X}^{j}_{conserv}$ en deux ensembles disjoints $\mathcal{X}^{j}_{explic}$ contenant les $x_{.}^{(1)}$ explicatives, au sens de *randomForest*,* et $\mathcal{X}^{j}_{elim}$ contenant celles de bruit ou faiblement efficientes.

Synoptique du processus

Construction des modèles de FA imbriqués à partir de $\mathcal{X}^{j}_{conserv}$, estimation des $\overline{R}^{OOB}_{(\mathcal{X}^{j}_{imbr.l})}$ et des $\widehat{\sigma}_{\left(R^{OOB}_{(imbr.l)}\right)}$ stabilisés sur 10 run, estimation de ψ^{j}_{explic} par la règle du *min + 1SD*, identification des paquets explicatifs de Génuer $\mathcal{X}^{j}_{explic}$ et de Bourrelly $\mathcal{X}^{j}_{0.explic}$.

Résultats renvoyés par MVG, analyse et remarques

Remarques liminaires

En ordonnée sont portées les $\bar{R}^{OOB}_{(.)}$. En abscisse le nombre de variables contenues dans $\mathcal{X}^{j}_{imbr.l}$. Et, sur le graphique la variable en vert est la dernière imbriquée.

ψ^{j}_{explic} est matérialisé par une ligne verte de pointillés. Le mîn$\left(\bar{R}^{OOB}_{(.)}\right)$ par une ligne verte claire de traits discontinus.

La coupure du modèle imbriqué formé de $\mathcal{X}^{j}_{explic}$ est une ligne blanche. Celle de $\mathcal{X}^{j}_{0.explic}$ est matérialisée par une ligne violette en pointillés.

Classification -ToyData

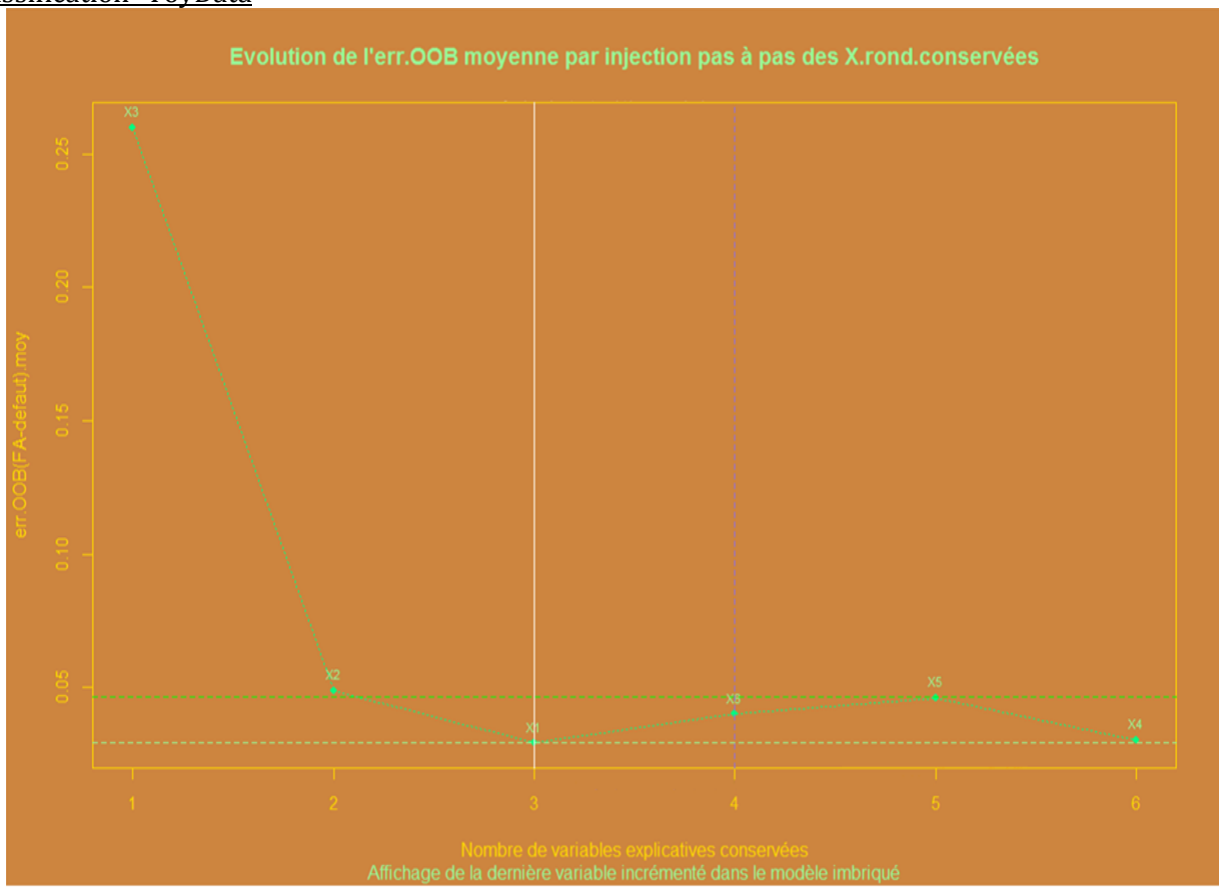


Figure 183 : Représentation de la variation de $\bar{R}^{OOB}_{(.)}$ en fonction du nombre de variables imbriquées $\mathcal{X}^{ToyData}_{imbr.l} \subset \mathcal{X}^{ToyData}_{conserv}$

Analyse

Le paquet de Génuer $\mathcal{X}_{\text{explic}}^{\text{ToyData}}$ contient $(x_{.}^{3}; x_{.}^{3};\ x_{.}^{1})$; les 3 variables interagissant fortement. Le paquet Bourrelly $\mathcal{X}_{0.\text{explic}}^{\text{j}} = (x_{.}^{3}; x_{.}^{3};\ x_{.}^{1}; x_{.}^{6})$ comporte les trois mêmes et $x_{.}^{6}$ qui interagit aussi mais plus faiblement. Dans cette configuration des choses $\mathcal{X}_{0.\text{explic}}^{\text{ToyData}}$ est plus pertinent que $\mathcal{X}_{\text{explic}}^{\text{ToyData}}$

Régression – Friedman.1

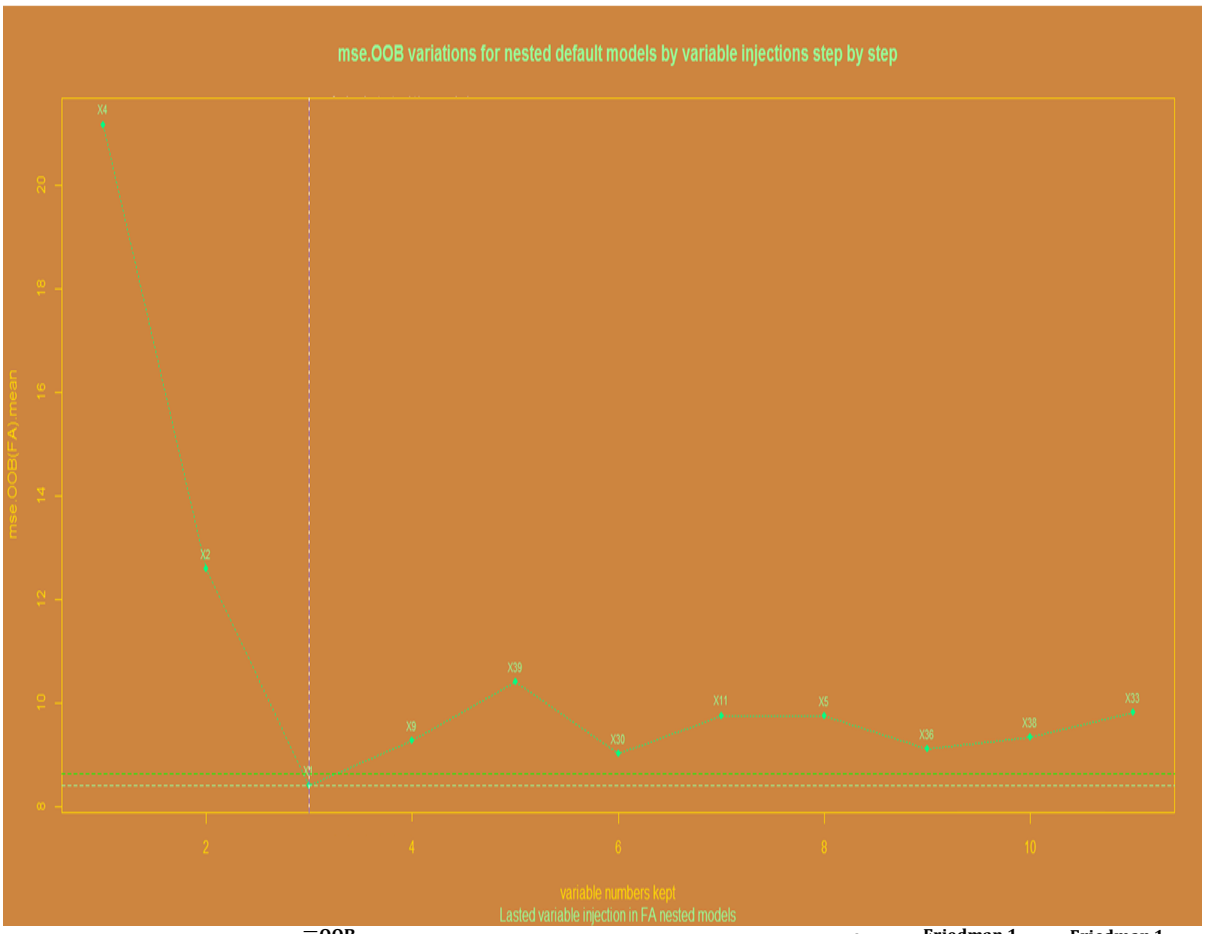


Figure 184 : Variation de $\bar{R}_{(.)}^{OOB}$ en fonction du nombre de variables imbriquées : $\mathcal{X}_{\text{imbr.l}}^{\text{Friedman.1}} \subset \mathcal{X}_{\text{conserv}}^{\text{Friedman.1}}$

Analyse

Le paquet de Génuer $\mathcal{X}_{\text{explic}}^{\text{Friedman1}}$ est identique à celui de Bourrelly $\mathcal{X}_{0.\text{explic}}^{\text{Friedman1}}$, Les deux contiennent les variables explicatives $(x_{.}^{4}; x_{.}^{2};\ x_{.}^{1})$ - trois variables qui interagissant fortement avec $y_{.}^{\text{Friedman1}}$.

Remarques

En classification c'est le paquet Bourrelly qui a été retenu et testé $\mathcal{X}_{0.\text{explic}}^{\text{j}} = (x_{.}^{3}; x_{.}^{3};\ x_{.}^{1}; x_{.}^{6})$.

Généralement le paquet de Génuer $\mathcal{X}_{\text{explic}}^{\text{j}}$ est plus restreint que le paquet de Bourrelly $\mathcal{X}_{0.\text{explic}}^{\text{j}}$. Cependant, $\mathcal{X}_{0.\text{explic}}^{\text{j}}$ n'est pas forcément plus pertinent que $\mathcal{X}_{\text{explic}}^{\text{j}}$. En revanche lorsqu'on applique MVG à des données réelles, dans le cadre d'une problématique de santé publique abordée par le prisme de la géographie, il convient d'identifier tous les leviers possibles permettant d'influencer le PM* étudiés. Quitte à discuter *a posteriori* à l'appui de connaissances expertes, de leur influence sanitaire potentielle et leur caractère onéreux. Avant de passer à la phase suivante il est nécessaire de choisir l'un ou l'autre de ces deux paquets

PHASE 3 : SELECTION DES VARIABLES PREDICTIVES

Objectif

Extraire de $\mathcal{X}^{j}_{explic}$ ou $\mathcal{X}^{j}_{0.explic}$ un paquet de variables *s'adaptant à la parcimonie prédictive* $\mathcal{X}^{j}_{pred}$

Synoptique du processus

Estimation du seuil d'élimination de bruit ψ^{j}_{pred}. Elimination par injection séquentielle des $x^{(l)}_{.}$ du paquet explicatif, dans des modèles de FA par défaut, Celles qui n'améliorent pas significativement l'erreur OOB $\overline{\Delta}(R^{OOB}_{.})$ sont éliminées, i.e. que $\overline{\Delta}(R^{OOB}_{.}) \leq \psi^{j}_{pred}$. Les valeurs sont stabilisées sur 50 forêts.

Résultats renvoyés par MVG, analyse et remarques

Remarques liminaires

Le seuil ψ^{j}_{pred} est matérialisé, sur le graphique, par une ligne rouge discontinue. Les $\overline{\Delta}(R^{OOB}_{l})$ sont représentés par des barres verticales vertes fluo. La $x^{(l)}_{.}$ testée est portée en abscisse. A chaque itération le modèle retenu $\mathcal{X}^{j}_{pred}$documente la barre par un vecteur inscrit en blanc

Classification : ToyData

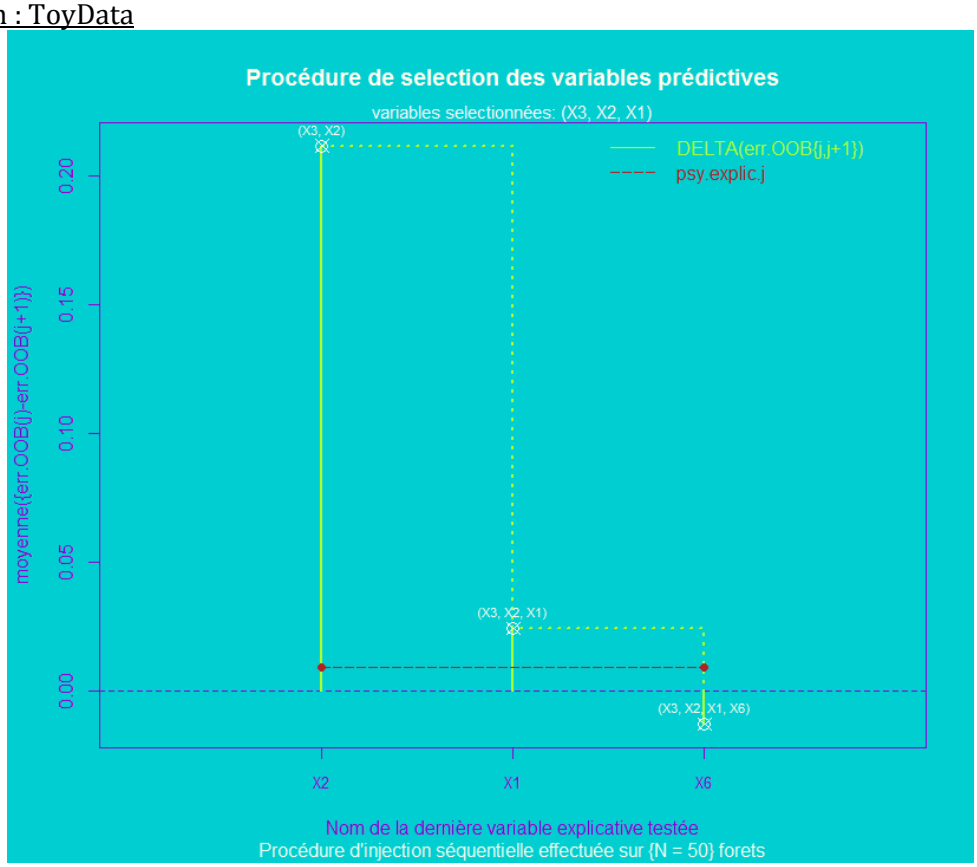


Figure 185 : Représentation graphique des valeurs des $\overline{\Delta}(R^{OOB}_{l})$ en fonction de la variable testée $x^{l} \subset \mathcal{X}^{ToyData}_{explic}$

Analyse et remarques:

le paquet : $\mathcal{X}^{ToyData}_{pred} = (x^{3}_{.}; x^{2}_{.}, x^{1}_{.})$, la variable $x^{6}_{.}$ est éliminée. Il est identique à $\mathcal{X}^{ToyData}_{explic}$, et contient toutes les variables interagissant fortement avec la réponse.

Régression : Friedman.1

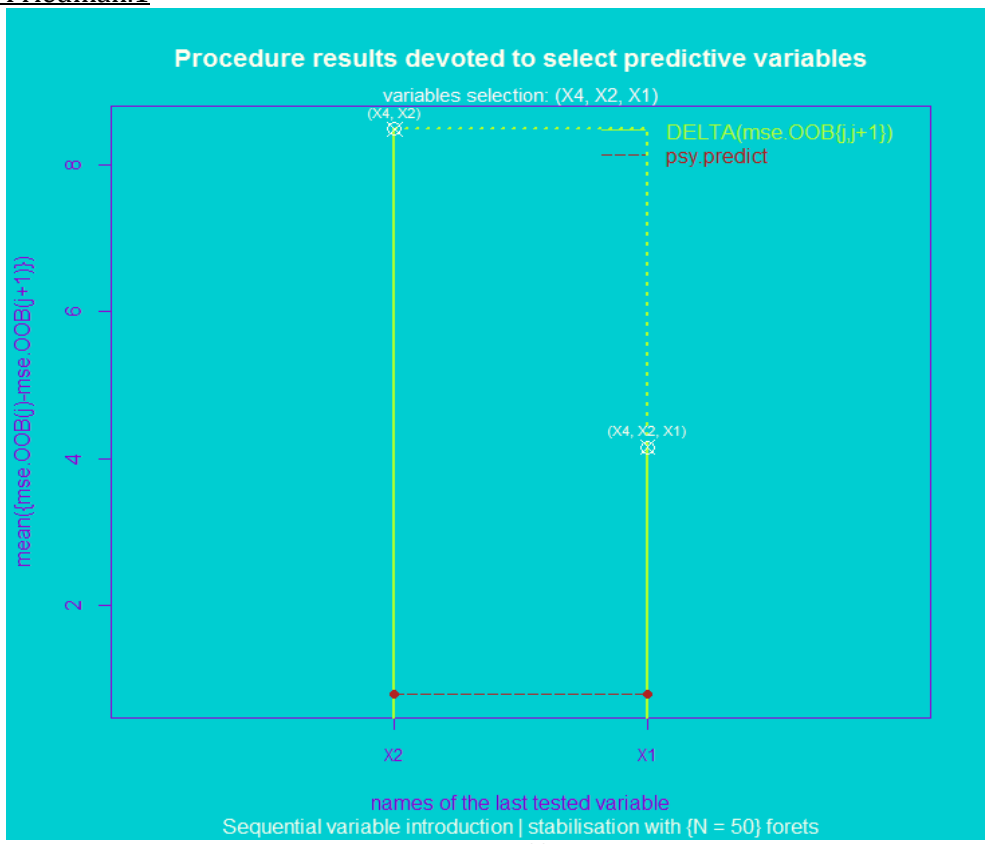


Figure 186 : Représentation graphique des valeurs des $\overline{\Delta}(\mathbf{R}_{\mathbf{l}}^{\mathbf{OOB}})$ en fonction de la variable testée $\mathbf{x}^{\mathbf{l}} \subset \mathcal{X}_{\mathbf{explic}}^{\mathbf{Friedman.1}}$

Analyse et remarques:
le paquet $\mathcal{X}_{\text{pred}}^{\text{Friedman1}} = (x_{\cdot}^{4};\ x_{\cdot}^{2}, x_{\cdot}^{1})$, toutes les variables sont conservées. Il s'agit des trois variables dont les interactions mathématiques sont les plus fortes.

La procédure de construction de $\mathcal{X}_{\text{pred}}^{\text{j}}$ est sensible aux variables contenues dans le paquet explicatif testé. Il suffit de modifier légèrement sa structure et les résultats sont complètement différents.
Aussi, comme il s'agit d'une procédure step by step, la valeur de $\psi_{\text{pred}}^{\text{j}}$ affecte fortement l'architecture de $\mathcal{X}_{\text{pred}}^{\text{j}}$. De fait l'utilisation de $c_{\psi.\text{pred}}^{\text{j}}$ est lourde de conséquences.

ESTIMATION DE LA QUALITE STATISTIQUE DES RESULTATS

Remarques liminaires
L'algorithme MVG permet de construire un modèle de FA à partir, soit de $\mathcal{X}_{0.\text{explic}}^{\text{j}}$, soit de $\mathcal{X}_{\text{explic}}^{\text{j}}$ permettant d'expliquer la variabilité d'une cible y_{i}^{j}. MVG permet aussi de créer un modèle de FA – à partir de$\mathcal{X}_{\text{pred}}^{\text{j}}$ permettant de les prédire par des $\hat{y}_{\text{i}}^{\text{j}}$.
Comme les données d'apprentissage - In-Sample (IS) - $\mathcal{L}_{n_i=50}^{\text{DataToy}(100)}$; $\mathcal{L}_{n_i=50}^{\text{Friedman1}(50)}$ sont construits à partir de générateurs stochastiques, alors des échantillons test – Out-Of-Sample (OOS) - : $\mathcal{L}'_{n_i=50}^{\text{DataToy}(100)}$; $\mathcal{L}'_{n_i=50}^{\text{Friedman1}(50)}$ peuvent être construits – ils contiennent de nouveaux individus qui sont générés à partir de règles stochastiques néanmoins identiques. Dans ce cas l'erreur de généralisation est estimée In-Bag (IB), i.e. sur tous les arbres de la FA puisque aucun des nouveaux individus n'a servi à les construire.

Objectif

Analyser la robustesse explicative et prédictive des modèles de FA afin d'évaluer la qualité des résultats de MVG, et par extension subodorer sa fiabilité dans le cadre de son application aux jeux de données géographiques morbides : $\mathcal{L}^{j}_{U_k}$

Synoptique d'un processus en temps

1. Estimation de la capacité explicative: la qualité des modèles de FA est évaluée OOB, sur $\mathcal{L}^{\mathrm{DataToy(100)}}_{n_i=50}$; $\mathcal{L}^{\mathrm{Friedman1(50)}}_{n_i=50}$- IS, par $\overline{R}^{OOB-IS}\left(\hat{f}^{j}_{FA}(\mathcal{X}^{j}_{.}|\Lambda_j)\right)$ et selon le contexte : En classification par l'err.OOB est la proportion d'individus mal classés. En régression par la mse.OOB - l'estimateur de la variance - convertie en pourcentage de variance expliquée pour faciliter l'interprétation : $\mathbb{var}.\mathbb{explain}^{OOB-IS}_{\hat{f}^{j}_{FA}(\mathcal{X}^{j}_{.}|\Lambda_j)}$.

2. Estimation de la qualité prédictive: La qualité des prédiction est évaluée IB sur: $\mathcal{L}^{\mathrm{DataToy.new(100)}}_{n_i=50\ |\ new}$; $\mathcal{L}^{\mathrm{Friedman1(50)}}_{n_i=50\ |\ new}$ - OOS, par $\overline{R}^{IB-OOS}\left(\hat{f}^{j}_{FA}(\mathcal{X}^{j}_{.}|\Lambda_j)\right)$, selon le contexte, en classification par l'err.IB, en régression par la mse.IB - convertie en $\mathbb{var}.\mathbb{explain}^{IB-OOS}_{FA.}$
Ces analyses sont effectuées sur les différents modèles de FA utilisés tel que :

Nom du modèle	FA.naïve	FA.opt	FA.explic	Bagging CART.explic	FA.pred	Bagging CART.préd
Paramètres	$\dot{\Lambda}_j$	$\widehat{\Lambda}_j$	$\dot{\Lambda}_j$	$\ddot{\Lambda}_j$	$\dot{\Lambda}_j$	$\ddot{\Lambda}_j$
Variables utilisées	$\left\{\mathcal{X}^{j}_{input} \cup \mathcal{X}^{j}_{new}\right\}$		$\left\{\mathcal{X}^{j}_{explic} \cup \mathcal{X}^{j}_{explic.new}\right\}$		$\left\{\mathcal{X}^{j}_{pred} \cup \mathcal{X}^{j}_{pred.new}\right\}$	

Tableau 37 : Noms, paramètres et paquets de variables utilisées pour les différents modèles de FA utilisés

Avec : $\dot{\Lambda}_j$ le vecteur des paramètres par défaut implémentés dans randomForest*, $\widehat{\Lambda}_j$ le vecteur des paramètres optimisés, $\ddot{\Lambda}_j$ le vecteur des paramètres par défaut sauf lorsque vaut mtry $= 1$, la valeur est portée à 2

Analyse et remarque sur les résultats renvoyés par MVG

Les résultats présentés ont été stabilisés sur 10 forêts. Dans la mesure où le Bagging-CART n'améliore pas l'erreur de généralisation, ils ne sont pas rappelés dans l'estimation OOS.

Classification -ToyData

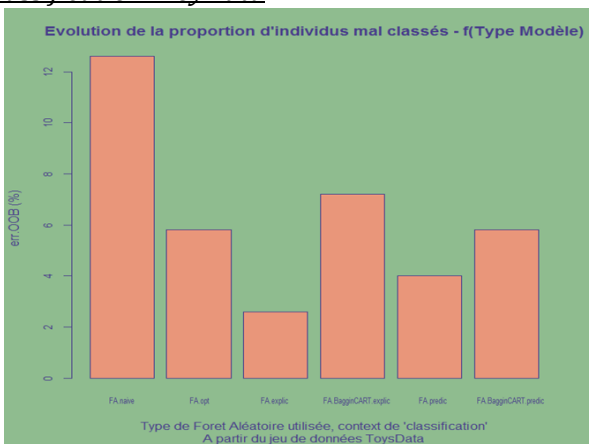


Figure 187 : Analyse OOB de la qualité explicative des modèles MSV estimée IS

Figure 188 : Analyse IB de la qualité prédictive des modèles MSV estimée OOS

Les FA.naïves fournissent des estimateurs assez bons puisque l'err.OOB=13%. Les FA.opt diminuent de moitié l'err.OOB=6%. Les FA explicatives performent en divisant encore par deux l'err.OOB (3%) par rapport aux FA optimisées. Comme l'analyse est IS, les FA.pred ont une qualité inférieure aux FA.explic avec err.OOB=4%. Les Bagging CART explicatifs ou prédictifs fournissent de moins bons résultats que les FA associées.

Les FA.naïves engendrent une err.IB=11%, curieusement inférieure à l'err.OOB estimée IS. les FA.opt font à peine mieux : err.IB=9%. Par contre, les FA.explic ont une err.IB=6%, presque deux fois plus faible que celle des FA.naïves. Finalement comme prévu les FA prédictives performent sur toutes les autres, avec une err.IB=4% - équivalente à celle estimée *Out-Of-Sample*.

Régression –Friedman.1

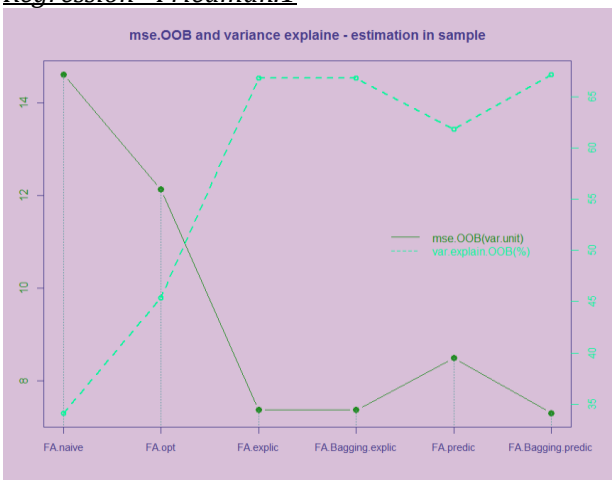


Figure 189 : Analyse OOB de la qualité explicative des modèles MSV estimée IS

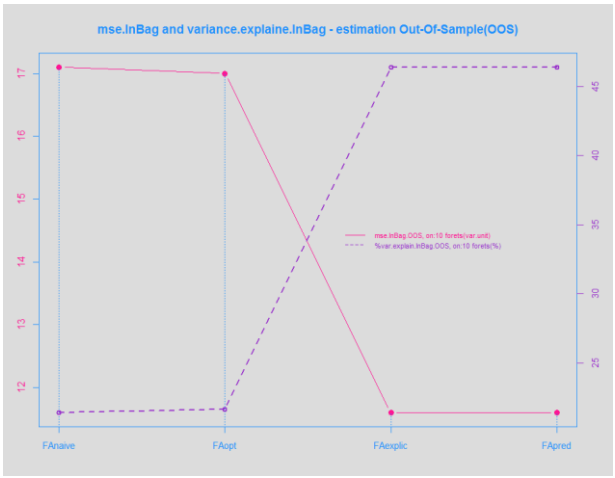


Figure 190 : Analyse IB de la qualité prédictive des modèles MSV estimée OOS

Les FA.naïves expliquent moins de 35% de la var.OOB. Quant aux FA.opt elles ont une var.explain.OOB qui atteint 45%. Les FA.explic, le Bagging.CART.explic, les FA.pred et le Bagging.CART.pred, expliquent environ 63% de la var.OOB, car $\mathcal{X}^{j}_{explic}$ et $\mathcal{X}^{j}_{pred}$ sont identiques. Le Bagging.CART. performe sur les FA car les paquets sont petits donc les paramètres engendrent des *FA complètement randomisées*, au sens ou l'entend Breiman.

Les FA.naïves expliquent moins de 20% de la var.IB. Les FA.opt, qui correspondent dans ce cas à un Bagging.CART atteignent tout juste une var.explain.IB=20%. Les modèles FA.explic et FA.pred se montrent concurrentiels avec une var.explain.IB qui vaut presque 50%, car $\mathcal{X}^{j}_{explic}$ et $\mathcal{X}^{j}_{pred}$ sont presque identiques. Les paramètres devraient engendrer des *FA complètement randomisées*, cependant mtry a été portée à 2.

REMARQUES GENERALES

Remarques heuristiques:
VSURF est une méthode mathématique aux performances prometteuses pour expliquer et prédire des phénomènes complexes sur lesquels il est possible d'acquérir des quantités d'informations très importantes, de nature hétéroclite, et que les modèles statistiques classiques sont incapables d'analyser (Genuer, 2010).

Des résultats analogues très encourageants ont été obtenus sur des jeux de données réelles dont les mécanismes sont connus (Biau, Lugosi et al., 2008)
VSURF présente l'avantage d'être à la fois très simple et très efficace. Elle a fait ses preuves et a séduit de nombreux experts dans le cadre de problématiques complexes d'apprentissage dans le champ de la santé et sur des jeux de données *en grandes dimensions* (Genuer, 2010).

Remarques et perspectives d'application en géographie de la santé:
L'algorithme MyVsurfGéo* (MVG) a été appliqué à un autre jeu de données de type DataToy $\mathcal{L}_{\mathrm{n}_i=50}^{\mathrm{DataToy\ (50)}}$ i.e. 50 individus pour 50 variables dont 6 explicatives. Le cadre statistique est plus proche de celui inhérent à la sélection des variables à partir des i.st.e* représentatifs des FE/FIM. Les résultats obtenus sont encore plus encourageants que ceux présentés pour $\mathcal{L}_{\mathrm{n}_i=50}^{\mathrm{DataToy\ (100)}}$. En effet le paquet $\mathcal{X}_{\mathrm{explic}}^{\mathrm{DataToy(50)}} = \left(x_{.}^{3}\,;x_{.}^{2}\,;x_{.}^{1}\,;x_{.}^{6}\,;x_{.}^{5}\right)$ compte cette fois cinq variables explicatives, non pas seulement trois.
S'agissant des capacités prédictives, le modèle de FApred performe sur tous les autres pour l'analyse de la qualité menée *Out-Of-Sample.* L'analyse des prédictions a conduit à une erreur IB estimée à 2% sur $\mathcal{L}_{\mathrm{n}_i=50|\mathrm{new}}^{\mathrm{DataToy\ (50)}}$ - donc diminuée de moitié par rapport à celle estimée sur un jeu en *grandes dimensions*.
Les résultats de l'analyse des modèles et plus particulièrement des FA.explic et des FA.pred offrent des perspectives très prometteuses de l'application de MVG à des données réelles.
L'algorithme MVG semble parfaitement adapté pour caractériser les interactions statistiques entre la géographie des séquelles, modélisée par les i.st.m* : $\mathrm{z'}^{\mathrm{j}}_{(\mathrm{U}_k),c}$ et $\mathrm{z'}^{\mathrm{j}}_{(\mathrm{U}_k),c}$ (chapitre.2) et celle des FE/FIM, modélisée par les i.st.e* $x^{\mathrm{l}}_{(\mathrm{U}_k)}$ et les $x'^{\mathrm{l}}_{(\mathrm{U}_k)}$ (chapitre.3).
L'algorithme MVG est adapté à la complexité des jeux de données géographiques. En effet les $\mathcal{L}_{\mathrm{U}_k}^{\mathrm{Sequelle.j}}$ mélangent des $x^{\mathrm{l}}_{(\mathrm{U}_k)}$ dont la nature, les unités, la précision et les incertitudes sont disparates, ce qui complexifie l'identification des DES*. Par conséquent, l'analyse des variables contenues dans le paquet Bourrelly : $\mathcal{X}^{\mathrm{j}}_{0.\mathrm{explic}}$ trouvera tout son sens. Cependant, celui de Génuer $\mathcal{X}^{\mathrm{j}}_{\mathrm{explic}}$ qui sera systématiquement utilisé lors du passage à la phase :3. En effet, il s'agira non pas tant d'inventer des i.st.e* explicatifs mais plutôt d'analyser les $x^{\mathrm{l}}_{(\mathrm{U}_k)}$ redondants supplétifs (section.B).
En outre l'analyse géographique des $\mathcal{X}^{\mathrm{j}}_{\mathrm{pred}}$ semble adaptée à l'identification des leviers permettant statistiquement d'interagir avec les PM, et par conséquent d'identifier des DES* prégnants. En contrepartie l'étude des i.st.e* contenus dans $\mathcal{X}^{\mathrm{j}}_{\mathrm{bruit}}$ permettra de spécifier la géographie des FE/FIM* sur lesquels il est inutile de se focaliser d'un point de vue médical (i.e. à l'échelle individuelle) et politique (i.e. à l'échelle communautaire).

Enfin, au regard des résultats obtenus, la stratégie de sélection semble être plus performante dans le contexte de la classification que dans celui de la régression. Ce qui signifie que les analyses géographiques des $\mathrm{z'}^{\mathrm{j}}_{(\mathrm{U}_k),q}$ seront censément plus fiables que celles effectuées sur les $\mathrm{z'}^{\mathrm{j}}_{(\mathrm{U}_k),c}$. Cependant, cette hypothèse ne peut pas être validée par cette unique application comparative.

SECTION B) APPROCHE GEOGRAPHIQUE

L'algorithme MyVsurfGeo (MVG) est appliqué aux i.st.e. $x^{l}_{(\mathrm{U}_k)}$ représentatifs de la géographie des FE/FIM* intégrés (chapitre.3). L'objectif est d'identifier ceux qui interagissent statistiquement avec : les prévalences spatiales patients-années pondérées EpiGéoStat $z^{j}_{(\mathrm{U}_k),c}$ – contexte de régression - et les propensions spatiales patients-années pondérées EpiGéoStat $z^{j}_{(\mathrm{U}_k),q}$ – contexte de classification. Dans un premier temps MVG est appliqué dans l'optique de confectionner des paquets de $x^{l}_{(\mathrm{U}_k)}$ explicatifs et prédictifs. Ils sont ensuite analysés conjointement afin d'identifier les : Déterminants Environnementaux de Santé* (DES), Facteurs de Risques Environnementaux Contributifs* (FREC), ou Probablement aggravants (FREPA) des états de santé observés : CATA, THYR, TUM2. Ensuite le modèle de FA.MVG prédit la géographie des séquelles, par le biais de : $\hat{z}^{j}_{(\mathrm{U}_k),c}$ et $\hat{z}^{j}_{(\mathrm{U}_k),q}$. L'analyse de la géographie : des résidus $\hat{\varepsilon}^{\mathrm{j}}_{(\mathrm{U}_k)}$ – en régression, et des confusions $\hat{\varsigma}^{\mathrm{j}}_{(\mathrm{U}_k)}$ - en classification, permet d'estimer la qualité des prédictions MVG et d'évaluer la pertinence des DES, FREC et FREPA liés aux séquelles.

STRATEGIE D'APPLICATION DE MYVSURFGEO AUX SEQUELLES

L'algorithme MVG est appliqué d'abord sur $z^{\mathrm{j}}_{(\mathrm{U}_k),c}$, puis sur $z^{\mathrm{j}}_{(\mathrm{U}_k),q}$ associées à chaque séquelle. Les $\mathcal{L}^{\mathrm{j}}_{\mathrm{U}_k}$ se compose d'un sous-ensemble de variables potentiellement explicatives $x^{\mathrm{l}}_{(\mathrm{U}_k)}$, spécifiques à chaque séquelle. L'analyse MVG dépend du Contexte statistique*. L'application VSURF est conditionnée par des paramètres. Ceux utilisés dans l'algorithme MVG sont spécifiés pour les différentes phases :

- Calibration : $\widehat{\Lambda}_{\mathrm{j}}$;
- Hiérarchisation - $\overline{\mathrm{VI}}_{\mathrm{j}}(\mathrm{x}^{\mathrm{l}})$ et élimination - $\mathcal{X}^{\mathrm{j}}_{\mathrm{bruit}}$.
- Identification des variables explicatives - $\mathcal{X}^{\mathrm{j}}_{\mathrm{explic}}$, $\mathcal{X}^{\mathrm{j}}_{0.\mathrm{explic}}$.
- Sélection des variables prédictives - $\mathcal{X}^{\mathrm{j}}_{\mathrm{pred}}$.
- Prédiction des i.st.m* : $\hat{\mathrm{z}}^{\mathrm{j}}_{(\mathrm{U}_k),\mathrm{c}}$ et $\hat{\mathrm{z}}^{\mathrm{j}}_{(\mathrm{U}_k),\mathrm{q}}$, à partir du modèle de FA.MVG

RAPPEL SUR LES INDICATEURS SPATIOTEMPORELS UTILISES

La géographie des FE/FIM* intégrés est modélisée par 64 i.st.e. qui représentent la variabilité spatiale des expositions à des *FE* pertinents** - discrétisés en 4 composantes, chacune associée à un code couleur les : FIM, 8 i.st.e* pondérés EpiGéoStat $\mathrm{x}'^{\mathrm{l:FIM}}_{(\mathrm{U}_k)}$ FE-SAN, 15 i.st.e* pondérés temporellement $\mathrm{x}'^{\mathrm{l:SAN}}_{(\mathrm{U}_k)}$; FE-SOCIO.ECO,17 i.st.e* $\mathrm{x}^{\mathrm{l:SOCIO.ECO}}_{(\mathrm{U}_k)}$; FE-PHY.CHIM, 23 i.ste $\mathrm{x}^{\mathrm{l:PHY.CHIM}}_{(\mathrm{U}_k)}$ dont un bruité par des informations spatiotemporelles : $\mathrm{x}''^{\mathrm{RADON}}_{(\mathrm{U}_k)}$. Par ailleurs, 6 $\mathrm{x}^{\mathrm{l}}_{(\mathrm{U}_k)}$ sont utilisés pour modéliser la variabilité des expositions à des *FE* Curieux* et* intégrés pour tester la robustesse de MVG, il s'agit pour les : FIM, $\mathrm{x}'^{\mathrm{ACPHY}}_{(\mathrm{U}_k)}$; FE-SAN, des $\mathrm{x}^{\mathrm{l:SAN}}_{(\mathrm{U}_k)}$ sans rapport *avec la séquelle* ; FE-SOCIO.ECO, $\mathrm{x}^{\mathrm{att.BIENS}}_{(\mathrm{U}_k)}$; $\mathrm{x}^{\mathrm{att.PHY}}_{(\mathrm{U}_k)}$; $\mathrm{x}^{\mathrm{INSECU}}_{(\mathrm{U}_k)}$; FE-PHY.CHIM, $\mathrm{x}^{\mathrm{FEFO}}_{(\mathrm{U}_k)}$; $\mathrm{x}^{\mathrm{PREV}}_{(\mathrm{U}_k)}$.

CHOIX DES* VARIABLES CONSTITUTIVES DE $\mathcal{X}^{\mathrm{j}}_{\mathrm{input}}$

Objectif

Constituer pour chaque séquelle un $\mathcal{X}^{\mathrm{j}}_{\mathrm{input}}$ pertinent et adapté à l'analyse MVG

Remarques liminaires

Notations : par la suite les i.st.e* bruités $\mathrm{x}'^{\mathrm{l}}_{(\mathrm{U}_k)}$ seront notés x._l ; Les autres x_l.

Certains i.st.e* environnementaux sont redondants, c'est le cas par exemple des FE-PHY.CHIM* météorologiques - x_RAY, x_TEMP, x_NJAP - d'exposition à la radioactivité environnementale x_238Pu,

x_137Cs-sol ; x_3H, x_ALPHA ; des FE-SAN* d'accès aux EML*, i.e. : x_IRM, x_SCAN… et aux Services hospitaliers : x._HEMAs ; x._ORLs… (chapitre.3).
Les i.st.e* redondants sont généralement sur-représentés : en l'occurrence les 9 i.st.e* $x_{(U_k)}^{l:RNM}$ et les 8 $x_{(U_k)}^{l:SAN}$ modélisant l'accès aux plateaux techniques des établissements de santé.
Certains i.st.e* ont été intégrés pour des séquelles spécifiques, i.e. : x_APL-OPHT, x_RAY, x_OPHT.s pour les CATA. Ou x_131i et x._ENDO.s pour les THYR…

Hypothèses et propositions heuristiques

Certains i.st.e* vraiment aberrants ou redondants, au regard de la séquelle considérée, peuvent être supprimés pour préserver *la puissance statistique* des FA (Saporta, 2006).
La suppression d'i.st.e, différents selon la séquelle, permet d'éluder le risque de dépendance du modèle à l'ordre des variables afin de garantir une meilleure consistance statistique des résultats (Somol, Pudil et al., 1999).
Cependant certains i.st.e* Curieux* de test doivent être conservés pour évaluer la capacité de MVG à éliminer les variables de bruit, sélectionner les variables explicatives, puis supprimer les redondances du modèle de FA.pred (Genuer, 2010).

Proposition heuristique

Le nombre d'i.st.e* introduits à l'initialisation dans les $\mathcal{L}_{U_k}^{j}$ est fixé à 62/70. Les $x_{(U_k)}^{l}$ diffère selon la séquelle mais reste identique quel que soit le Contexte statistique*

PHASE D'ANALYSE DES* DONNEES

En régression - sur $z_{(U_k),c}^{j}$ - MVG renvoie : trois graphiques (section.A), deux items statistiques : môy $\left(z_{(\cdot),c}^{j} \right)$ et $\widehat{\sigma}\left(z_{(\cdot),c}^{j} \right)$. Afin d'augmenter l'intelligibilité des résultats les $z_{(U_k),c}^{j}$ sont multipliés par 1000.

En classification - sur $z_{(U_k),q}^{j}$ - MVG renvoie : deux graphiques (section.A), deux items, i.e. le nombre des valeurs associé à chaque modalité - card $\left(z_{(U_k),q}^{j} = c_j \right)$, et la proportion moyenne - môy $\left(z_{(U_k),q}^{j} = c_j \right)$. Aucune transformation n'est appliquée aux $z_{(U_k),q}^{j}$.

PHASE 0 - CALIBRATION:

Objectif : optimiser le vecteur $\Lambda_j = (\text{ntree} ; \text{mtry}; \text{nodesize})$ en minimisant la $\overline{R}_{\cdot}^{OOB}(FA)$ stabilisée sur 10 forêts. L'estimation de $\widehat{\Lambda}_j$ permet la transformation : FA.naïve → FA.opt
Paramètres spécifiés à partir de connaissances expertes :

$$\text{nodesize} = \begin{cases} 3 & \text{régression} \\ 1 & \text{classification} \end{cases}$$

Paramètres spécifiés par calibration numérique :

$$\text{ntree} = (\{250 ; 500 ; 100\}) \; \text{mtry} = \left\{ \left[1; \frac{\sqrt{p_l}}{2}; \sqrt{p_l}; 2\sqrt{p_l} ; 4\sqrt{p_l} ; \; \frac{p_l}{4}; \frac{p_l}{3}; \frac{p_l}{2}; \frac{3}{4} p_l; p_l \right] \right\}$$

PHASE 1 – HIERARCHISATION & ELIMINATION

Objectif : Estimer les $\overline{VI}_j(\cdot)$, hiérarchiser les $z_{(\cdot),\cdot}^{j}$, disjoindre de $\mathcal{X}_{input}^{j}$ en : $\mathcal{X}_{bruit}^{j}$ et $\mathcal{X}_{conserv}^{j}$

Hiérarchisation descendante

Estimation des $\overline{VI}_j \left(z_{(\cdot),\cdot}^{j} \right)$ stabilisés sur 50 forêts. Diagramme de tukey des $VI_{j,r} \left(z_{(\cdot),\cdot}^{j} \right)$, $\overline{VI}_j \left(z_{(\cdot),\cdot}^{j} \right)$ et $\overline{\overline{VI}}_j \left(z_{(\cdot),\cdot}^{j} \right)$. Les variables sont groupées par composante environnementale et repérables par une couleur : FIM* ; FE-SAN* ; FE-SOCIO.ECO* ; FE-PHY.CHIM*, afin identifier les groupes qui se détachent. Puis, une hiérarchisation descendante des $x_{(\cdot)}^{l}$ en fonction des $\overline{VI}_j\left(x_{(\cdot)}^{l} \right)$ est effectuée. Les 10 scores les plus forts ainsi que leurs écarts-types sont déclinés.

Elimination des variables de bruit

Spécification de : $c^{j}_{\psi.bruit} \in \{1; \dots; 15\}$ – ce coefficient expert subjectif prend des valeurs fortes car : $\widehat{T}^{j}_{MVG}((l))$ est plus profond que $\widehat{T}^{j}_{VSURF}((l))$.

Estimation de ψ^{j}_{bruit} de façon à éliminer au moins la moitié des $z^{j}_{\cdot,\cdot}$ pour lesquelles $\overline{VI}_j(\cdot) \approx 0$.

Graphiques renvoyés : un archétype de $\widehat{T}^{j}_{MVG}((l))$, et un graphique des $\overline{VI}_j(\cdot)$en fonction de $x^{(l)}_{(\cdot)}$avec le seuil d'élimination ψ^{j}_{bruit}.

PHASE 2 - SELECTION DES* VARIABLES EXPLICATIVES

Objectif : Disjoindre de $\mathcal{X}^{j}_{conserv}$, les paquets de Génuer $\mathcal{X}^{j}_{explic}$ et de Bourrelly $\mathcal{X}^{j}_{0.explic}$

Principe : spécification de : $c^{j}_{\psi.explic} \in \{1; 2\}$, estimation de ψ^{j}_{explic} – stabilisé sur 10 runs - afin de ne retenir environ qu'une dizaine de $x^{(l)}_{(U_k)}$ explicatives et limiter grossièrement les redondances. Un graphique des $\overline{R}^{OOB}_{(\mathcal{X}^{j}_{imbr})}$ en fonction de la dernière $x^{(l)}_{(U_k)}$ imbriquée dans $\mathcal{X}^{j}_{imbr}$, est renvoyé.

Remarque : $\mathcal{X}^{j}_{explic}$ est retenu, $\mathcal{X}^{j}_{0.explic}$ est utilisé pour l'interprétation uniquement.

PHASE 3 - SELECTION DES* VARIABLES PREDICTIVES

Objectif : Identifier à partir du paquet explicatif retenu, les $x^{l}_{(U_k)}$ prédictifs : $\mathcal{X}^{j}_{pred}$

Principe : Spécification de $c^{j}_{\psi.pred} \in \{1; 4\}$ en régression , de $c^{j}_{\psi.pred} = 1$ en classification de sorte que ψ^{j}_{pred} – stabilisé sur 10 runs – joue correctement son rôle d'élimination des redondances - la stratégie est moins efficace en régression. Ensuite, élimination séquentielle des $x^{l}_{(U_k)}$ qui n'améliorent pas le gain d'erreur OOB stabilisé, i.e. lorsque $\overline{\Delta}(R^{OOB}_{l}) \leq \psi^{j}_{pred}$. Le graphique renvoyé est celui des valeurs : de $\overline{\Delta}(R^{OOB}_{l})$ exprimées en fonction de la $x^{l}_{(U_k)}$ testée, et de $\psi^{j}_{pred.def}$ – par défaut, et· ψ^{j}_{pred} – utilisé.

ANALYSE IS DE LA QUALITE DES* MODELES MVG UTILISABLES

Objectif : évaluer la qualité et le gain relatif de $\overline{R}^{OOB}_{(\cdot)}$ des modèles de FA utilisables dans MVG.

Remarques liminaires : la taille de la cohorte est petite. Les interactions entre $z'^{j}_{\cdot,\cdot}$ et $x^{l}_{\cdot}$ sont complexes. Il n'est statistiquement pas pertinent de construire un échantillon test sur $\mathcal{L}^{j}_{U_k}$. L'analyse doit forcément être effectuée In-Sample (IS).

Principe : sur chaque modèle FA.MVG l'erreur OOB, stabilisée sur 10 forêts, est évaluée IS $\overline{R}^{OOB-IS}\left(\hat{f}^{j}_{FA}(\mathcal{X}^{j}_{\cdot}|\Lambda_j)\right)$, et la $\mathbb{var}.\mathbb{explain}^{OOB-IS}_{\hat{f}^{j}_{FA}(\mathcal{X}^{j}_{\cdot}|\Lambda_j)}$ est spécifiée en régression. Les Modèles testés sont :

Modèle MVG	FA.naïve	FA.opt	FA.explic	Bagging.explic	FA.pred	Bagging.préd
Paramètres	$\dot{\Lambda}_j$	$\widehat{\Lambda}_j$	$\dot{\Lambda}_j$	$\ddot{\Lambda}_j$	$\dot{\Lambda}_j$	$\ddot{\Lambda}_j$
Variables	$\mathcal{X}^{j}_{input}$		$\{\mathcal{X}^{j}_{explic} \cup \mathcal{X}^{j}_{0.explic}\}$		$\mathcal{X}^{j}_{pred}$	

Tableau 38 : Noms, paramètres et variables utilisées dans modèles de FA proposés par MVG

Avec : $\dot{\Lambda}_j$ les paramètres *randomForest** par défaut, $\ddot{\Lambda}_j$ les paramètres par défaut mais mtry ≥ 2, Les graphiques renvoient : $\overline{R}^{OOB-IS}(\cdot)$ en fonction de $\hat{f}^{j}_{FA}(\mathcal{X}^{j}_{\cdot}|\Lambda_j)$ - et la $\mathbb{var}.\mathbb{explain}^{OOB-IS}$.

PREDICTIONS GEOGRAPHIQUES MVG

Objectif : Spécifier le modèle $\hat{f}^{j}_{MVG}\left(\mathcal{X}^{j}_{MVG}\middle|\Lambda_{j-MVG}\right)$ permettant d'estimer : $z^{j}_{(U_k),c}$ et $z^{j}_{(U_k),q}$

Remarques liminaires : Cette phase succède à l'analyse des variables explicatives et à la qualité des modèles rencontrés. L'analyse des variables explicatives montre que sur les $z^{j}_{(U_k),c}$ les $x^{l:FIM}_{(U_k)}$ sont déterminantes, par conséquent il est impossible d'effectuer des prévisions dans les communes où aucun patient n'est spatialisé. En revanche, des prédictions OOS sont envisageables sur $z^{j}_{(U_k),q}$. Mais comme les $\hat{z}^{j}_{(U_k),q}$ ont pour objet de décrire la qualité spatiotemporelle des $\hat{z}^{j}_{(U_k),c}$ elles ne présentent aucun intérêt seules.

Modèles MVG utilisés : FA.explic, variables $\left\{\mathcal{X}^{j}_{explic} \cup \mathcal{X}^{j}_{0.explic}\right\}$, paramètres *randomForest** :

$$\Lambda_{j-MVG} = \left(\text{ntree} = 100\ ;\ \text{mtry} = \begin{cases}(\lceil p_l/3 \rceil \vee 2) & \text{régression}\\ (\lfloor \sqrt{p_l} \rfloor \vee 2) & \text{classification}\end{cases};\ \text{nodesize} = \begin{cases}3 & \text{régression}\\ 1 & \text{classification}\end{cases}\right)$$

METHODE D'ANALYSE ET D'INTERPRETATION DES RESULTATS MVG

Il s'agit de proposer une méthode d'interprétation des résultats obtenus par suite de l'application de MVG aux $\mathcal{L}^{j}_{U_k}$. Dans un premier temps, une stratégie d'interprétation conjointe des $x^{l}_{(U_k)}$ simultanément contenus dans les paquets explicatifs et prédictifs de $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$, est proposée. Dans un second temps, une analyse de la robustesse des $\hat{z}'^{j}_{(U_k),c}$ et $\hat{z}'^{j}_{(U_k),q}$ est présentée ainsi que le principe de documentation des résultats statistiques et cartographiques.

PRINCIPE DE CARACTERISATION DES FE/FIM* A PARTIR DES VARIABLES EXPLICATIVES ET PREDICTIVES

L'application de MVG aux $\mathcal{L}^{j}_{U_k}$ permet pour chaque $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$ de confectionner deux paquets de variables $x^{l}_{(U_k)}$ explicatives, celui de Génuer $\mathcal{X}^{j}_{explic}$ et celui de Bourrelly $\mathcal{X}^{j}_{0.explic}$,. et un paquet de $x^{l}_{(U_k)}$ prédictives : $\mathcal{X}^{j}_{pred}$.

Remarques liminaires : Les $z'^{j}_{Uk,c}$ sont les i.st.m* principaux. Les $z'^{j}_{Uk,q}$ caractérisent, entre autres, la certitude spatiotemporelle des $z'^{j}_{Uk,c}$. Compte tenu des nombreuses incertitudes de la modélisation géographique des séquelles, seule l'analyse conjointe des $z'^{j}_{Uk,c}$ et $z'^{j}_{Uk,q}$ est pertinente (chapitre.2). Cette assertion peut être étendue aux résultats obtenus par MVG.

Proposition : *la Grille de lecture MVG* permet de déterminer les FE/FIM* géographiques qui interagissent conjointement avec $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$ à l'aide des $x^{l}_{(U_k)}$ présentes dans $\mathcal{X}^{j}_{explic}$, $\mathcal{X}^{j}_{0.explic}$ et dans $\mathcal{X}^{j}_{pred}$. Cette lecture croisée des résultats MVG fait aussi intervenir les $\overline{VI}_j(\cdot)$ afin de prendre en compte l'intensité des interactions statistiques entre les $x^{l}_{(U_k)}$ et les $z'^{j}_{(U_k),c}$ et dans l'optique de caractériser – globalement l'efficience morbide - des FE/FIM* intégrés.

Les Déterminants Environnementaux de Santé* (DES) : sont des FE/FIM* qui ont une influence forte sur les PM. Ils sont identifiés soit à partir des $x^{l}_{(U_k)}$, qui modélisent la géographie de FE/FIM, dont les $\overline{VI}_j(\cdot)$ obtenus en régression, i.e. sur $z'^{j}_{(U_k),c}$, sont très élevés et qui se trouvent à la fois dans $\mathcal{X}^{j}_{explic}$ et dans $\mathcal{X}^{j}_{pred}$. Ou encore, par des $x^{l}_{(U_k)}$ qui sont simultanément présents dans les paquets $\mathcal{X}^{j}_{explic}$ et $\mathcal{X}^{j}_{pred}$ associés aux $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$.

Les Facteurs de Risques Environnementaux Contributifs* (FREC) : sont des FE/FIM* qui ont une influence notable sur les PM, surtout lorsqu'ils se combinent. Il s'agit des $x^{l}_{(U_k)}$ qui caractérisent directement la géographie de certains FE/FIM, dont les $\overline{VI}_j(\cdot)$ de régression sont plus faibles que ceux caractérisant les DES, mais qui sont simultanément présentes dans les paquets $\mathcal{X}^{j}_{explic}$ associés aux $z'^{j}_{(U_k),c}$ et aux $z'^{j}_{(U_k),q}$ – et qui sont aussi représentés dans $\mathcal{X}^{j}_{0.explic}$ – ou qui se retrouvent au moins dans l'un des paquets $\mathcal{X}^{j}_{pred}$.

Les Facteurs de Risques Environnementaux Probablement Aggravants* (FREPA) : sont des FE/FIM* qui peuvent être soupçonnés, en se combinant, d'aggraver les effets délétères des expositions aux DES* et FREC. Il s'agit des $x^{l}_{(U_k)}$ qui caractérisent directement ou indirectement, i.e. par redondances répétées, la géographie de FE/FIM* particuliers, dont les $\overline{VI}_j(\cdot)$ peuvent être parfois faibles, mais qui sont simultanément représentés dans les paquets $\mathcal{X}^{j}_{explic}$ et $\mathcal{X}^{j}_{0.explic}$ associés aux $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$

En admettant que les $x^{l}_{(U_k)}$ caractérisent bien les DES, les FREC et les FREPA identifiés, il est désormais possible de prédire la géographie des états de santé des patients spatialisés.

PRINCIPE D'EVALUATION DE LA ROBUSTESSE DES PREDICTIONS GEOGRAPHIQUES

L'analyse géographique des $\hat{z}'^{j}_{(U_k),c}$ et $\hat{z}'^{j}_{(U_k),q}$ ne présente aucun intérêt sur le plan épidémiologique puisqu'elle a déjà été effectuée sur les i.st. observés : $z'^{j}_{(U_k),c}$ et $z'^{j}_{(U_k),q}$ (chapitre.2). En revanche l'analyse de la robustesse statistique des prédictions géographique IS trouve tout son sens puisqu'elle permet d'évaluer la pertinence de la dialectique MVG.

Objectif : évaluer la qualité des états de santé prédits par l'estimateur MVG : $\hat{f}^{j}_{MVG}\left(\mathcal{X}^{j}_{MVG}\middle|\Lambda_{j-MVG}\right)$ à partir des $x^{l}_{(U_k)}$ représentatifs des DES, des FREC et des FREPA.

Contexte de la régression : $\hat{z}'^{j}_{(U_k),c}$

Les prévalences spatiales *patients-années* EpiGéoStat prédites sont des i.st. quantitatifs.

Analyse des extrema géographiques prédits

Principe : comparer les valeurs extrêmes de $\hat{z}'^{j}_{(U_k),c}$ qui sont fortement dissemblantes des $z'^{j}_{(U_k),c}$, ce qui est particulièrement préjudiciable pour les U_k concernées.

Documentation cartographique : Les noms de communes affichés concernent les U_k caractérisées d'outliers au regard de la distribution statistique spatiale (Saporta, 2006), i.e. celles pour lesquelles :

$$\hat{z}'^{j}_{(U_k),c} > \hat{Q}_3\left(\hat{z}'^{j}_{(U_k),c}\right) + 1{,}5 \times \widehat{IQR}\left(\hat{z}'^{j}_{(U_k),c}\right)$$

Analyse spatiale de la distribution globale des prédictions

Principe : La mise en perspective des histogrammes de la distribution spatiale des $z'^{j}_{(Uk),c}$ et $\hat{z}'^{j}_{(U_k),c}$ permet d'évaluer la façon dont les valeurs sont globalement prédites, i.e. si elles sont plutôt sous-estimées ou surestimées.

Représentations statistiques : Les valeurs prises par les i.st.m* sont présentées par des histogrammes. Les valeurs situées à l'extrémité droite de la queue de la distribution spatiale sont agrégées dans la dernière classe. Les graphiques sont complétés par un tableau contenant les principaux paramètres *statistiques de position et de dispersion* (Saporta, 2006).

Analyse des résidus géographiques

Définition : les résidus géographiques des prédictions sont définis conventionnellement (Saporta, 2006) par :

$$\hat{\varepsilon}^{j}_{(U_k)} = \left(z'^{j}_{(U_k),c} - \hat{z}'^{j}_{(U_k),c} \right)$$

Principe : lorsque les prédictions $\hat{z}'^{j}_{(U_k),c}$ sont de bonne qualité les indicateurs subséquents doivent tendre vers la valeur spécifiée :

Le pourcentage de variance OOB expliquée :

$$\text{var. explain. OOB}\left(\hat{z}'^{j}_{(U_k),c} \right) \overset{\text{def}}{\cong} \left(1 - \frac{\hat{\text{var}}\left(\hat{\varepsilon}^{j}_{(U_k)} \right)}{\hat{\text{var}}\left(z'^{j}_{(U_k),c} \right)} \right) \geq 60\%$$

Le pourcentage de variance expliquée du modèle et estimée IS :

$$\text{var. explain. IS}\left(\hat{z}'^{j}_{(U_k),c} \right) \overset{\text{def}}{\cong} \left(1 - \frac{\hat{\text{var}}\left(z'^{j}_{(U_k),c} - \hat{z}'^{j}_{(U_k),c} \right)}{\hat{\text{var}}\left(z'^{j}_{(U_k),c} \right)} \right) \geq \text{var. explain. OOB}\left(\hat{z}'^{j}_{(U_k),c} \right)$$

Le coefficient de corrélation entre les valeurs prédites et estimées :

$$\hat{\rho}\left(\hat{z}'^{j}_{(U_k),c} \,;\, z'^{j}_{(U_k),c} \right) \overset{\text{def}}{\cong} \frac{\hat{\text{cov}}\left(z'^{j}_{(U_k),c};\, \hat{z}'^{j}_{(U_k),c} \right)}{\hat{\sigma}_{\left(z'^{j}_{(\cdot),c} \right)} \cdot \hat{\sigma}_{\left(\hat{z}'^{j}_{(\cdot),c} \right)}} \approx \mp 1$$

Analyse des résidus géographiques standardisés et sémiologie cartographique utilisée

Remarques liminaires : La compatibilité des $\hat{\varepsilon}^{j}_{(U_k)}$ avec une Loi normale a systématiquement été vérifiée - pour un niveau de risque $\alpha = 5\%$ - avec le test de Shapiro-Wilk (Saporta, 2006), auquel a été appliquée la correction d'adaptabilité aux effectifs intermédiaires de type : AS R94 puisque $\left\{ n_{(U_k)} = 421 \right\} \gg 50$ (Royston, 1995). La statistique de Shapiro-Wilk correspondait à une p. value $\gg 5\%$. De fait, les $\hat{\varepsilon}^{j}_{(U_k)}$ ont pu être standardisés

Définition : Les résidus géographiques standardisés permettent d'uniformiser l'analyse de la qualité des $\hat{z}'^{j}_{(U_k),c}$ pour toutes les séquelles (Saporta, 2006), ils sont définis par :

$$\dot{\varepsilon}^{j}_{(U_k)} = \frac{\hat{\varepsilon}^{j}_{(U_k)}}{\hat{\sigma}\left(\hat{\varepsilon}^{j}_{(U_k)} \right)}$$

Principe : La qualité statistique des $\hat{z}'^{j}_{(U_k),c}$ s'évalue par le biais des éléments suivants :

La moyenne spatiale des $\dot{\varepsilon}^{j}_{(U_k)}$ doit être approximativement nulle :

$$\hat{\text{moy}}\left(\dot{\varepsilon}^{j}_{(U_k)} \right) \overset{\text{def}}{\cong} \bar{\dot{\varepsilon}}^{j}_{(U_k)} \approx 0$$

La qualité des $\hat{z}'^{j}_{(U_k),c}$ peut être caractérisée par l'appartenance à des Intervalles de Confiance – Gaussiens Bilatères – définis par des niveaux : rg de risques adaptés :

$$\text{IC. géo}^{j}_{\dot{\varepsilon}} = \left[\!\left[t_{rg};\, t_{(1-rg)} \right]\!\right]$$

Avec : $t_{rg} \sim \mathcal{N}(0\,;1)$. Cette stratégie consiste à rejeter à tort rg % des valeurs prédites qui du simple fait du hasard outrepassent les valeurs des bornes t_{rg} et $t_{(1-rg)}$ et, de fait, mettre en exergue des niveaux qualitatifs caractérisant les prédictions géographiques, tel que :

Les $\hat{z}'^{j}_{(U_k),c}$ sont particulièrement robustes lorsque

$$\dot{\varepsilon}^{j}_{(U_k)} \in \left[\!\left[t_{\{\varpi\}};\, t_{(1-\varpi)} \right]\!\right] ; \text{avec: } \{\varpi = 25\%\}$$

Les valeurs restent correctement prédites mais néanmoins maculées d'incertitudes, lorsque

$$\left| \dot{\varepsilon}^{j}_{(U_k)} \right| \in \left[\!\left[t_{(1-\varpi)} \,;\, t_{(1-\alpha)} \right]\!\right] ; \text{avec: } \{\alpha = 10\%\}$$

Les prédictions sont fortement incertaines et ne correspondent pas à la réalité observée –l'erreur commise est supérieure à deux écarts-types, car $\hat{\sigma}\left(\dot{\varepsilon}^{j}_{(U_k)}\right) = 1$, lorsque :

$$\left|\dot{\varepsilon}^{j}_{(U_k)}\right| \geq t_{(1-\tau)} \text{ ; avec: } \{\tau = 2\%\}$$

Représentation statistique graphique

Sur le graphique des valeurs de $\dot{\varepsilon}^{j}_{(U_k)}$, la ligne : continue rouge matérialise $\{t_{(\alpha)} \approx -1{,}28\}$, continue bleue matérialise $\{t_{(1-\alpha)} \approx 1{,}28\}$, discontinue rose $\{t_{(\varpi)} \approx -0{,}670\}$, discontinue bleue clair $\{t_{(1-\varpi)} \approx 0{,}670\}$

Représentation et interprétation cartographique

Lorsque $\dot{\varepsilon}^{j}_{(U_k)} < \{t_{(\alpha)} \approx -1{,}28\}$ le modèle surestime fortement la réalité géographique des $z'^{j}_{(U_k),c}$ et ces communes sont affichées en rouge. Quand $\dot{\varepsilon}^{j}_{(U_k)} > \{t_{(1-\alpha)} \approx 1{,}28\}$, le modèle sous-estime de façon importante la morbidité géographique observée les U_k concernées sont en bleu. Dans les communes où $\left|\dot{\varepsilon}^{j}_{(U_k)}\right| \in [\![\{t_{(\varpi)} \approx 0{,}670\}\,;\,\{t_{(1-\alpha)} \approx 1{,}27\}]\!]$, le modèle sous-estime ou surestime, un peu, les $z'^{j}_{(U_k),c}$., les communes concernées sont représentées respectivement en rose et en bleu clair. Là où $\left|\dot{\varepsilon}^{j}_{(U_k)}\right| \in [\![\{t_{(\varpi)} \approx -0{,}669\};\,\{t_{(1-\varpi)} \approx 0{,}669\}]\!]$, le modèle prédit particulièrement bien les $z'^{j}_{(U_k),c}$, les U_k sont en blanc. Enfin lorsque les prédictions MVG sont médiocres, i.e. $\left|\dot{\varepsilon}^{j}_{(U_k)}\right| > \{t_{(1-\tau)} \approx 2{,}05\}$, le nom des communes concernées est affiché en noir.

Contexte de la Classification $\hat{z}'^{j}_{(U_k),q}$

Les propensions spatiales patients-années EpiGéoStat prédites sont des i.st. qualitatifs.

Analyse visuelle des dissemblances prédictives

Principe : repérer les U_k pour lesquelles les modalités de $\hat{z}'^{j}_{(U_k),q}$ et $z'^{j}_{(U_k),q}$.sont différentes.

Documentation cartographique : La règle d'affichage des labels pour $\hat{z}'^{j}_{(U_k)q}$ est identique à celle utilisée pour $z'^{j}_{(U_k)q}$, soit :

$$\text{label}_{(U_k)} = \begin{cases} \text{Display} & \text{lorsque:} & \{z'^{\text{CATA}}_{(U_k)q} = \text{OUI}\} \\ \text{Display} & \text{lorsque:} & \{z'^{\text{THYR}}_{(U_k)q} = \text{OUI}\} \\ \text{Display} & \text{lorsque:} & \{z'^{\text{TUM2}}_{(U_k)q} \in \{\text{OUI} \cup \text{INCERTAIN}\}\} \end{cases}$$

Analyse spatiale de la distribution globale des prédictions

Principe : La mise en perspective des histogrammes des fréquences d'apparition des valeurs $z'^{j}_{(Uk),q}$ et $\hat{z}'^{j}_{(U_k),q}$ permet d'évaluer la qualité globale des modalités prédites, par comparaison du nombre de $\mathcal{C}_j$.

Représentations statistiques : *Les histogrammes des fréquences* empiriques ne sont pas assortis d'un tableau récapitulatif car ils sont *correctement documentés* (Saporta, 2006).

Analyse des confusions géographiques

Définition : *les confusions* géographiques effectuées sur les prédictions sont notées : $\hat{\varsigma}^{j}_{(U_k)}$. Elles sont évaluées par convention, par :

$$\hat{\varsigma}^{j}_{(U_k)} = \begin{cases} \{1 = \text{CONFUSION}\} & \text{lorsque:} & \left(z'^{j}_{(U_k),q} \neq \hat{z}'^{j}_{(U_k),q}\right) \\ \{0 = \text{EQUIVALENCE}\} & \text{lorsque:} & \left(z'^{j}_{(U_k),q} = \hat{z}'^{j}_{(U_k),q}\right) \end{cases}$$

Principe : lorsque les prédictions $\hat{z}'^{j}_{(U_k),q}$ sont de bonne qualité les indicateurs subséquents doivent tendre vers la valeur spécifiée:
Le Nombre Total des Confusions commises, sur chacune des modalités :

$$NTC^{c_j}_{géo}\left(\hat{z}'^{j}_{q}\right) \stackrel{def}{=} \sum_{k=1}^{N(U_k)} \mathbb{1}_{\left\{z'^{j}_{(U_k),q} \neq \hat{z}'^{j}_{(U_k),q}\right\} \cap \left\{z'^{j}_{(U_k),q} = c_j\right\}} \to 0^{+}$$

La Proportion Relative de Confusions commises, sur chacune des modalités :

$$PRC^{c_j}_{géo}\left(\hat{z}'^{j}_{q}\right) \stackrel{def}{=} \frac{1}{N\left(\left\{z'^{j}_{(U_k),q} = c_j\right\}\right)} \cdot \sum_{k=1}^{N(U_k)} \mathbb{1}_{\left\{z'^{j}_{(U_k),q} \neq \hat{z}'^{j}_{(U_k),q}\right\} \cap \left\{z'^{j}_{(U_k),q} = c_j\right\}} \to 0^{+}$$

La Proportion Globale de Confusions Commises les modalités :

$$PGC^{c_j}_{géo}\left(\hat{z}'^{j}_{q}\right) \stackrel{def}{=} \frac{1}{N(U_k)} \cdot \sum_{k=1}^{N(U_k)} \mathbb{1}_{\left\{z'^{j}_{(U_k),q} \neq \hat{z}'^{j}_{(U_k),q}\right\}} \to 0^{+}$$

Avec : $c_j \in \mathcal{C}^j = \{\text{OUI; NON; INCERTAIN}\}$; $\forall\, j = \{\text{CATA; THYR; TUM2}\}$

Représentation statistique graphique
Le tableau de synthèse des confusions : il simplifie la lecture de la *matrice des confusions* en déclinant les indicateurs statistiques : $NTC^{c_j}_{géo}\left(\hat{z}'^{j}_{q}\right)$; $PRC^{c_j}_{géo}\left(\hat{z}'^{j}_{q}\right)$ et $PGC^{c_j}_{géo}\left(\hat{z}'^{j}_{q}\right)$, en fonction des c_j. Il est adapté aux variables multi-classes. *La matrice des confusions géographiques* décline pour chaque modalité c_j : sur la diagonale le nombre de $z'^{j}_{(U_k),q}$ bien prédites, et le reste de la matrice dénombre les confusions avec les modalités concernées (Saporta, 2006).

Représentations cartographiques
Les modalités de $\hat{\varsigma}^{j}_{(U_k)}$ sont projetées dans les U_k - la variable est booléenne.
Code couleur associé :

$$\hat{\varsigma}^{j}_{(U_k)} = \begin{cases} \text{ORANGE} & \text{lorsque:} \quad \left\{\hat{\varsigma}^{j}_{(U_k)} = \text{CONFUSION}\right\} \\ \text{CYAN} & \text{lorsque:} \quad \left\{\hat{\varsigma}^{j}_{(U_k)} = \text{EQUIVALENCE}\right\} \end{cases}$$

Nom des communes affichées :

$$\text{label}_{(U_k)} = \begin{cases} \text{Display} & \text{lorsque:} \quad \left\{\hat{\varsigma}^{j}_{(U_k)} = \text{CONFUSION}\right\} \\ \text{None} & \text{lorsque:} \quad \left\{\hat{\varsigma}^{j}_{(U_k)} = \text{EQUIVALENCE}\right\} \end{cases}$$

PRESENTATION DES RESULTATS MVG

La méthode MVG est appliquée à chaque séquelle : CATA, THYR et TUM2, conformément aux propositions énoncées. Les résultats sont présentés d'abord sur l'i.st.m* quantitatif $z'^{j}_{(Uk),c}$ puis sur les i.st.m* qualitatifs $z'^{j}_{(Uk),q}$. Pour chaque Contexte statistique* : les i.st.e* $x^{l}_{(Uk)}$ actifs; les coefficients experts : $c^{j}_{\psi.}$ appliqués aux seuils : $\psi^{j}_{.}$., Les items statistiques renvoyés par MVG ; Ainsi que les représentations cartographiques et leur analyse spatiale et statistique, sont déclinées dans cet ordre. Comme l'analyse géographique des prédictions ne présente pas d'intérêt les cartographies sont présentées en petit format et uniquement pour les U_k situées en PACA et aux alentours, où 50% de la cohorte LEA est spatialisée (chapitre.1).

SEQUELLE : CATARACTES

JEU DE DONNEES D'APPRENTISSAGE UTILISE : $\mathcal{L}_{U_k}^{CATA}$

Les variables réponses sont les i.st.m* $z'^{CATA}_{(U_k),c}$ ou $z'^{CATA}_{(U_k),q}$ qui modélisent la géographie des cataractes. Les i.st.e* : $x'^{l}_{(U_k)}$ et $x^{l}_{(U_k)}$, écartés sont ceux qui modélisent les expositions spatiotemporelles aux : FE-SAN* par des temps d'accès bruités, aux item sanitaires : x_ORL, x_ORL.S, x._HEMAs et x._TEP. Ils ont été choisis aléatoirement parmi tous les i.st.e* Curieux* de cette composante, largement représentée. Les FE_PHY.CHIM des expositions à la radioactivité environnementale sont surreprésentés mais actuellement suspectés d'être des facteurs aggravants du risque de CATA - pour chaque milieu environnemental, un i.st.e* a été retiré aléatoirement : x_ALPHA, x_131i, x_125Sb. Quant aux FE_SOCIO-ECO, ils sont très représentés, x_attBIEN a été retiré arbitrairement. Cet i.st.e* conjoncturel curieux* est redondant avec x_attPHY.

Les i.st.e* inclus dans $\mathcal{X}_{input}^{CATA}$ sont déclinés avec le code couleur environnemental associé :

x._SEXE x._AGE_DIAG x._DSUIVI x._TYPLEUC x._PROTOC x._RECHUT x._GREF x._IRACT x._ACPHY x._GENE x._OPHT x._PEDIA x._RADIO x._NEURs x._HEMAs x._OPHTs x._SCAN x._CAME x._IRM x_APL_GENE x_APL_OPHT x_txCHOM x_txFoyFisc x_txOUV x_GINIp x_GINIm x_RevMedm x_RevMedp x_txMORT x_txAccPOP x_txAccNAT x_txBAC x_txEAR x_int.AGRI x_tx.SAU x_FDep99 x_FDep09 x_FDepXX x_attPHY x_INSECU x_RADON x.._RADON x_TEMP x_RAY x_NJAP x_FEFO x_PEST x_URIN x_PREV x_EGRA x_GAMMA x_BETA x_3H x_238Pu x_137Cs_SOL x_90Sr x_137Cs_BIOL x_TOPO x_irsCd x_irsPb x_irsNi x_DJECr x_PERC (62/62)

REGRESSION : IDENTIFICATION MVG DES* VARIABLES EXPLICATIVES ET PREDICTIVES

Cible : $z'^{CATA}_{Uk,c}$ - les prévalences spatiales pondérées EpiGéoStat converties en patients-années.

PHASE D'ANALYSE DES* DONNEES

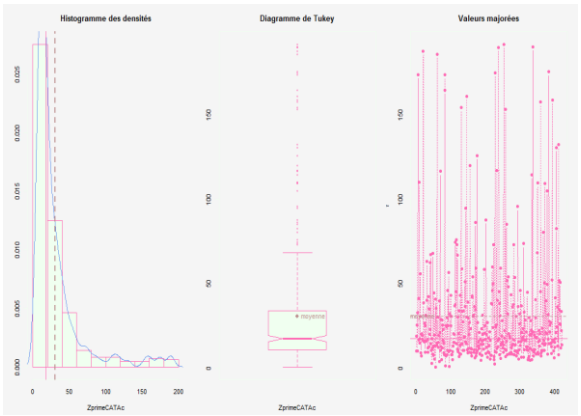


Figure 191 : Histogramme, boite à moustache, valeurs de $z'^{CATA}_{Uk,c}$

Statistiques MVG

Moyenne spatiale : $\bar{z}'^{CATA}_{Uk,c} = 0{,}03089$

Ecart-type biaisé : $\hat{\sigma}\left(z'^{CATA}_{Uk,c}\right) = 0{,}03567$

PHASE 0 - CALIBRATION:

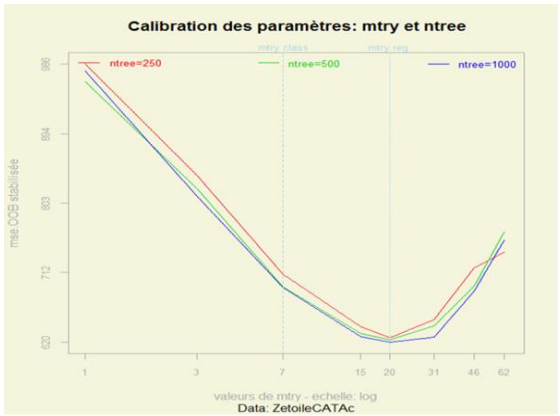


Figure 192 : $\bar{R}^{OOB}_{\cdot}(\cdot)$ en fonction de ntree et mtry

Objectif :

FA.naïve → FA.opt

Paramètres MVG optimisés:

$\hat{\Lambda}_j = (\text{ntree} = 1000\,;\, \text{mtry} = 20;\, \text{nodesize} = 3)$

PHASE 1 – HIERARCHISATION & ELIMINATION DES* VARIABLES DE BRUIT

Objectif : Estimer les $\overline{VI}_j(\cdot)$; Hiérarchiser les $z^{CATA}_{(U_k),q}$; Disjoindre de $\mathcal{X}_{input}^{CATA}$ en : $\mathcal{X}_{bruit}^{CATA}$ et $\mathcal{X}_{conserv}^{CATA}$

Estimation et présentation des scores *randomForest**

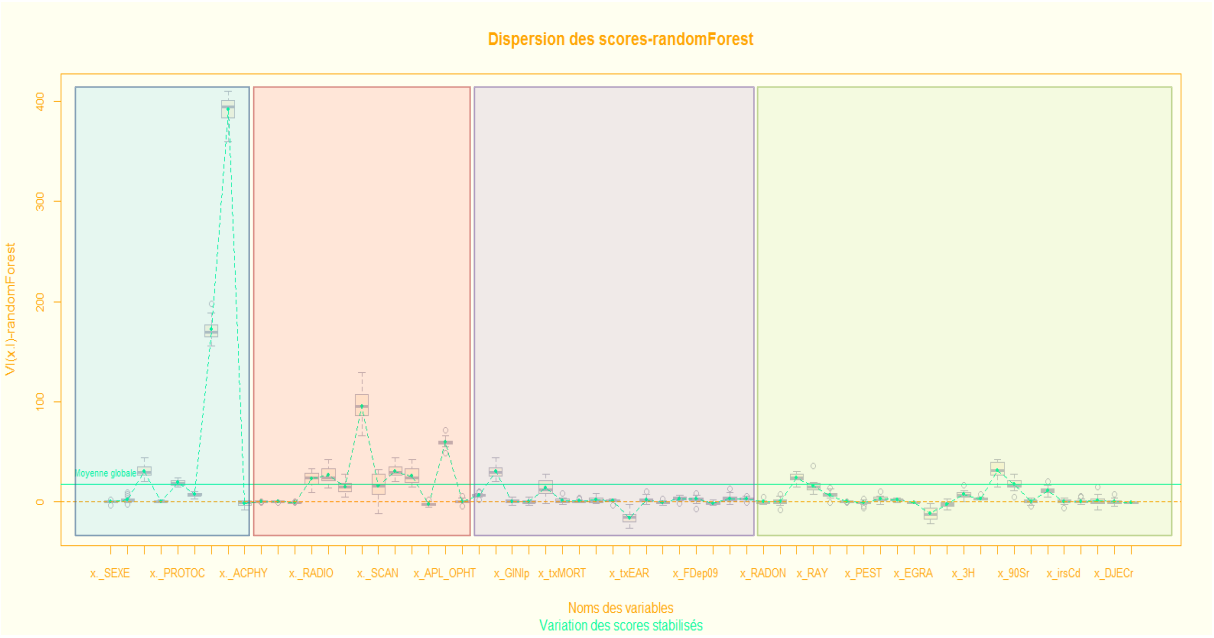


Figure 193 : Diagramme de Tukey des $VI_{j,r}\left(z^{j}_{(\cdot),.}\right)$, $\overline{VI}_j\left(z^{j}_{(\cdot),.}\right)$ et $\overline{\overline{VI}}_j\left(z^{j}_{(\cdot),.}\right)$ - estimation stabilisée sur 50 forêts.

Remarque : l'influence morbide des composantes environnementales peut être pré-classée à partir des scores des i.st.e, tel que : (i) FIM (ii) FE-SAN*;(iii ex-aequo) FE-PHY.CHIM* et FE-SOCIO.ECO*.*

Hiérarchisation descendante

Valeurs des $\overline{VI}_j(\cdot)$ et des $\hat{\sigma}\left(VI_j(\cdot)\right)$ pour les 10 i.st.e* : $x^{(l)}_{(U_k)}$ les plus importants sont :

	VI.moy	SD.VI.moy
x._IRACT	391.02588	20.57560
x._GREF	171.69346	17.18099
x._OPHTs	95.83150	22.68264
x_APL_OPHT	59.79005	10.53837
x_137Cs_SOL	31.92062	5.30037
x._DSUIVI	30.82624	4.69365
x._CAME	30.82624	10.76628
x_txOUV	30.37801	5.01021
x._NEURs	26.84550	10.02981
x._IRM	26.30998	9.05481

Elimination des variables de bruit

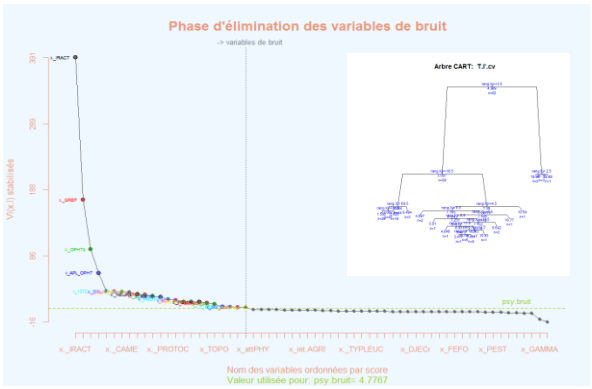


Figure 194 : Valeurs des $\overline{VI}_j(\cdot)$ associés aux $x^{(l)}_{(U_k)}$ et $\hat{T}^{j}_{MVG}((l))$

Estimation de ψ^{CATA}_{bruit} :

Coefficient subjectif expert $c^{CATA}_{\psi.bruit} = 3$

Disjonction des variables

$\mathcal{X}^{j}_{bruit}$: 39/62 et $\mathcal{X}^{j}_{conserv}$:23/62

Variables éliminées identifiées comme du bruit

x_attPHY x_FDep99 x_238Pu x_URIN x_INSECU x_FDep09 x._AGE_DIAG x_int.AGRI* x_txAccNAT x_PREV x_irsNi x_txBAC x_txMORT x_txAccPOP x._TYPLEUC x_txCHOM x_irsCd x_irsPb x_GINIp x.._RADON x_137Cs_BIOL x_DJECr x_GINIm x._OPHT x_RADON x._GENE x_FEFO x._SEXE x._PEDIA x_tx.SAU x_EGRA x_PEST x_PERC x_FDepXX x._ACPHY x_BETA x_APL_GENE x_GAMMA x_txEAR (39/62)

PHASE 2 - SELECTION DES* VARIABLES EXPLICATIVES

Estimation de $\psi_{\text{explic}}^{\text{CATA}}$:

Coefficient subjectif expert : $c_{\psi.\text{explic}}^{\text{THYR}} = 2$

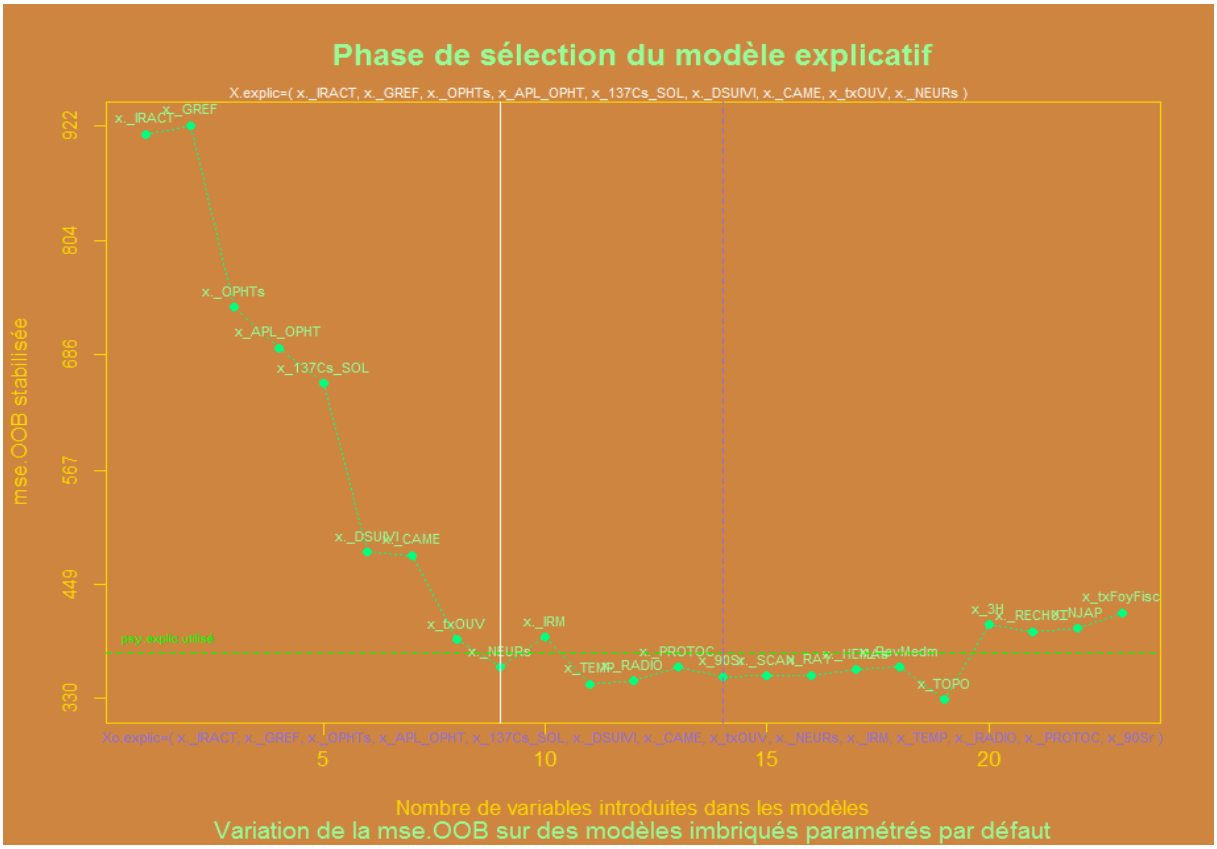


Figure 195 : Variation de $\overline{\mathbf{R}}_{(.)}^{\mathbf{OOB}}$ en fonction du nombre de variables imbriquées $\boldsymbol{\mathcal{X}}_{\mathbf{imbr.l}}^{\mathbf{CATA}} \subset \boldsymbol{\mathcal{X}}_{\mathbf{conserv}}^{\mathbf{CATA}}$;

Résultats - Paquets retenus : Génuer $\mathcal{X}_{\text{explic}}^{\text{CATA}}$

x._IRACT x._GREF x._OPHTs x_APL_OPHT x_137Cs_SOL x._DSUIVI x._CAME x_txOUV x._NEURs

(9/62)

*Remarque : le paquet Bourrelly $\mathcal{X}_{0.explic}^{CATA}$ contient aussi les i.st.e**

[...] x._IRM x_TEMP x._RADIO x._PROTOC x_90Sr

PHASE 3 - VARIABLES PREDICTIVES

Estimation de $\psi_{\text{pred}}^{\text{CATA}}$:

Coefficient subjectif expert : $c_{\psi.\text{pred}}^{\text{CATA}} = 1$

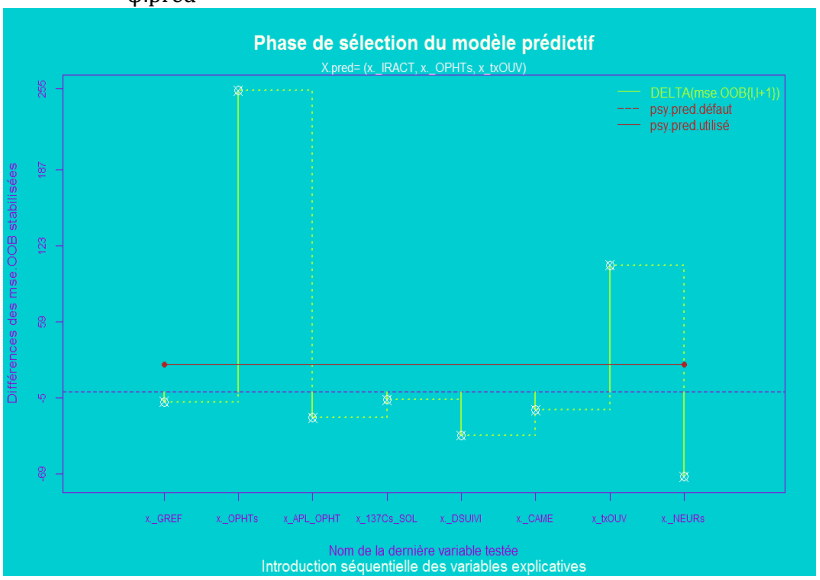


Figure 196 : Valeur de $\overline{\Delta}(R_l^{OOB})$ en fonction de $\mathbf{x}_{(U_k)}^{l}$ testé

Résultats - $\mathcal{X}_{\text{pred}}^{\text{CATA}}$ contient :

x._IRACT x._OPHTs x_txOUV (3/62)

ANALYSE DE LA QUALITE DES* MODELES MVG

Estimation OOB de qualité des modèles MVG utilisables $\hat{f}_{FA}^{CATA}(\mathcal{X}_{.}^{j}|\Lambda_j)$:

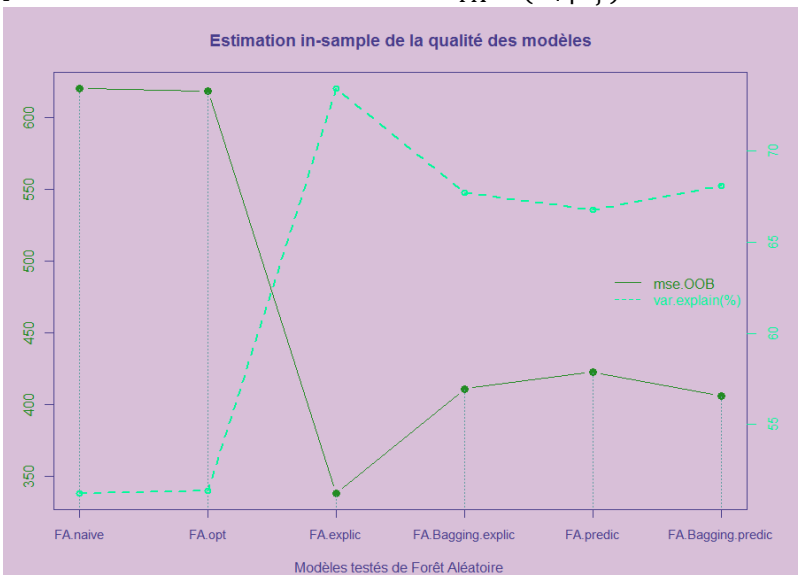


Figure 197 : Variation de la mse.OOB et de la var.explain.OOB en fonction du modèle MVG testé

Remarque : les valeurs de $\overline{R}^{OOB-IS}(\cdot)$ et de $\mathbb{var.explain}_{f_{FA}^{j}.}^{OOB-IS}$ sont estimées *In-Sample*

ANALYSE DE LA QUALITE PREDICTIVE

MODELE:	FA.naive	FA.opt	FA.explic	FA.pred
var.explain.OOB	51,24%	51,38%	73,41%	66,76%
gain.OOB.absolu	0,00%	0,14%	22,17%	15,52%
gain.OOB.relatif	0,00%	0,29%	45,47%	31,83%

Tableau 39 : Analyse des gains absolus et relatifs de mse.OOB et de var.explain.OOB par rapport à une FA.naïve

REGRESSION : PREDICTIONS GEOGRAPHIQUES MVG

L'application de MVG à $z'^{CATA}_{(U_k),c}$ permet d'obtenir le prédicteur MVG : $\hat{f}^{CATA}_{MVG}(\mathcal{X}^{j}_{MVG}|\Lambda_{j-MVG})$. Le vecteur des paramètres spécifiés est : $\Lambda_{j-MVG} = (\text{ntree} = 1000\,;\text{mtry} = ([p_1/3] \vee 2)\,;\text{nodesize} = 3)$ et $\{\mathcal{X}^{CATA}_{MVG} = \mathcal{X}^{CATA}_{explic}\}$. Les résultats présentés sont conformes aux propositions de début de section :

Cartographie des prévalences spatiales EpiGéoStat converties patients-années

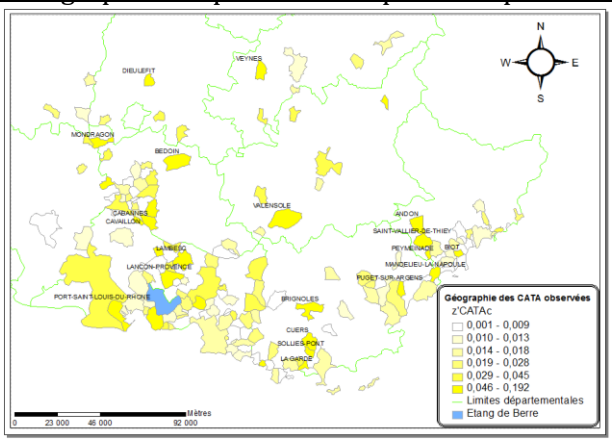


Figure 198 : Valeurs observées prises par $z'^{CATA}_{(U_k),c}$

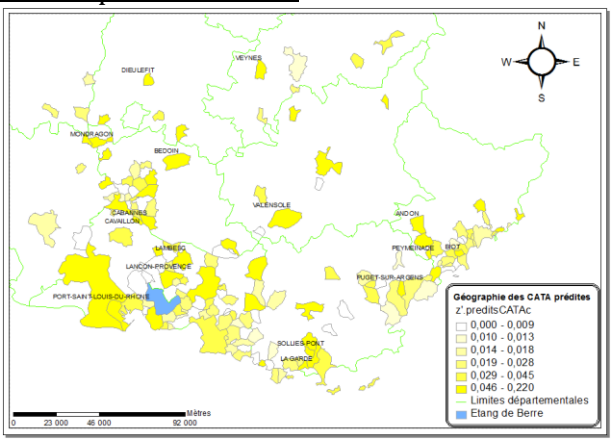
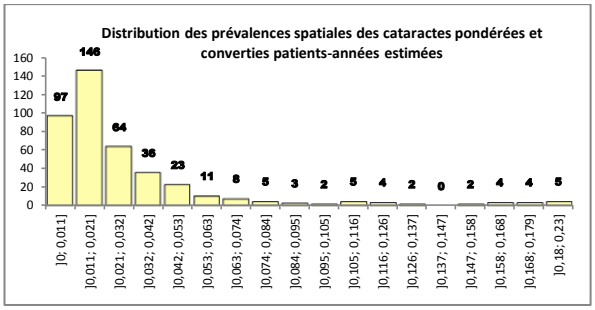


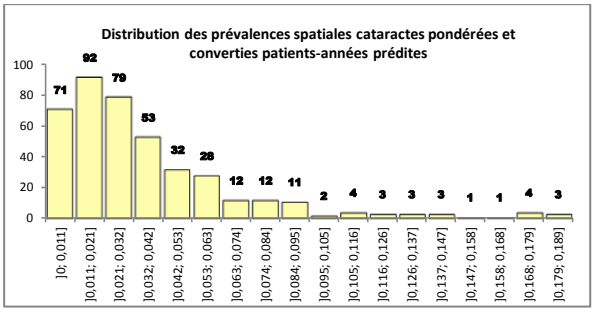
Figure 199 : Valeurs MVG prédites prises par $\hat{z}'^{CATA}_{(U_k),c}$

Distribution statistique des prévalences spatiales EpiGéoStat converties patients-années



421 - Uk	Paramètres de dispersion & de position						
Estimateur	min(.)	Q1^(.)	mêd()	Q3^(.)	max(.)	môy(.)	σ^(.)
Estimation	0,001	0,011	0,0182	0,034	0,192	0,0308	0,0357

Figure 200 : Distribution et tableau statistique des estimés $z'^{CATA}_{(U_k),c}$



421 - Uk	Paramètres de dispersion & de position						
Estimateur	min(.)	Q1^(.)	mêd()	Q3^(.)	max(.)	môy(.)	σ^(.)
Estimation	0,000	0,015	0,026	0,049	0,219	0,0392	0,0400

Figure 201 : Distribution et tableau statistique des prédictions $\hat{z}'^{CATA}_{(U_k),c}$

Cartographie, graphique et synthèse statistique des résidus géographiques standardisés

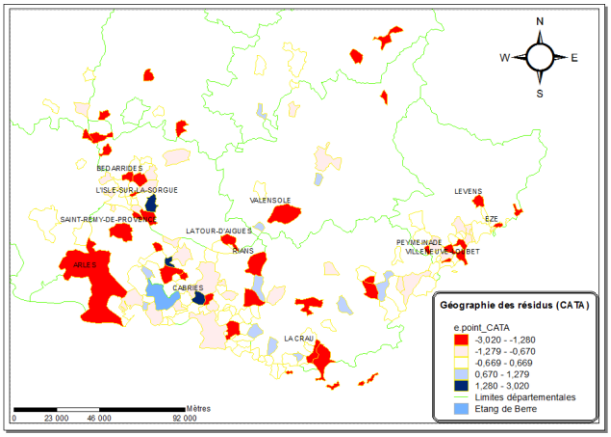
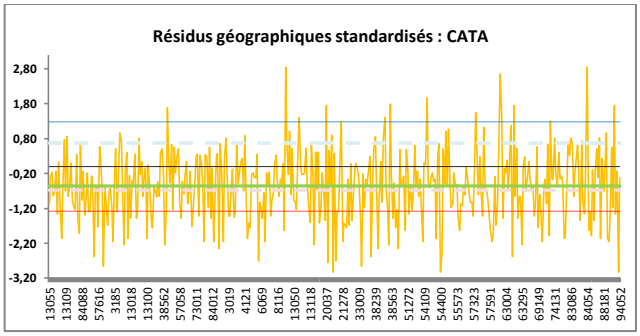


Figure 202 : Valeurs observées prises par $\dot{\varepsilon}^{CATA}_{(U_k)}$ dans les U_k sises en région PACA et aux alentours



Estimateur de qualité statistique	Estimation
var.explain.OOB($\hat{z}'^{j}_{(U_k),c}$)	73,42%
var.explain.IS($\hat{z}'^{j}_{(U_k),c}$)	74,30%
môy($\dot{\varepsilon}^{j}_{(U_k)}$)	-0,558
corr.($\hat{z}'^{j}_{(U_k),c}$; $z'^{j}_{(U_k),c}$)	92,79%

Figure 203 : Valeurs et synthèse statistique des $\dot{\varepsilon}^{CATA}_{(U_k)}$

CLASSIFICATION : IDENTIFICATION MVG DES* VARIABLES EXPLICATIVES ET PREDICTIVES

Cible : $z'^{CATA}_{Uk,q}$ propension spatiale pondérée EpiGéoStat et par le ratio participation/exposition

PHASE D'ANALYSE DES* DONNEES

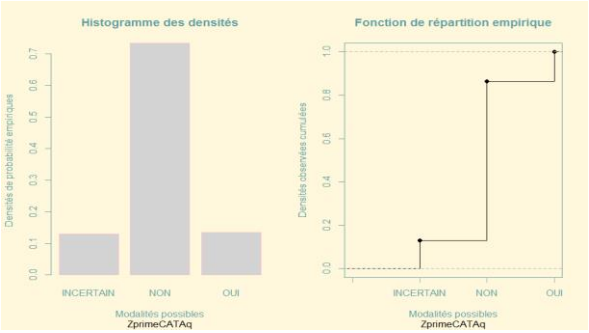


Figure 204 : Histogramme et Fonction de répartition, de $z'^{CATA}_{Uk,q}$

Statistiques MVG

Nombre des modalités card $\left(z^{j}_{(U_k),q} = c_j \right)$

INCERTAIN	NON	OUI
55	307	57

Proportion des modalités môy $\left(z^{j}_{(U_k),q} = c_j \right)$

INCERTAIN	NON	OUI
13,06	73,40	13,54

PHASE 0 - CALIBRATION:

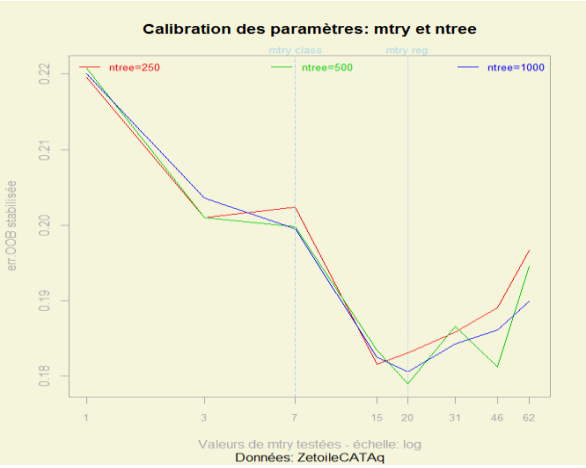


Figure 205 : $\bar{R}^{OOB}_{\cdot}(\cdot)$ en fonction de ntree et mtry

Objectif

FA.naïve → FA.opt

Paramètres MVG optimisés:

$\hat{\Lambda}_{THYR} = (ntree = 500\,; mtry = 20; nodesize = 1)$

PHASE 1 – HIERARCHISATION & ELIMINATION DES* VARIABLES DE BRUIT

Objectif : estimer les $\overline{VI}_j(\cdot)$, hiérarchiser les $z^{CATA}_{(U_k),q}$, disjoindre de $\mathcal{X}^{CATA}_{input}$ en : $\mathcal{X}^{CATA}_{bruit}$ et $\mathcal{X}^{CATA}_{conserv}$

Estimation et présentation des scores *randomForest*

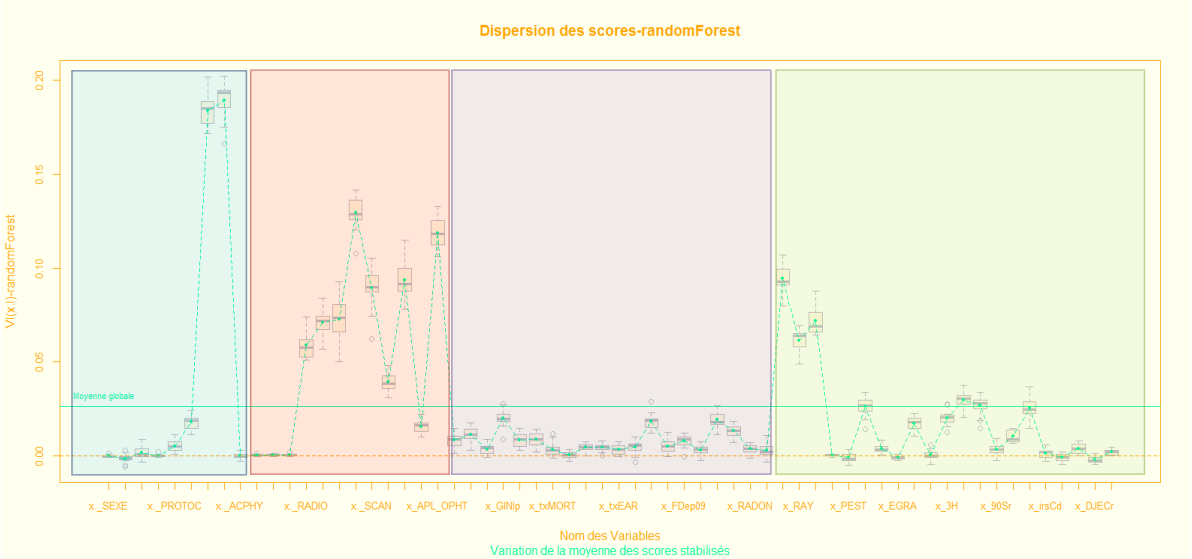


Figure 206 : Diagramme de Tukey des $VI_{j,r}\left(z^{j}_{(\cdot),\cdot} \right)$, $\overline{VI}_j\left(z^{j}_{(\cdot),\cdot} \right)$ et $\overline{\overline{VI}}_j\left(z^{j}_{(\cdot),\cdot} \right)$ - estimation stabilisée sur 50 forêts.

Remarque : l'influence morbide des composantes environnementales peut être pré-classée à partir des scores des i.st.e, tel que : (i) FIM, (ii) FE-SAN* (iii), FE-PHY.CHIM*, (iv) FE-SOCIO.ECO*.*

Hiérarchisation descendante

Valeurs des $\overline{VI}_j(\cdot)$ et des $\hat{\sigma}\left(VI_j(\cdot)\right)$ pour les 10 i.st.e* : $x_{(U_k)}^{(l)}$ les plus importants :

	VI.moy	SD.VI.moy
x._IRACT	0.18914	0.01085
x._GREF	0.18364	0.01088
x._OPHTs	0.12960	0.01099
x_APL_OPHT	0.11840	0.00973
x_TEMP	0.09446	0.00884
x._IRM	0.09374	0.00959
x._SCAN	0.08957	0.00898
x._HEMAs	0.07260	0.00809
x_NJAP	0.07174	0.00766
x._NEURs	0.07095	0.00782

Elimination des variables de bruit

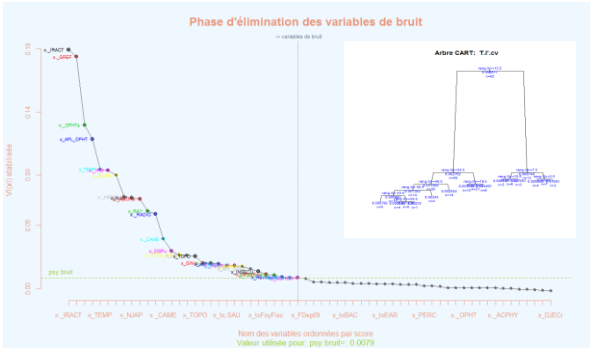


Figure 207 : Valeurs des $\overline{VI}_j(\cdot)$ associés aux $x_{(U_k)}^{(l)}$, et $\hat{T}_{MVG}^{j}((I))$

Estimation de ψ_{bruit}^{CATA}

Coefficient subjectif expert : $c_{\psi.bruit}^{CATA} = 5$

Disjonction des variables

$\mathcal{X}_{bruit}^{CAT}$: 32/62 et $\mathcal{X}_{conserv}^{CATA}$:30/62

Variables éliminées identifiées comme du bruit

x_FDep09 x._PROTOC x_FDep99 x_int.AGRI* x_txAccNAT x_txBAC x_PREV x_RADON x_irsNi x_txMORT x_txEAR x_txOUV x_90Sr x_FDepXX x.._RADON x_PERC x._DSUIVI x_irsCd x._PEDIA x_txAccPOP x._OPHT x._GENE x_FEFO x_BETA x._TYPLEUC x._ACPHY x._SEXE x_EGRA x_irsPb x_PEST x._AGE_DIAG x_DJECr (32/62)

PHASE 2 - SELECTION DES* VARIABLES EXPLICATIVES

Estimation de $\psi_{explic.}^{CATA}$:

Coefficient subjectif expert : $c_{\psi.explic}^{CATA} = 2$

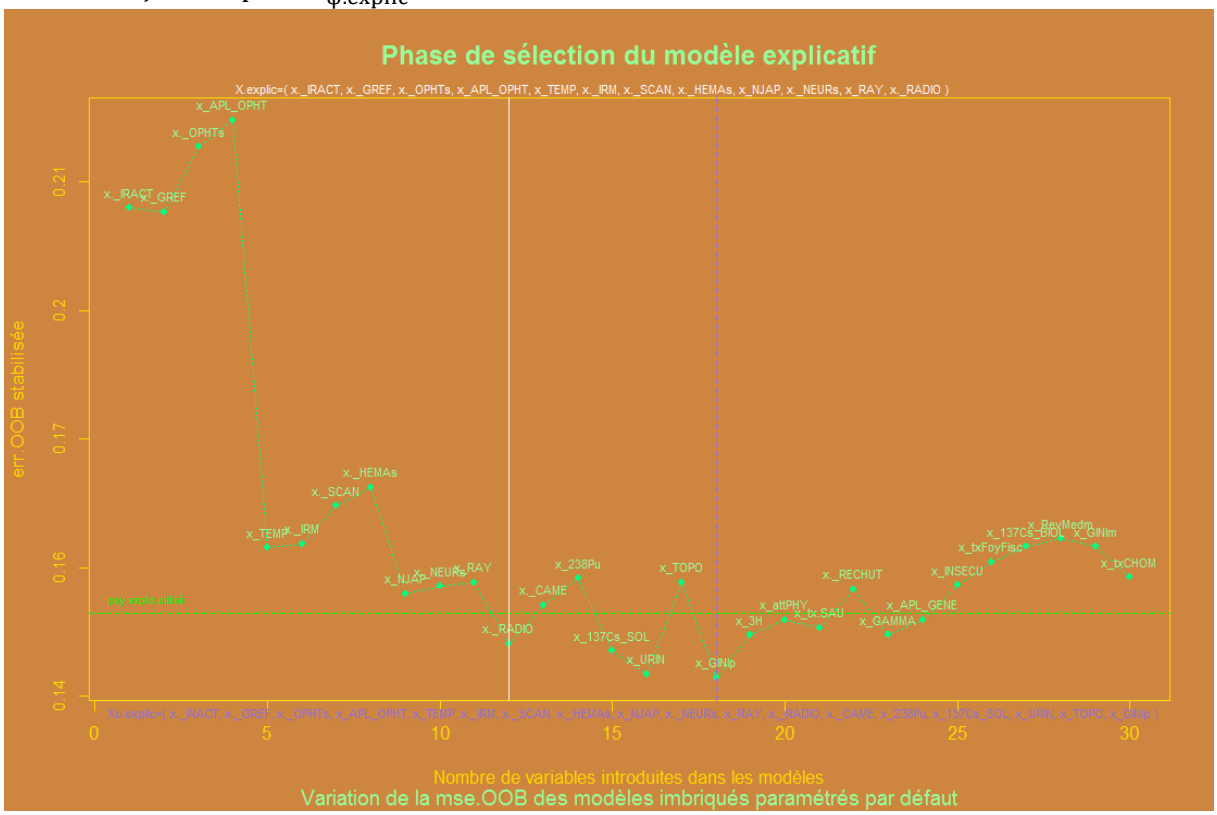


Figure 208 : Variation de $\overline{R}_{(.)}^{OOB}$ en fonction du nombre de variables imbriquées $\mathcal{X}_{imbr.l}^{CATA} \subset \mathcal{X}_{conserv}^{CATA}$;

Résultats - Paquets retenus : Génuer $\mathcal{X}_{\text{explic}}^{\text{CATA}}$

x._IRACT x._GREF x._OPHTs x_APL_OPHT x_TEMP x._IRM
x._SCAN x._HEMAs x_NJAP x._NEURs x_RAY x._RADIO (12/62)

*Remarque : le paquet Bourrelly $\mathcal{X}_{0.explic}^{CATA}$ contient aussi les i.st.e**

[...] x._CAME x_238Pu x_137Cs_SOL x_URIN x_TOPO x_GINIp (18/62)

PHASE 3 - VARIABLES PREDICTIVES

Estimation de $\psi_{\text{pred}}^{\text{CATA}}$:Coefficient subjectif expert : $c_{\psi.\text{pred}}^{\text{CATA}} = 1$

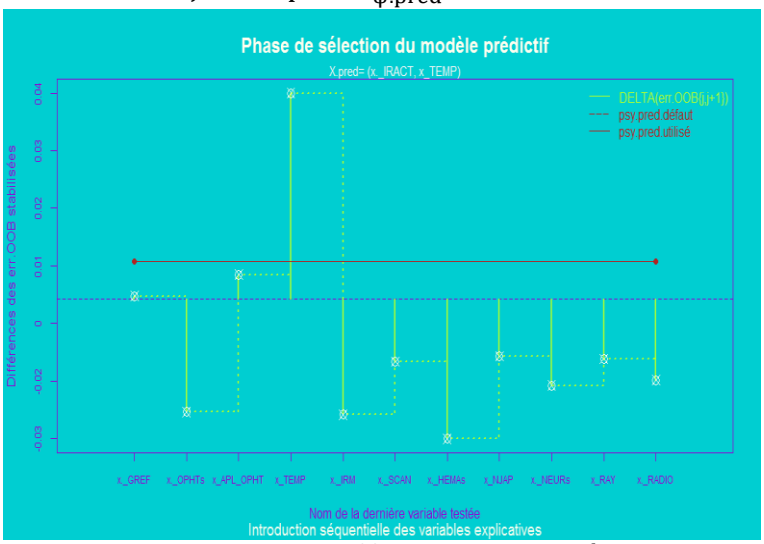


Figure 209 : Valeur de $\overline{\Delta}(R_l^{OOB})$ en fonction de $x_{(U_k)}^{l}$ testé

Résultats - $\mathcal{X}_{\text{pred}}^{\text{CATA}}$ contient :

x._IRACT x_TEMP (2/62)

ANALYSE DE LA QUALITE DES MODELES MVG

Estimation OOB de qualité des modèles MVG utilisables $\hat{f}_{\text{FA}}^{\text{CATA}}(\mathcal{X}_{\cdot}^{j}|\Lambda_j)$:

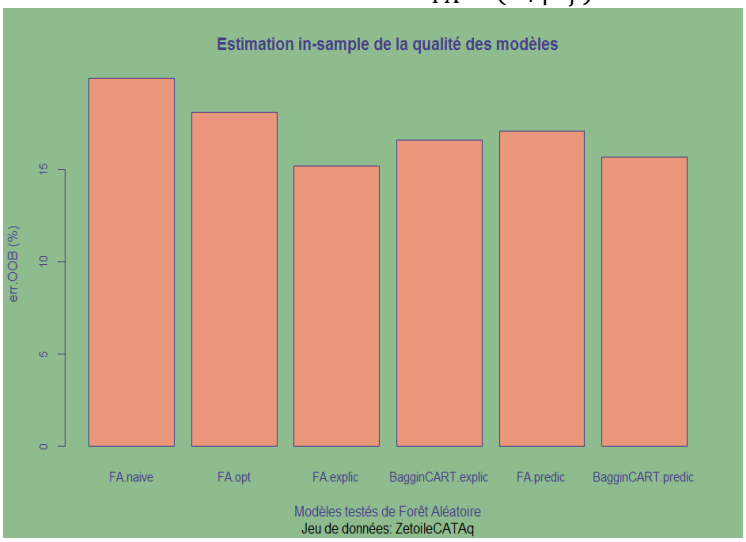


Figure 210 : Variation de l'err.OOB en fonction du modèle MVG testé

Remarque : les valeurs de $\overline{R}^{\text{OOB}-\text{IS}}(\cdot)$ sont estimées *In-Sample* (IS)

ANALYSE DE LA QUALITE PREDICTIVE

MODELE:	FA.naive	FA.opt	FA.explic	FA.pred
err.OOB généralisée :	19,95%	18,05%	15,20%	17,10%
gain.OOB.absolu :	0,00%	1,90%	4,75%	2,85%
gain.OOB.relatif :	0,00%	9,52%	23,81%	14,29%

Tableau 40 : Analyse des gains absolus et relatifs d'err.OOB généralisées par rapport à une FA.naïve

Classification : PREDICTIONS GEOGRAPHIQUES MVG

L'application de MVG à $z'^{CATA}_{(U_k),q}$ permet d'obtenir le prédicteur MVG : $\hat{f}^{CATA}_{MVG}(x^{j}_{MVG}|\Lambda_{j-MVG})$. Le vecteur des paramètres spécifiés est : $\Lambda_{j-MVG} = (ntree = 1000\,; mtry = ([\sqrt{p_l}] \vee 2)\,; nodesize = 1)$ et $\{x^{CATA}_{MVG} = x^{CATA}_{explic}\}$. Les résultats présentés sont conformes aux propositions de cette section :

Cartographie des propensions spatiales EpiGéoStat répétées au prorata du ratio participation/exposition

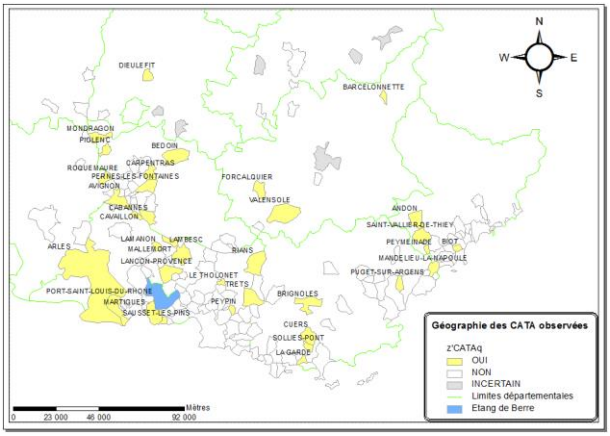


Figure 211 : Valeurs observées prises par $z'^{CATA}_{(U_k),q}$

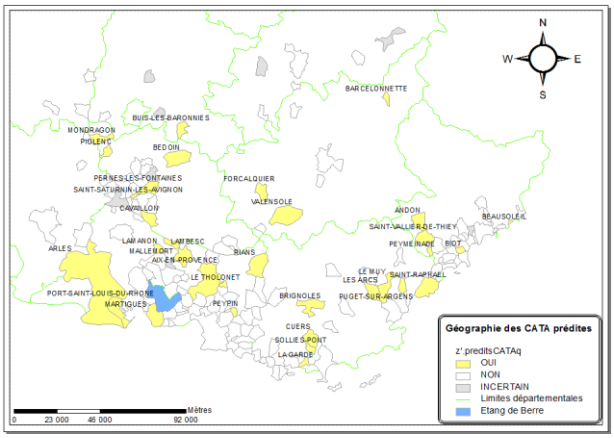


Figure 212 : Valeurs MVG prédites prises par $\hat{z}'^{CATA}_{(U_k),q}$

Fréquence d'apparition des propensions spatiales EpiGéoStat répétées au prorata du ratio participation/exposition

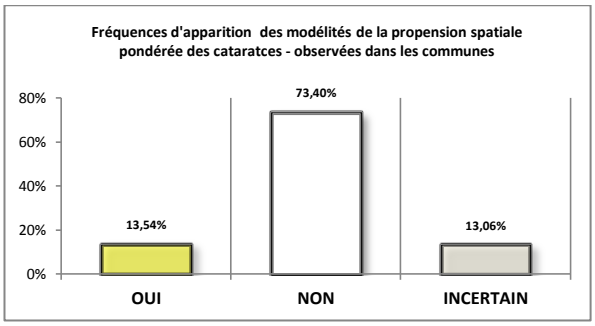


Figure 213 : Fréquences spatiales estimées des modalités de $z'^{CATA}_{(U_k),q}$

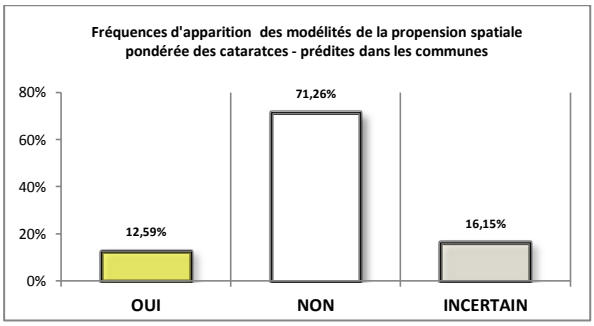


Figure 214 : Fréquences spatiales prédites des modalités de $\hat{z}'^{CATA}_{(U_k),q}$

Cartographie, graphique et synthèse statistique des confusions géographiques

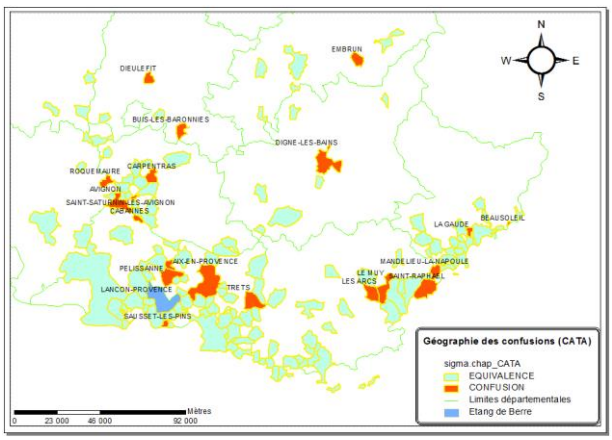
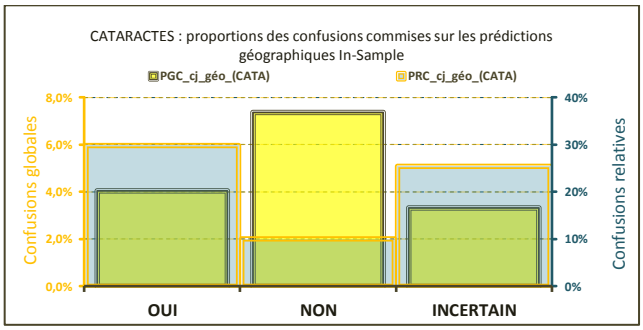


Figure 215 : Valeurs prises par $\hat{\varsigma}^{CATA}_{(U_k)}$ dans les U_k situées en région PACA et aux alentours



Synthèse : CONFUSIONS	OUI	NON	INCERTAIN	TOTAL
NTC_cj_géo_($\hat{z}'^{j}_{(U_k),q}$)	17	31	14	62
PRC_cj_géo_($\hat{z}'^{j}_{(U_k),q}$)	29,82%	10,03%	25,45%	-
PGC_cj_géo_($\hat{z}'^{j}_{(U_k),q}$)	4,04%	7,36%	3,33%	14,73%

Figure 216 : Synthèse et représentation graphique des $\hat{\varsigma}^{CATA}_{(U_k)}$

SEQUELLE : TUMEURS THYROÏDIENNES

JEU DE DONNEES D'APPRENTISSAGE UTILISE : $\mathcal{L}_{U_k}^{THYR}$

Les variables réponses sont les i.st.m* $z'^{THYR}_{(U_k),c}$ ou $z'^{THYR}_{(U_k),q}$. Ils modélisent la géographie des tumeurs thyroïdiennes. Les i.st.e* : $x'^{l}_{(U_k)}$ et $x^{l}_{(U_k)}$ écartés sont ceux qui modélisent les expositions spatiotemporelles aux FE-SAN* Curieux* les plus aberrants : x_OPHTs et x_APL.OPHT et x._CAM - ce dernier aurait peut-être été plus pertinent que x._TEP. Les FE_PHY.CHIM d'expositions à la radioactivité environnementale sont surreprésentés mais très discutés dans la littérature des THYR, un seul parmi les moins congrus a été écarté aléatoirement : x_BETA, et avec lui ceux des expositions aux paramètres météo : x_RAY et x_TEMP qui ont été intégrés pour les CATA. FE_SOCIO-ECO - très représentés - de fait x_attPHY et x_GINI.m ont été retirés aléatoirement parmi les i.st.e* conjoncturels redondants et peu pertinents*.

Les i.st.e* inclus dans $\mathcal{X}_{input}^{THYR}$ sont déclinés avec le code couleur environnemental associé :

x._SEXE x._AGE_DIAG x._DSUIVI x._TYPLEUC x._PROTOC x._RECHUT x._GREF x._IRACT x._ACPHY
x._GENE x._ORL x._PEDIA x._RADIO x._NEURs x._ORLs x._ENDOs x._HEMAs x._TEP
x._SCAN x._IRM x_APL_GENE x_txCHOM x_txFoyFisc x_txOUV x_GINIm x_RevMedp
x_RevMedm x_txMORT x_txAccPOP x_txAccNAT x_txBAC x_txEAR x_int.AGRI* x_tx.SAU
x_FDep99 x_FDep09 x_FDepXX x_attBIENS x_INSECU x_RADON x.._RADON x_NJAP x_FEFO
x_PEST x_URIN x_PREV x_EGRA x_GAMMA x_ALPHA x_3H x_238Pu x_125Sb
x_137Cs_SOL x_131i x_90Sr x_TOPO x_irsCd x_irsPb x_irsNi x_DJECr x_PERC (62/62)

REGRESSION : IDENTIFICATION MVG DES* VARIABLES EXPLICATIVES ET PREDICTIVES

Cible : $z'^{THYR}_{Uk,c}$ les prévalences spatiales pondérées EpiGéoStat converties en *patients-années.*

PHASE D'ANALYSE DES* DONNEES

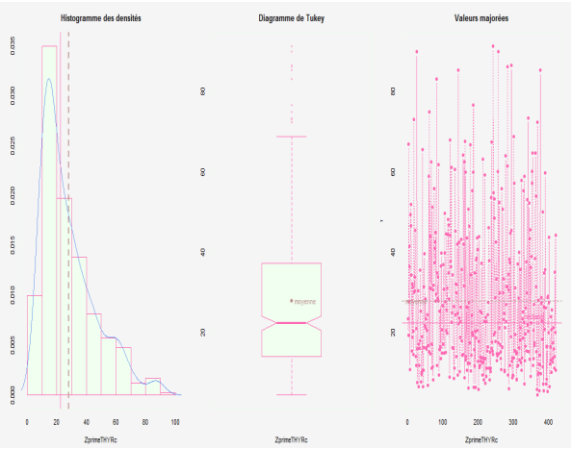


Figure 217 : Histogramme, Boite à moustache, valeurs de $z'^{THYR}_{Uk,c}$

Statistiques MVG

Moyenne spatiale : $\bar{z}'^{THYR}_{Uk,c} = 0{,}0279$

Ecart-type biaisé : $\hat{\sigma}(z'^{THYR}_{Uk,c}) = 0{,}01838$

PHASE 0 - CALIBRATION:

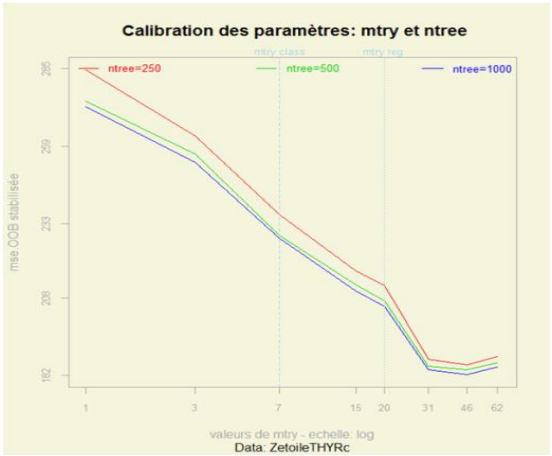


Figure 218 : $\bar{R}^{OOB}(\cdot)$ en fonction de ntree et mtry

Objectif

FA.naïve → FA.opt

Paramètres MVG optimisés:

$\hat{\Lambda}_j = (\text{ntree} = 1000\,;\, \text{mtry} = 46;\, \text{nodesize} = 3)$

PHASE 1 – HIERARCHISATION & ELIMINATION DES* VARIABLES DE BRUIT

Objectif : Estimer les $\overline{VI}_j(\cdot)$, hiérarchiser les $z^{THYR}_{(U_k),q}$, disjoindre de $\mathcal{X}_{input}^{THYR}$ en : $\mathcal{X}_{bruit}^{THYR}$ et $\mathcal{X}_{conserv}^{THYR}$

Estimation et présentation des scores *randomForest**

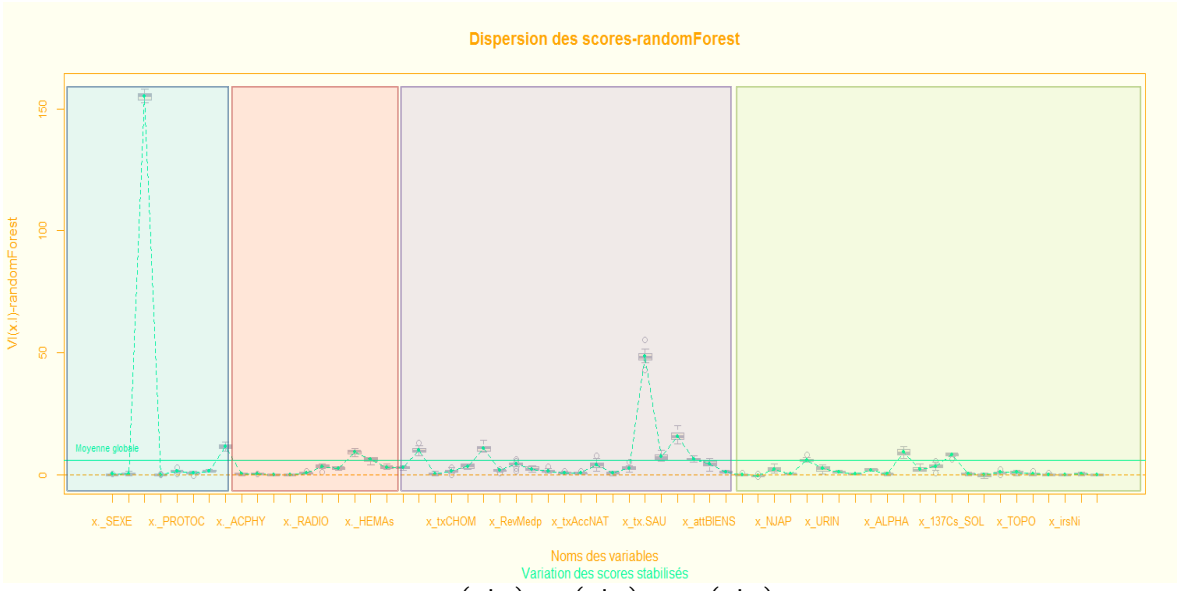


Figure 219 : Diagramme de Tukey des $VI_{j,r}\left(z^{j}_{(\cdot),.}\right)$, $\overline{VI}_j\left(z^{j}_{(\cdot),.}\right)$ et $\overline{\overline{VI}}_j\left(z^{j}_{(\cdot),.}\right)$ - estimation stabilisée sur 50 forêts.

Remarque : l'influence morbide des composantes environnementales peut être pré-classée à partir des scores des i.st.e, tel que : (i) FIM (ii) FE-SOCIO.ECO, (iii ex-aequo) FE-PHY.CHIM* et FE-SAN*

Hiérarchisation descendante

Valeurs des $\overline{VI}_j(\cdot)$ et des $\hat{\sigma}\left(VI_j(\cdot)\right)$ pour les 10 i.st.e* $x^{(I)}_{(U_k)}$ les plus importants sont :

	VI.moy	SD.VI.moy
x._DSUIVI	154.96870	3.04792
x_tx.SAU	48.51685	2.85778
x_FDep09	15.71430	2.51974
x._IRACT	11.37742	1.21350
x_txOUV	10.73310	1.49553
x._IRM	9.86013	1.34246
x_3H	9.06409	1.62006
x._ENDOs	9.00856	1.48597
x_137Cs_SOL	8.04113	1.07256
x_FDep99	7.24328	1.29537

Elimination des variables de bruit

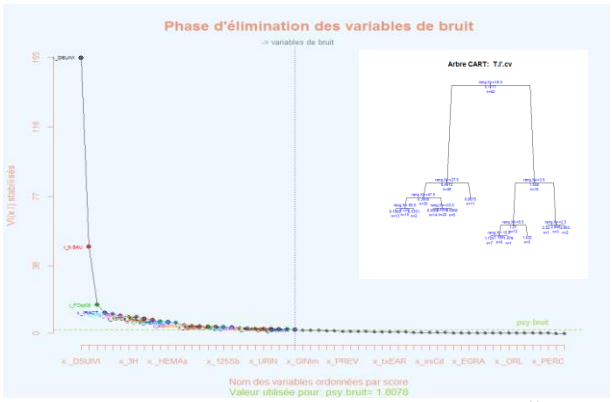


Figure 220 : Valeurs des $\overline{VI}_j(\cdot)$ associés aux $x^{(I)}_{(U_k)}$, et $\hat{T}^{j}_{MVG}((I))$

Estimation de ψ^{THYR}_{bruit}

Coefficient subjectif expert : $c^{THYR}_{\psi.bruit} = 10$

Disjonction des variables

$\mathcal{X}^{j}_{bruit}$: 34/62 et $\mathcal{X}^{j}_{conserv}$: 28/62

Variables éliminées identifiées comme du bruit

x_GINIm x._GREF x_txMORT x_txCHOM x._PROTOC x_PREV x_INSECU x_TOPO x_137Cs_BIOL x_txAccPOP x._RADIO x_txEAR x_txAccNAT x._RECHUT x._AGE_DIAG x_APL_GENE x_irsCd x_ALPHA x_131i x_DJECr x._ACPHY x_EGRA x_FEFO x._SEXE x._GENE x_irsPb x._ORL x._TYPLEUC x_irsNi x._PEDIA x_RADON x_PERC x_90Sr x.._RADON (34/62)

PHASE 2 SELECTION DES* VARIABLES EXPLICATIVES

Estimation de $\psi_{\text{explic}}^{\text{THYR}}$

Coefficient subjectif expert : $c_{\psi.\text{explic}}^{\text{THYR}} = 1$

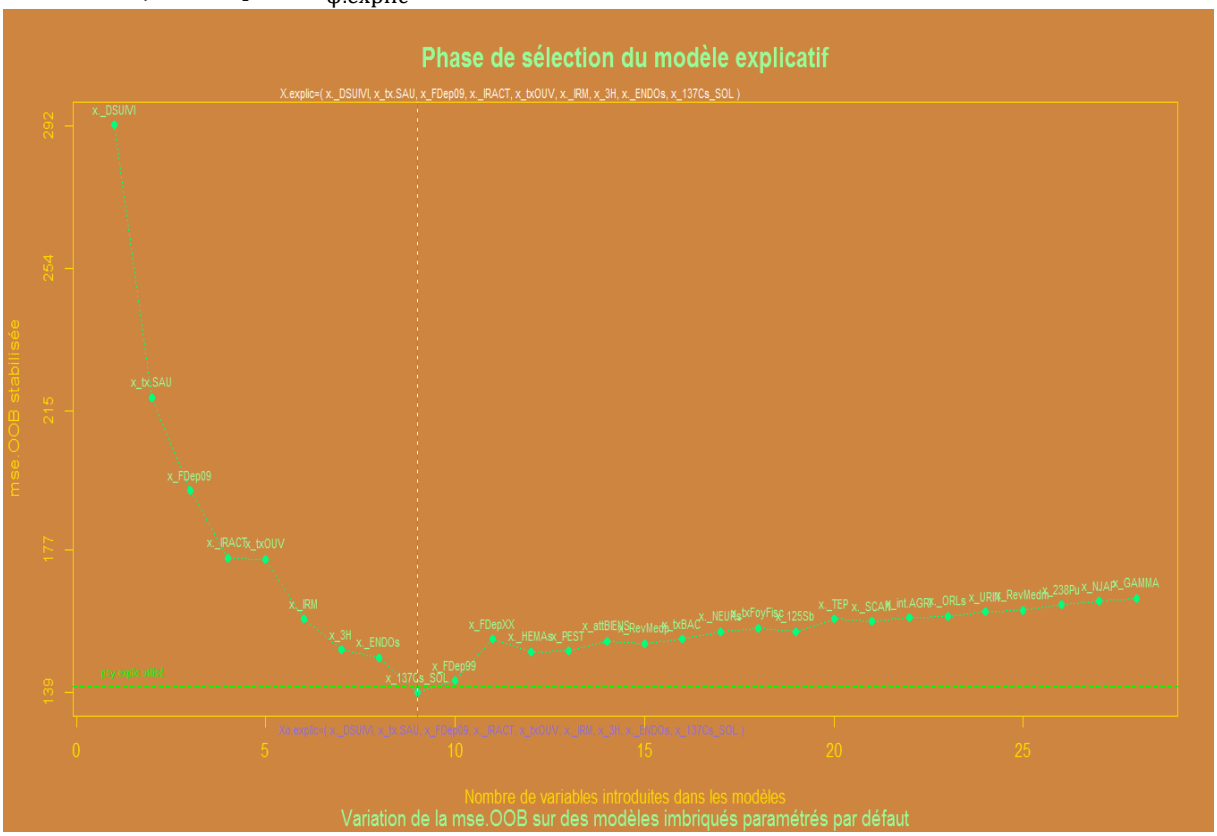


Figure 221 : Variation de $\bar{R}_{(.)}^{\text{OOB}}$ en fonction du nombre de variables imbriquées $\mathcal{X}_{\text{imbr.l}}^{\text{THYR}} \subset \mathcal{X}_{\text{conserv}}^{\text{THYR}}$;

Résultats - Paquets retenus : Génuer $\mathcal{X}_{\text{explic}}^{\text{THYR}}$

x._DSUIVI x_tx.SAU x_FDep09 x._IRACT x_txOUV x._IRM x_3H x._ENDOs x_137Cs_SOL (9/62)

Remarque : le paquet Bourrelly $\mathcal{X}_{0.explic}^{THYR}$ est identique au paquet Génuer $\mathcal{X}_{explic}^{THYR}$.

PHASE 3 - VARIABLES PREDICTIVES

Estimation de ψ_{pred}^{THYR}

Coefficient subjectif expert : $c_{\psi.pred}^{THYR} = 4$

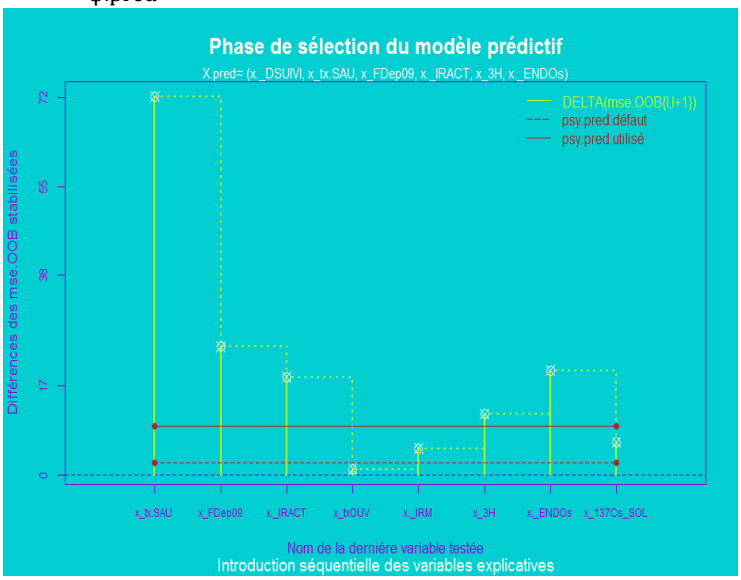


Figure 222 Valeur de $\overline{\Delta}(R_l^{OOB})$ en fonction de $x_{(U_k)}^{l}$ testé

Résultats - $\mathcal{X}_{pred}^{THYR}$ contient :

x._DSUIVI x_tx.SAU x_FDep09 x._IRACT x_3H x._ENDOs (6/62)

ANALYSE DE LA QUALITE DES* MODELES MVG

Estimation OOB de qualité des modèles MVG utilisables $\hat{f}_{FA}^{THYR}(\mathcal{X}_{\cdot}^{j}|\Lambda_j)$:

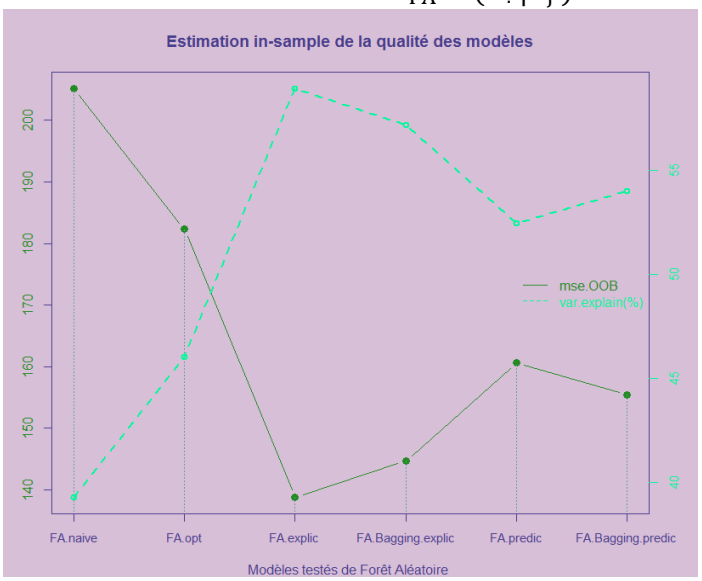


Figure 223 : Variation de la mse.OOB et de la var.explain.OOB en fonction du modèle MVG testé

Remarque : les valeurs de $\overline{R}^{OOB-IS}(\cdot)$ et de $\mathbb{var.explain}_{\hat{f}_{FA}^{j}.}^{OOB-IS}$ sont estimées *In-Sample*

ANALYSE DE LA QUALITE PREDICTIVE

MODELE:	FA.naive	FA.opt	FA.explic	FA.pred
var.explain.OOB	39,28%	46,04%	58,91%	52,45%
gain.OOB.absolu	0,00%	5,61%	18,45%	14,64%
gain.OOB.relatif	0,00%	11,21%	36,86%	29,24%

Tableau 41 : Analyse des gains absolus et relatifs de mse.OOB et de var.explain.OOB par rapport à une FA.naïve

REGRESSION : PREDICTIONS GEOGRAPHIQUES MVG

L'application de MVG à $z'^{THYR}_{(U_k),c}$ permet d'obtenir le prédicteur MVG : $\hat{r}^{THYR}_{MVG}(x^{j}_{MVG}|\Lambda_{j-MVG})$. Le vecteur des paramètres spécifiés est : $\Lambda_{j-MVG} = (ntree = 1000\,;\, mtry = ([p_1/3] \vee 2)\,;\, nodesize = 3)$ et $\{x^{THYR}_{MVG} = x^{THYR}_{explic}\}$. Les résultats présentés sont conformes aux propositions de début de section :

Cartographie des prévalences spatiales EpiGéoStat converties en *patients-années*

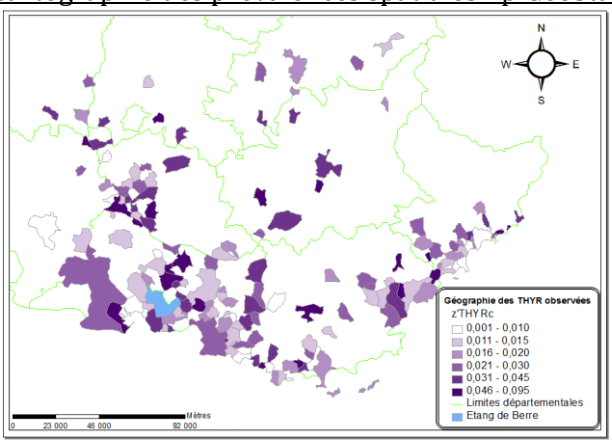


Figure 224 : Valeurs observées prises par $z'^{THYR}_{(U_k),c}$

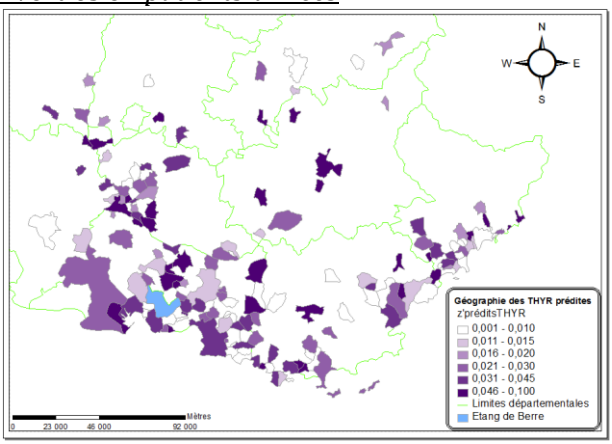
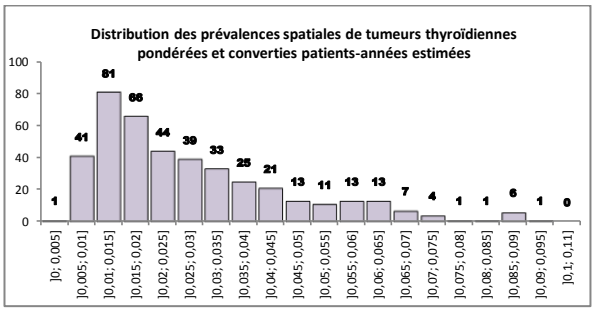


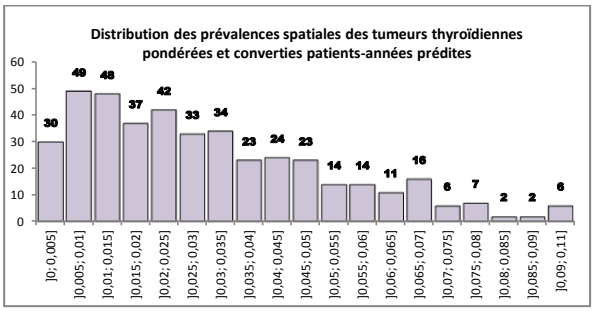
Figure 225 : Valeurs MVG prédites prises par $\hat{z}'^{THYR}_{(U_k),c}$

Distribution statistique des prévalences spatiales EpiGéoStat converties en *patients-années*



421 - Uk	Paramètres de dispersion & de position						
Estimateur	min(.)	Q1^(.)	mêd()	Q3^(.)	max(.)	môy(.)	σ^(.)
Estimation	0,005	0,014	0,022	0,037	0,091	0,0279	0,0184

Figure 226 : Distribution et tableau statistique des estimés $z'^{THYR}_{(U_k),c}$



421 - Uk	Paramètres de dispersion & de position						
Estimateur	min(.)	Q1^(.)	mêd()	Q3^(.)	max(.)	môy(.)	σ^(.)
Estimation	0,000	0,013	0,025	0,044	0,100	0,0303	0,0217

Figure 227 : Distribution et tableau statistique des prédictions $\hat{z}'^{THYR}_{(U_k),c}$

Cartographie, graphique et synthèse statistique des résidus géographiques standardisés

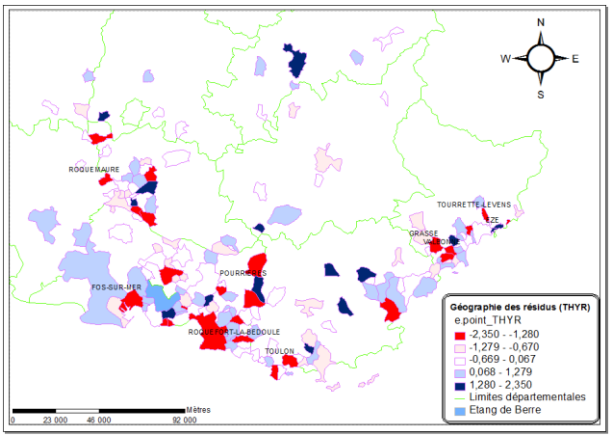
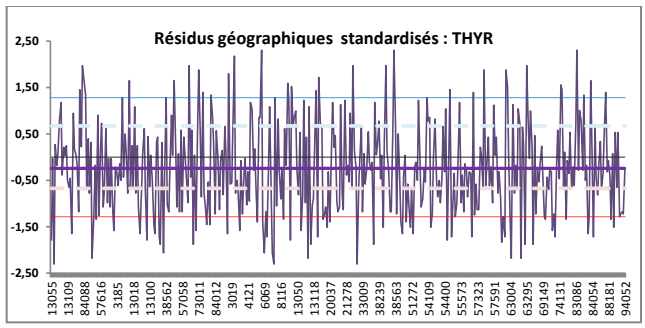


Figure 228 : Valeurs observées prises par $\hat{\varepsilon}^{THYR}_{(U_k)}$ dans les U_k situées en région PACA et aux alentours



Estimateur de qualité statistique	Estimation
var.explain.OOB($\hat{z}'^{j}_{(U_k),c}$)	58,91%
var.explain.IS($\hat{z}'^{j}_{(U_k),c}$)	60,36%
môy($\hat{\varepsilon}^{j}_{(U_k)}$)	-0,239
corr.($\hat{z}'^{j}_{(U_k),c}$; $z'^{j}_{(U_k),c}$)	88,55%

Figure 229 : Valeurs et synthèse statistique des $\hat{\varepsilon}^{THYR}_{(U_k)}$

CLASSIFICATION : IDENTIFICATION MVG DES* VARIABLES EXPLICATIVES ET PREDICTIVES

Cible : $z'^{THYR}_{Uk,q}$ propension spatiale pondérée EpiGéoStat et par le ratio participation/exposition

PHASE D'ANALYSE DES* DONNEES

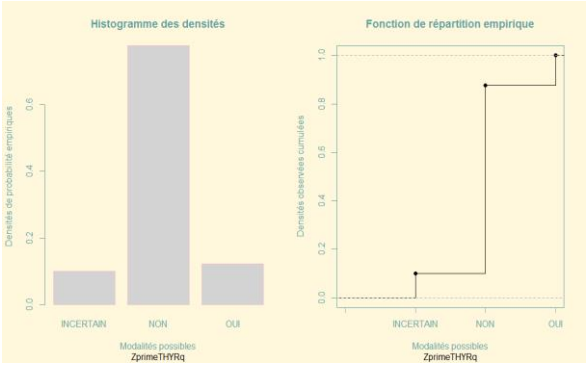


Figure 230 : Histogramme, fonction de répartition, de $z'^{THYR}_{Uk,q}$

Statistiques MVG

Nombre de modalités $\text{card}\left(z^{j}_{(U_k),q} = c_j \right)$

INCERTAIN	NON	OUI
42	327	52

Proportions des modalités $\text{môy}\left(z^{j}_{(U_k),q} = c_j \right)$

INCERTAIN	NON	OUI
9.98	77,67	12.35

PHASE 0 - CALIBRATION:

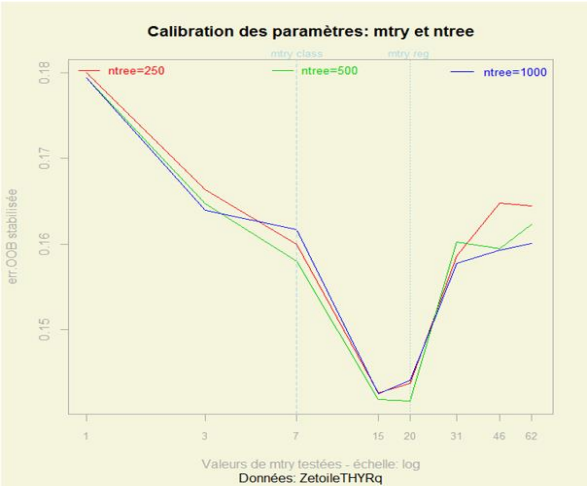


Figure 231 $\bar{R}^{OOB}_{\cdot}(\cdot)$ en fonction de ntree et mtry

Objectif

FA.naïve → FA.opt

Paramètres MVG optimisés:

$\hat{\Lambda}_{THYR} = (\text{ntree} = 1000\,; \text{mtry} = 20; \text{nodesize} = 1)$

PHASE 1 – HIERARCHISATION & ELIMINATION DES* VARIABLES DE BRUIT

Objectif : Estimer les $\overline{VI}_j(\cdot)$, hiérarchiser les $z^{THYR}_{(U_k),q}$, disjoindre de $\mathcal{X}^{THYR}_{input}$ en : $\mathcal{X}^{THYR}_{bruit}$ et $\mathcal{X}^{THYR}_{conserv}$

Estimation et présentation des scores *randomForest**

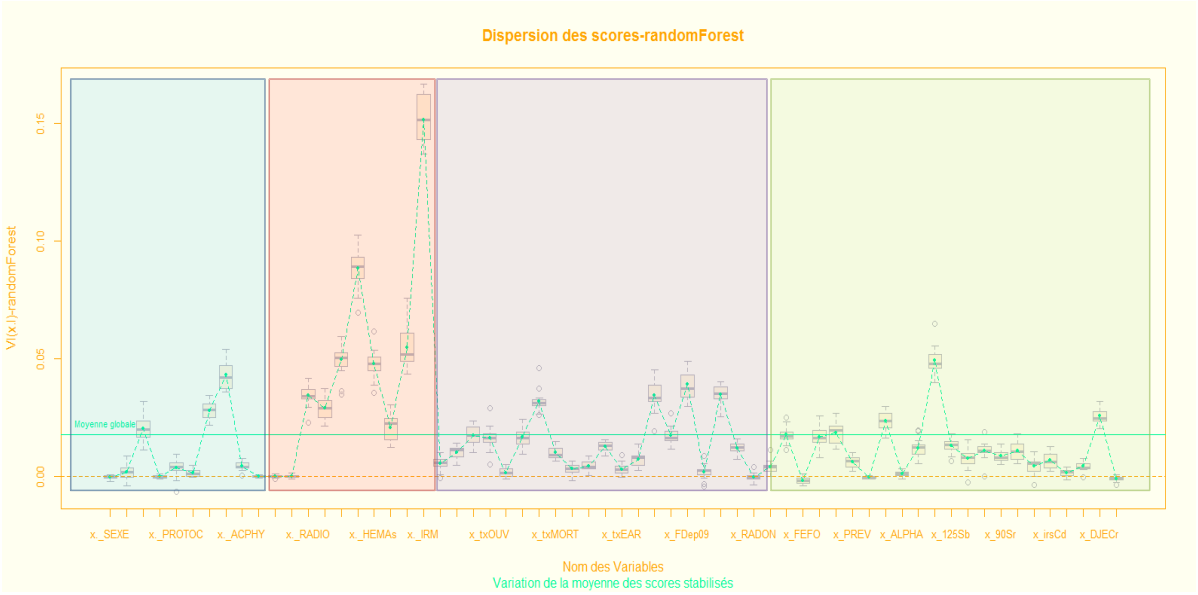


Figure 232 : Diagramme de Tukey des $VI_{j,r}\left(z^{j}_{(\cdot),\cdot} \right)$, $\overline{VI}_j\left(z^{j}_{(\cdot),\cdot} \right)$ et $\overline{\overline{VI}}_j\left(z^{j}_{(\cdot),\cdot} \right)$ - estimation stabilisée sur 50 forêts.

Remarque : l'influence morbide des composantes environnementales peut être pré-classée à partir des scores des i.st.e, tel que (i) FE-SAN, (ii exequo) FE-PHY.CHIM* et FIM*, (iii) FE-SOCIO.ECO*.*

Hiérarchisation descendante

Valeurs des $\overline{\mathrm{VI}}_j(\cdot)$ et des $\widehat{\sigma}\left(\mathrm{VI}_j(\cdot)\right)$ pour les 10 i.st.e* : $x^{(l)}_{(U_k)}$ les plus importants

	VI.moy	SD.VI.moy
x._IRM	0.15152	0.01078
x._ENDOs	0.08849	0.00819
x._SCAN	0.05491	0.00711
x._ORLs	0.04961	0.00542
x_238Pu	0.04915	0.00585
x._HEMAs	0.04778	0.00646
x._IRACT	0.04303	0.00536
x_FDep09	0.03904	0.00711
x_attBIENS	0.03468	0.00505
x_tx.SAU	0.03437	0.00543

Elimination des variables de bruit

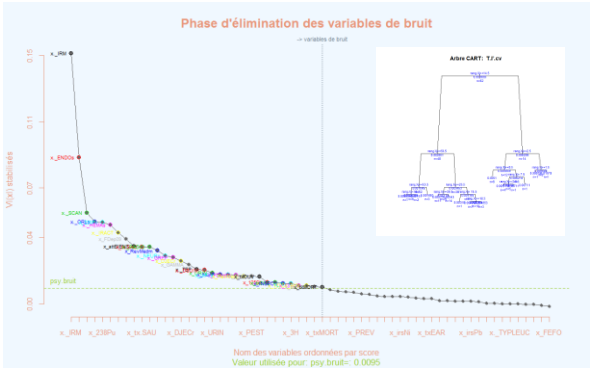


Figure 233 : Valeurs des $\overline{\mathrm{VI}}_j(\cdot)$ associés aux $x^{(l)}_{(U_k)}$, et $\widehat{\mathbf{T}}^{j}_{\mathbf{MVG}}((\mathbf{l}))$

Estimation de $\psi^{\mathrm{THYR}}_{\mathrm{bruit}}$

Coefficient subjectif expert : $c^{\mathrm{THYR}}_{\psi.\mathrm{bruit}} = 10$

Disjonction des variables

$\mathcal{X}^{j}_{\mathrm{bruit}}$: 29/62 et $\mathcal{X}^{j}_{\mathrm{conserv}}$:33/62

Variables éliminées identifiées comme du bruit

x_90Sr x_137Cs_SOL x_int.AGRI* x_irsCd x_PREV x_APL_GENE x_txAccNAT x._ACPHY x_TOPO x_irsNi x.._RADON x._PROTOC x_txAccPOP x_txEAR x_FDepXX x._AGE_DIAG x_GINIm x._RECHUT x_irsPb x_ALPHA x._ORL x._PEDIA x._GENE x._TYPLEUC x_RADON x._SEXE x_EGRA x_PERC x_FEFO (29/62)

PHASE 2 - SELECTION DES* VARIABLES EXPLICATIVES

Estimation de $\psi^{\mathrm{THYR}}_{\mathrm{explic}}$

Coefficient subjectif expert : $c^{\mathrm{THYR}}_{\psi.\mathrm{explic}} = 2$

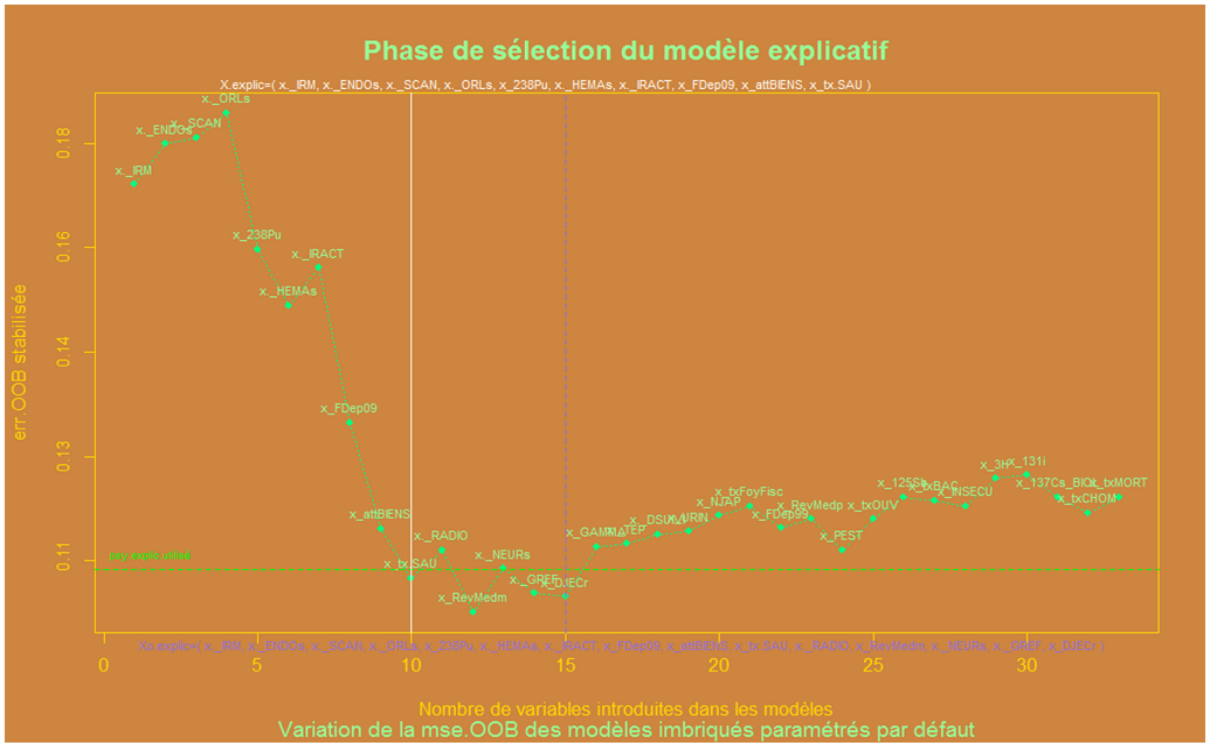


Figure 234 : Variation de $\overline{\mathrm{R}}^{\mathrm{OOB}}_{(.)}$ en fonction du nombre de variables imbriquées $\mathcal{X}^{\mathrm{THYR}}_{\mathrm{imbr.l}} \subset \mathcal{X}^{\mathrm{THYR}}_{\mathrm{conserv}}$;

Résultats - Paquets retenus : Génuer $\mathcal{X}_{\text{explic}}^{\text{THYR}}$

x._IRM x._ENDOs x._SCAN x._ORLs x_238Pu x._HEMAs x._IRACT x_FDep09 x_attBIENS x_tx.SAU (10/62)

*Remarque : le paquet Bourrelly $\mathcal{X}_{0.explic}^{j}$ contient aussi les i.st.e**

[…] x._RADIO x_RevMedm x._NEURs x._GREF x (15/62)

PHASE 3 - VARIABLES PREDICTIVES

Estimation de $\psi_{\text{pred}}^{\text{THYR}}$: Coefficient subjectif expert : $c_{\psi.\text{pred}}^{\text{THYR}} = 1$

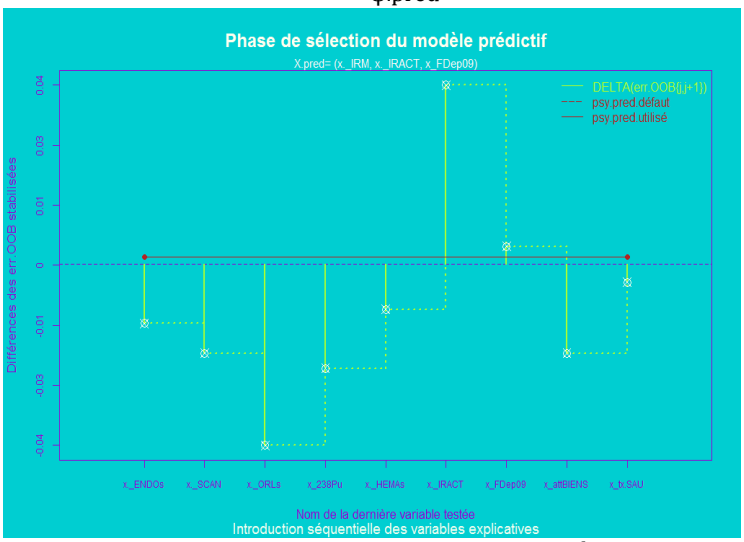


Figure 235 : Valeur de $\overline{\Delta}(R_l^{OOB})$ en fonction de $x_{(U_k)}^{l}$ testé

Résultats - $\mathcal{X}_{\text{pred}}^{\text{THYR}}$ contient : x._IRM x._IRACT FDep09 (3/62)

ANALYSE DE LA QUALITE DES* MODELES MVG

Estimation OOB de qualité des modèles MVG utilisables $\hat{f}_{\text{FA}}^{\text{THYR}}(\mathcal{X}_{\cdot}^{j}|\Lambda_j)$:

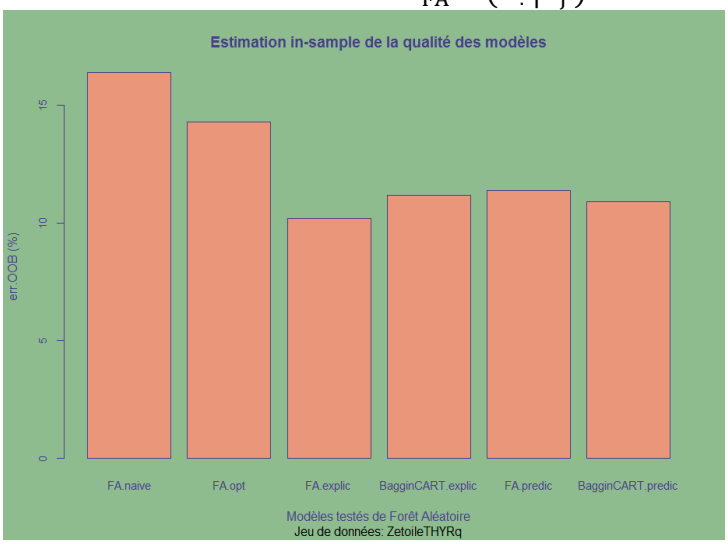


Figure 236 : Variation de l'err.OOB en fonction du modèle MVG testé

Remarque : les valeurs de $\overline{\text{R}}^{\text{OOB}-\text{IS}}(\cdot)$ sont estimées *In-Sample* (IS)

ANALYSE DE LA QUALITE PREDICTIVE

MODELE:	FA.naive	FA.opt	FA.explic	FA.pred
err.OOB généralisée :	**16,17%**	**14,25%**	**10,21%**	**11,40%**
gain.OOB.absolu :	0,00%	1,92%	5,96%	4,77%
gain.OOB.relatif :	0,00%	11,86%	36,84%	29,49%

Tableau 42 : Analyse des gains absolus et relatifs d'err.OOB généralisée par rapport à une FA.naïve

Classification : PREDICTIONS GEOGRAPHIQUES MVG

L'application de MVG à $z'^{THYR}_{(U_k),q}$ permet d'obtenir le prédicteur MVG : $\hat{f}^{THYR}_{MVG}(x^{j}_{MVG}|\Lambda_{j-MVG})$. Le vecteur des paramètres spécifiés est : $\Lambda_{j-MVG} = (ntree = 1000\,; mtry = (\lfloor\sqrt{p_l}\rfloor \vee 2)\,; nodesize = 1)$ et $\{x^{THYR}_{MVG} = x^{THYR}_{explic}\}$. Les résultats présentés sont conformes aux propositions de cette section :

Cartographie des propensions spatiales EpiGéoStat répétées au prorata du ratio participation/exposition

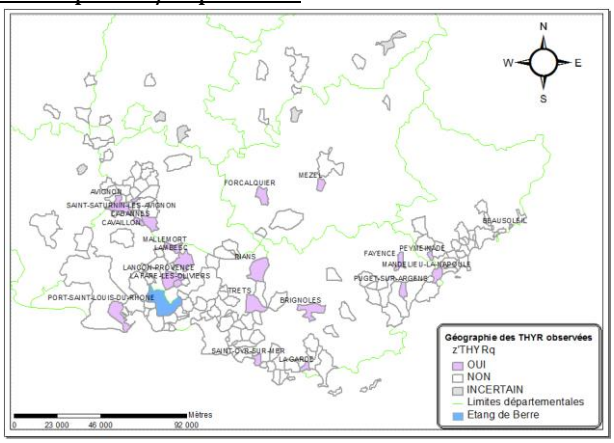


Figure 237 : Valeurs observées prises par $z'^{THYR}_{(U_k),q}$

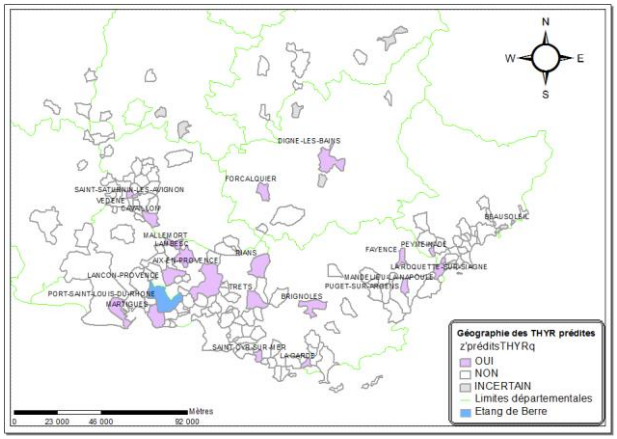


Figure 238 : Valeurs MVG prédites prises par $\hat{z}'^{THYR}_{(U_k),q}$

Distribution des propensions spatiales pondérées EpiGéoStat répétées au prorata du ratio participation/exposition

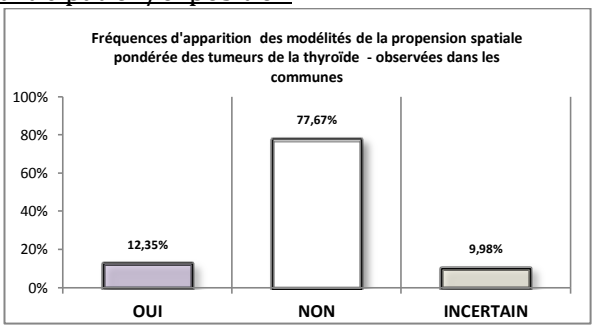


Figure 239 : Fréquences spatiales estimées des modalités de $z'^{THYR}_{(U_k),q}$

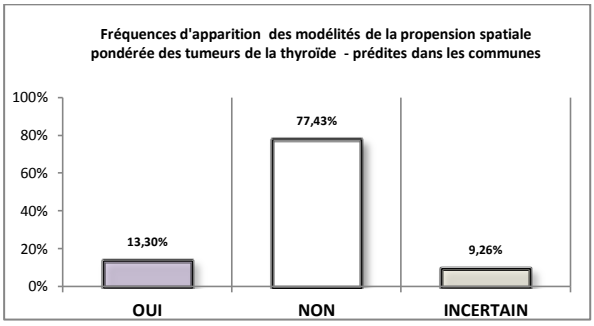


Figure 240 : Fréquences spatiales prédites des modalités de $\hat{z}'^{THYR}_{(U_k),q}$

Cartographie, graphique et synthèse statistique des confusions géographiques

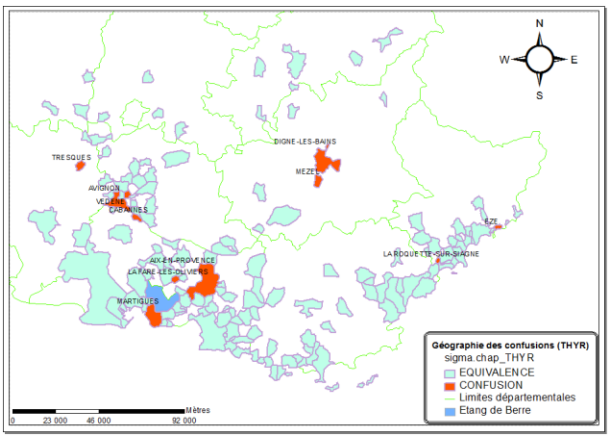
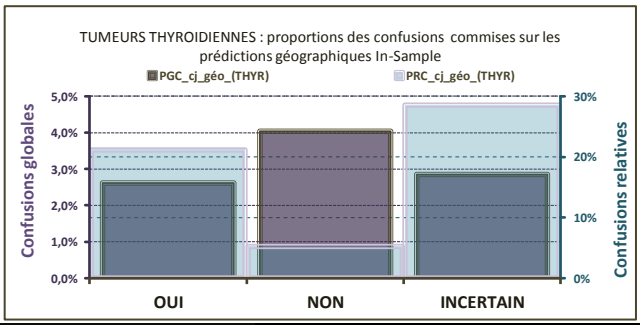


Figure 241 : Valeurs prises par $\hat{\varsigma}^{THYR}_{(U_k)}$ dans les U_k sises en région PACA et aux alentours



Synthèse : CONFUSIONS	OUI	NON	INCERTAIN	TOTAL
NTC_cj_géo_($\hat{z}'^{j}_{(U_k),q}$)	11	17	12	40
PRC_cj_géo_($\hat{z}'^{j}_{(U_k),q}$)	21,15%	5,20%	28,57%	-
PGC_cj_géo_($\hat{z}'^{j}_{(U_k),q}$)	2,61%	4,04%	2,85%	9,50%

Figure 242 : Synthèse et représentation graphique des $\hat{\varsigma}^{THYR}_{(U_k)}$

SEQUELLE : TUMEURS SECONDAIRES

JEU DE DONNEES D'APPRENTISSAGE UTILISE : $\mathcal{L}_{U_k}^{TUM2}$

Les variables réponses sont les i.st.m* $z'^{TUM2}_{(U_k),c}$ ou $z'^{TUM2}_{(U_k),q}$ - ils modélisent la géographie des tumeurs secondaires. Les i.st.e* : $x'^{l}_{(U_k)}$ et $x^{l}_{(U_k)}$ écartés sont ceux qui modélisent les expositions spatiotemporelles aux : FE-SAN* Curieux* les plus aberrants : x_OPHTs et x_APL.OPHT et x._TEP – bien que x._TEP soit peut être plus pertinent que x._SCAN. Les FE_PHY.CHIM caractérisant les expositions à la radioactivité environnementale sont surreprésentés mais particulièrement discutés pour tous les types de cancer, de fait, seules les expositions aux paramètres météo : x_RAY et x_TEMP ont été écartées - puisque intégrées pour les CATA. Les FE_SOCIO-ECO redondants : x_attBIEN et x_GINI.p. par symétrie à ce qui a été fait pour les THYR.

Les i.st.e* inclus dans $\mathcal{X}_{input}^{TUM2}$ sont déclinés avec le code couleur environnemental associé :

x._SEXE x._AGE_DIAG x._DSUIVI x._TYPLEUC x._PROTOC x._RECHUT x._GREF x._IRACT x._ACPHY x._GENE x._ORL x._PEDIA x._NEURs x._ORLs x._ENDOs x._HEMAs x._SCAN x._IRM x_CAM x_APL_GENE x_txCHOM x_txFoyFisc x_txOUV x_GINIm x_RevMedm x_RevMedp x_txMORT x_txAccPOP x_txAccNAT x_txBAC x_txEAR x_int.AGRI* x_tx.SAU x_FDep99 x_FDep09 x_FDepXX x_attPHY x_INSECU x_RADON x.._RADON x_NJAP x_FEFO x_PEST x_URIN x_PREV x_EGRA x_GAMMA x_ALPHA x_BETA x_3H x_238Pu x_125Sb x_137Cs_SOL x_131I x_90Sr x_137Cs_BIOL x_TOPO x_irsCd x_irsPb x_irsNi x_DJECr x_PERC (62/62)

REGRESSION : IDENTIFICATION MVG DES* VARIABLES EXPLICATIVES ET PREDICTIVES

Cible : $z'^{TUM2}_{Uk,c}$ les prévalences spatiales pondérées EpiGéoStat converties en *patients-années*

PHASE D'ANALYSE DES* DONNEES

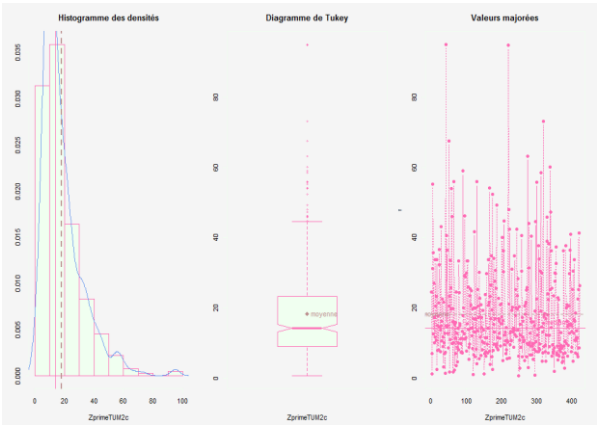


Figure 243 : Histogramme, boite à moustache, valeurs de $z'^{TUM2}_{Uk,c}$

Statistiques MVG

Moyenne spatiale : $\bar{z}'^{TUM2}_{Uk,c} = 0{,}01826$

Ecart-type biaisé : $\hat{\sigma}\left(z'^{TUM2}_{Uk,c}\right) = 0{,}01372$

PHASE 0 - CALIBRATION:

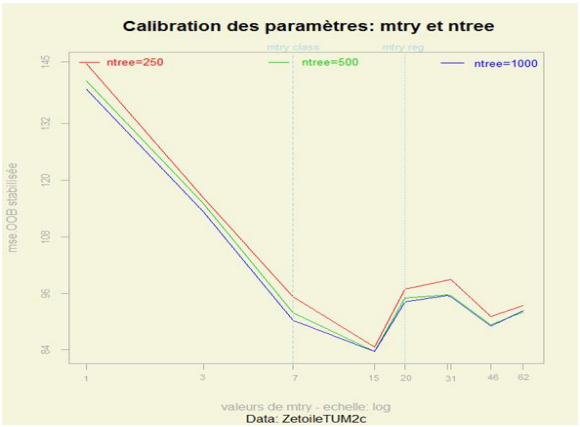


Figure 244 : $\overline{R}^{OOB}_{\cdot}(\cdot)$ en fonction de ntree et mtry

Objectif

FA.naïve → FA.opt

Paramètres MVG optimisés:

$\widehat{\Lambda}_j = (\text{ntree} = 500\,;\, \text{mtry} = 15;\, \text{nodesize} = 3)$

PHASE 1 – HIERARCHISATION & ELIMINATION DES* VARIABLES DE BRUIT

Objectif : Estimer les $\overline{VI}_j(\cdot)$, hiérarchiser les $z^{TUM2R}_{(U_k),c}$, disjoindre de $\mathcal{X}_{input}^{TUM2}$ en : $\mathcal{X}_{bruit}^{TUM2}$ et $\mathcal{X}_{conserv}^{TUM2}$

Estimation et présentation des scores *randomForest**

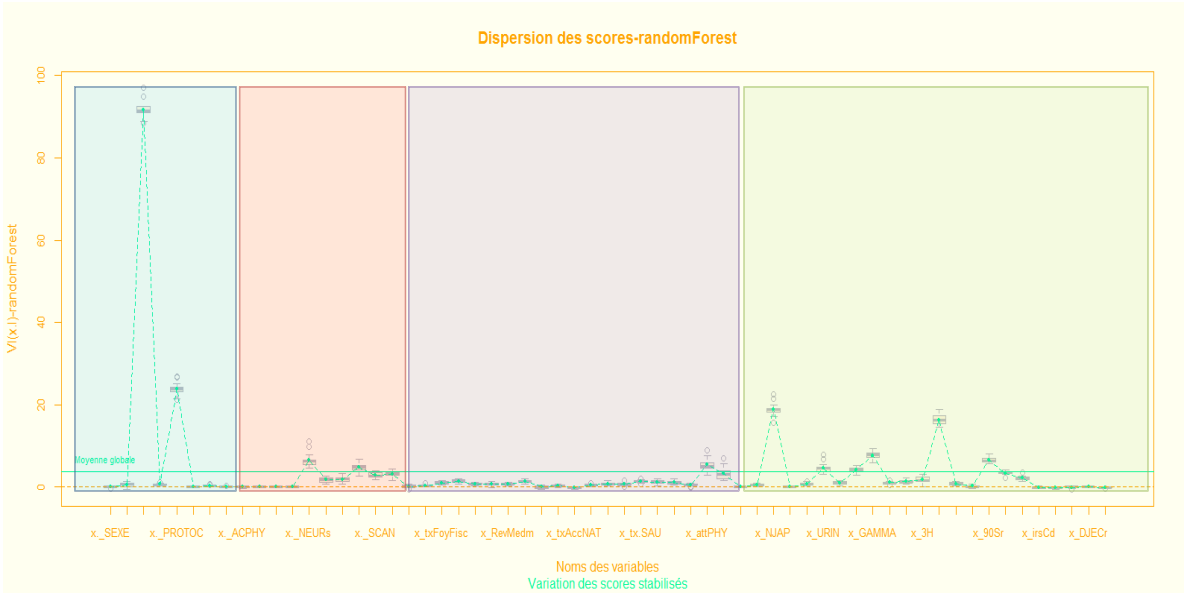


Figure 245 : Diagramme de Tukey des $VI_{j,r}\left(z^{j}_{(\cdot),.}\right)$, $\overline{VI}_j\left(z^{j}_{(\cdot),.}\right)$ et $\overline{\overline{VI}}_j\left(z^{j}_{(\cdot),.}\right)$ - estimation stabilisée sur 50 forêts.

Remarque : l'influence morbide des composantes environnementales peut être pré-classée à partir des scores des i.st.e, tel que : (i) FIM (ii) FE-PHY.CHIM*, (iii ex-aequo) FE-SOCIO.ECO* et FE-SAN*

Hiérarchisation descendante

Valeurs des $\overline{VI}_j(\cdot)$ et des $\hat{\sigma}\left(VI_j(\cdot)\right)$ pour les 10 i.st.e* : $x^{(l)}_{(U_k)}$ les plus importants :

	VI.moy	SD.VI.moy
x._DSUIVI	91.74778	2.39392
x._PROTOC	23.85804	1.39719
x_NJAP	18.82897	1.57253
x_238Pu	16.34323	1.59018
x_GAMMA	7.71471	1.10406
x._NEURs	6.54465	1.36086
x_90Sr	6.53623	0.93773
x_attPHY	5.37593	1.04316
x._HEMAs	4.80027	1.00791
x_URIN	4.62491	0.88775

Elimination des variables de bruit

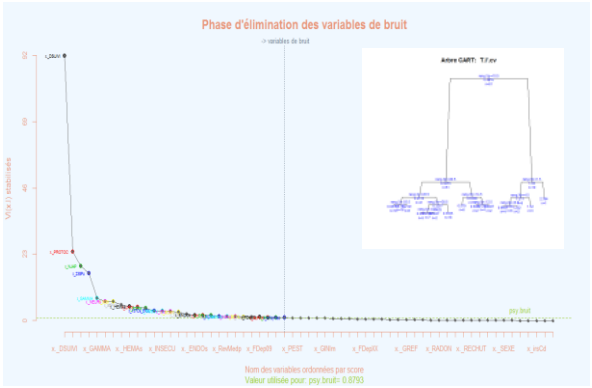


Figure 246 : Valeurs des $\overline{VI}_j(\cdot)$ associés aux $x^{(l)}_{(U_k)}$, et $\hat{T}^{j}_{MVG}((l))$

Estimation de ψ^{TUM2}_{bruit}

Coefficient subjectif expert : $c^{TUM2}_{\psi.bruit} = 10$

Disjonction des variables

$\mathcal{X}^{j}_{bruit}$: 32/62 et $\mathcal{X}^{j}_{conserv}$:30/62

Variables éliminées identifiées comme du bruit

x_PEST x_RevMedm x_CAME x_125Sb x_GINIm x_txEAR x_int.AGRI* x._TYPLEUC x._AGE_DIAG x_FDepXX x.._RADON x_txBAC x_txCHOM x_131I x_txAccPOP x._GREF x_137Cs_SOL x_APL_GENE x._IRACT x_RADON x_FEFO x_DJECr x._ACPHY x._RECHUT x._ORL x._GENE x._PEDIA x._SEXE x_txMORT x_irsNi x_PERC x_irsCd x_irsPb txAccNAT (32/62)

PHASE 2 - SELECTION DES* VARIABLES EXPLICATIVES

Estimation de $\psi_{\text{explic}}^{\text{TUM2}}$

Coefficient subjectif expert : $c_{\psi.\text{explic}}^{\text{TUM2}} = 1$

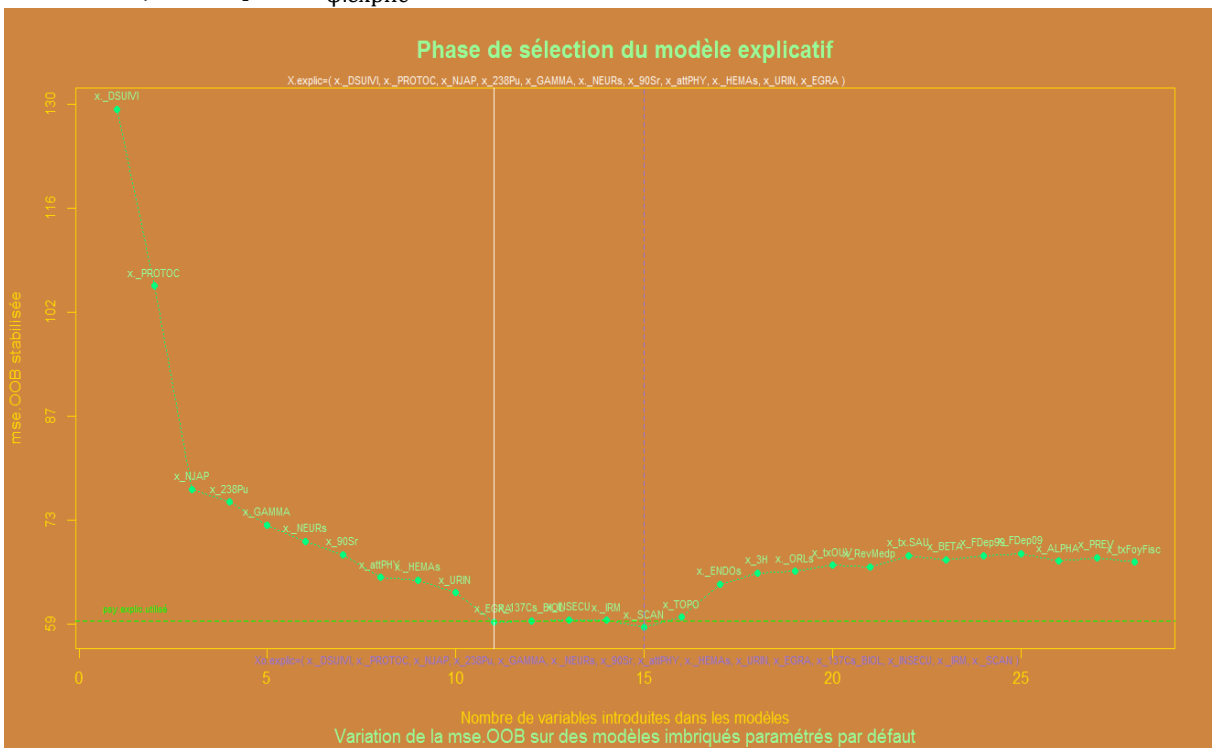


Figure 247 : Variation de $\overline{\text{R}}_{(.)}^{\text{OOB}}$ en fonction du nombre de variables imbriquées $\mathcal{X}_{\text{imbr.l}}^{\text{TUM2}} \subset \mathcal{X}_{\text{conserv}}^{\text{TUM2}}$;

Résultats - Paquets retenus : Génuer $\mathcal{X}_{\text{explic}}^{\text{TUM2}}$

x._DSUIVI x._PROTOC x_NJAP x_238Pu x_GAMMA x._NEURs x_90Sr x_attPHY x._HEMAs x_URIN x_EGRA (11/62)

Remarque : le paquet Bourrelly $\mathcal{X}_{0.explic}^{TUM2}$ contient aussi les i.st.e :*

[…] x_137Cs_BIOL x_INSECU x._IRM x._SCAN (15/62)

PHASE :3 VARIABLES PREDICTIVES

Estimation de ψ_{pred}^{TUM2}

Coefficient subjectif expert : $c_{\psi.pred}^{TUM2} = 2$

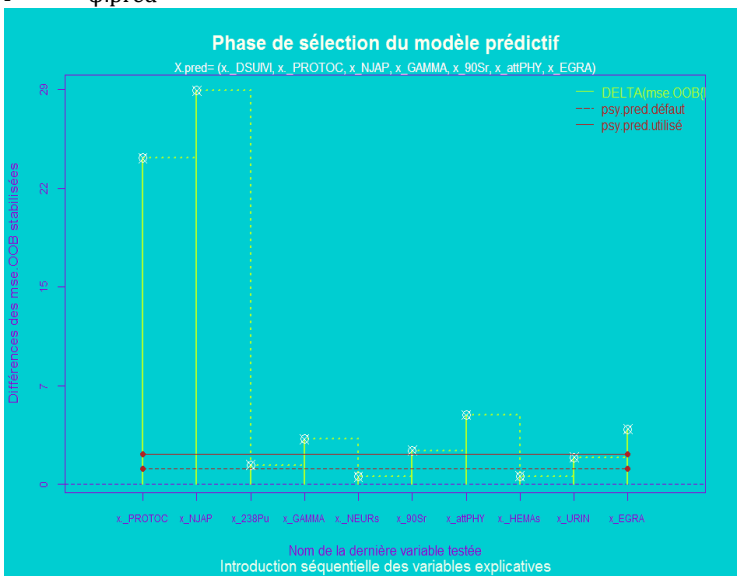


Figure 248 : Valeur de $\overline{\Delta}(R_l^{OOB})$ en fonction de $x_{(U_k)}^{l}$ testé

Résultats - $\mathcal{X}_{pred}^{TUM2}$ contient :

x._DSUIVI x._PROTOC x_NJAP x_GAMMA x_90Sr x_attPHY x_EGRA (7/62)

ANALYSE DE LA QUALITE DES* MODELES MVG

Estimation OOB de qualité des modèles MVG utilisables $\hat{f}_{FA}^{TUM2}(\mathcal{X}_{.}^{j}|\Lambda_j)$:

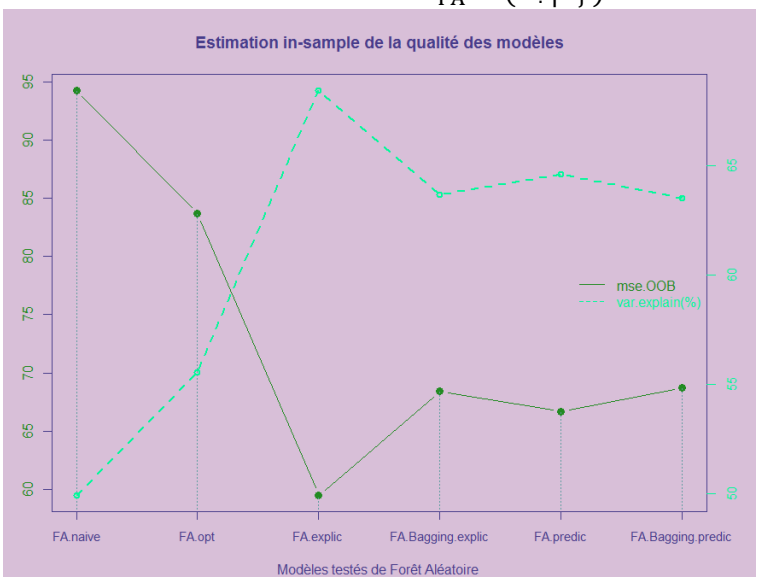


Figure 249 : Variation de la mse.OOB et de la var.explain.OOB en fonction du modèle MVG testé

Remarque : les valeurs de $\overline{R}^{OOB-IS}(\cdot)$ et de $\mathbb{var.explain}_{\hat{f}_{FA}^{j}}^{OOB-IS}$ sont estimées *In-Sample*

ANALYSE DE LA QUALITE PREDICTIVE

MODELE:	FA.naive	FA.opt	FA.explic	FA.pred
var.explain.OOB	49,94%	55,55%	68,39%	64,58%
gain.OOB.absolu	0,00%	5,61%	18,45%	14,64%
gain.OOB.relatif	0,00%	11,21%	36,86%	29,24%

Tableau 43 : Analyse des gains absolus et relatifs de mse.OOB et de var.explain.OOB par rapport à une FA.naïve

REGRESSION : PREDICTIONS GEOGRAPHIQUES MVG

L'application de MVG à $z'^{TUM2}_{(U_k),c}$ permet d'obtenir le prédicteur MVG : $\hat{f}^{j}_{MVG}(x^{j}_{MVG}|\Lambda_{j-MVG})$. Le vecteur des paramètres spécifiés est : $\Lambda_{j-MVG} = (\text{ntree} = 1000\,;\text{mtry} = ([p_1/3] \vee 2)\,;\text{nodesize} = 3)$ et $\{x^{TUM2}_{MVG} = x^{TUM2}_{explic}\}$. Les résultats présentés sont conformes aux propositions de début de section :

Cartographie des prévalences spatiales EpiGéoStat converties en *patients-années*

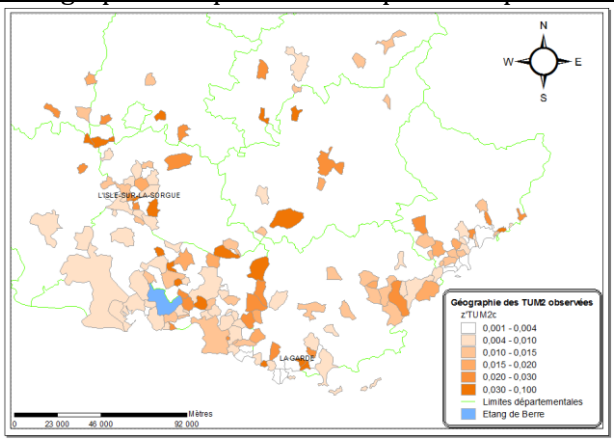


Figure 250 : Valeurs observées prises par $z'^{TUM2}_{(U_k),c}$

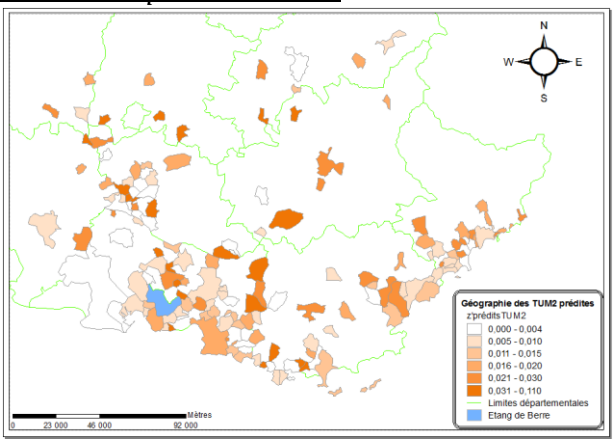
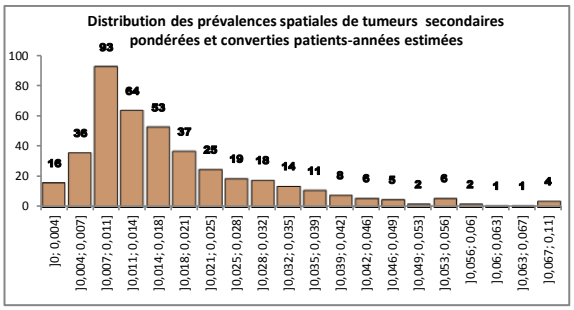


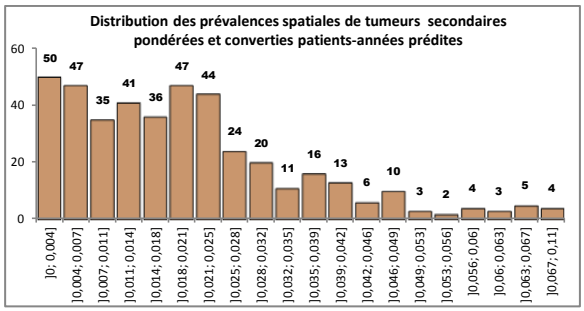
Figure 251 : Valeurs MVG prédites prises par $\hat{z}'^{TUM2}_{(U_k),c}$

Distribution statistique des prévalences spatiales EpiGéoStat converties en *patients-années*



421 - Uk	Paramètres de dispersion & de position						
Estimateur	min(.)	Q1^(.)	mêd()	Q3^(.)	max(.)	môy(.)	σ^(.)
Estimation	0,001	0,009	0,014	0,023	0,095	0,0183	0,0137

Figure 252 : Distribution et tableau statistique des estimés $z'^{TUM2}_{(U_k),c}$



421 - Uk	Paramètres de dispersion & de position						
Estimateur	min(.)	Q1^(.)	mêd()	Q3^(.)	max(.)	môy(.)	σ^(.)
Estimation	0,000	0,008	0,018	0,026	0,106	0,0199	0,0157

Figure 253 : Distribution et tableau statistique des prédictions $\hat{z}'^{TUM2}_{(U_k),c}$

Cartographie, graphique et synthèse statistique des résidus géographiques standardisés

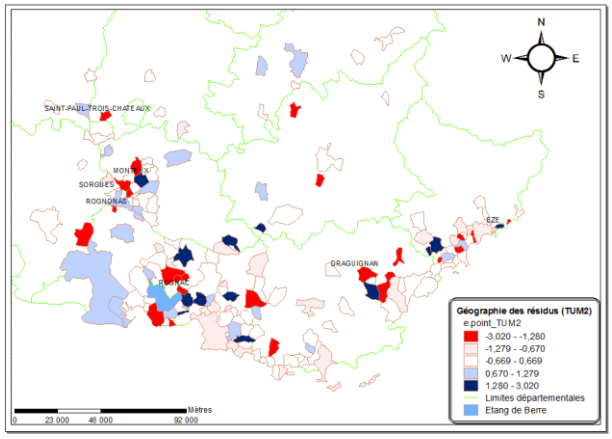
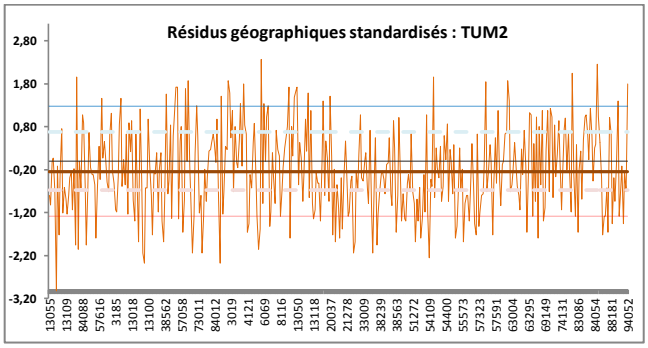


Figure 254 : Valeurs observées prises par $\dot{\varepsilon}^{TUM2}_{(U_k)}$ dans les U_k sises en région PACA et aux alentours



Estimateur de qualité statistique	Estimation
var.explain.OOB($\hat{z}'^{j}_{(U_k),c}$)	68,41%
var.explain.IS($\hat{z}'^{j}_{(U_k),c}$)	69,50%
môy($\dot{\varepsilon}^{j}_{(U_k)}$)	-0,240
corr.($\hat{z}'^{j}_{(U_k),c}$; $z'^{j}_{(U_k),c}$)	89,75%

Figure 255 : Valeurs et synthèse statistique des $\dot{\varepsilon}^{TUM2}_{(U_k)}$

CLASSIFICATION : IDENTIFICATION MVG DES* VARIABLES EXPLICATIVES ET PREDICTIVES

Cible : $z'^{TUM2}_{Uk,q}$ propension spatiale pondérée EpiGéoStat et par le ratio participation/exposition

PHASE D'ANALYSE DES* DONNEES

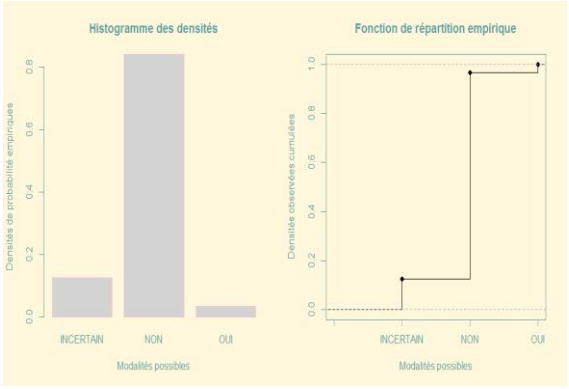


Figure 256 : Histogramme, fonction de répartition, de $z'^{TUM2}_{Uk,q}$

Statistiques MVG

Nombre des modalités : card $\left(z^{j}_{(U_k),q} = c_j \right)$

INCERTAIN	NON	OUI
53	354	14

Prportion des modalités môy $\left(z^{j}_{(U_k),q} = c_j \right)$

INCERTAIN	NON	OUI
12.59	84.09	3.33

PHASE 0 - CALIBRATION:

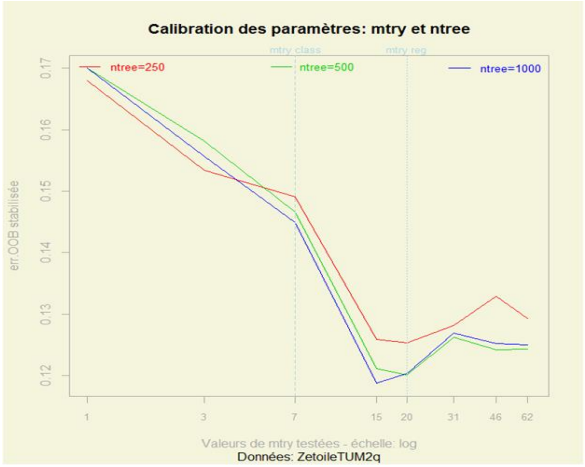


Figure 257 :. $\bar{R}^{OOB}(\cdot)$ en fonction de ntree et mtry

Objectif

FA.naïve → FA.opt

Paramètres MVG optimisés:

$\hat{\Lambda}_j = (\text{ntree} = 1000\,; \text{mtry} = 15; \text{nodesize} = 1)$

PHASE 1 – HIERARCHISATION & ELIMINATION DES* VARIABLES DE BRUIT

Objectif : Estimer les $\overline{VI}_j(\cdot)$, hiérarchiser les $z^{TUM2}_{(U_k),q}$, disjoindre de $\mathcal{X}^{TUM2}_{input}$ en : $\mathcal{X}^{TUM2}_{bruit}$ et $\mathcal{X}^{TUM2}_{conserv}$

Estimation et présentation des scores *randomForest**

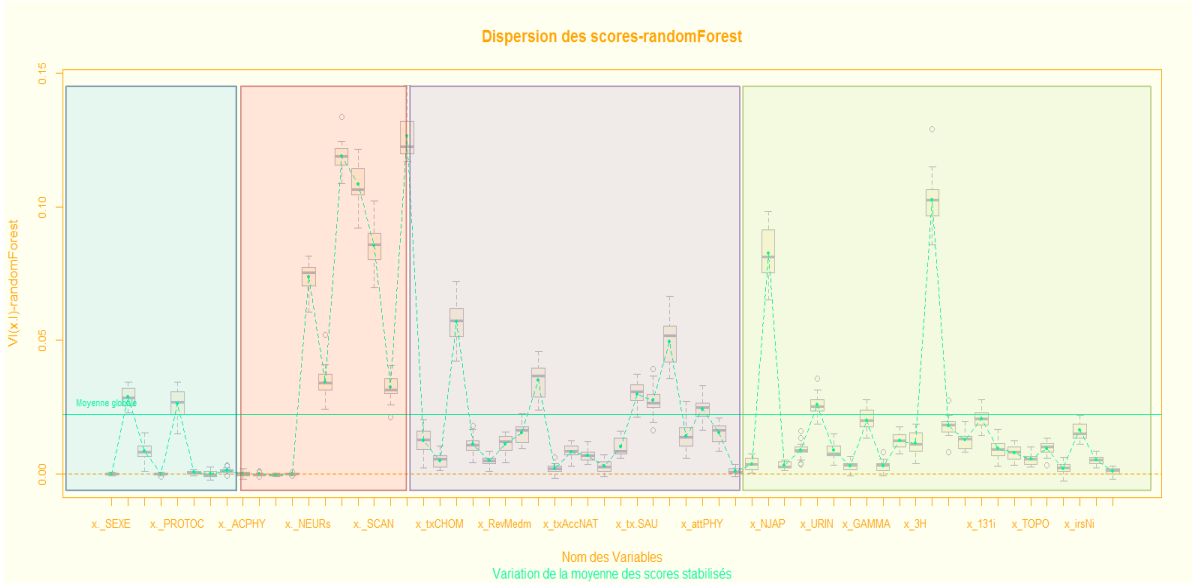


Figure 258 : Diagramme de Tukey des $VI_{j,r}\left(z^{j}_{(\cdot),.} \right)$, $\overline{VI}_j\left(z^{j}_{(\cdot),.} \right)$ et $\overline{\overline{VI}}_j\left(z^{j}_{(\cdot),.} \right)$ - estimation stabilisée sur 50 forêts.

Remarque : l'influence morbide des composantes environnementales peut être pré-classée à partir des scores des i.st.e, tel que (i) FE-SAN, (ii) FE-PHY.CHIM*, (iii) FE-SOCIO.ECO* et (iv) FIM*

Hiérarchisation descendante

Valeurs des $\overline{VI}_j(\cdot)$ et des $\hat{\sigma}\left(VI_j(\cdot)\right)$ pour les 10 i.st.e* : $x_{(U_k)}^{(l)}$ les plus importants

	VI.moy	SD.VI.moy
x._IRM	0.12661	0.01050
x._ENDOs	0.11903	0.01045
x._HEMAs	0.10848	0.01086
x_238Pu	0.10267	0.00856
x._SCAN	0.08544	0.01007
x_NJAP	0.08246	0.00875
x._NEURs	0.07353	0.00888
x_txFoyFisc	0.05692	0.00957
x_FDep09	0.04954	0.00793
x_txMORT	0.03503	0.00626

Elimination des variables de bruit

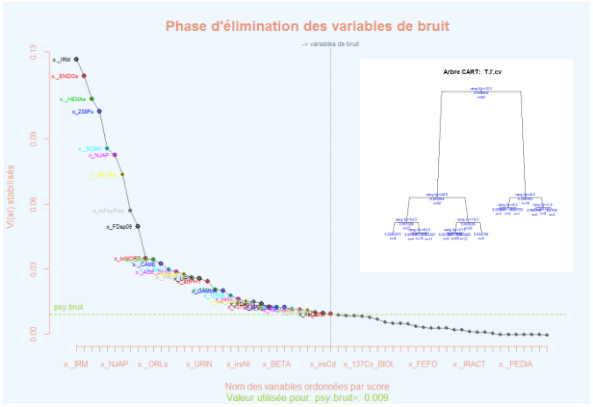


Figure 259 : Valeurs des $\overline{VI}_j(\cdot)$ associés aux $x_{(U_k)}^{(l)}$, et $\hat{T}_{MVG}^{j}((I))$

Estimation de ψ_{bruit}^{TUM2}

Coefficient subjectif expert : $c_{\psi.bruit}^{TUM2} = 15$

Disjonction des variables

$\mathcal{X}_{bruit}^{TUM2}$: 28/62 et $\mathcal{X}_{conserv}^{TUM2}$:34/62

Variables éliminées identifiées comme du bruit

x_PEST x_PREV x._DSUIVI x_txAccNAT x_137Cs_BIOL x_txBAC x_TOPO x_DJECr x_txCHOM x_GINIm x.._RADON x_FEFO x_ALPHA x_txEAR x_EGRA x_irsPb x_txAccPOP x._IRACT x_RADON x_PERC x._RECHUT x._TYPLEUC x._GREF x._PEDIA x._SEXE x._ACPHY x._GENE x._ORL (28/62)

PHASE 2 - SELECTION DES* VARIABLES EXPLICATIVES

Estimation de ψ_{explic}^{TUM2}

Coefficient subjectif expert : $c_{\psi.explic}^{TUM2} = 1$

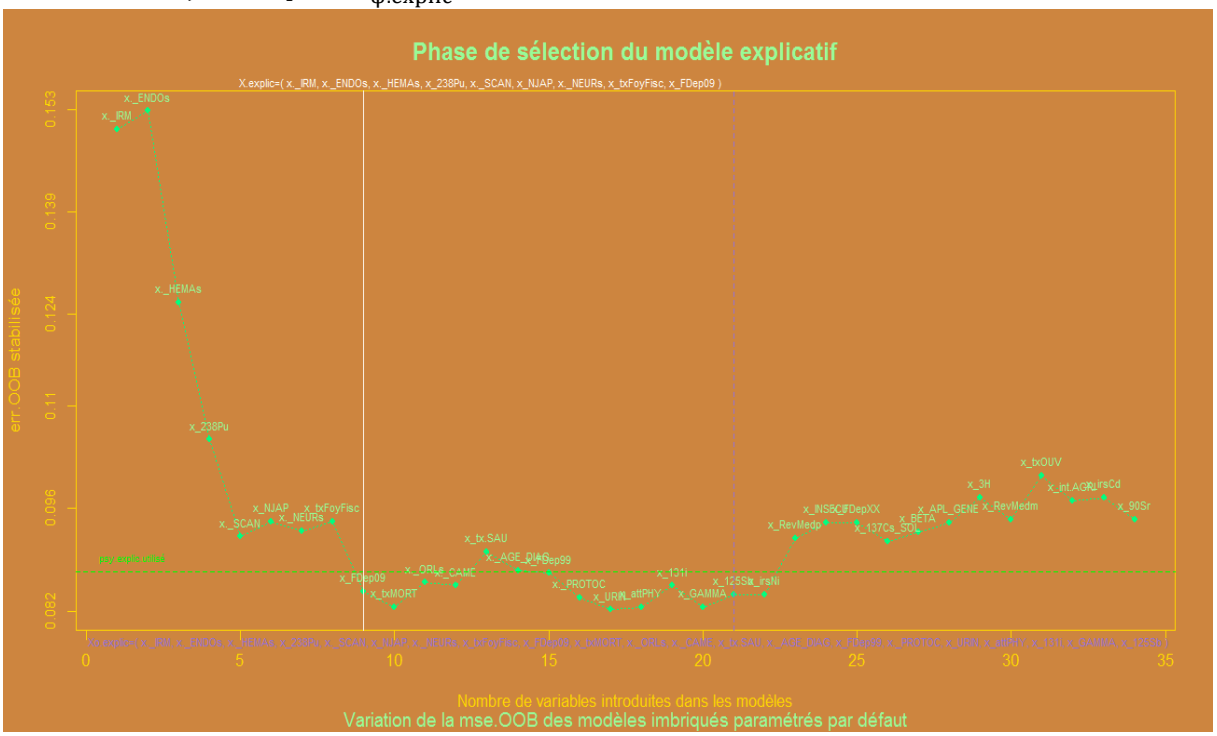


Figure 260 : Variation de $\overline{R}_{(.)}^{OOB}$ en fonction du nombre de variables imbriquées $\mathcal{X}_{imbr.l}^{TUM2} \subset \mathcal{X}_{conserv}^{TUM2}$;

Résultats - Paquets retenus : Génuer $\mathcal{X}_{\text{explic}}^{\text{TUM2}}$

x._IRM x._ENDOs x._HEMAs x_238Pu x._SCAN x_NJAP x._NEURs x_txFoyFisc x_FDep09 (9/62)

*Remarque : le paquet Bourrelly $\mathcal{X}_{0.explic}^{TUM2}$ contient aussi les i.st.e**

[…] x_txMORT x._ORLs x._CAME x_tx.SAU x._AGE_DIAG x_FDep99 x._PROTOC x_URIN x_attPHY x_131i x_GAMMA x_125Sb (21/62)

PHASE 3 - VARIABLES PREDICTIVES

Estimation de $\psi_{\text{pred}}^{\text{TUM2}}$: coefficient subjectif expert : $c_{\psi.\text{pred}}^{\text{TUM2}} = 1$

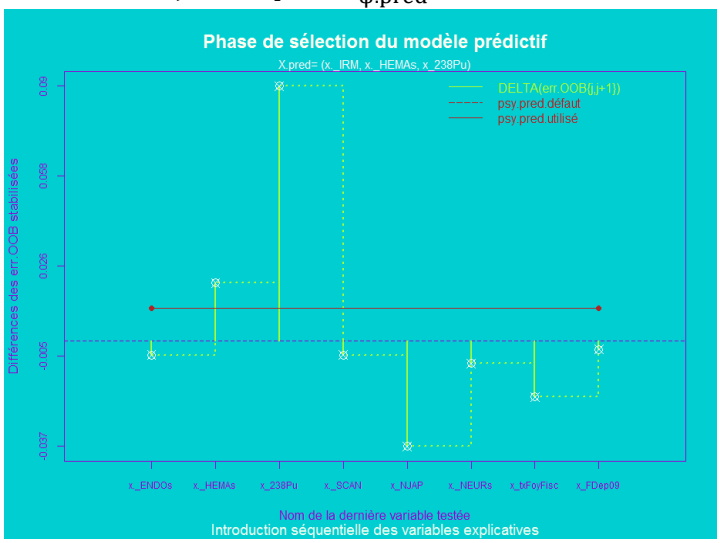


Figure 261 : Valeur de $\overline{\Delta}(R_l^{OOB})$ en fonction de $\mathbf{x}_{(U_k)}^{l}$ testé

Résultats - $\mathcal{X}_{\text{pred}}^{\text{TUM2}}$ contient : x._IRM x._HEMAs x_238Pu (3/62)

ANALYSE DE LA QUALITE DES* MODELES MVG

Estimation OOB de la qualité des modèles MVG utilisables $\hat{f}_{\text{FA}}^{\text{TUM2}}(\mathcal{X}_{\cdot}^{j}|\Lambda_j)$:

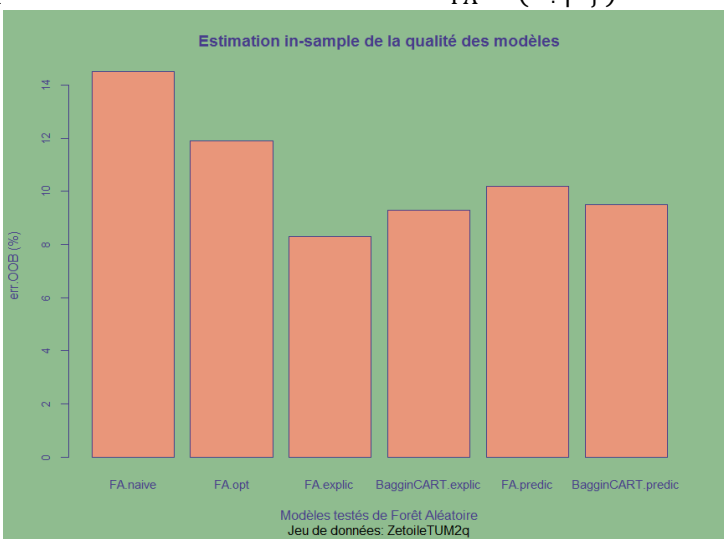


Figure 262 : Variation de l'err.OOB en fonction du modèle MVG testé

Remarque : les valeurs de $\overline{R}^{\text{OOB}-\text{IS}}(\cdot)$ sont estimées *In-Sample* (IS)

ANALYSE DE LA QUALITE PREDICTIVE

MODELE:	FA.naive	FA.opt	FA.explic	FA.pred
err.OOB généralisée :	**14,49%**	**11,88%**	**8,31%**	**10,21%**
gain.OOB.absolu :	0,00%	2,61%	6,18%	4,28%
gain.OOB.relatif :	0,00%	18,03%	42,62%	29,51%

Tableau 44 : Analyse des gains absolus et relatifs d'err.OOB généralisées par rapport à une FA.naïve

Classification : PREDICTIONS GEOGRAPHIQUES MVG

L'application de MVG à $z'^{TUM2}_{(U_k),q}$ permet d'obtenir le prédicteur MVG : $\hat{f}^{j}_{MVG}(x^{j}_{MVG}|\Lambda_{j-MVG})$. Le vecteur des paramètres spécifiés est : $\Lambda_{j-MVG} = (ntree = 1000\,; mtry = (\lfloor\sqrt{p_l}\rfloor \vee 2)\,; nodesize = 1)$ et $\{x^{TUM2}_{MVG} = x^{TUM2}_{explic}\}$; Les résultats présentés sont conformes aux propositions de cette section :

Cartographie des propensions spatiales EpiGéoStat répétées au prorata du ratio participation/exposition

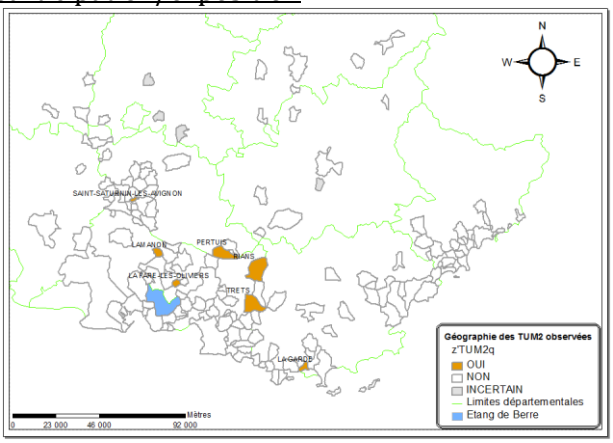


Figure 263 : Valeurs observées prises par $z'^{TUM2}_{(U_k),q}$

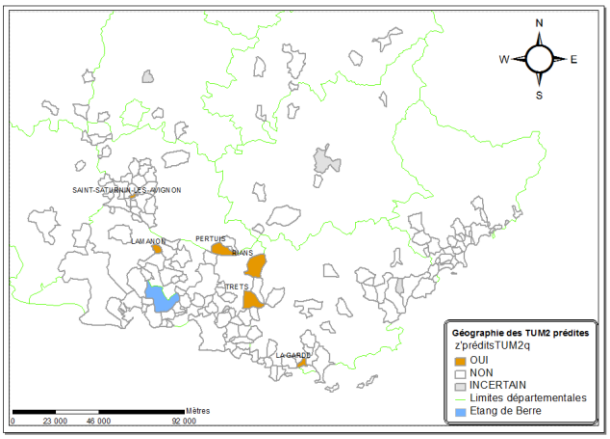


Figure 264 : Valeurs MVG prédites prises par $\hat{z}'^{TUM2}_{(U_k),q}$

Distribution des propensions spatiales EpiGéoStat répétées au prorata du ratio participation/exposition

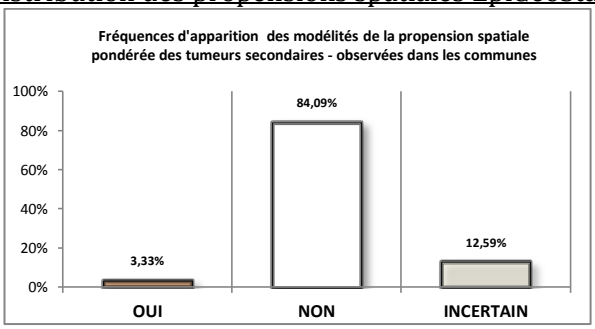


Figure 265 : Fréquences spatiales estimées des modalités de $z'^{TUM2}_{(U_k),q}$

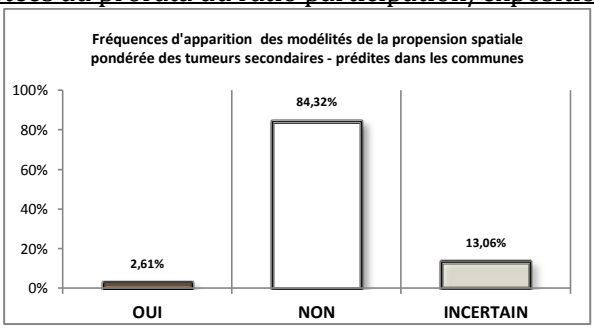


Figure 266 : Fréquences spatiales prédites des modalités de $\hat{z}'^{TUM2}_{(U_k),q}$

Cartographie, graphique et synthèse statistique des confusions géographiques

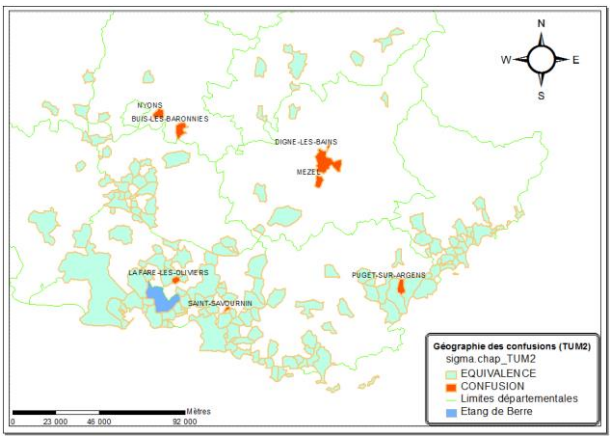
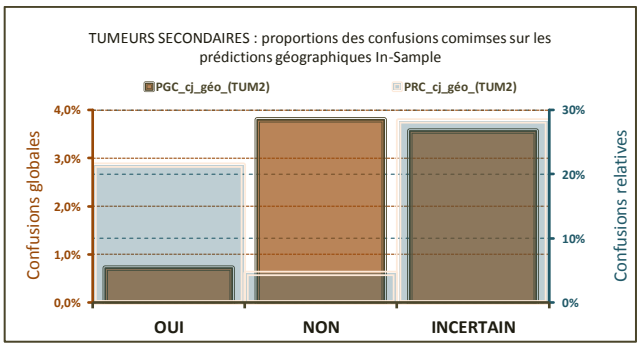


Figure 267 : Valeurs prises par $\hat{\varsigma}^{TUM2}_{(U_k)}$ dans les U_k sises en région PACA et aux alentours



Synthèse : CONFUSIONS	OUI	NON	INCERTAIN	TOTAL
NTC_cj_géo_$(\hat{z}'^{j}_{(U_k),q})$	3	16	15	34
PRC_cj_géo_$(\hat{z}'^{j}_{(U_k),q})$	21,43%	4,52%	28,30%	-
PGC_cj_géo_$(\hat{z}'^{j}_{(U_k),q})$	0,71%	3,80%	3,56%	8,08%

Figure 268 : Synthèse et représentation graphique des $\hat{\varsigma}^{TUM2}_{(U_k)}$

ANALYSE DE L'INFLUENCE DES FACTEURS ENVIRONNEMENTAUX ET DE LA QUALITE DES MODELISATIONS

L'algorithme MVG a été appliqué à la géographie des séquelles étudiées : CATA, THYR, TUM2. Il s'agit désormais d'analyser les résultats obtenus.

Dans un premier temps l'objectif est de caractériser l'influence des FE/FIM* modélisés sur les états de santé étudiés. L'idée est de procéder à une analyse simultanée des paquets $\mathcal{X}^{j}_{explic}$ et $\mathcal{X}^{j}_{pred}$, menée conjointement sur $z'^{j}_{(Uk),c}$ et $z'^{j}_{(Uk),q}$ et en tenant compte des scores *randomForest** obtenus en régression. La stratégie d'analyse s'opère conformément à *la grille de lecture MVG** proposée, au regard des connaissances acquises dans l'état de l'art (chapitre.1) et des caractéristiques granulaires des $x^{l}_{(Uk)}$ (chapitre3). Les objectifs sont : identifier les Déterminant Environnementaux de Santé (DES), Facteurs de Risques Environnementaux Contributifs* (FREC), Facteurs de Risques Environnementaux Potentiellement Aggravants (FREPA), et ensuite de hiérarchiser et spécifier les types d'expositions environnementales associées.

Dans un second temps la qualité statistique des prédictions $\hat{z}'^{j}_{(U_k),c}$ et $\hat{z}'^{j}_{(U_k),q}$ est analysée. Le but est de déterminer la qualité de l'estimateur MVG : $\hat{f}^{j}_{MVG}\left(\mathcal{X}^{j}_{MVG}\middle|\Lambda_{j-MVG}\right)$ en fonction de sa capacité à prédire la géographie des états de santé observés : $z'^{j}_{(Uk),c}$ et les $z'^{j}_{(Uk),q}$; et conséquemment, de conjecturer la pertinence des i.st.e* $x^{l}_{(Uk)}$ explicatifs, contenus dans $\mathcal{X}^{j}_{MVG}$, et donc implicitement la pertinence DES, FREC et FREPA qu'ils modélisent.

SEQUELLE : CATARACTES

REMARQUES LIMINAIRES :

Le nombre d'extrema contenus dans $z'^{CATA}_{Uk,c}$ est important et leurs valeurs sont volatiles. Cette caractéristique statistique est due : à la conversion en *patients-années* - les CATA ont une latence précoce, à son incidence élevée sur les patients spatialisés, aux distorsions spatiotemporelles induites par les pondérations EpiGéoStat - puisque les CATA ont un niveau d'incertitude médian par rapport aux THYR et TUM2 (chapitre.2).

L'analyse conjointe des $\mathcal{X}^{CATA}_{bruit}$ des $z'^{CATA}_{Uk,c}$ et $z'^{CATA}_{Uk,q}$ permet, d'un point de vue statistique, de caractériser de bruit environnemental les expositions géographiques : aux ETM*, au radon, à la quasi intégralité des FE/FIM curieux*, à la spécialisation économique des espaces et à la défaveur sociale. Les caractéristiques intrinsèques des individus qui n'ont *a priori* aucune influence statistique sur les cataractes sont : x_.SEXE, x._TYPELEUC et x_.AGE_DIAG.

DETERMINANTS ENVIRONNEMENTAUX DE SANTE (DES)

i. **FIM* -** <u>les effets secondaires liés à la lourdeur des traitements au long cours :</u>

Les scores extrêmes obtenus en régression par la géographie des irradiations corporelles totales (x._IRACT) et celle des patients greffés (x._GREF) propulsent les FIM* au rang de DES. Ces derniers mettent en cause les effets secondaires à moyen et long termes des irradiations corporelles – qui de surcroit peuvent exposer accidentellement les patients à des doses exagérément élevées liées à une défaillance du système médical – et aux effets toxiques des médicaments anti-rejets prescrits suite à une greffe de moelle. L'effet temporel des traitements reçus est renforcé par la géographie des protocoles prescrits (x._PROTOC) et celle de la durée du suivi (x._DSUIVI) qui modélise, entre autres, les effets secondaires au long cours.

<u>Remarque :</u> Les valeurs des $\overline{VI_j}(\cdot)$ prises par les autres i.st.e, en régression et aussi en classification, ne permettent pas d'identifier d'autres DES.

FACTEURS DE RISQUES ENVIRONNEMENTAUX CONTRIBUTIFS* (FREC)

i. **FE-SAN* -** L'accès aux soins lié à la qualité du tissu sanitaire territorial - modélisé par:

L'accès géographique à des ophtalmologues soit à partir des distances temps bruitées d'accès à un service d'ophtalmologie (x._OPHTs), soit par l'Accès Potentiel Localisé* aux ophtalmologues libéraux (x_APL-OPHT). La mise en évidence de la qualité du tissu sanitaire territorial sur l'incidence, de la prise en charge et de la latence des CATA est sans équivoque. Elle est même amplifiée par la présence de FE-SAN Curieux* caractérisant les distances temps bruitées d'accès à tous les items sanitaires*, i.e. aux praticiens libéraux (x._RADIO), aux plateaux techniques des établissements de santé qu'il s'agisse de services spécialisés (x._NEURs, x._HEMAs) ou de leurs EML* (x._IRM, x._CAME, x._SCAN).

FACTEURS DE RISQUES ENVIRONNEMENTAUX PROBABLEMENT AGGRAVANTS* (FREPA)

i. **FE-PHY.CHIM* -** Les expositions géographiques aux paramètres météo - modélisées par :

La variabilité spatiotemporelle des doses journalières de rayonnement induites par la globalité des émissions solaires (x_RAY), des températures mensuelles moyennes (x_TEMP) et du nombre de jours de pluie mensuels (x_NJAP). L'influence des conditions météorologiques, au moins comme facteur de risque aggravant pour les CATA, est largement documentée et elle est renforcée par la variabilité des niveaux altimétriques x (x_TOPO) puisque l'altitude expose les yeux à des rayonnements solaires plus intenses.

ii. FE-SOCIO.ECO* - Les conjonctures sociales induisant des prédispositions géographiques morbides – modélisées par :

Les niveaux spatiotemporels d'emplois ouvriers territoriaux (x_txOUV) caractérisent des catégories socio-professionnelles généralement plus pauvres, plus enclines à des conduites à risques vis-à-vis du recours aux soins, de l'exposition professionnelle à des substances toxiques et à la consommation de psychotropes. L'effet néfaste de l'augmentation de la *distance a-spatiale* à l'offre de soins est renforcé par la disparité de la répartition des richesses (x_GINIp).

iii. FE-PHY.CHIM* - Les expositions géographiques potentielles à la radioactivité environnementale - essentiellement artificielle - modélisées par :

La variabilité spatiotemporelle de l'activité volumique: du Plutonium 238 (x_238Pu) et du Césium 137 (x_137Cs-sol) dans les sols, et par du strontium 90 (x_90Sr) dans les denrées alimentaires, dont la présence dans l'environnement est presque exclusivement due à des rejets liés à l'exploitation des INB. L'hypothèse de l'effet aggravant de la radioactivité environnementale sur l'augmentation du risque de CATA est assez récente.

ANALYSE DE LA QUALITE DES MODELES MVG ET DES PREDICTIONS GEOGRAPHIQUES

REGRESSION - Qualité des modèles MVG :

La phase d'optimisation des paramètres permettant de passer des FA.naïves aux FA.opt n'améliore par la var.explain.OOB, estimée à 51,3%, car les paramètres optimisés sont ceux implémentés par Breiman. Cette phase n'est pas nécessaire à l'estimation des $\overline{VI}_j(\cdot)$. En revanche, la qualité des prédictions géographiques des FA.MVG est excellente avec : une var.explain.MVG de 74,30% et un coefficient de corrélation de 92,79 %.

REGRESSION - Qualité des prédictions géographiques :

Les prédictions de FA.MVG sont fiables puisque la majorité des écarts standardisés sont, tel que : $\dot{\varepsilon}^{j}_{(U_k)} \in [\![t_{(\varpi)} ; t_{(1-\varpi)}]\!]$.

Cependant les $\hat{z}'^{CATA}_{Uk,c}$ surestiment presque systématiquement la réalité géographique car $\bar{\varepsilon}^{CATA}_{(U_k)} \ll 0$. De fait, la distribution spatiale des valeurs intermédiaires de $\hat{z}'^{CATA}_{Uk,c}$ est décalée vers la droite par rapport à celle des $z'^{CATA}_{Uk,c}$.

Le nombre d'écarts significatifs entre prédictions et observations, tel que $\dot{\varepsilon}^{j}_{(U_k)} > t_{(1-\alpha)}$, est anecdotique. Par conséquent, les valeurs de $\hat{z}'^{CATA}_{Uk,c}$ ne sous-estiment que rarement, de façon exagérée, la réalité géographique observée et décrite par les $z'^{CATA}_{(U_k),c}$. En revanche, le nombre de $\dot{\varepsilon}^{j}_{(U_k)} < t_{(1-\alpha)}$ est

beaucoup plus important. Le modèle prédictif MVG a donc tendance, lorsqu'il se fourvoie, à surestimer fortement les valeurs de $z'^{CATA}_{(U_k),c}$.

L'analyse conjointe des distributions spatiales statistiques et des cartographies des : $z'^{CATA}_{Uk,c}$, $\hat{z}'^{CATA}_{Uk,q}$ et $\dot{\varepsilon}^{CATA}_{(U_k)}$ permet d'aborder les parties de l'échantillon les plus complexes :

Les minima géographiques sont relativement biens prédits par $\hat{z}'^{CATA}_{Uk,c}$. En PACA les communes pour lesquelles un minima est prédit à tort, i.e $\dot{\varepsilon}^{CATA}_{(U_k)} \geq t_{(1-\tau)}$, sont l'Isle sur la Sorgue et Cabries.

Les maxima géographiques sont moins bien prédits. La région PACA est particulièrement affectée – Le modèle MVG commet une erreur *anormalement* forte, i.e. lorsque $\dot{\varepsilon}^{j}_{(U_k)} < t_{(\tau)}$, sur 11 communes en PACA : Levens ,Eze, Peymeinade, Villeneuve-Loubet, La Crau, Rians, Valensole, La Tour d'Aigues, Arles, Saint-Rémy-De-Provence, Bedarrides.

Cependant, les $\hat{z}'^{CATA}_{Uk,c}$ commettent peu d'erreurs sur la caractérisation des communes en tant qu'outliers statistiques. En PACA, 17 communes sont considérées comme telles par $z'^{CATA}_{Uk,c}$ contre 14 par $\hat{z}'^{CATA}_{Uk,c}$ et la majeure partie d'entre elles sont identiques.

CLASSIFICATION - Qualité des modèles MVG

La phase d'optimisation des paramètres permet de passer d'une FA.naïve, dont l'err.OOB est de 19,95%, à une FA.opt où elle n'est plus que de 18,05%. Cette phase est nécessaire à l'estimation de $\overline{VI}_j(\cdot)$ robuste. Le modèle FA.explic est assez performant puisque l'err.OOB est diminuée à 15,20%, soit un gain relatif 23,81% par rapport à la FA.naïve.

CLASSIFICATION - Qualité des prédictions géographiques

Le modèle FA.MVG utilisé pour effectuer les prédictions géographiques est particulièrement fiable. L'err.MVG globale est de 14,73%. Quant aux confusions relatives elles s'élèvent seulement à 29,82% sur « OUI », à 10,03% sur « NON », et à 25,45% sur « INCERTAIN ». Les confusions géographiques portent essentiellement sur les communes dans lesquelles les patients spatialisés sont associés à une forte incertitude EpiGéoStat.

SEQUELLE : TUMEURS THYROÏDIENNES

REMARQUES LIMINAIRES

Le nombre d'extrema contenus dans $z'^{THYR}_{Uk,c}$ est très faible mais les valeurs prises sont particulièrement volatiles. Cette caractéristique statistique est engendrée par : la conversion en *patients-années* liée à une latence importante, une incidence élevée de THYR sur les patients spatialisés puisque presque équivalente à celle des CATA. Et surtout, du fait que THYR est la séquelle la plus incertaine donc la plus impactée par les pondérations EpiGéoStat (chapitre.2).

L'analyse conjointe des $\mathcal{X}^{THUR}_{bruit}$ de $z'^{THYR}_{Uk,c}$ et de $z'^{THYR}_{Uk,q}$ met en évidence le manque de puissance explicative des expositions géographiques aux : paramètres météo, radon, l'éloignement temporel aux praticiens de santé libéraux, FE/FIM* Curieux* en quasi-totalité. Et étonnamment, à certains FIM* modélisés comme : x_.SEXE, x._TYPELEUC – pourtant théoriquement attendus.

DETERMINANTS ENVIRONNEMENTAUX DE SANTE (DES)

i. **FIM* -** Les effets secondaires des traitements et plus particulièrement de ceux agressifs :

Le score extrême obtenu en régression par la géographie de la durée du suivi des patients (x._DSUIVI) propulse les FIM* au rang de DES. L'influence temporelle du traitement reçu pour la LA semble évidente. Cependant la prévalence spatiale pondérée des THYR est aussi conditionnée par la géographie de la lourdeur des traitements reçus. En l'occurrence par le fait que les patients aient subi une irradiation corporelle totale (x._IRACT) – avec un risque augmenté par l'exposition médicale à des doses accidentellement élevées - ou, avec une efficience moindre, une greffe de moelle (x.GREF).

Remarque : La valeur des $\overline{VI_j}(\cdot)$ obtenus en régression par un i.st.e contextuel aurait pu permettre de qualifier les FE-SOCIO.ECO* de DES* si seulement l'importance des i.st.e* associés à celui-ci n'avait pas été cannibalisée par tous les autres FE* en classification.*

FACTEURS DE RISQUES ENVIRONNEMENTAUX CONTRIBUTIFS (FREC)

i. **FE-SOCIO.ECO* -** La spécialisation économique des espaces prédisposant aux phénomènes morbides - modélisée par :

L'exposition géographique à des substances toxiques diffuses induite par la spécialisation des territoires dans le domaine de l'agriculture. Le score obtenu par la géographie des Surfaces Agricoles Utilisées (x_txSAU) met en évidence un risque contextuel de contamination des milieux « eau, air, sol et biologique » par les pesticides. Cette hypothèse est renforcée la géographie des doses journalières ingérées de chrome (x_DJE.Cr) - ETM présent dans les pesticides. Aussi, l'effet conjoncturel des expositions à des substances toxiques est amplifié par des risques d'exposition professionnelle dus à l'impact de l'agriculture sur les activités socio-économiques locales comme l'augmentation d'emplois peu qualifiés dans le domaine agro-industriel – supposés plus risqués – et suggérés indirectement par (x_.txOUV ; x._FDep09).

ii. **FE-SAN* -** L'accès aux soins lié à la qualité du tissu sanitaire territorial - modélisé par :

Des distances temps bruitées d'accès qui touchent d'abord aux Equipements Matériels Lourds (EML*) inhérents au diagnostic et au traitement des THYR tels que les IRM (x._IRM), les Scanner*s (x._SCAN) – l'influence de la qualité du diagnostic est amplifiée par la présence de l'accès aux radiologues libéraux (x._RADIO) dans $\mathcal{X}^{j}_{0.explic}$ en classification - et ensuite, à des services spécialisés dans la prise en charge des THYR comme ceux d'endocrinologie (x._ENDOs), d'otorhinolaryngologie (x._ORLs), d'hématologie (x._HEMAs) et de neurologie (x._NEURs).

FACTEURS DE RISQUES ENVIRONNEMENTAUX PROBABLEMENT AGGRAVANTS* (FREPA) :

i. **FE-PHY.CHIM* -** Les expositions géographiques potentielles à la radioactivité environnementale artificielle - modélisées par :

La variabilité spatiotemporelle des niveaux d'activité volumique des eaux douces par les isotopes radioactifs du Tritium (x_3H) et des sols par le Plutonium 238 (x_238Pu) ainsi que le Césium 137 (x_137Cs-sol), dont l'accumulation dans les compartiments environnementaux est exclusivement liée aux émissions autorisées ou accidentelles des INB* en fonctionnement normal – et de façon résiduelle aux catastrophes et aux essais atomiques pour le Plutonium 238.

ii. FE-SOCIO.ECO* - L'effet contextuel de défaveur induisant des prédispositions géographiques morbides - modélisé par :

La *défaveur sociale* qui suggère des niveaux accrus de chômage, des catégories professionnelles plus pauvres, des niveaux culturels plus faibles et une répartition des richesses - fiscales et salariales - peu équitable (x_FDep09, x_txOUV; x_RevMed.m), et par l'insécurité territoriale qui favorise l'exposition sociale au stress psychologique (x_attBIENS). Ces caractéristiques conjoncturelles augmentent la distance *a-spatiale* du recours aux soins, les conduites à risques et la consommation de psychotropes.

ANALYSE DE LA QUALITE DES* MODELES MVG ET DES* PREDICTIONS GEOGRAPHIQUES

REGRESSION - Qualité des modèles MVG :

La phase d'optimisation des paramètres permet de passer des FA.naïves dont la var.explain.OOB est de 39% aux FA.opt qui expliquent 46 % de la variance. Cette phase est donc nécessaire à l'estimation de $\overline{VI}_j(\cdot)$ robustes. La qualité des prédictions géographiques des FA.MVG est excellente avec : une var.explain.MVG de 60,36% et un coefficient de corrélation de 88,55 %.

REGRESSION - Qualité des prédictions géographiques :

Les prédictions du modèle FA.MVG sont assez fiables, bien qu'il s'agisse du modèle le moins robuste comparativement aux autres séquelles, puisque la majorité des écarts standardisés sont tels que : $\dot{\varepsilon}^{j}_{(U_k)} \in [\![t_{(\varpi)} ; t_{(1-\varpi)}]\!]$.

Les $\hat{z}'^{THYR}_{Uk,c}$ surestiment légèrement la réalité géographique, car $\bar{\dot{\varepsilon}}^{THYR}_{(U_k)} < 0$ et la distribution spatiale des $\hat{z}'^{THYR}_{Uk,c}$ est décalée vers la droite par rapport à celle des $z'^{THYR}_{Uk,c}$.

Le nombre d'écarts significativement important entre prédiction et observation, i.e. lorsque $\left|\dot{\varepsilon}^{j}_{(U_k)}\right| > t_{\alpha}$, reste raisonnable. Le modèle MVG ne présente pas une tendance particulière à surestimer ou à sous-estimer fortement les valeurs de $z'^{TUM2}_{(U_k),c}$.

L'analyse conjointe des distributions spatiales statistiques et des cartographies des : $z'^{THYR}_{Uk,c}$, $\hat{z}'^{THYR}_{Uk,q}$ et $\dot{\varepsilon}^{THYR}_{(U_k)}$ permet d'appréhender les parties de l'échantillon les plus difficiles à prédire :

Les minima géographiques sont relativement biens prédits par $\hat{z}'^{THYR}_{Uk,c}$. En PACA les communes pour lesquelles un minima est prédit à tort, i.e. $\dot{\varepsilon}^{j}_{(U_k)} \geq t_{(1-\tau)}$, sont : Eze et Pourrieres.

Les maxima géographiques sont moins bien prédits que les minima. La région PACA est particulièrement affectée. Les surestimations anormalement fortes, i.e. lorsque $\dot{\varepsilon}^{j}_{(U_k)} < t_{(\tau)}$, concernent les sept communes suivantes : Tourrette-Levens, Grasse, Valbonne, Toulon, Roquefort-La-Bedoule, Fos-Sur-Mer et Roquemaure.

Cependant, les $\hat{z}'^{THYR}_{Uk,c}$ ne commettent aucune erreur sur la caractérisation des communes en tant qu'outliers statistiques par comparaison aux $z'^{THYR}_{Uk,c}$. En PACA, aucun outlier n'est observé et aucun n'est prédit.

CLASSIFICATION - Qualité des modèles MVG

La phase d'optimisation des paramètres permet de passer des FA.naïves, dont l'err.OOB est de 16,17%, à une FA.opt où elle n'est plus que de 14,25%. Cette phase est donc nécessaire à l'estimation de $\overline{VI}_j(\cdot)$ robuste. Le modèle FA.explic est particulièrement performant puisque l'err.OOB ne vaut plus que 10,21%, soit un gain relatif 36,84% par rapport aux FA.naïves

CLASSIFICATION - Qualité des prédictions géographiques

Le modèle FA.MVG utilisé pour effectuer les prédictions géographiques est fiable. L'err.MVG globale est de 9,5%. Quant aux confusions relatives elles s'élèvent seulement à : 21,15% sur « OUI », 5,2% sur « NON », et à 28,57% sur « INCERTAIN ». Les confusions géographiques portent essentiellement sur les communes dans lesquelles les patients spatialisés sont associés à une forte incertitude EpiGéoStat.

SEQUELLE : TUMEURS SECONDAIRES

REMARQUES LIMINAIRES

Les TUM2 englobent toutes les tumeurs secondaires majeures, donc une partie des THYR. De fait, il est normal de retrouver des analogies.

Les valeurs extrêmes contenues dans $z'^{TUM2}_{Uk,c}$ sont peu nombreuses ce qui est dû à : la conversion en *patients-années* d'une séquelle dont la latence est longue, une incidence particulièrement faible et à un niveau d'incertitude EpiGéoStat, le plus faible de tous.

L'analyse conjointe des $\mathcal{X}^{TUM2}_{bruit}$, des $z'^{TUM2}_{Uk,c}$ et $z'^{TUM2}_{Uk,q}$ montre l'inanité statistique des expositions géographiques aux métalloïdes, radon, éloignements temporels aux praticiens de santé libéraux. Le manque d'efficience de la majeure partie des FE-SOCIO.ECO, de la quasi-totalité des FE/FIM* Curieux*, et plus étonnamment, en classification, aux FIM* modélisés par les i.st.e* : x_.RECHUT, x._IRACT et x._GREFF.

DETERMINANTS ENVIRONNEMENTAUX DE SANTE (DES)

i. FIM* - Les effets temporels délétères des traitements et le caractère chronique des leucémies :

Le score extrême obtenu en régression par la géographie de la durée du suivi des patients (x._DSUIVI) propulse les FIM* au rang de DES. L'influence des FIM* est corroborée par le rôle du protocole de traitement (x._PROTOC). Ces i.st.e* représentent les effets secondaires des traitements utilisés. Ils peuvent caractériser de façon indirecte leur agressivité. L'effet temporel des FIM* est accentué, en classification, par **la géographie des âges au diagnostic (x._AGE_DIAG).**

Remarque : *L'écart sur les $\overline{VI}_j(\cdot)$ en régression ne permet pas d'identifier d'autres DES. Cependant les i.st.e* présents conjointement dans les paquets explicatifs et prédictifs pour les différents contextes statistiques dénotent un effet environnement évident.*

FACTEURS DE RISQUES ENVIRONNEMENTAUX CONTRIBUTIFS* (FREC)

i. FE-PHY.CHIM* - Les expositions géographiques à la radioactivité environnementale et plus particulièrement à celle d'origine artificielle sont modélisées par :

Des doses de rayons γ induites par la dégradation des radionucléides présents dans l'air qui sont essentiellement liés à des apports cosmiques ou telluriques et à l'activité des INB* (x_GAMMA). L'accumulation du Plutonium 238 dans les sols (x_238Pu) - un radionucléide artificiel fortement cancérigène - est liée à l'activité des INB* et de façon résiduelle aux catastrophes nucléaires et aux essais atomiques historiquement effectués dans l'atmosphère. L'impact des radionucléides artificiels présents dans l'environnement, qui semblent contribuer au risque de TUM2, est renforcé par la variabilité géographique du Strontium 90 dans le lait (x_90Sr) et du risque d'Exposition Géographique à des Radionucléides Artificiels liée à la proximité spatiale d'INB* (x_EGRA) – en régression, ainsi qu'à celle de l'iode 131 dans le lait (x_131i) et de l'Antimoine 125 dans le sol (x_125Sb) présent dans : $\mathcal{X}^{j}_{0.explic}$ - en classification

ii. FE-SAN* - L'accès géographique aux services et aux équipements médicaux des établissements de santé est modélisé par :

Des distances temps bruitées d'accès à des services spécialisés dans le traitement des cancers - i.e. d'hématologie (x._HEMAs), de neurologie (x._NEURs), d'endocrinologie (x._ENDOs), et à des Equipements Matériels Lourds (EML*) inhérents au diagnostic et au traitement des TUM2 - tels que les IRM (x._IRM) et les Scanner*s (x._SCAN). Ces i.st.e* caractérisent la défaillance globale des tissus sanitaires* territoriaux. Ils sont largement représentés en régression et *a fortiori* en classification.

FACTEURS DE RISQUES ENVIRONNEMENTAUX PROBABLEMENT AGGRAVANTS* (FREPA)

i. FE-SOCIO.ECO* - L'effet de conjonctures défavorables induisant des prédispositions géographiques morbides est modélisé par :

L'insécurité territoriale qui engendre des expositions contextuelles au stress psychologique d'origine sociale (x_attPHY, x_attBIENS, x_INSECU), et par la *défaveur sociale composite* qui suggère des niveaux accrus de chômage, des catégories socio-professionnelles plus pauvres et plus exposées à des substances nocives, des niveaux culturels plus faibles et une répartition des richesses peu équitable – notamment sur le plan fiscal (x_FDep09, x_txFoyFisc ; x_FDep99)

ii. FE-PHY.CHIM* - Les effets indirects de la pluviométrie et les affres de l'urbanisation – modélisés par :

Les effets de la pluviométrie sont protéiformes. Ici, elle est considérée pour son rôle épurateur de la radioactivité présente dans l'air, principal vecteur de contamination des compartiments environnementaux par des radionucléides. Les eaux de pluie sont les plus chargées en particules radioactives et autres substances chimiques nocives (x_NJAP). Les effets délétères des substances toxiques comme les HAP, les métalloïdes, les MES, les ondes magnétiques… induites par l'urbanisation et l'industrialisation des espaces, et qui conjointement engendrent des nuisances sociales : surconcentration de population, misère, insécurité, dégradation des paysages, bruits… - sont aussi suggérées (x_URIN).

ANALYSE DE LA QUALITE DES* MODELES MVG ET DES* PREDICTIONS GEOGRAPHIQUES

REGRESSION - Qualité des modèles MVG :

La phase d'optimisation des paramètres permet de passer d'une FA.naïve dont la var.explain.OOB est de 50 % aux FA.opt qui expliquent environ 55,5 % de la variance. Cette phase est donc nécessaire à l'estimation de $\overline{VI}_j(\cdot)$ robustes. La qualité des prédictions géographiques des FA.MVG est excellente avec une var.explain de 69,50% et un coefficient de corrélation de 89,75 %.

REGRESSION - Qualité des prédictions géographiques :

Les prédictions du modèle FA.MVG sont fiables puisque un nombre important d'écarts standardisés est tel que : $\dot{\varepsilon}^{j}_{(U_k)} \in [\![t_{(\varpi)} ; t_{(1-\varpi)}]\!]$.

Les $\hat{z}'^{TUM2}_{Uk,c}$ surestiment très légèrement la réalité géographique, puisque $\bar{\dot{\varepsilon}}^{j}_{(U_k)} < 0$ et que la distribution spatiale des $\hat{z}'^{TUM2}_{Uk,c}$ est décalée vers la droite par rapport à celle des $z'^{TUM2}_{Uk,c}$

Le nombre d'écarts significatifs entre prédiction et observation, i.e. lorsque $\left|\dot{\varepsilon}^{j}_{(U_k)}\right| > t_{\alpha}$, reste raisonnable et le modèle MVG ne présente pas une tendance particulière à surestimer plus fortement les valeurs de $z'^{TUM2}_{(U_k),c}$.

L'analyse conjointe des distributions spatiales statistiques et des cartographies des : $z'^{TUM2}_{Uk,c}$, $\hat{z}'^{TUM2}_{Uk,q}$ et des $\dot{\varepsilon}^{TUM2}_{(U_k)}$ permet d'aborder les parties de l'échantillon les plus difficiles à prédire :

Les minima géographiques sont occasionnellement erronés par les $\hat{z}'^{TUM2}_{Uk,c}$, de façon très significative. En PACA les communes pour lesquelles un minimun est prédit à tort, i.e. $\dot{\varepsilon}^{j}_{(U_k)} \geq t_{(1-\tau)}$, sont : Eze et Monteux

Les maxima géographiques sont plus fortement surestimés et plus nombreux. En région PACA, i.e. les surestimations anormalement fortes des $z'^{TUM2}_{(U_k),c}$concernent les communes de : Sorgues, Rognac, Rognonas, Draguignan. Une surestimation curieusement extrême est prédite pour la commune de Montluçon.

En somme, les $\hat{z}'^{TUM2}_{Uk,c}$ commettent quelques erreurs notables sur les prédictions extrêmes de $z'^{TUM2}_{Uk,c}$. En PACA les communes qualifiées d'outliers statistiques sur $z'^{TUM2}_{Uk,c}$ sont : l'Isle-Sur-La-Sorgue et La-Garde - or aucune n'est caractérisable comme telle à partir de : $\hat{z}'^{j}_{(U_k),c}$.

CLASSIFICATION - Qualité des modèles MVG :

La phase d'optimisation des paramètres permet de passer des FA.naïves, dont l'err.OOB est de 14,5%, à une FA.opt où elle n'est plus que de 11,88%. Cette phase est donc nécessaire à l'estimation de $\overline{VI}_j(\cdot)$ robustes. Le modèle FA.explicatif est particulièrement performant puisque l'err.OOB ne vaut plus que 8.31%, soit un gain relatif 42,62% par rapport aux FA.naïves

CLASSIFICATION - Qualité des prédictions géographiques :

Le modèle FA.MVG utilisé pour effectuer les prédictions géographiques est particulièrement fiable. L'err.MVG globale est de 8.8%. Quant aux confusions relatives elles s'élèvent seulement à 21,5% sur « OUI », 4,5% sur « NON », et 28,3% sur « INCERTAIN ». Les confusions géographiques portent essentiellement sur les communes dans lesquelles les patients spatialisés sont associés à une forte incertitude EpiGéoStat.

REMARQUES PARTICULIERES SUR MVG

Remarque sur la capacité de MVG à prédire et à expliquer les états de santé

La qualité des modèles MVG - $\hat{f}^{j}_{MVG}\left(\mathcal{X}^{j}_{MVG}\middle|\Lambda_{j-MVG}\right)$ est particulièrement fiable. Il perme det prédire, avec une grande acuité, uniquement à partir des i.st.e* $x^{l}_{(U_k)}$ représentatifs des DES, des FREC et des FREPA, la géographie des états de santé étudiés.

Contexte de la régression

Les prédictions $\hat{z}'^{j}_{Uk,c}$ ont toujours tendance à surestimer la valeur des $z'^{j}_{Uk,c}$. Cependant, La capacité des $\hat{z}'^{j}_{Uk,c}$ à caractériser les outliers statistiques est globalement convenable.

S'agissant des extrêma géographiques observés les minima sont généralement correctement prédits. Par contre, les maxima le sont beaucoup moins.

Enfin, les écarts extrêmes entre observations $z'^{j}_{Uk,c}$ et prédictions $\hat{z}'^{j}_{Uk,c}$, touchent généralement des surestimations – anormalement fortes – ce qui est particulièrement préjudiciable en matière de qualité sanitaire environnementale - pour les communes concernées.

Contexte de classification

Les $\hat{z}'^{j}_{Uk,q}$ ont tendance à sous-estimer un temps la modalité $c_j = OUI$. Ils surestiment légèrement la modalité $c_j = NON$. Mais ils caractérisent globalement assez bien les Uk pour lesquelles $c_j =$ INCERTAIN.

L'amplitude des erreurs commises par $\hat{f}^{j}_{MVG}\left(\mathcal{X}^{j}_{MVG}\middle|\Lambda_{j-MVG}\right)$, dans les différents contextes statistiques, est d'autant plus grande que la séquelle est incertaine d'un point de vue EpiGéoStat, sa latence est courte et son incidence petite. Il convient de rappeler que l'incidence spatiale des CATA (13%) est proche de celle des THYR (10,2%) à l'aune de celle des TUM2 (4%), et aussi, que l'incertitude spatiotemporelle EpiGéoStat des séquelles peut être hiérarchisée de la façon suivante : THYR, CATA, TUM2.

Globalement

La qualité des opérateurs $\hat{f}^{j}_{MVG}\left(\mathcal{X}^{j}_{MVG}\middle|\Lambda_{j-MVG}\right)$ est particulièrement bonne. Ce modèle prédictif MVG utilise tous les i.st.e* explicatifs et représentatifs des Déterminent Environnementaux De Sante (DES), Facteurs de Risques Environnementaux Contributifs* (FREC) et Facteurs de Risques Environnementaux Probablement Aggravants* (FREPA). Par conséquent, l'hypothèse que ces FE/FIM* soient effectivement les facteurs qui conditionnent la géographie des PM* étudiés – séquelles : CATA, THYR et TUM2 – semble très prégnante.

Conclusion sur la pertinence prédictive et sur la qualité explicative de MVG

Dans la mesure où l'opérateur : $\hat{f}^{j}_{MVG}\left(\mathcal{X}^{j}_{MVG}\middle|\Lambda_{j-MVG}\right)$ permet de prédire, avec une grande acuité, la géographie d'états de santé à partir d'i.st.e* : $x^{l}_{(U_k)}$ qui les caractérisent.

Et puisque la proposition MVG se fonde sur des modélisations environnementales robustes (chapitre3) et une méthode d'apprentissage statistique multidimensionnelle très puissante : VSURF (chapitre.4).

Alors : la dialectique MVG permet en effet d'identifier des DES, des FREC et des FREPA que l'on peut supposer fiables et à même d'être pris en considération sur le plan médical et politique, et à partir desquels il est possible d'imaginer des mesures afin d'améliorer l'accès à une bonne santé environnementale* individuelle, i.e. pour les patients, et collective, i.e. adaptées aux besoins des populations locales.

Remarque sur l'interprétation explicative des résultats MVG

La dialectique MVG se décompose en trois phases: Hiérarchisation et élimination des variables de bruits, sélection des variables explicatives, et identification des variables qui s'adaptent à la parcimonie prédictive. Chaque phase donne lieu à la confection d'un paquet d'i.st.e* $x^{l}_{(U_k)}$: $\mathcal{X}^{j}_{bruit}$; $\mathcal{X}^{j}_{explic}$; $\mathcal{X}^{j}_{pred}$;.
L'analyse conjointe de ces paquets permet d'identifier les DES, les FREC et les FREPA. Il convient cependant de faire quelques remarques quant à l'analyse et l'interprétation :

Phase 1 – étape de hiérarchisation - interprétation des scores associés aux i.st.e. $x^{l}_{(U_k)}$

Les scores randomForest*, à l'instar des scores CART (non présentés), sont suspectés d'être biaisés en faveur des variables qualitatives - booléennes ou multiclasses. Or, quel que soit le contexte, les FE/FIM* Curieux* de test – modélisés par des $x^{l}_{(U_k)}$ mutilasses (x_RADON ; x_ACPHY) ou des FE/FIM* qui n'ont manifestement aucune influence - modélisés par des $x^{l}_{(U_k)}$ binaires (x._SEXE ; x._TYPELEUC) montrent, empiriquement, que cette hypothèse est vraisemblablement fausse.

Phase 1 – étape d'élimination - interprétation des i.st.e. $x^{l}_{(U_k)}$ identifiés comme du bruit $\mathcal{X}^{j}_{bruit}$

Les i.st.e* $x^{l}_{(U_k)}$ modélisant la géographie des FE/FIM* Curieux* de test ont parfaitement joué leur rôle. Ce qui est le cas pour : les FIM* – x._ACPHY, les FE-SAN* – modélisés par des DTA aux praticiens libéraux - à l'exception de x._RADIO la seule DTA consistante, et les FE-PHY.CHIM* – x_RADON, x_FEFO et x_PREV, qui ont systématiquement été éliminés.
En contrepartie, certains i.st.e* $x^{l}_{(U_k)}$ modélisant des FE* connotés comme Curieux* de test se sont avérés avoir un pouvoir explicatif important, c'est le cas des FE-SAN* modélisés par l'accès à des items sanitaires* qui n'ont *a priori* rien à voir avec la séquelle analysée, et des FE-SOCIO.ECO* modélisés par x_attBIEN, x_attPHY et x_INSECU (chapitre3).

Phase 2 : interprétation des i.st.e. $x^{l}_{(U_k)}$ contenus dans le paquet explicatif $\mathcal{X}^{j}_{explic}$

Concernant ceux modélisant des FIM

L'i.st.e* x._DSUIVI a un pouvoir explicatif particulièrement fort. En contrepartie, son interprétation dans le cadre de l'analyse géographique est ambivalente comme l'a fait remarquer le Professeur Auquier. En effet, cet i.st.e* modélise à la fois une sorte de risque associé à la durée d'expositions combinées à des FE* nocifs, et en même temps, les effets délétères au long cours des traitements anti-cancers. Il convient de remarquer que par définition x_DSUIVI est différente de x_tee (chapitre.2) (chapitre.3). En revanche c'est tout le problème de l'interprétation géographique des i.st.e* qui se pose et qui est discuté plus en détail dans la conclusion générale.

Concernant ceux modélisant des FE-SAN Curieux* de test*

Certains i.st.e* FE-SAN* introduits pour tester MVG s'avèrent avoir un pouvoir explicatif particulièrement important en dépit du fait qu'ils n'ont strictement aucun rapport avec l'état de santé. Les temps d'accès bruité aux services hospitaliers et aux EML* sont caractérisés des i.st.e* redondants (chapitre.3) il convient de les interpréter pour leur capacité globale à modéliser l'accès au plateau technique des établissements de santé.

Concernant ceux modélisant des FE-SOCIO.ECO Curieux* de test*

Le pouvoir explicatif de : x_att.BIENS ; x_att.PHY ; x_INSECU est particulièrement fort. Mais ces i.st.e* sont corrélés. Ils ont été qualifiés d'i.st.e* Curieux* de test à cause de leur granularité* grossière et de leur Distance a-spatiale morbide* *a priori* éloignée des états de santé étudiés (chapitre.3). Or, ces faiblesses semblent plutôt se présenter comme une force et il convient de les interpréter non pas pour ce à quoi ils étaient initialement destinés, i.e. une exposition contextuelle au stress psychologique aux effets néfaste sur la santé (Inserm - Expertise collective, 2011) mais plus globalement, i.e. pour des

carences politiques en matière de lutte contre : *la pauvreté, la délinquance, l'exclusion, l'équité d'accès à l'éducation et la culture* (Godin, 2007).

Concernant ceux modélisant des FE-PHY.CHIM

Le pouvoir explicatif des expositions aux paramètres géophysiques pour les CATA, i.e à ceux météorologiques (x_RAY, x_TEMP, x_NJAP) et morphologique (x_TOPO), permet de les élever au rang de FREPA. Il s'agit de facteurs documentés. Cependant, par construction topologique ou par nature, ils sont fortement auto-corrélés ou corrélés (chapitre.3).

Le pouvoir explicatif des expositions potentielles à la radioactivité environnementale semble particulièrement probant, ce qui les promulgue au rang de FREC. Cependant les x_l:RNM* sont surreprésentés et sont presque tous spatialement auto-corrélés - car intimement liés à la nature et à l'intensité de l'activité des INB* situées à proximité. Leur statut de FREC fait que lorsque l'i.st.e* x_NJAP est qualifié d'explicatif, il doit être interprété comme le principal vecteur de propagation des radionucléides dans les compartiments environnementaux.

Phase 3 : interprétation des i.st.e. $x^{l}_{(U_k)}$ contenus dans le paquet prédictif $\mathcal{X}^{j}_{pred}$

En régression, comme en classification, l'utilisation judicieuse de $c^{j}_{\psi.pred}$ permet d'éliminer avec une grande précision la redondance des i.st.e* contenus dans le paquet explicatif et de former un ensemble de variables parfaitement adaptées à la parcimonie prédictive OOS.

Conclusion sur les perspectives de modélisation

Au sujet de la robustesse de la dialectique MVG

L'importance des FIM, notamment en régression: $z'^{j}_{U_k,c}$, sur la géographie des séquelles empêche d'effectuer des prédictions OOS, i.e. dans les communes où aucun patient n'est spatialisé. Cette remarque n'est pas forcément extensible au contexte de la classification, i.e. sur aux $z'^{j}_{U_k,q}$, car selon la séquelle, ils peuvent être prédits sans avoir recours à la géographie des CIM. Cependant comme $z'^{j}_{U_k,q}$ permet, entre autres, de caractériser la qualité spatiotemporelle de $z'^{j}_{U_k,c}$ il n'y a aucun intérêt à le prédire indépendamment de ce dernier (chapitre.2).

La qualité prédictive et explicative de la dialectique MVG a été évaluée IS. Elle semble particulièrement robuste. Cependant il serait intéressant de l'évaluer OOS, i.e. à l'aide des nouveaux patients inclus dans la dernière base LEA disponible

Au sujet de l'interprétation des prédictions de l'opérateur MVG

En régression : Lorsque le modèle MVG surestime fortement la valeur de $z'^{j}_{Uk,c}$ cela présente l'inconvénient d'engager la qualité en santé environnementale* des communes concernées. Ce qui est préjudiciable puisque cela suggère une mauvaise accessibilité aux soins, des expositions à des substances toxiques (pesticides ou radionucléides artificiels) ou à des contextes socio-économiques prédisposant aux phénomènes morbides (stress, insécurité, pauvreté). En contrepartie, cette surestimation présente l'avantage de donner plus de vigueur aux politiques territoriales dans leur lutte contre les expositions aux DES, FREC et FREPA. Le risque étant évidemment de mettre en place des mesures exorbitantes car inadaptées à la réalité.

A contrario lorsque les $\hat{z}'^{j}_{Uk,c}$ sous-estiment la réalité géographique territoriale, ceci est aussi préjudiciable. Minimiser les risques liés aux expositions environnementales morbides engendre un marasme des politiques de lutte contre les inégalités de santé environnementale* et donc, par extension, diminue la qualité de vie des populations *in-situ.* Par exemple, les populations à risques ne seront pas forcément bien informées des risques d'exposition aux DES, FREC et FREPA auxquels elles doivent être attentives.

En classification : il est impératif de tenir compte conjointement des informations apportées par $\hat{z}'^{j}_{Uk,q}$ afin d'évaluer la qualité spatiotemporelle des $\hat{z}'^{j}_{Uk,c}$. Aussi, une attention particulière doit être portée aux cartographies des confusions géographiques : $\hat{\varsigma}^{j}_{(U_k)}$. Elles apportent une information complémentaire à celle des $\hat{z}'^{j}_{(U_k),q}$, quant au degré de certitude des prédictions environnementales des $\hat{z}'^{j}_{(U_k),c}$.

REMARQUES GENERALES

Les résultats obtenus de l'application de MyVsurfGéo* (MVG) aux états de santé étudiés sont robustes d'un point de vue statistique. Les x^{l}_{Uk} explicatifs permettent de prédire de façon particulièrement fiable les i.st.m* $z'^{j}_{Uk,c}$ et $z'^{j}_{Uk,q}$ observés.

L'analyse conjointe des i.st.e* x^{l}_{Uk} contenus dans les paquets explicatifs et prédictifs permet d'identifier et de hiérarchiser les FE/FIM* en fonction de leur efficience morbide et de les discrétiser en DES, FREC et FREPA.

S'agissant des prédictions géographiques :

En régression, les $\hat{z}'^{j}_{Uk,c}$ sont relativement robustes bien qu'ils surestiment toujours légèrement la réalité géographique décrite par les $z'^{j}_{Uk,c}$. Cette surestimation est d'autant plus forte que les séquelles sont incertaines d'un point de vue EpiGéoStat.

En classification les modalités de $z'^{j}_{Uk,c}$ sont prédites avec une bonne qualité par les $\hat{z}'^{j}_{Uk,q}$. La modalité OUI est toujours légèrement sous-estimée. Globalement les prédictions effectuées IS par le modèle FA.MVG sont très fiables.

Par conséquent, les $\hat{z}'^{j}_{Uk,c}$ et les $\hat{z}'^{j}_{Uk,q}$peuvent être fusionnés afin de caractériser les espaces territoriaux des Risques d'Expositions Géographiques (REG) morbides - mais uniquement au vu des caractéristiques environnementales des communes et décrites par les x^{l}_{Uk} et non plus à partir des états de santé observés.

Cependant au vu des nombreuses incertitudes qui maculent l'identification DES, des FREC et des FREPA, il convient de corroborer les résultats obtenus par une approche *individus-centrée** avant de caractériser les communes par des REG prédits à partir des caractéristiques environnementales qui leurs sont associées.

SECTION C) APPROCHE INDIVIDUS-CENTREE ET RISQUES D'EXPOSITIONS GEOGRAPHIQUES MORBIDES

Il s'agit de proposer des indicateurs géographiques adaptés à la gestion durable des territoires, caractérisant les espaces en fonction de Risques d'Expositions Géographiques (REG) morbides - prédits à partir de leurs caractéristiques environnementales. Les i.st.m* $z'^{REG,j}_{(U_k)}$ sont adpatés à cet objectif (chapitre.2). Ils peuvent être prédits à partir des i.st.e* $x^{l}_{(U_k)}$, contenus dans $\mathcal{X}^{j}_{MVG}$, modélisant la géographie des DES, des FREC et des FREPA.
Préalablement, il convient de valider la capacité des $x^{l}_{(U_k)}$ à prédire la géographie des états de santé étudiés. La robustesse prédictive de MVG a été évaluée In-Sample et ne peut pas être évaluée Out-Of-Sample (OOS). Toutefois l'approche individus-centrée* constitue une alternative intéressante pour valider cet objectif. Cependant, il convient de garder à l'esprit qu'en géographie il est difficile - voire impossible - d'établir *des relations causales* au sens épidémiologique *du terme* (Hill, 1965).

L'approche individus-centrée* suppose de travailler directement sur les variables épidémiologiques : séquelles y^{j}_{i}, en appariant aux individus les i.st.e* $x^{l}_{(U_k)}$ modélisant la géographie des FE/FIM* de leur milieu de vie. Mais la caractérisation statistique des interactions santé environnement s'en trouve complexifiée.

De fait, la capacité d'apprentissage de l'algorithme MVG doit être *boostée* avant d'être appliquée à ces jeux de données de validation, notés : $\Omega^{j}_{n_i}$. La démarche est interdisciplinaire, à mi-chemin entre l'épidémiologie et la géographie. L'idée consiste à comparer l'efficience morbide des DES, des FREC et des FREPA identifiés à partir des $x^{l}_{\cdot}$ explicatifs obtenus, dans le cadre de ces deux approches bien distinctes.

Hypothèse : Si les DES, FREC et FREPA de l'approche individus-centrée* peuvent être confondus avec ceux de l'approche géographique alors les prédictions géographiques MVG $\hat{z}'^{j}_{Uk,c}$ et $\hat{z}'^{j}_{Uk,q}$ sont robustes.
Conséquence : Il suffira d'adapter la stratégie d'estimation des REG à la logique environnementale prédictive, i.e. de fusionner $\hat{z}'^{j}_{Uk,c}$ et $\hat{z}'^{j}_{Uk,q}$ (chapitre.2) pour obtenir : $\hat{z}'^{REG,j}_{(U_k)}$. La robustesse de ces indicateurs géographiques composites $\hat{z}'^{REG,j}_{(U_k)}$ peut être évaluée au regard de ceux estimés sur les états de santé observés : $z'^{REG,j}_{(U_k)}$. Si elle s'avère consistante l'objectif est atteint.

Intérêt : Ainsi les espaces pourront être caractérisés par des REG morbides à partir de leurs caractéristiques environnementales. L'information colportée par les $\hat{z}'^{REG,j}_{(U_k)}$ donne aux acteurs de santé publique un moyen d'évaluer la qualité spatiale sanitaire territoriale, l'opportunité d'imaginer des mesures visant à réduire les expositions géographiques aux DES, FREC, FREPA tout en évaluant l'impact espéré sur la morbidité., et par conséquent, de garantir aux populations locales l'accès à une bonne santé environnementale* (Salem, 1995).

FACTEURS ENVIRONNEMENTAUX ET RISQUES MORBIDES - APPROCHE INDIVIDUS CENTREE

L'intérêt est d'évaluer la capacité de MVG à prédire la géographie des PM* étudiés à partir des DES, des FREC et des FREPA qui les conditionnent. Or comme cette thèse arrivait à son terme, la capacité prédictive MVG n'a pas pu être évaluée sur une partie de la Cohorte LEA non utilisée.
De plus, la modélisation géographique des séquelles par les i.st.m* : $z'^{j}_{Uk,c}$ et $z'^{j}_{Uk,q}$ souffre d'un problème d'inconsistance* statistique, atténué par un facteur de certitude π_i^{stat} discutable. L'approche individus-centrée* permet d'éluder ce biais statistique.

Par ailleurs, [en géographie l'exposition au risque] *dans l'espace et le temps est exprimée sous forme d'une relation possible entre le milieu et l'homme* (Bailly et Beguin, 2005) - et cette approche épidémiologique de la problématique donne plus de consistance aux DES, FREC et FREPA qui conditionnent les séquelles développées par les patients.

En contrepartie, l'analyse statistique des interactions entre les i.st.e* appariés x_i^l et les séquelles y_i^j est complexifiée. L'application de l'algorithme MVG aux jeux de validations $\mathfrak{Q}_{n_i}^j$ est incapable de les apréhender correctement. Par conséquent, une stratégie de *boosting** est intégrée afin d'améliorer les capacités d'apprentissage de MVG. Son implémentation permet de focaliser l'attention du processus sur les patients les mieux spatialisés, i.e. ceux dont les $x_{(U_k)}^l$ décrivent avec le plus de certitude l'environnement géographique de leur lieu de vie.
La stratégie BoostMyVsurfGéo* (BMVG) est destinée à valider les résultats de la dialectique géographique. Elle se limite à l'analyse de l'efficience des DES, FREC et FREPA sur les séquelles. L'interprétation des $x_{(U_k)}^l$ qui les caractérisent, a déjà longuement été discutée (section.B).

STRAGIE BOOSTMYVSURFGEO (BMVG)

BoostMyVsurfGéo* (BMVG) est une version améliorée de MVG adaptée à la complexité des jeux de validation $\mathfrak{Q}_{n_i}^j$ de cette approche mêlant la dialectique géographique à celle de l'épidémiologie.

Objectif

Adapter les capacités d'apprentissage de l'algorithme MVG à la complexité des données de validation : $\mathfrak{Q}_{n_i}^j$, de façon à construire des paquets $\mathcal{X}_{\mathbb{B}.explic}^j$, contenant les i.st.e* $x_{(U_k)}^l$permettant d'expliquer les séquelles y_i^j développées par les patients. L'identification des DES, des FREC des FREPA s'effectue conditionnellement à $\mathcal{X}_{\mathbb{B}.explic}^j$ et à partir de *la grille de lecture BVMG*. Ils seront ensuite comparés à ceux obtenus dans le cadre de l'approche géographique.

Proposition heuristique

Intégrer en amont de l'algorithme MVG un processus de *boosting** afin d'améliorer ses capacités statistiques d'apprentissage (Han et Kamber, 2006). La loi de probabilité spécifiée doit focaliser l'attention du processus sur la partie de $\mathfrak{Q}_{n_i}^j$: la plus complexe à appréhender - i.e. sur les patients atteints (Michel, Auquier et al., 2007), et la plus fiable - i.e. les patients spatialisés avec la plus grand certitude – en injectant de la *connaissance géographique experte* dans l'idée de la théorie *des ensembles flous* (Dubois Didier, Prade Henri, 2004).

Jeux de données de validation - BVMG

Les données utilisées dans le processus de validation individus-centrée* : $\Omega_{n_i}^{j}$ se constituent des $n_i = 747$ patients spatialisés. A chaque individu est appariée la variable séquelle y_i^j de la base consolidée ainsi que les i.st.e* $x_{(U_k)}^{l}$ caractérisant l'environnement géographique de sa commune de résidence, tel que : .

$$\Omega_{n_i}^{j} = \bigcup_{i=1}^{n_i} \left(y_i^j ; \bigcup_{k=1}^{n_{(U_k)}} \left(x_{(U_k)}^{l:FIM} ; x_{(U_k)}^{l:SAN} ; x_{(U_k)}^{l:SOC.ECO} ; x_{(U_k)}^{l:PHYCHIM} \right) \middle| \{V_{i,1} \neq \phi\} \cap \{V_{i,1} = V_{(U_k)}\} \right)$$

Remarque

Les i.st.e* $x_{(U_k)}^{l:FIM}$ caractérisant les FIM* sont intégrés pour leurs effets documentés sur l'état de santé. Ils sont plus pertinents* que les variables épidémiologiques $x_i^{l:CIM}$ car ils ont été confectionnés dans une logique géographique et en vue d'harmoniser la distances *a-spatiales morbides* à celle des autres i.st.e* à connotation sanitaire, socio-économique et physicochimique - empreints d'incertitudes par des bruits de fond environnementaux (chapitre.3).

EXPRESSION DE LA PROBLEMATIQUE D'APPRENTISSAGE ET COMPLEXITE

Remarques liminaires

A l'aune des i.st.m* $z'^{j}_{(U_k),q}$ utilisés dans le cadre de l'approche géographique les variables séquelles LEA y_i^j sont booléennes (chapitre.1).

L'incidence globale des séquelles sur la population spatialisée est estimée pour les : CATA à 12,99%, THYR à 10,17%. TUM à 2, 4.02% (chapitre.1).

Problématique d'apprentissage

L'application de MVG aux $\Omega_{n_i}^{j}$ est spécieuse, sur le plan statistique. L'expérience montre qu'en dépit d'une err.OOB de généralisation très satisfaisante en apparence, les err.OOB relatives commises sur la modalité $y_i^j = \text{OUI}$ sont médiocres.

Illustration pratique

L'application de MVG à la séquelle : TUM2 conduit aux résultats suivants :

Synthèse : CONFUSIONS	OUI	NON	TOTAL
NTC_cj_géo_($\hat{\mathbf{z}}'^{j}_{(U_k),q}$)	29	1	**30**
PRC_cj_géo_($\hat{\mathbf{z}}'^{j}_{(U_k),q}$)	96,67%	0,14%	-
PGC_cj_géo_($\hat{\mathbf{z}}'^{j}_{(U_k),q}$	3,88%	0,13%	**4,02%**

Tableau 45 : Résultats des err.OOB relatives et globales de l'application de MVH aux TUM2

Alors que l'err.OOB de généralisation semble satisfaisante, $\bar{R}_{\cdot}^{OOB}\left(\hat{f}_{MVG}^{TUME}(\mathcal{X}_{MVG}^{j}|\Lambda_j)\right) \approx 4\%$, Le modèle se contente de prédire systématiquement une absence de séquelle. De fait l'err.OOB relative est catastrophique : $\bar{R}_{\cdot}^{OOB}\left(\hat{f}_{MVG}^{TUME}(\cdot)\middle|\{y_i^j = \text{OUI}\}\right) \approx 100\%$.

Conséquence :

L'algorithme MVG est inadapté à la logique individus-centrée* et ce constat est d'autant plus vrai que l'incidence de la séquelle est faible.

Le paquet explicatif MVG : $\mathcal{X}_{explic}^{j}$ est inconsistant et ne permet pas l'identification des DES, FREC et FREPA pertinents*.

Propositions

Afin de pallier la complexité de l'approche individus-centrée*, l'idée est d'appliquer itérativement MVG à des sous-échantillons de $\Omega_{n_i}^{j}$ en focalisant le processus d'apprentissage sur : Les patients atteints -

$y_i^j = 1$, et les plus fiables - i.e. ceux pour lesquels les caractéristiques environnementales appariées x_i^l décrivent leur lieu de résidence avec le plus de certitude.
Il s'agit de garantir à la fois une bonne qualité globale et relative du modèle explicatif et le boosting* est justement une stratégie vouée à satisfaire cet objectif.

PRINCIPE DE CONSTRUCTION DES ECHANTILLONS BMVG

La procédure génère $B_{\mathbb{B}}$ échantillons BVMG $\mathfrak{L}^{j,\Theta_b^{\mathbb{B}}}_{n_{\mathbb{B}}^j}$, dans une logique de randomisation boosting* à connotation géographique, afin d'augmenter la puissance statistique d'apprentissage de MVG.

Focalisation du processus d'apprentissage sur les patients atteints

Chacun des b-échantillons BVMG : $\mathfrak{L}^{j,\Theta_b^{\mathbb{B}}}_{n_{\mathbb{B}}^j}$ contient $n_{\mathbb{B}}^j$ patients parmi lesquels :

Le nombre d'individus atteints s'élève à :

$$\left\{ n^j_{\left\{\mathbb{B}\middle|y_i^j=OUI\right\}} = \left[pia^j_{\mathbb{B}} \cdot card\left(y_\cdot^j\right)\right]\right\}$$

Le nombre d'individus sains s'élève à :

$$\left\{ n^j_{\left\{\mathbb{B}\middle|y_i^j=NON\right\}} = \left[pis^j_{\mathbb{B}} \cdot n^j_{\left\{\mathbb{B}\middle|y_i^j=OUI\right\}}\right]\right\}$$

Avec : $pia^j_{\mathbb{B}}$ et $pis^j_{\mathbb{B}}$ des paramètres BMVG fixés expérimentalement sur un compromis visant à optimiser les prédictions modales les plus difficiles à appréhender et à préserver de la puissance statistique du processus.

Focalisation du processus d'apprentissage sur les patients spatialisés les plus fiables

Les $\mathfrak{L}^{j,\Theta_b^{\mathbb{B}}}_{n_i}$ sont confectionnés de façon à utiliser les x_i^l qui modélisent la géographie des FE/FIM* des communes de résidence, avec la meilleure qualité spatiotemporelle possible. Autrement dit, il s'agit d'utiliser les patients dont la spatialisation est plus sûre.

Pour ce faire, les individus sont tirés conditionnellement à une loi de probabilité empirique $\mathbb{p}_i\left(\varrho^{géo}\right)$ qui est inversement proportionnelle au facteur d'incertitude géographique π_i^{geo} (chapitre.2).

La loi de probabilité géographique qui détermine la contribution des patients à l'analyse BVMG est soumise à une contrainte de *non biais* (Saporta, 2006), et s'estime empiriquement :

$$\mathbb{p}_i\left(\varrho^{géo}\right) = \varrho_i^{geo} \cdot \left(\sum_{i=1}^{n} \varrho_i^{geo}\right)^{-1}$$

Avec : $\varrho_i^{geo} = \left(1 - \pi_i^{géo}\right),\ \forall\, i = \{1, \dots, n_i\}$, Les échantillons d'apprentissage BVMG se constitue donc de la façon suivante :

$$\mathfrak{L}^{j,\Theta_b^{\mathbb{B}}}_{n_{\mathbb{B}}^j} = \left\{ \mathfrak{L}^{j,\Theta_b^{\mathbb{B}}}_{n_i}\left\{\Theta_b^{\mathbb{B}}; n_{\mathbb{B}}^j\right\} \middle| \Theta_b^{\mathbb{B}} \sim \mathcal{M}\left((1, \dots, n); \left(\mathbb{p}_1\left(\varrho^{géo}\right), \dots, \mathbb{p}_n\left(\varrho^{géo}\right)\right)\right)\right\}, \qquad \forall\, b \in \{1; \dots; B_{\mathbb{B}}\}$$

Remarque

Le facteur d'incertitude statistique π_i^{stat} n'est pas pris en compte. La dialectique individus-centrée* permet d'éluder le problème d'inconsistance* statistique.

Le facteur d'incertitude épidémiologique : π_i^{epi} n'est pas utilisé non plus car les données morbides LEA sont sûres et qu'il est déjà utilisé dans l'estimation des $x_{(U_k)}^{l:FIM}$

Le vecteur ϱ_i^{geo} ne contient pas de valeurs aberrantes, i.e. telles que $\varrho_i^{geo} \leq 0{,}5$.

Les ϱ_i^{geo} les plus faibles correspondent à des patients dont la spatialisation cumule plusieurs incertitudes spatiotemporelles géographiques.

Plus de la moitié des patients spatialisés a un $\varrho_i^{geo} > 0{,}9$, et va contribuer fortement à l'analyse BMVG.

En particulier, 19 patients ont un $\varrho_i^{geo} > 0{,}99$ qui outrepasse la borne supérieure des outliers (chapitre.2).

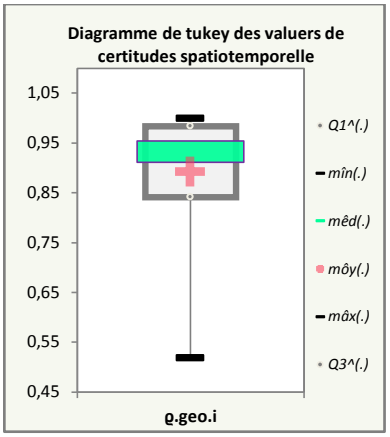


Figure 269 : Récapitulatif statistique des valeurs prises par ϱ_i^{geo}

En somme l'algorithme BVMG consiste à confectionner $B_{\mathbb{B}}$ échantillons et à leur appliquer MVG. Cependant, $B_{\mathbb{B}}$ paquets d'i.st.e* explicatifs : $\mathcal{X}_{explic}^{j,b}$ sont générés et autant d'err.OOB globales et relatives sont estimées. Une stratégie d'agrégation ensembliste permet d'obtenir les résultats BVMG définitifs.

PRINCIPE D'APPRENTISSAGE BMVG

Principe d'application de MVG aux échantillons BMVG

A l'initialisation $B_{\mathbb{B}}$-échantillons d'apprentissage BVMG sont générés, tel que :

$$\mathfrak{Q}_{n_{\mathbb{B}}^j}^{j,\Theta_{\cdot}^{\mathbb{B}}} = \left(\mathfrak{Q}_{n_{\mathbb{B}}^j}^{j,\Theta_{b=1}^{\mathbb{B}}}, \dots, \mathfrak{Q}_{n_{\mathbb{B}}^j}^{j,\Theta_{b=B_{\mathbb{B}}}^{\mathbb{B}}}\right)$$

Sur chacun d'eux l'algorithme MVG est appliqué itérativement de la façon suivante :

1. Optimisation numérique des paramètres $\widehat{\Lambda}_{j,b} = \{\{ntree\,;mtry\}|nodesize = 1\}$;
2. Estimation, stabilisée sur 50 forêts, des scores d'importance des variables : $\overline{VI}_j^b(x_{\cdot}^{l})$;
3. Hiérarchisation descendante des variables $x^{(l)}$, en fonction de leur score ;
4. Estimation de $\psi_{bruit}^{j,b}$ pondéré par $c_{\mathbb{B}.bruit}^{\psi}$, coefficient expert de bruit fixé empiriquement ;
5. Disjonction des variables de bruit éliminées $\mathcal{X}_{bruit}^{j,b}$ et des variables conservées $\mathcal{X}_{conserv}^{j,b}$;
6. Estimation de $\psi_{explic}^{j,b}$ pondéré par $c_{\mathbb{B}.explic}^{\psi}$, coefficient expert fixé expérimentalement ;
7. Sélection des variables explicatives ; Le paquet Bourrelly : $\mathcal{X}_{0.explic}^{j,b}$ est retenu ;
8. Calcul du nombre de variables contenues dans les paquets de Génuer : p_{explic}^{b} et de Bourrelly $p_{0.explic}^{b}$;
9. La procédure est interrompue avant l'identification des variables prédictives de $\mathcal{X}_{pred}^{j,b}$;
10. Prédiction *In-Sample* $\hat{y}_i^{j,b}$ parl'opérateur MVG : $\hat{f}_{MVG}^{j}(\mathcal{X}_{MVG}^{j,b}|\Lambda_{j.MVG})$.
11. Les éléments MVG sont stockés dans une matrice information BVMG: $BMVG_{info}^{\mathbb{B}}$

Un processus d'agrégation ensembliste est implémenté, en aval de l'algorithme MVG modifié, afin de fusionner les informations contenues dans $BMVG_{info}^{\mathbb{B}}$. Il permet de confectionner le paquet explicatif BVMG : $\mathcal{X}_{\mathbb{B}.explic}^{j}$, et d'estimer la qualité prédictive de l'opérateur BVMG : $\hat{f}_{\mathbb{B}MVG}^{j}(\cdot)$.

Principe de confection du paquet explicatif géographique BMVG

a. Estimation des scores BVMG par la moyenne empirique des $\overline{\mathrm{VI}}_j^b\left(x_{\cdot}^l\right)$ des variables contenues dans les paquets Bourrelly : $\mathcal{X}_{0.\mathrm{explic}}^{\mathrm{j,b}}$. Les autres sont portés à zéro, de sorte que :

$$\overline{\overline{\mathrm{VI}}}_{\mathrm{j}}^{\mathbb{B}}\left(x_{\cdot}^l\right)=\frac{1}{\mathrm{B}_{\mathbb{B}}}\sum_{b=1}^{\mathrm{B}_{\mathbb{B}}}\left(\overline{\mathrm{VI}}_{\mathrm{j}}^{b}\left(x_{\cdot}^{(l),b}\right)\cdot \mathbb{1}_{\left\{x_{\cdot}^{(l),b}\subset \mathcal{X}_{0.\mathrm{explic}}^{\mathrm{j,b}}\right\}}\right)$$

b. Estimation du nombre moyen de variables explicatives contenues dans les $\mathrm{B}_{\mathbb{B}}$ paquets de Genuer :-$\mathcal{X}_{\mathrm{explic}}^{\mathrm{j,b}}$, tel que :

$$\overline{\mathrm{p}}_{\mathbb{B}.\mathrm{explic}}^{\mathrm{j}}=\left[\frac{1}{\mathrm{B}_{\mathbb{B}}}\sum_{b=1}^{\mathrm{B}_{\mathbb{B}}}\mathrm{p}_{\mathrm{explic}}^{b}\right]$$

c. Hiérarchisation dépendante des $x_{\cdot}^l$ au regard des scores BMVG : $\overline{\overline{\mathrm{VI}}}_{\mathrm{j}}^{\mathbb{B}}\left(x_{\cdot}^l\right)$.

d. Sélection des $\overline{\mathrm{p}}_{\mathbb{B}.\mathrm{explic}}^{\mathrm{j}}$ premières $x_{\cdot,\mathbb{B}}^{(l)}$ les plus importantes et confection du paquet explicatif BMVG :

$$\mathcal{X}_{\mathbb{B}.\mathrm{explic}}^{\mathrm{j}}=\bigcup_{l=1}^{\overline{\mathrm{p}}_{\mathbb{B}.\mathrm{explic}}^{\mathrm{j}}}\left(x_{\cdot,\mathbb{B}}^{(l)}\middle|\left\{\overline{\overline{\mathrm{VI}}}_{\mathrm{j}}^{\mathbb{B}}\left(x_{\cdot}^{(1)}\right)\geq\cdots\geq\overline{\overline{\mathrm{VI}}}_{\mathrm{j}}^{\mathbb{B}}\left(x_{\cdot}^{(l)}\right)\geq\cdots\geq\overline{\overline{\mathrm{VI}}}_{\mathrm{j}}^{\mathbb{B}}\left(x_{\cdot}^{(p_l)}\right)\right\}\right)$$

Principe prédictif de l'opérateur BMVSG

L'opérateur prédictif : $\hat{\mathrm{f}}_{\mathrm{BMVG}}^{\mathrm{j}}(\cdot)$ est adapté au contexte de la classification binaire. Il s'effectue par *un vote majoritaire* au regard des prédictions $\hat{\mathrm{y}}_{\mathrm{n}_{\cdot}^{\mathrm{j},\mathbb{B}}}^{\mathrm{j,b}}$ stcockées dans $\mathrm{BMVG}_{\mathrm{info}}^{\mathbb{B}}$. En cas d'équiprobabilité la règle suivante a été fixée :

$$\hat{\mathrm{y}}_{\cdot}^{\mathrm{j},\mathbb{B}}=\begin{cases}\mathrm{môd}\left(\bigcup_{b=1}^{\mathrm{B}_{\mathbb{B}}}\left(\hat{\mathrm{y}}_{\mathrm{n}_{\mathbb{B}}^{\mathrm{j}}}^{\mathrm{j,b}}=\phi\right)\right) & \text{lorsque:}\quad \mathbb{P}_{F_{\mathrm{n}}}\left(\left\{\bigcup_{b=1}^{\mathrm{B}_{\mathbb{B}}}\left(\hat{\mathrm{y}}_{\mathrm{n}_{\mathbb{B}}^{\mathrm{j}}}^{\mathrm{j,b}}=\phi\right)=c_j\right\}\right)\neq\frac{1}{2}\\ \operatorname{argmin}\left(\bigcup_{i=1}^{\mathrm{n}}\left(\mathrm{y}_i^{\mathrm{j}}\right)\right) & \text{lorsque:}\quad \mathbb{P}_{F_{\mathrm{n}}}\left(\left\{\bigcup_{b=1}^{\mathrm{B}_{\mathbb{B}}}\left(\hat{\mathrm{y}}_{\mathrm{n}_{\mathbb{B}}^{\mathrm{j}}}^{\mathrm{j,b}}=\phi\right)=c_j\right\}\right)=\frac{1}{2}\end{cases}$$

Principe d'estimation de la qualité prédictive de BMVSG

Estimation In-Sample de l'erreur de généralisation globale du modèle prédictif BVMG :

$$\overline{\mathrm{R}}_{\mathbb{B}.\mathrm{g}}^{\mathrm{OOB}}\left(\hat{\mathrm{f}}_{\mathrm{BMVG}}^{\mathrm{j}}(\cdot)\right)=\frac{1}{\mathrm{n}}\cdot\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbb{1}_{\left\{\hat{\mathrm{y}}_{\mathrm{i}}^{\mathrm{j},\mathbb{B}}\neq y_{\mathrm{i}}^{j}\right\}}$$

Estimation In-Sample des erreurs de généralisation modales du modèle prédictif BVMG :

$$\overline{\mathrm{R}}_{\mathbb{B}.\mathrm{m}}^{\mathrm{OOB}}\left(\hat{\mathrm{f}}_{\mathrm{BMVG}}^{\mathrm{j}}(\cdot)\middle|\hat{\mathrm{y}}_{\mathrm{i}}^{\mathrm{j},\mathbb{B}}=\mathrm{c}_{\mathrm{j}}\right)=\frac{1}{\mathrm{n}}\cdot\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbb{1}_{\left\{\hat{\mathrm{y}}_{\mathrm{i}}^{\mathrm{j},\mathbb{B}}\neq y_{\mathrm{i}}^{j}\right\}\cap\left\{\hat{\mathrm{y}}_{\mathrm{i}}^{\mathrm{j},\mathbb{B}}=c_j\right\}}$$

Estimation In-Sample des erreurs de généralisation relatives du modèle prédictif BVMG :

$$\overline{\mathrm{R}}_{\mathbb{B}.\mathrm{r}}^{\mathrm{OOB}}\left(\hat{\mathrm{f}}_{\mathrm{BMVG}}^{\mathrm{j}}(\cdot)\middle|\hat{\mathrm{y}}_{\mathrm{i}}^{\mathrm{j},\mathbb{B}}=\mathrm{c}_{\mathrm{j}}\right)=\frac{1}{\operatorname{card}\left(\left\{\hat{\mathrm{y}}_{\mathrm{i}}^{\mathrm{j},\mathbb{B}}=\mathrm{c}_{\mathrm{j}}\right\}\right)}\cdot\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbb{1}_{\left\{\hat{\mathrm{y}}_{\mathrm{i}}^{\mathrm{j},\mathbb{B}}\neq y_{\mathrm{i}}^{j}\right\}\cap\left\{\hat{\mathrm{y}}_{\mathrm{i}}^{\mathrm{j},\mathbb{B}}=c_j\right\}}$$

Avec : $c_j = \{\text{OUI} ; \text{NON}\}$ l'ensemble des modalités associées à la variable séquelle $y_{\cdot}^{j}$, les paramètres utilisés dans l'algorithme BVMG et la façon dont ils ont été spécifiés pour son application aux données LEA sont déclinés dans le paragraphe suivant.

PRINCIPE D'APPLICATION DE BMVG

Dans cette sous-section sont déclinées les paramètres utilisés, les remarques expérimentales, les items statistiques renvoyés par BVMG.

Phase de construction des échantillons d'apprentissage BMVG

Paramètres :

$\{\mathbf{B}_{\mathbb{B}} = \mathbf{300}\}$**,** est le nombre d'échantillons BVMG générés. Il permet de garantir qu'une prédiction : $\hat{y}_i^{j,\mathbb{B}}$ puisse être associée à chaque patient. Les échantillons contiennent un nombre réduit de patients $n_{\mathbb{B}}^{j} \leq n_i$ et la probabilité qu'une $\hat{y}_i^{j,b}$ leur soit associée est inversement proportionnelle à leur facteur d'incertitude spatiotemporelle : $\pi_i^{géo}$

$\left\{\mathbf{pia}_{\mathbb{B}}^{\mathbf{j}} \in [\![\mathbf{80}\%; \mathbf{90}\%]\!]\right\}$, fixe le nombre d'individus inclus qui ont développé la séquelle, de façon à optimiser la qualité prédictive OOB de l'opérateur BMVG sur la modalité : $y_i^j = \text{OUI}$. Sa valeur est spécifiée inversement proportionnelle à l'incidence de la séquelle

$\left\{\mathbf{pis}_{\mathbb{B}}^{\mathbf{j}} = \mathbf{3}\right\}$, détermine le nombre de sujets sains. Il ne dépend pas de l'incidence des séquelles. Il a pour dessein de préserver la puissance statistique, i.e. d'intégrer un maximum d'individus tout en conservant une qualité prédictive OOB convenable sur : $y_i^j = \text{NON}$

Remarque :

Les valeurs prises par $\mathbf{B}_{\mathbb{B}}$**,** $pia_{\mathbb{B}}^{j}$ et $pis_{\mathbb{B}}^{j}$ conditionnent le nombre d'individus : $n_{\mathbb{B}}^{j}$ des échantillons BMVG. Leurs valeurs ont été évaluées expérimentalement après plusieurs essais.

Valeurs utilisées pour l'application de BMVG aux données LEA :

749 - patients	CATA	THYR	TUME
Incidence	12,99%	10,17%	4,02%
$\mathbf{pia}_{\mathbb{B}}^{\mathbf{j}}$	80%	85%	90%
$\mathbf{n}_{\mathbb{B}\|\mathbf{OUI}}^{\mathbf{j}}$	78	65	27
$\mathbf{n}_{\mathbb{B}\|\mathbf{NON}}^{\mathbf{j}}$	234	195	81
$\mathbf{n}_{\mathbb{B}}^{\mathbf{j}}$	**312**	**260**	**108**

Tableau 46 : Nombre de patients atteints et sains inclus dans les échantillons BVMG, en fonction de la séquelle considérée

Phase d'application itérative de MGV modifié aux échantillons BMVG

Paramètres

Le seuil d'élimination des variables de bruit $\psi_{\text{bruit}}^{j,b}$ est pondéré par $c_{\mathbb{B}.\text{bruit}}^{\psi} = 10$;

Le seuil de sélection des variables explicatives $\psi_{\text{explic}}^{j,b}$ est pondéré par $c_{\mathbb{B}.\text{explic}}^{\psi} = 1$;

Le paquet de variables explicatives retenu est celui de Bourrelly : $\mathcal{X}_{0.\text{explic}}^{j,b}$

L'opérateur prédictif MVG utilise : Le paquet $\mathcal{X}_{\text{MVG}}^{j,b} = \mathcal{X}_{\text{explic}}^{j,b}$ et le vecteur : $\Lambda_{\text{j.MVG}} = \dot{\Lambda}$

Remarques

Les valeurs de $c_{\mathbb{B}.\text{bruit}}^{\psi}$ et $c_{\mathbb{B}.\text{explic}}^{\psi}$ ont été fixées au regard de celles utilisées dans l'approche géographique en classification (section.B).

Le paquet de variables explicatives Genuer étant plus frustre, il n'est pas adapté à la phase d'agrégation ensembliste BVMG. Il est utilisé par le prédicteur $\hat{f}_{\text{MVG}}^{j}(\cdot)$ mais n'est pas conservé.

Phase d'agrégation ensembliste BMVG

Items statistiques renvoyés par l'algorithme

Le paquet BMVG de variables explicatives : $\mathcal{X}^{\mathrm{j}}_{\mathbb{B}.\mathrm{explic}}$;

Les scores BMVG associés aux i.st.e* explicatifs : $\overline{\overline{\mathrm{VI}}}^{\mathbb{B}}_{\mathrm{j}}\left(x^{l}_{\cdot}\right)$, tableaux et graphiques radars

Les prédictions BMVG : $\hat{\mathrm{y}}^{\mathrm{j},\mathbb{B}}_{\cdot}$;

Les err.OOB globale $\overline{\mathrm{R}}^{\mathrm{OOB}}_{\mathbb{B}.\mathrm{g}}\left(\hat{\mathrm{f}}^{\mathrm{j}}_{\mathrm{BMVG}}(\cdot)\right)$ et relatives $\overline{\mathrm{R}}^{\mathrm{OOB}}_{\mathbb{B}.\mathrm{r}}\left(\hat{\mathrm{f}}^{\mathrm{j}}_{\mathrm{BMVG}}(\cdot)\middle|\hat{\mathrm{y}}^{\mathrm{j},\mathbb{B}}_{\mathrm{i}}=\mathrm{c}_{\mathrm{j}}\right)$, ainsi que le nombre de confusions modales $\mathrm{N}^{\mathrm{OOB}}_{\mathbb{B}.\mathrm{m}}\left(\hat{\mathrm{f}}^{\mathrm{j}}_{\mathrm{BMVG}}(\cdot)\middle|\hat{\mathrm{y}}^{\mathrm{j},\mathbb{B}}_{\mathrm{i}}=\mathrm{c}_{\mathrm{j}}\right)$ et leur proportion $\mathrm{N}^{\mathrm{OOB}}_{\mathbb{B}.\mathrm{m}}\left(\hat{\mathrm{f}}^{\mathrm{j}}_{\mathrm{BMVG}}(\cdot)\middle|\hat{\mathrm{y}}^{\mathrm{j},\mathbb{B}}_{\mathrm{i}}=\mathrm{c}_{\mathrm{j}}\right)$, et les tableaux et histogrammes de synthèse de la *matrice des confusions.*

Remarque

BMVG ne renvoie pas de paquet prédictif : $\mathcal{X}^{\mathrm{j}}_{\mathbb{B}.\mathrm{pred}}$ car la procédure d'injection séquentielle est une opération très instable, il suffit de changer légèrement la valeur de $\psi^{j,b}_{\mathrm{pred}}$ ou de permuter ne serait-ce qu'une variable dans $\mathcal{X}^{\mathrm{j,b}}_{\mathrm{explic}}$ – à l'initialisation de la phase 3 VSURF – pour modifier complètement le contenu de $\mathcal{X}^{\mathrm{j,b}}_{\mathrm{pred}}$ (Genuer, 2010). Par conséquent, la confection d'un $\mathcal{X}^{\mathrm{j}}_{\mathbb{B}.\mathrm{pred}}$ par le processus d'agrégation BVMG proposé n'a aucun sens.

Application et perspectives

L'algorithme BMVG a été appliqué aux données LEA. Les résultats obtenus sont déclinés dans la sous-section suivante. Si la qualité statistique de la procédure BMVG est adéquate, alors l'analyse des i.st.es appariés aux patient x^{l}_{i} contenus dans $\mathcal{X}^{\mathrm{j}}_{\mathbb{B}.\mathrm{explic}}$ permet d'identifier des DES, des FREC et des FREPA. Il conviendra d'établir au préalable une *grille de lecture BVMG* adaptée à l'approche individus-centrée.

PRESENTATION DES RESULTATS BMVG AUX SEQUELLES

L'algorithme BMVG est appliqué aux patients de la Cohorte LEA - i.e. aux variables séquelles : cataractes y_i^{CATA}, tumeurs thyroïdiennes y_i^{THYR} et tumeurs secondaires : y_i^{TUM2} - Conditionnellement aux i.st.e* appariés $x_i^{\dagger}$ et conformément aux principes énoncés auparavant.

Variables utilisées

A l'initialisation les i.st.e* x_i^l actifs dans les échantillons d'apprentissage BMVG sont spécifiés selon la séquelle, de la même façon que dans l'approche géographique (section.B).

Fiabilité des résultats présentés

Afin de garantir la consistance de la proposition heuristique BMVG, l'algorithme a été relancé 10 fois : Le paquet de variables explicatives BMVG : $\mathcal{X}^{j}_{\mathbb{B}.explic}$ est resté inchangé. Les err.OOB globale et relatives n'ont pas varié de façon significative, du moins pas au-delà de l'instabilité de l'algorithme *randomForest**.

SEQUELLE : CATARACTES

L'application de BMVG à la séquelle : cataractes : y_i^{CATA} a conduit aux résultats suivants :

Variables explicatives BMVG contenues dans $\mathcal{X}^{CATA}_{\mathbb{B}.explic}$:

$x^{(l)}_{U_k}$	x._IRACT	x._GREF	x_GAMMA	x._RECHUT	x._DSUIVI	x_APL-OPHT	x._90Sr	x_RAY
$\overline{\overline{VI}}^{\mathbb{B}}_j(x^l)$	0,1662	0,1449	0,0861	0,0736	0,0756	0,0634	0,0563	0,0363

Tableau 47 : des i.st.e* $x^{(l)}$ ordonnés et valeurs des $\overline{\overline{VI}}^{\mathbb{B}}_{CATA}(x^l)$ associées

Estimation : Visuelle de l'importance des $x^{(l)}$ et qualité prédictive IS de l'opérateur BMVG :

Importance explicative BVMG des $x_i^{(l)}$

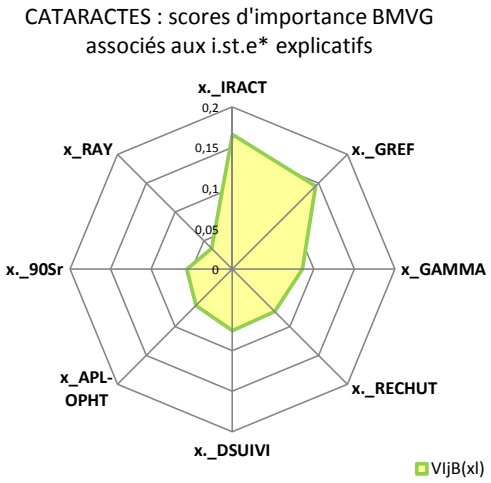
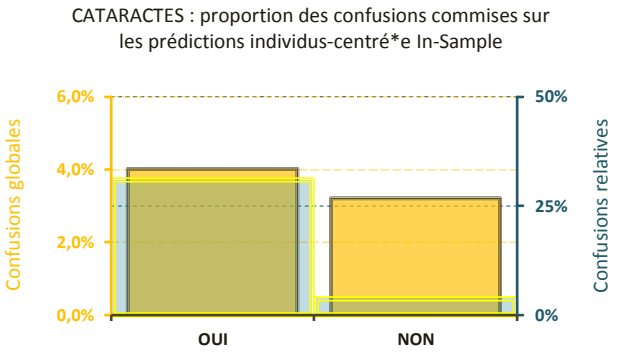


Figure 270 : Graphique radar des $\overline{\overline{VI}}^{\mathbb{B}}_j(x^l)$ associés aux $x^{(l)}_{U_k}$

Synthèse de la matrice des confusions MVG



BMVG(CATA)	OUI	NON	TOTAL
$N^{OOB}_{\mathbb{B}.\{m\cup g\}}\left(\hat{f}^j_{BMVG}(\cdot)\middle\vert\hat{y}_i^{j,\mathbb{B}}=c_j\right)$	30	24	54
$\overline{R}^{OOB}_{\mathbb{B}.m}\left(\hat{f}^j_{BMVG}(\cdot)\middle\vert\hat{y}_i^{j,\mathbb{B}}=c_j\right)$	30,93%	3,69%	-
$\overline{R}^{OOB}_{\mathbb{B}.\{m\cup g\}}\left(\hat{f}^j_{BMVG}(\cdot)\right)$	4,02%	3,21%	7,23%

Figure 271 : Histogramme et Tableau des err.OOB globales et relatives

SEQUELLES : TUMEURS THYROÏDIENNES

L'application de BMVG aux tumeurs thyroïdiennes : y_i^{THYR} a conduit aux résultats suivants :
Variables explicatives BMVG contenues dans $\mathcal{X}_{\mathbb{B}.explic}^{THYR}$:

$x_{U_k}^{(l)}$	x._IRACT	x._DSUIVI	x._GREF	x_238Pu	x_txCHOM	x_txBAC
$\overline{\overline{VI}}_j^{\mathbb{B}}(x_\cdot^l)$	0,1324	0,0753	0,0625	0,0591	0,0302	0,0225

Tableau 48 : i.st.e* $x_\cdot^{(l)}$ ordonnés et des valeurs $\overline{\overline{VI}}_{THYR}^{\mathbb{B}}(x_\cdot^l)$ associées

Estimation : Visuelle de l'importance des $x_{U_k}^{(l)}$, et qualité prédictive IS de l'opérateur BMVG

Importance explicative BVMG des $x_\cdot^{(l)}$

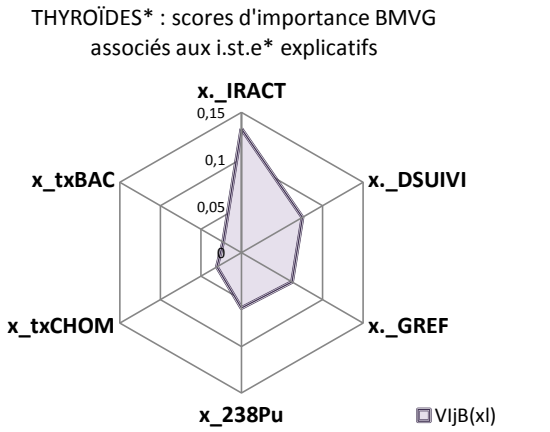
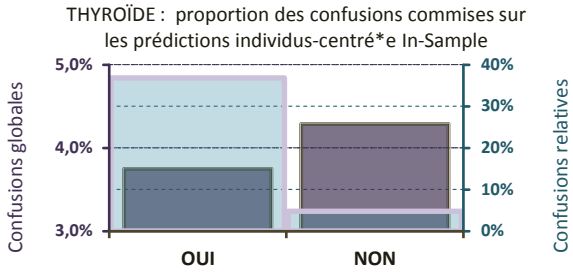


Figure 272 : Graphique radar des $\overline{\overline{VI}}_j^{\mathbb{B}}(x_\cdot^l)$ associés aux $x_\cdot^{(l)}$

Synthèse de la matrice des confusions MVG



BMVG(THYR)	OUI	NON	TOTAL
$N_{\mathbb{B}.\{mUg\}}^{OOB}\left(\hat{f}_{BMVG}^{j}(\cdot)\middle\vert\hat{y}_i^{j,\mathbb{B}}=c_j\right)$	28	32	60
$\overline{R}_{\mathbb{B}}^{OOB}\left(\hat{f}_{BMVG}^{j}(\cdot)\middle\vert\hat{y}_i^{j,\mathbb{B}}=c_j\right)$	36,84%	4,77%	-
$\overline{R}_{\mathbb{B}.\{mUg\}}^{OOB}\left(\hat{f}_{BMVG}^{j}(\cdot)\right)$	3,75%	4,28%	8,03%

Figure 273 : Histogramme et Tableau des err.OOB globales et relatives

SEQUELLES : TUMEURS SECONDAIRES

L'application de BMVG aux tumeurs secondaires : y_i^{TUM2} a conduit aux résultats suivants :
Variables explicatives BMVG contenues dans $\mathcal{X}_{explic}^{TUM2}$:

$x_{U_k}^{(l)}$	x._GREF	x._IRACT	x_238Pu	x._HEMAs	x._PROTOC	x_FDep99	x_INSECU
$\overline{\overline{VI}}_j^{\mathbb{B}}(x_\cdot^l)$	0,1234	0,1104	0,08231	0,0723	0,0531	0,0402	0,0236

Tableau 49 : i.st.e* $x_\cdot^{(l)}$ ordonnés et des valeurs $\overline{\overline{VI}}_{TUM2}^{\mathbb{B}}(x_\cdot^l)$ associées

Estimation : Visuelle de l'importance des $x_i^{(l)}$, et qualité prédictive IS de l'opérateur BMVG

Importance explicative BVMG des $x_\cdot^{(l)}$

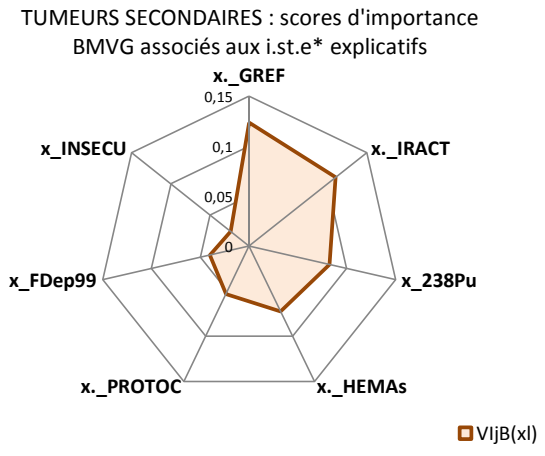
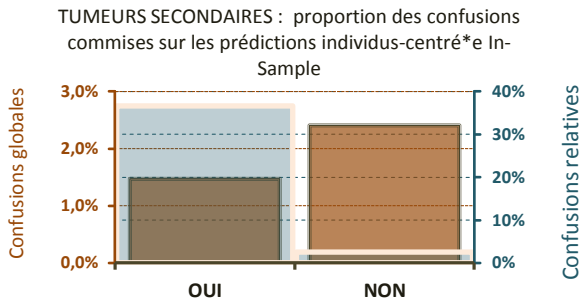


Figure 274 : Graphique radar des $\overline{\overline{VI}}_j^{\mathbb{B}}(x_\cdot^l)$ associés aux $x_\cdot^{(l)}$

Synthèse de la matrice des confusions MVG



BMVG(TUM2)	OUI	NON	TOTAL
$N_{\mathbb{B}.\{mUg\}}^{OOB}\left(\hat{f}_{BMVG}^{j}(\cdot)\middle\vert\hat{y}_i^{j,\mathbb{B}}=c_j\right)$	11	18	29
$\overline{R}_{\mathbb{B}}^{OOB}\left(\hat{f}_{BMVG}^{j}(\cdot)\middle\vert\hat{y}_i^{j,\mathbb{B}}=c_j\right)$	36,67%	2,51%	-
$\overline{R}_{\mathbb{B}.\{mUg\}}^{OOB}\left(\hat{f}_{BMVG}^{j}(\cdot)\right)$	1,47%	2,41%	3,88%

Figure 275 Histogramme et Tableau des err.OOB globales et relatives

ANALYSE DES RESULTATS BMVG

L'application de BMVG aux variables morbides LEA : y_i^j représentatives des séquelles développées par les patients permet de confectionner un paquet $\mathcal{X}^j_{\mathbb{B}.\text{explic}}$ contenant les i.st.e* susceptibles de les expliquer. Il s'agit désormais de proposer une stratégie d'analyse des $x_.^l$ qu'ils contiennent, adaptée à la dialectique individus-centrée*, afin d'identifier des DES, des FREC et des FREPA. Et par suite, d'évaluer la pertinence heuristique des résultats obtenus dans le cadre de l'approche géographique.

PRINCIPE D'IDENTIFICATION BMVG DES FACTEURS ENVIRONNEMENTAUX EXPLICATIFS

Le principe se limite à la confrontation les résultats obtenus par MVG et BMVG. L'analyse spatiale des $x_.^l$ contenus dans $\mathcal{X}^j_{\mathbb{B}.\text{explic}}$ a déjà largement été établie dans l'analyse géographique.

Objectifs

Proposer *une grille de lecture* des x_i^l contenues dans les $\mathcal{X}^j_{\mathbb{B}.\text{explic}}$ afin d'identifier des DES, FREC et des FREPA potentiellement explicatifs des séquelles développées par les patients. Ensuite, la pertinence heuristique de MVG, dans le cadre de l'approche géographique, est évaluée à partir des DES, FREC et FREPA identifiés par BMVG – i.e. par une approche individus-centrée* - en confrontant les résultats obtenus.

Hypothèse

Si les dialectiques MVG et BMVG sont adaptées à l'identification de Facteurs Environnementaux* (FE) liés à des risques morbides, alors les $x_.^l$ contenues dans les paquets $\mathcal{X}^j_{\mathbb{B}.\text{explic}}$ et $\mathcal{X}^j_{\text{explic}}$ devraient caractériser des DES, des FREC et des FREPA à peu près semblables.

Proposition d'une démarche analytique

(i) Analyser la qualité prédictive IS du modèle BMVG à partir des err.OOB stabilisées : globale $\overline{R}^{OOB}_{\mathbb{B}.\{m\cup g\}}\left(\hat{f}^j_{BMVG}(\cdot)\right)$, et relatives $\overline{R}^{OOB}_{\mathbb{B}.m}\left(\hat{f}^j_{BMVG}(\cdot)\middle|\hat{y}_i^{j,\mathbb{B}}=c_j\right)$.

(ii) Identifier des DES, des FREC et des FREPA en fonction des scores $\overline{\overline{VI}}^{\mathbb{B}}_j(x_.^l)$ et du nombre de $x_.^l$, inclus dans $\mathcal{X}^j_{\mathbb{B}.\text{explic}}$ caractérisant un même type d'exposition environnementale géographique - *Grille de lecture BMVG*.

Les Déterminants Environnementaux de Santé* (DES) : ont une efficience forte sur les séquelles. Ils sont identifiés soit par : un $x_.^l$ dont le $\overline{\overline{VI}}^{\mathbb{B}}_j(x^l)$ est considérablement élevé, ou un nombre important de $x_.^l$ aux $\overline{\overline{VI}}^{\mathbb{B}}_j(x^l)$ élevés qui modélisent l'exposition à des FE/FIM* analogues.

Les Facteurs de Risques Environnementaux Contributifs* (FREC) : ont une influence notable sur les séquelles, surtout lorsqu'ils se combinent. Ils sont caractérisés par au moins deux $x_.^l$ aux $\overline{\overline{VI}}^{\mathbb{B}}_j(x^l)$ plutôt élevés et qui modélisent l'exposition à des FE/FIM* analogues.

Les Facteurs de Risques Environnementaux Probablement Aggravants* (FREPA) : ont une influence combinée soupçonnée d'aggraver l'effet délétère des expositions aux DES* et FREC. Il s'agit de tous les autres $x_.^l$ contenues dans $\mathcal{X}^j_{\mathbb{B}.\text{explic}}$, caractérisant des expositions environnementales géographiques non éligibles au rang de DES* et de FREC.

(iii) Comparer les x_i^l des $\mathcal{X}^j_{\mathbb{B}.\text{explic}}$ et leur statut de DES, de FREC ou de FREPA de l'approche individus-centrée*, avec les $x^l_{U_k}$ contenus dans les $\mathcal{X}^j_{\text{explic}}$ et le statut environnemental morbide qui les caractérisent dans l'approche géographique afin de corroborer ou d'infirmer la pertinence heuristique des deux approches.

APPLICATION AUX SEQUELLES

Les principes d'analyse et de caractérisation BMVG sont appliqués aux résultats déclinés dans le cadre de l'approche individus-centrée*. Il s'agit : d'évaluer la pertinence heuristique du modèle prédictif BVMG, de caractériser l'efficience des FE/FIM* géographiques sur les pathologies étudiées, et enfin de comparer les résultats obtenus avec ceux de la dialectique géographique.

SEQUELLE : CATARACTES

Analyse statistique de la qualité des prédictions IS des modèles explicatifs BMVG

L'err.OOB globale est de 7,23% ce qui est acceptable pour un modèle FA.explic prédit IS. Les confusions relatives du modèle BMVG sur la modalité $\hat{y}_i^{j,\mathbb{B}} = \text{NON}$ sont correctes car estimées à 3,69%. La modalité $\hat{y}_i^{j,\mathbb{B}} = \text{OUI}$ s'élèvent à 30,93%, l'incidence des CATA sur les patients spatialisés étant de 12,99%, ce qui associé à la complexité de la problématique d'apprentissage, permet de valider la capacité dialectique de BMVG à mettre en exergue des i.st.e* $x_.^l$ explicatifs.

Analyse des variables explicatives BMVG

Rappel i.st.e explicatifs appariés aux patients x_i^l contenus dans $\mathcal{X}^{\text{CATA}}_{\mathbb{B}.\text{explic}}$:*

x_IRACT x_GREF x_GAMMA x_RECHUT x_DSUIVI x_APL_OPHT x_90Sr x_RAY

Caractérisation des FE/FIM par la grille de lecture BMVG* - des scores et des analogies :*

DES

Les FIM : Les effets secondaires de traitements agressifs au long cours*

- Scores explicatifs BMVG extrêmes obtenus par (x._IRACT ; x._GREF)
- Analogie entre 4/8 i.st.e* x_i^l explicatifs, dont : (x._DSUIVI ; x._RECHUT)

FREC

FE-PHY.CHIM : L'exposition géographique à la radioactivité environnementale et plus particulièrement à celle d'origine artificielle*

- Scores explicatifs BMVG élevés obtenus par (x_GAMMA)
- Analogie entre 2/8 i.st.e* x_i^l explicatifs, dont : (x_90Sr)

FREPA

FE-SAN : L'accès aux soins lié à la qualité du tissu sanitaire territorial et en particulier aux praticiens de santé libéraux*

- 1/8 i.st.e* x_i^l explicatifs : (x_APL-OPHT)

FE-PHY.CHIM : Les expositions géographiques aux paramètres météorologiques et en particulier aux rayonnements solaires – qui suppose indirectement des apports cosmiques de radionucléides*

- 1/8 i.st.e* x_i^l explicatifs : (x_RAY)

Comparaison heuristique des résultats obtenus pour les deux approches:

APPROCHE GEOGRAPHIQUE			APPROCHE INDIVIDUS-CENTREE		
Déterminants Environnementaux de Santé* (DES)					
(i) FIM	La lourdeur des traitements et de leurs effets secondaires temporels	x._IRACT ; x._GREF ; x._DSUIVI ; x._PROTOC	(i) FIM	La lourdeur des traitements et de leurs effets secondaires temporels	x._IRACT ; x._GREF ; x._DSUIVI ; x._RECHUT
Facteurs de Risques Environnementaux Contributifs* (FREC)					
(i) FE-SAN	L'accès aux soins : qualité du tissu sanitaire territorial (x._HEMAs ; x._IRM, x._CAME x._SACN)	x._OPHTs ; x_APL-OPHT; x._RADIO; x._NEURs,;	(i) FE-PHY.CHIM	Exposition à la radioactivité environnementale ; artificielle	x_GAMMA x_90Sr
Facteurs de Risques Environnementaux Probablement Aggravants* (FREPA)					
(i) FE-PHY.CHIM	Exposition : paramètres météo	x_RAY ; x_TEMP ; x_NJAP; x_TOPO	(i) FE-SAN	L'accès aux soins : qualité du tissu sanitaire territorial	x_APL-OPHT
(ii) FE-SOCIO.ECO	Les conjonctures sociales induisant des prédispositions morbides	x_txOUV ; x_GINIp	(ii) FE-PHY.CHIM	Exposition : paramètres météo	x_RAY
(iii) FE-PHY.CHIM	Exposition : radioactivité environnementale ; artificielle	x_238Pu ; x_137Cs-sol ; x_90Sr			

Tableau 50 : Synthèse des DES, FREC et FREPA identifiés pour les CATA, des i.st.e* associés, dans les approches : Géographique et Individus-centrée

Interprétation

Les résultats obtenus par l'analyse individus-centrée* BMVG permettent de corroborer ceux obtenus par l'approche géographique MVG, avec **une évidence forte.**

SEQUELLE : TUMEURS THYROÏDIENNES

Analyse statistique de la qualité des prédictions IS des modèles explicatifs BMVG

L'err.OOB globale est de 8,03% ce qui est très confortable pour un modèle FA.explic prédit IS. Les confusions relatives du modèle BMVG sur : la modalité $\hat{y}_i^{j,\mathbb{B}} = \text{NON}$ sont plus que correctes puisque estimées à 4,28%, la modalité $\hat{y}_i^{j,\mathbb{B}} = \text{OUI}$ s'élèvent à 36,84%, et l'incidence des THYR sur les patients spatialisés étant de 10,17%. Ceci, associé à la complexité de la problématique d'apprentissage, permet de valider la capacité dialectique de BMVG à mettre en exergue des i.st.e* $x_{\cdot}^{l}$ explicatifs.

Analyse des variables explicatives BMVG

Rappel i.st.e explicatifs appariés aux patients x_i^l contenus dans $\mathcal{X}_{\mathbb{B}.\text{explic}}^{\text{THYR}}$:*

x_IRACT x_DSUIVI x_GREF x_238Pu x_txCHOM x_txBAC

Caractérisation des FE/FIM par la grille de lecture BMVG* - des scores et des analogies :*

DES

Les FIM : Les effets secondaires de traitements délétères au long cours*

- Scores explicatifs extrêmes BMVG obtenus par (x._IRACT)
- Analogie entre 3/6 i.st.e* x_i^l explicatifs, dont : (x._DSUIVI ; x._GREF)

FREC

FE-PHY.CHIM : L'exposition géographique à la radioactivité environnementale d'origine artificielle liée à l'activité des INB* et de façon résiduelle aux catastrophes et aux essais nucléaires historiques*

- Score explicatif BMVG élevé obtenu par (x_238Pu)

FREPA

FE-SOCIO.ECO : L'effet de conjonctures défavorables induisant des prédispositions morbides – essentiellement liées à la pauvreté et au niveau de culture*

- Analogie entre 2/6 i.st.e* x_i^l explicatifs, dont : (x_txCHOM ; x_txBAC)

Comparaison heuristique des résultats obtenus pour les deux approches:

APPROCHE GEOGRAPHIQUE			APPROCHE INDIVIDUS-CENTREE		
Déterminants Environnementaux de Santé* (DES)					
(i) FIM	La lourdeur des traitements et leurs effets secondaires temporels	x._DSUIVI ; x._IRACT ; x.GREF	(i) FIM	La lourdeur des traitements et leurs effets secondaires temporels	x._IRACT ;; x._DSUIVI ; x._GREF
Facteurs de Risques Environnementaux Contributifs* (FREC)					
(i) FE-SOCIO.ECO	Spécialisation économique des espaces induisant des prédispositions morbides	x_txSAU ; (x_DJE.Cr) x_.txOUV ; x._FDep09	(i) FE-PHY.CHIM	Exposition à la radioactivité environnementale ; artificielle	x_238Pu
(ii) FE-SAN	L'accès aux soins : qualité du tissu sanitaire territorial (x._ORLs; x._HEMAs, x._NEURs)	x._IRM; x._SCAN; x._RADIO; x._ENDOs ;			
Facteurs de Risques Environnementaux Probablement Aggravants* (FREPA)					
(i) FE-PHY.CHIM	Exposition à la radioactivité environnementale ; artificielle	x_3H; x_238Pu ; x_137Cs-sol	(i) FE-SOCIO.ECO	Les conjonctures sociales induisant des prédispositions morbides	x_txCHOM ; x_txBAC
(ii) FE-SOCIO.ECO	Les conjonctures sociales induisant des prédispositions morbides	x_FDep09, x_txOUV; x_RevMed.m ; x_attBIENS			

Tableau 51 : Synthèse des DES, FREC et FREPA identifiés pour les THYR, des i.st.e* associés, dans les approches : Géographique et Individus-centrée

Interprétation

Les résultats obtenus par l'analyse individus-centrée* BMVG permettent de corroborer ceux obtenus par l'approche géographique MVG, avec **une évidence modérée.**

SEQUELLE : TUMEURS SECONDAIRES

Analyse statistique de la qualité des prédictions IS des modèles explicatifs BMVG

L'err.OOB globale est de 3,88% ce qui est ambivalent pour le modèle FA.explic prédit IS. Cependant les confusions relatives du modèle BMVG sur : la modalité $\hat{y}_i^{j,\mathbb{B}} =$ NON sont plus que correctes puisque estimées à 2,51%, la modalité $\hat{y}_i^{j,\mathbb{B}} =$ OUI valent 36,67%, ce qui est un exploit au vu de complexité qui caractérise ce jeu de données d'apprentissage. En revanche, comme l'incidence des TUM2 sur les patients spatialisés est de 4,02%, la capacité de BMVG à identifier convenablement les i.st.e* $x_{\cdot}^{l}$ explicatifs est envisageable mais néanmoins discutable

Analyse des variables explicatives BMVG

Rappel i.st.e explicatifs appariés aux patients x_i^l contenus dans $\mathcal{X}_{\mathbb{B}.\text{explic}}^{\text{TUM2}}$:*

x_GREF x_IRACT x_238Pu x._HEMAs x_PROTOC x_FDep99 x_INSECU

Caractérisation des FE/FIM par la grille de lecture BMVG* - des scores et des analogies :*

DES

Les FIM : Les effets secondaires délétères de traitements agressifs*

- Scores explicatifs BMVG extrêmes obtenus par (x._GREF ; x._IRACT)
- Analogie entre 3/7 i.st.e* x_i^l explicatifs, dont : (x._PROTOC)

FREC

FE-PHY.CHIM : L'exposition géographique à la radioactivité environnementale d'origine artificielle liée à l'activité des INB* et de façon résiduelle aux catastrophes et aux essais nucléaires historiques*

- Score explicatif BMVG élevé obtenu par (x_238Pu)

FE-SAN : L'accès aux soins territoriaux et en particulier aux services des établissements de santé spécialisés dans la prise en charge des cancers de type séquelle*

- Score explicatif BVMG élevé obtenu par (x._HEMAs)

FREPA

FE-SOCIO.ECO : L'effet combiné de la défaveur sociale et d'expositions contextuelles au stress lié à des nuisances sociales, comme l'insécurité territoriale.*

- Analogie entre 2/7 i.st.e* x_i^l explicatifs : (x_FEep99 ; x_INSECU)

Comparaison heuristique des résultats obtenus pour les deux approches:

APPROCHE GEOGRAPHIQUE			APPROCHE INDIVIDUS-CENTREE		
Déterminants Environnementaux de Santé* (DES)					
(i) FIM	Les effets secondaires délétères des traitements au long cours	x._DSUIVI; x._PROTOC; x._AGE_DIAG	(i) FIM	La lourdeur des traitements et leurs effets secondaires temporels	x._GREF ; x._IRACT; x._PROTOC
Facteurs de Risques Environnementaux Contributifs* (FREC)					
(i) FE-PHY.CHIM	Exposition à la radioactivité environnementale ; (i) artificielle, (ii) naturelle x_131i; x_125Sb	x_GAMMA ; x_238Pu ; x_90Sr ; x_EGRA ;	(i) FE-PHY.CHIM	Exposition à la radioactivité environnementale artificielle	x_238Pu
(ii) FE-SAN	L'accès aux soins : qualité du tissu sanitaire territorial (x._IRM, x._SCAN)	x._HEMAs; x._NEURs; x._ENDOs;	(ii) FE-SAN	L'accès aux soins : qualité du tissu sanitaire territorial	x._HEMAs
Facteurs de Risques Environnementaux Probablement Aggravants* (FREPA)					
(i) FE-SOCIO.ECO	Les conjonctures sociales induisant des prédispositions morbides (x_txFoyFisc ; x_FDep99)	x_attPHY, x_attBIENS, x_INSECU ; x_FDep09,	(i) FE-SOCIO.ECO	Les conjonctures sociales induisant des prédispositions morbides	x_FEep99 ; x_INSECU
(ii) FE-PHY.CHIM	Exposition : paramètres météo et leurs effets catalyseurs indirects sur la radioactivité environnementale	x_NJAP			

Tableau 52 : Synthèse des DES, FREC et FREPA identifiés pour les TUM2, des i.st.e* associés, dans les approches : Géographique et Individus-centrée

Interprétation

Les résultats obtenus par l'analyse individus-centrée* BMVG permettent de corroborer ceux obtenus par l'approche géographique MVG, avec **une évidence très forte.**

REMARQUES, CONCLUSIONS ET PERSPECTIVES

L'analyse des err.OOB globales et relatives des prédictions In-Sample a permis de démonter la qualité prédictive de l'opérateur BMVG. La méthode BMVG est une alternative intéressante à VSURF pour analyser l'efficience morbide de Facteurs Environnementaux* (FE) géographiques.

L'intégration d'un Boosting* à connotation géographique dans la dialectique MVG est une stratégie d'apprentissage adaptée à la complexité de l'approche individus-centrée*. Cependant, sa fiabilité est inversement proportionnelle à l'incidence des états de santé étudiés. Plus l'incidence d'une séquelle est faible et moins les i.st.e* $x_{\cdot}^{l}$ de $\mathcal{X}^{j}_{\mathbb{B}.\text{explic}}$ sont à même de prédire les variables morbides : y_{i}^{j}.

Les principes d'analyse proposés, fondés sur *la grille de lecture BMVG** des scores $\overline{\overline{\text{VI}}}_{j}^{\mathbb{B}}(x_{\cdot}^{l})$ et des analogies entre i.st.e* $x_{\cdot}^{l}$ contenus dans $\mathcal{X}^{j}_{\mathbb{B}.\text{explic}}$, sont facilités par l'utilisation des graphiques radars, et semblent parfaitement adaptés à l'identification des DES, des FREC et des FREPA.

La mise en perspective des résultats MVG - de l'approche géographique - et des résultats BMVG - de l'approche individus-centrée* - ont permis de montrer que :

*Les FIM** sont systématiquement éligibles au rang de DES. En particulier pour ce qui de la géographie des traitements agressifs liés à des irradiations corporelles ou des greffes. Aussi, les effets secondaires temporels sont rappelés par : x._DSUIVI qui a un pouvoir explicatif fort. Il convient de rappeler que son interprétation est ambivalente, puisque la géographie des durées de suivi modélise à la fois les effets au long cours des traitements et ceux d'expositions environnementales combinées.

*Les FE-PHY.CHIM** obtiennent systématiquement un statut de FREC, en particulier l'exposition à la radioactivité environnementale d'origine artificielle liée à la diffusion de radionucléides par les INB* en activité, et naturelle induite par le rayonnement cosmique. Ce constat avait déjà été établi par l'approche géographique, avec une prégnance cependant moins forte. L'effet des expositions climatiques est aussi mis en exergue par l'approche individus-centrée*. Toutefois, il demeure difficile de distinguer les effets directs des paramètres météorologiques de leurs effets indirects, i.e. contributifs à l'augmentation de la radioactivité environnementale.

Les FE-SAN, dans le cadre de l'approche individus-centrée*, sont éligibles au rang soit de FREC soit de FREPA. La qualité des tissus sanitaires* territoriaux et en particulier l'accès aux praticiens de santé libéraux et aux services spécialisés des établissements de santé ont une efficience morbide explicative notable. Or dans le cadre de l'approche géographique ces expositions semblent avoir une influence morbide plus forte puisqu'elles sont, le plus souvent, éligibles au rang de FREC.

*Les FE-SOCIO.ECO** sont généralement qualifiés de FREPA, à l'instar de ce qui est observé dans l'approche géographique. Cependant, le rôle des conjonctures de défaveur sociale induisant des prédispositions géographiques morbides – et plus particulièrement de la pauvreté, de la précarité et des défaillances d'accès à la culture, et en dépit de son statut Curieux*, celle du sentiment de stress lié au niveau géographique de nuisances sociales, comme l'insécurité.

Les résultats obtenus avec BMVG ont été validés statistiquement par l'analyse des confusions. La capacité de MVG à identifier des DES, des FREC et des FREPA explicatifs et à prédire, à partir des i.st.e* $x_{\cdot}^{l}$ qui les caractérisent, la géographie des états de santé étudiés a pu être validée par l'approche individus-centrée*. Par conséquent il est désormais possible de caractériser les communes par des REG à partir les prédictions MVG, au vu des caractéristiques environnementales des territoires.

MODELISATION ENVIRONNEMENTALE DES RISQUES D'EXPOSITIONS GEOGRAPHIQUES MORBIDES

L'observation des états de santé d'une population, en l'occurrence, ceux des individus spatialisés de la Cohorte LEA, permet de caractériser les territoires par une typologie des Risques d'Expositions Géographiques (REG) morbides par des i.st.m* $z_{(U_k)}^{REG.j}$. Ces indicateurs sont assez synthétiques et suffisamment pertinents* pour caractériser les REG morbides auxquels sont assujetties les populations locales.

Ils sont adaptés à la gestion des espaces. De plus, leur analyse spatiale a permis de subodorer un effet environnement évident (chapitre.2). Mais ces derniers ne permettent pas d'identifier des leviers environnementaux (ou facteurs) permettant de limiter, par des mesures médicales ou politiques, les risques d'expositions morbides.

Cependant, la méthode MyVsurfGéo* (MVG) permet d'atteindre cet objectif en identifiant des Déterminants Environnementaux de Santé* (DES), des Facteurs de Risques Environnementaux Contributifs* (FREC) et des Facteurs de Risques Environnementaux Probablement Aggravants* (FREPA), à partir des caractéristiques environnementales des espaces, modélisées par des i.st.e* $x_{(U_k)}^{l}$ (chapitre.3).

La robustesse de MVG a été évaluée statistiquement. Sa capacité à identifier des DES, des FREC et des FREPA, à partir d'indicateurs géographiques, a été validée par une approche individus-centrée*, i.e. directement sur les variables épidémiologiques.

Ainsi, le géographe fournit aux médecins et aux politiques, des informations géographiques opérationnelles leur permettant d'identifier des leviers afin de garantir la santé publique. Les médecins peuvent préconiser des mesures individuelles préventives au regard des caractéristiques environnementales du lieu de vie de leurs patients.
Et les politiques disposent des connaissances nécessaires pour mettre en place des mesures collectives visant à garantir l'accès à une bonne santé environnementale*. Le coût de la mise en œuvre de ces mesures peut être estimé par des scenarii économiques. Leur efficience sur les DES, les FREC ou les FREPA peut être évaluée directement à partir des i.st.e* $x_{(U_k)}^{l}$. Et conséquemment, l'impact espéré sur la géographie des états de santé ciblés peut être estimée par l'opérateur $\hat{f}_{MVG}^{j}(\cdot)$.

Malheureusement, les indicateurs MVG prédits : $\hat{z}'^{j}_{(U_k),c}$ et $\hat{z}'^{j}_{(U_k),q}$ sont difficiles à interpréter. Ils sont entachés d'incertitudes et inadaptés *au contrôle territorial*. En revanche les $z_{(U_k)}^{REG.j}$ le sont. Ils permettent, par un processus de fusion, de caractériser l'impact de ces mesures par un indicateur spatial unique, synthétique et robuste, à partir duquel il est possible de prédire le gain espéré en matière de réduction des Risques d'Expositions Géographiques morbides.

En conséquence de quoi, les politiques disposent de moyens simples permettant d'évaluer, d'adapter et de justifier auprès des populations, le coût des mesures prises en santé environnementale* et de s'assurer qu'elles soient en adéquation avec leurs besoins et leurs attentes – puisque *le contrôle social* est *garant de la gestion durable des espaces* (Salem, 1995).

STRATEGIE PREDICTIVE DES RISQUES D'EXPOSITION GEOGRAPHIQUES MORBIDES A PARTIR DES CARACTERISTIQUES ENVIRONNEMENTALES GEOGRAPHIQUES

La méthode MVG permet d'identifier, avec une grande acuité, des DES, des FREC et des FREPA à partir desquels il est possible de prédire la géographie d'états de santé par les i.st.m* $\hat{z}'^{j}_{(U_k),c}$ et $\hat{z}'^{j}_{(U_k),q}$. Ces derniers ne sont pas adaptés à la gestion durable des espaces géographiques. Cependant, en les fusionnant, il est possible de caractériser les espaces par des Risques d'Expositions Géographiques morbides à partir d'un i.st.m* composite adéquate.

OBJECTIF

Proposer une méthode robuste de fusion des prédictions environnementales MVG, i.e. des $\hat{z}'^{j}_{(U_k),c}$ et $\hat{z}'^{j}_{(U_k),q}$, afin de caractériser les REG morbides territoriaux par un i.st.m* composite, noté $\hat{z}^{REG.j}_{(U_k)}$, uniquement à partir de la géographie des DES, FREC et FREPA associés au PM* ciblé.

REMARQUES LIMINAIRES

Les $z^{REG.j}_{(U_k)}$ sont estimés conditionnellement aux états de santé observés sur les populations locales et permettent de caractériser les espaces territoriaux par une typologie de REG à quatre modalités : PROBABLE ; POSSIBLE ; INDEMONTRABLE ; FAIBLE (chapitre.2).
L'opérateur MVG permet de prédire, à partir des caractéristiques environnementales des unités géographiques, l'état de santé des populations locales par le biais de $\hat{z}'^{j}_{(U_k),c}$ et $\hat{z}'^{j}_{(U_k),q}$. En adaptant la stratégie d'estimation des $z^{REG.j}_{(U_k)}$ à la logique prédictive, il est possible de supputer $\hat{z}^{REG.j}_{(U_k)}$ et par extension de valider l'objectif fixé.

L'analyse spatiale des REG a déjà largement été discutée (chapitre.2). De plus, l'analyse des REG prédits, par les $\hat{z}^{REG.j}_{(U_k)}$, est nécessairement moins fiable que celle effectuée sur les estimés : $z^{REG.j}_{(U_k)}$. Elle ne présente donc aucun intérêt. Par conséquent, l'analyse spatiale se limite à l'évaluation de la robustesse de la stratégie de fusion des REG – adaptée à la logique prédictive. Autrement dit, à une analyse des différenciations spatiales entre les $\hat{z}^{REG.j}_{(U_k)}$ et les $z^{REG.j}_{(U_k)}$.

PROPOSITION METHODOLOGIQUE

Le principe d'estimation des prédictions environnementales des REG morbides à partir de la modélisation géographique des DES, FREC et FREPA est le suivant :

$$\hat{z}^{REG.j}_{(U_k)} = \begin{cases} \text{PROBABLE} \quad \text{Lorsque:} \quad \left\{\left\{\hat{z}'^{j}_{(U_k),c} \geq \hat{\varphi}^{*}_{j}\right\} \cap \left\{\hat{z}'^{j}_{(U_k),q} = \text{"OUI"}\right\}\right\} \\ \text{POSSIBLE} \quad \text{Lorsque:} \quad \left\{\left\{\hat{z}'^{j}_{(U_k),c} \geq \hat{\varphi}^{*}_{j}\right\} \cap \left\{\hat{z}'^{j}_{(U_k),q} = \text{"NON"}\right\}\right\} \\ \text{INDEMONTRABLE} \quad \text{Lorsque:} \quad \left\{\left\{\hat{z}'^{j}_{(U_k),c} \in \mathbb{R}\right\} \cap \left\{\hat{z}'^{j}_{(U_k),q} = \text{"INCERTAIN"}\right\}\right\} \\ \text{FAIBLE} \quad \text{Lorsque:} \quad \left\{\left\{\hat{z}'^{j}_{(U_k),c} < \hat{\varphi}^{*}_{j}\right\} \cap \left\{\hat{z}'^{j}_{(U_k),q} = \text{"NON"}\right\}\right\} \end{cases}$$

Les valeurs des $\hat{z}^{REG.j}_{(U_k)}$ sont fortement conditionnées par celle de $\hat{\varphi}^{*}_{j}$, dont le calcul fait intervenir deux paramètres à l'instar de celle de φ^{*}_{j} (chapitre.2).

Estimation du seuil prédictif d'élasticité géographique

$$\hat{\varphi}_j^* = \left\lceil \hat{\xi}_{\text{géo}}^{\text{j}} \cdot \left(\bar{\hat{z}}'^{\text{j}}_{(U_k),c} + \hat{b}_{1-\alpha}^{*,\text{j}} \cdot \frac{\hat{\sigma}\left(\hat{z}'^{\text{j}}_{(U_k),c}\right)}{\sqrt{N(U_k)}} \right) \right\rceil$$

Estimation des paramètres prédictifs du seuil

La variable $\hat{b}_{1-\alpha}^{*,\text{j}}$, à l'instar de $b_{1-\alpha}^{*,\text{j}}$, s'estime par *bootstrap* conditionnellement à un niveau prédictif de risque choisi (Chernick, 1999). Le paramètre d'intérêt est donc $\hat{\alpha}^{\text{j}}$. Il correspond au niveau de risque admis dans l'estimation d'une borne supérieure de la moyenne spatiale EpiGéoStat prédite $\hat{z}'^{\text{j}}_{(U_k),c}$ qui ne devrait pas être dépassée du simple fait du hasard.

Le second paramètre est *le coefficient d'élasticité géographique prédictif :* $\hat{\xi}_{\text{géo}}^{\text{j}}$, qui, à l'instar de $\xi_{\text{géo}}^{j}$, consiste à injecter de la connaissance géographique experte – par *la théorie des ensembles flous* (Dubois Didier, Prade Henri, 2004) – afin de rationaliser le nombre de modalités $c_{\text{REG}}^{\text{j}} = \text{PROBABLE}$ et $c_{\text{REG}}^{\text{j}} = \text{PROBABLE}$, prises par $\hat{z}_{(U_k)}^{\text{REG.j}}$.

Remarque : Les valeurs des paramètres $\xi_{\text{géo}}^{j}$ et α^{j} ne sont pas aléatoires et peuvent être estimées (chapitre.2).

HYPOTHESE SUR LES PARAMETRES

Si les valeurs prises par $\xi_{\text{géo}}^{j}$ et α^{j} ne sont pas aléatoires alors celles de $\hat{\xi}_{\text{géo}}^{\text{j}}$ $\hat{\alpha}^{\text{j}}$ le sont encore moins et peuvent être calibrées en vertu des *principes de conservation partielle ou totale de leurs propriétés géographiques.*

SPECIFICATIONS METHODOLOGIQUES SUR L'ESTIMATION DES PARAMETRES

Les paramètres $\hat{\alpha}^{\text{j}}$ et $\hat{\xi}_{\text{géo}}^{j}$ permettent de rationaliser le nombre de $\hat{z}_{(U_k)}^{\text{REG.j}} = \{\text{PROBABLE} \cup \text{POSSIBLE}\}$. L'action de $\hat{\alpha}^{\text{j}}$ est limitée et se borne à des considérations statistiques. Celle du paramètre $\hat{\xi}_{\text{géo}}^{\text{j}}$ est beaucoup plus importante et permet aussi d'introduire des connaissances expertes sur la nature du phénomène tout en tenant compte des distorsions géographiques induites par les pondérations spatiales : $\pi_{\text{i}}^{\text{j}}$ et temporelles : $\text{tee}_{\text{i}}^{\text{j}}$ (chapitre.2).

CALIBRATION DU NIVEAU DE RISQUE PREDICTIF

Remarques liminaires

Plus la valeur de $\hat{\alpha}^{\text{j}}$ est grande et plus celle de $\hat{b}_{1-\alpha}^{*,\text{j}}$ sera statistiquement *petite*. Le choix d'un niveau de risque fait allégeance aux prénotions établies dans le domaine des statistiques. Par conséquence : $\hat{\alpha}^{\text{j}} \in [\![5\% ; 10\%]\!]$ (Saporta, 2006).

Les valeurs de $\hat{\alpha}^{\text{j}}$, à l'instar de celles prises par α^{j}, doivent être adaptées à l'action des pondérations EpiGéoStat et à l'injection des temps d'exposition à l'environnement $\text{tee}_{\text{i}}^{\text{j}}$ sur les caractéristiques statistiques de la variable $\hat{z}'^{\text{j}}_{(U_k),c}$ (chapitre.2).

Stratégie de calibration proposée

Comme α^{j} est fixé conditionnellement aux caractéristiques statistiques de $z'^{j}_{(U_k),c}$, et en supposant *un principe de conservation totale* des propriétés géographiques de l'i.st.m* $\hat{z}'^{j}_{(U_k),c}$, alors il est manifeste que $\{\widehat{\alpha}^{j} = \alpha^{j}\}$.

CALIBRATION DU COEFFICIENT D'ELASTICITE GEOGRAPHIQUE PREDICTIF

Remarques liminaires

Les $\hat{z}'^{j}_{(U_k),c}$ parviennent difficilement à prédire les extrema de $z'^{j}_{(U_k),c}$. Le nombre de $z'^{j}_{(U_k),q} = \text{OUI}$ prédits par $\hat{z}'^{j}_{(U_k),q}$ est systématiquement sous-estimé (chapitre.2). Les conséquences sur les propriétés statistiques de $\hat{z}^{REG.j}_{(U_k)}$ sont les suivantes : Le nombre de $\hat{z}^{REG.j}_{(U_k)} = \text{PROBABLE}$ est nécessairement sous-estimé. Le nombre de $\hat{z}^{REG.j}_{(U_k)} = \text{POSSIBLE}$ est fortement surestimé - au regard des modalités attendues et observées sur $z^{REG.j}_{(U_k)}$.

L'action de $\{\widehat{\alpha}^{j} = \alpha^{j}\}$ sur les $\hat{z}^{REG.j}_{(U_k)}$ est assez limitée (chapitre.2). De fait la calibration de $\hat{\xi}^{j}_{géo}$ doit permettre de juguler, de façon pertinente, le nombre de $\hat{z}^{REG.j}_{(U_k)} = \text{POSSIBLE}$, sans pour autant affecter le nombre de $\hat{z}^{REG.j}_{(U_k)} = \text{PROBABLE}$ qui est sous-estimé.

L'usage a montré que $\xi^{j}_{géo} \in [\![1\ ;\ 2]\!]$ (chapitre.2), et il n'y a aucune raison pour que les valeurs de $\hat{\xi}^{j}_{géo}$ diffèrent singulièrement de celles prises par $\xi^{j}_{géo}$.

Stratégie de calibration proposée

Puisque $\hat{\xi}^{j}_{géo}$ est lié à la fois à la nature du REG et aux caractéristiques statistiques des $\hat{z}^{REG.j}_{(U_k)}$, en supposant *un principe de conservation partielle* des composantes : phénoménologique et statistique des $z^{REG.j}_{(U_k)}$, alors les valeurs de $\hat{\xi}^{j}_{géo}$ sont identiques à celles de $\xi^{j}_{géo}$ à une constante près : $\aleph^{j}_{géo}$, tel que :

$$\hat{\xi}^{j}_{géo} = \left(\xi^{j}_{géo} + \aleph_{géo}\right) \in [\![1\ ;\ 2]\!]\ ;\quad \text{avec: } \{\aleph_{géo} = \text{"petit"}\}$$

Un *seuil d'élasticité géographique prédictif :* $\widehat{\varphi}^{*}_{j}$ robuste peut désormais être estimé. Les espaces peuvent donc être caractérisés par des Risques d'Expositions Géographiques morbides à partir de leurs caractéristiques environnementales.

PRINCIPES D'APPLICATION ET D'ANALYSE DES RISQUES PREDITS

La stratégie prédictive vouée à la caractérisation environnementale des espaces géographiques par des REG peut être appliquée aux états de santé étudiés – séquelles CATA, THYR et TUM2. Les paramètres utilisés, la stratégie d'analyse des prédictions, ainsi que les caractéristiques des résultats statistiques et cartographiques présentés, sont les suivants :

PARAMETRES SPECIFIES POUR LA MODELISATION

Les valeurs des $\widehat{\alpha}^{j}$ et $\hat{\xi}^{j}_{géo}$ sont calibrées en vertu *des principes de conservation partielle ou totale* énoncés auparavant. Les valeurs spécifiées sont les suivantes :

Le niveau de risque du seuil prédictif

$$\{\hat{\alpha}^j = \alpha^j\}$$

Le coefficient d'élasticité géographique prédictif

$$\hat{\xi}^{j}_{\text{géo}} = \left(\xi^{j}_{\text{géo}} + \aleph_{\text{géo}}\right) \in [\![1\ ;\ 2]\!]$$

Avec : $\{\aleph_{\text{géo}} = 0.05\}$, cette valeur a été fixée empiriquement en fonction de sa capacité à augmenter le nombre de $\hat{z}^{\text{REG.j}}_{(U_k)} = \text{PROBABLE}$ tout en veillant à juguler le nombre de $\hat{z}^{\text{REG.j}}_{(U_k)} = \text{POSSIBLE}$.

Paramètres utilisés pour chacune des séquelles

Séquelle :	CATA	THYR	TUM2
$\hat{\alpha}^j$	10%	5%	5%
$\hat{\xi}^{j}_{\text{géo}}$	1,15	1,55	1,95

Tableau 53 : Synthèse des paramètres spécifiés pour la prédiction des REG à chaque séquelle

ANALYSE DE LA QUALITE DES PREDICTIONS GEOGRAPHIQUES

Les $\hat{z}^{\text{REG.j}}_{(U_k)}$ sont des i.st. qualitatifs multi-classes

Analyse spatiale visuelle des dissemblances

Principe : repérer les U_k pour lesquelles les modalités de $z^{\text{REG.j}}_{(U_k)}$ et $\hat{z}^{\text{REG.j}}_{(U_k)}$.sont différentes.

Documentation cartographique : La règle d'affichage du nom des communes utilisée pour $z^{\text{REG.j}}_{(U_k)}$ est identique à celle spécifiée pour $\hat{z}^{\text{REG.j}}_{(U_k)}$, soit :

$$\text{label}_{(U_k)} = \begin{cases} \text{Display} & \text{lorsque:} \quad \left\{\hat{z}^{\text{CATA}}_{(U_k)} = \text{PROBABLE}\right\} \\ \text{Display} & \text{lorsque:} \quad \left\{\hat{z}^{\text{THYR}}_{(U_k)} = \text{PROBABLE}\right\} \\ \text{Display} & \text{lorsque:} \quad \left\{\hat{z}^{\text{TUM2}}_{(U_k)} \in \{\text{PROBABLE} \cup \text{POSSIBLE}\}\right\} \end{cases}$$

Analyse de la distribution spatiale globale des REG observés et prédits

Principe : La mise en perspective des histogrammes des fréquences d'apparition des modalités : r_j pour $z^{\text{REG.j}}_{(U_k)}$ et $\hat{z}^{\text{REG.j}}_{(U_k)}$ permet d'évaluer, par différenciation, la qualité globale des prédictions.

Représentations statistiques : Les histogrammes des fréquences empiriques sont agrémentés de tableaux déclinant le nombre total de communes associées à chaque modalité : r_j (Saporta, 2006).

Analyse des confusions géographiques

Définition : les confusions géographiques effectuées sur les prédictions sont notées : $\varsigma^{\text{REG.j}}_{(U_k)}$. Elles sont évaluées par convention (Saporta, 2006), par :

$$\varsigma^{\text{REG.j}}_{(U_k)} = \begin{cases} \{1 \overset{\text{def}}{\triangleq} \text{CONFUSION}\} & \text{lorsque:} \quad \left(\hat{z}^{\text{REG.j}}_{(U_k)} \neq z^{\text{REG.j}}_{(U_k)}\right) \\ \{0 \overset{\text{def}}{\triangleq} \text{EQUIVALENCE}\} & \text{lorsque:} \quad \left(\hat{z}^{\text{REG.j}}_{(U_k)} = z^{\text{REG.j}}_{(U_k)}\right) \end{cases}$$

Principe : Lorsque les prédictions : $\hat{z}^{\text{REG.j}}_{(U_k)}$ sont de bonne qualité les indicateurs subséquents doivent tendre vers la valeur spécifiée:

Le Nombre Total des Confusions commises, sur chacune des modalités :

$$\mathrm{NTC}_{\mathrm{geo}}^{\mathrm{r}_j}\left(\hat{z}_{(U_k)}^{REG.j}\right) = \sum_{k=1}^{N(U_k)} \mathbb{1}_{\left\{\hat{z}_{(U_k)}^{REG.j} \neq z_{(U_k)}^{REG.j}\right\} \cap \left\{z_{(U_k)}^{REG.j} = \mathrm{r}_j\right\}} \to 0^+$$

La Proportion Relative de Confusions commises, sur chacune des modalités :

$$\mathrm{PRC}_{\mathrm{geo}}^{\mathrm{r}_j}\left(\hat{z}_{(U_k)}^{REG.j}\right) = \frac{1}{N\left(\left\{z'^{REG.j}_{(U_k)} = \mathrm{r}_j\right\}\right)} \cdot \sum_{k=1}^{N(U_k)} \mathbb{1}_{\left\{\hat{z}_{(U_k)}^{REG.j} \neq z_{(U_k)}^{REG.j}\right\} \cap \left\{z_{(U_k)}^{REG.j} = \mathrm{r}_j\right\}} \to 0^+$$

La Proportion Globale de Confusions commises, sur chacune des modalités :

$$\mathrm{PGC}_{\mathrm{geo}}^{\mathrm{r}_j}\left(\hat{z}_{(U_k)}^{REG.j}\right) = \frac{1}{N(U_k)} \cdot \sum_{k=1}^{N(U_k)} \mathbb{1}_{\left\{\hat{z}_{(U_k)}^{REG.j} \neq z_{(U_k)}^{REG.j}\right\} \cap \left\{z_{(U_k)}^{REG.j} = \mathrm{r}_j\right\}} \to 0^+$$

La Proportion Globale de Confusions commises sur l'ensemble des modalités :

$$\mathrm{PGC}_{\mathrm{geo}}\left(\hat{z}_{(U_k)}^{REG.j}\right) = \frac{1}{N(U_k)} \cdot \sum_{k=1}^{N(U_k)} \mathbb{1}_{\left\{\hat{z}_{(U_k)}^{REG.j} \neq z_{(U_k)}^{REG.j}\right\}} \to 0^+$$

Avec : $\mathrm{r_j} \in \mathrm{REG^j} = \{\text{PROBABLE; POSSIBLE; INDEMONTRABLE ; POSSIBLE}\}, \forall\, \mathrm{j} \in \{\text{CATA; THYR; TUM2}\}$

Représentation statistique graphique

Le tableau de synthèse des confusions simplifie la lecture de la *matrice des confusions* en déclinant les indicateurs statistiques : $\mathrm{NTC}_{\mathrm{geo}}^{\mathrm{r}_j}\left(\hat{z}_{(U_k)}^{REG.j}\right)$, $\mathrm{PRC}_{\mathrm{geo}}^{\mathrm{r}_j}\left(\hat{z}_{(U_k)}^{REG.j}\right)$ et $\mathrm{PGC}_{\mathrm{geo}}^{\mathrm{r}_j}\left(\hat{z}_{(U_k)}^{REG.j}\right)$, en fonction des r_j. Il est adapté aux variables multi-classes. *La matrice des confusions géographiques* décline pour chaque modalité r_j sur sa diagonale le nombre de $z_{(U_k)}^{REG.j}$ bien prédits. Le reste de la matrice dénombre les confusions et les modalités concernées.

Représentations cartographiques

Les modalités de $\varsigma_{(U_k)}^{REG.j}$ sont projetées dans les U_k - la variable est booléenne :

Code couleur :

$$\varsigma_{(U_k)}^{REG.j} = \begin{cases} \text{Couleur séquelle terne, hachurée de vert} & \text{lorsque: } \left\{\varsigma_{(U_k)}^{REG.j} = \text{EQUIVALENCE}\right\} \\ \text{Couleur séquelle terne, hachurée de rouge} & \text{lorsque: } \left\{\varsigma_{(U_k)}^{REG.j} = \text{CONFUSION}\right\} \end{cases}$$

Documentation :

Les noms des communes affichés font allégeance à la règle suivante :

$$\mathrm{label}_{(U_k)} = \begin{cases} \text{None} & \text{lorsque: } \left\{\varsigma_{(U_k)}^{REG.j} = \text{EQUIVALENCE}\right\} \\ \text{Display} & \text{lorsque: } \left\{\varsigma_{(U_k)}^{REG.j} = \text{CONFUSION}\right\} \end{cases}$$

Remarques

Les estimations statistiques, en particulier celles des $\hat{\varphi}_j^*$, ont été effectuées avec le logiciel R (Institute for Statistics and Mathematics, 1997). Les représentations graphiques avec Excel (Microsoft, 2013). Les illustrations cartographiques avec ArcGis.10 (ESRI, 2013).

Les cartographies sont présentées en petit format et uniquement dans les U_k de PACA et aux alentours, où 50% des individus de la cohorte LEA 2009 sont spatialisés (chapitre.1).

APPLICATION AUX SEQUELLES ET PRESENTATION DES RESULTATS

La stratégie de prédiction environnementale des REG a été appliquée conformément aux principes énoncés auparavant et pour chaque séquelle : CATA, THYR et TUM2. Les résultats statistiques et cartographiques obtenus sont présentés subséquemment.

SEQUELLE : CATARACTES

Risques d'Expositions Géographiques aux cataractes

Les indicateurs $\hat{z}_{(U_k)}^{REG.(CATA)}$ $\hat{z}_{(U_k)}^{REG.(CATA)}$ et $\varsigma_{(U_k)}^{REG.(CATA)}$ des illustrations cartographiques sont présentés pour les U_k, sises en région PACA et aux alentours, où des patients sont spatialisés.

Cartographie des Risques d'Expositions Géographiques pour la population LEA

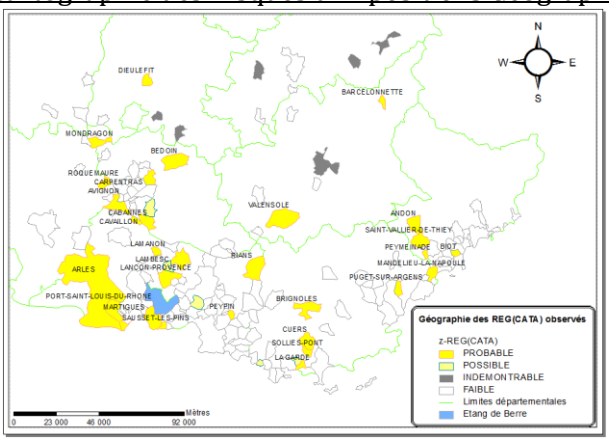


Figure 276 : Valeurs observées prises par $z_{(U_k)}^{REG.(CATA)}$

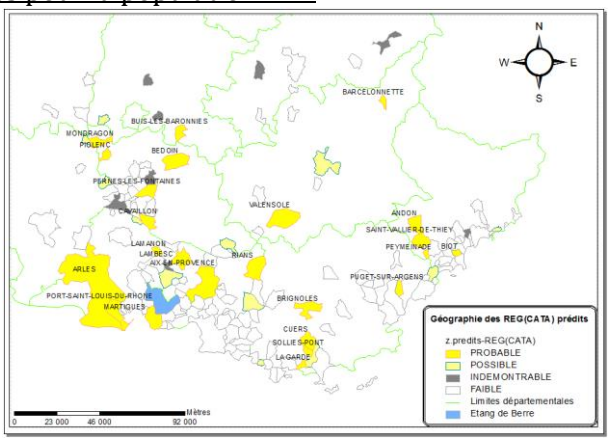
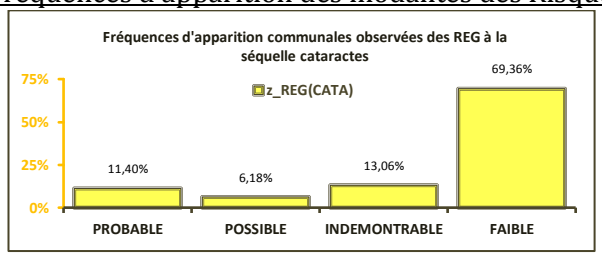


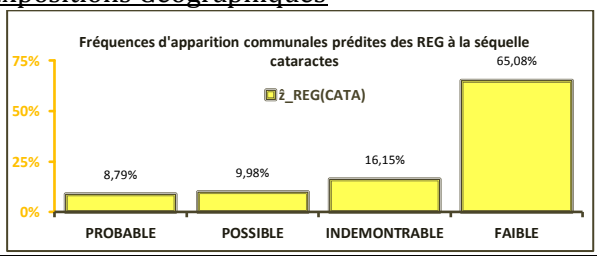
Figure 277 : Valeurs prédites prises par $\hat{z}_{(U_k)}^{REG.(CATA)}$

Fréquences d'apparition des modalités des Risques d'Expositions Géographiques



MODALITE	PROBABLE	POSSIBLE	INDEMONTRABLE	FAIBLE
Fréquences sur $\hat{z}_{(U_k)}^{REG.(CATA)}$	11,40%	6,18%	13,06%	69,36%
Nombre de communes	48	26	55	292

Figure 278 : Fréquences spatiales estimées sur les $z_{(U_k)}^{REG.(CATA)}$



MODALITE	PROBABLE	POSSIBLE	INDEMONTRABLE	FAIBLE
Fréquences sur $\hat{z}_{(U_k)}^{REG.(CATA)}$	8,79%	9,98%	16,15%	65,08%
Nombre de communes	37	42	68	274

Figure 279 : Fréquences spatiales estimées sur les $\hat{z}_{(U_k)}^{REG.(CATA)}$

Cartographie, graphique et synthèse statistique des confusions géographiques des REG

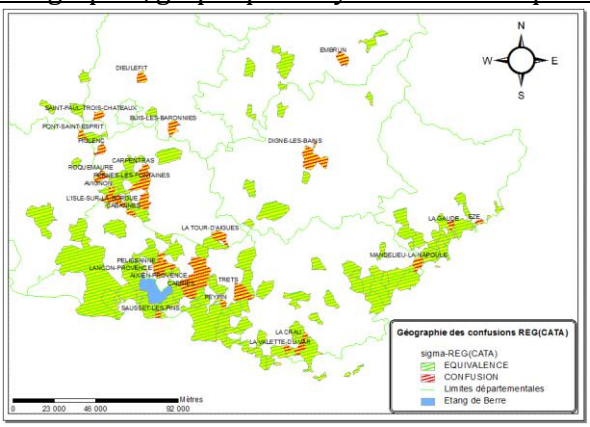
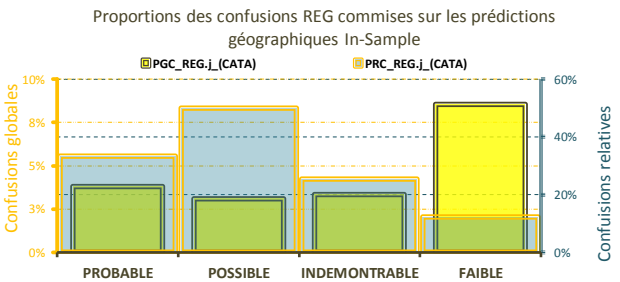


Figure 280 Valeurs prises par $\varsigma_{(U_k)}^{REG(CATA)}$ dans les U_k sises en région PACA et aux alentours



CATARACTES	PROBABLE	POSSIBLE	INDEMONTRABLE	FAIBLE	TOTAL
NTC_cj_géo_($\hat{z}_{(U_k)}^{REG.j}$)	16	13	14	36	79
PRC_rj_géo_($\hat{z}_{(U_k)}^{REG.j}$)	33,33%	50,00%	25,45%	12,33%	-
PGC_r.j_géo_($\hat{z}_{(U_k)}^{REG.j}$)	3,80%	3,09%	3,33%	8,55%	18,76%

Figure 281 : Synthèse et représentation graphique des $\varsigma_{(U_k)}^{REG(CATA)}$

SEQUELLES : TUMEURS THYROÏDIENNES

Risques d'Expositions Géographiques aux tumeurs thyroïdiennes

Les indicateurs $\hat{z}^{REG.(THYR)}_{(U_k)}$ $\hat{z}^{REG.(THYR)}_{(U_k)}$ et $\varsigma^{REG.(THYR)}_{(U_k)}$ des illustrations cartographiques sont présentés pour les U_k, de PACA et aux alentours, où des patients sont spatialisés.

Cartographie des Risques d'Expositions Géographiques pour la population LEA

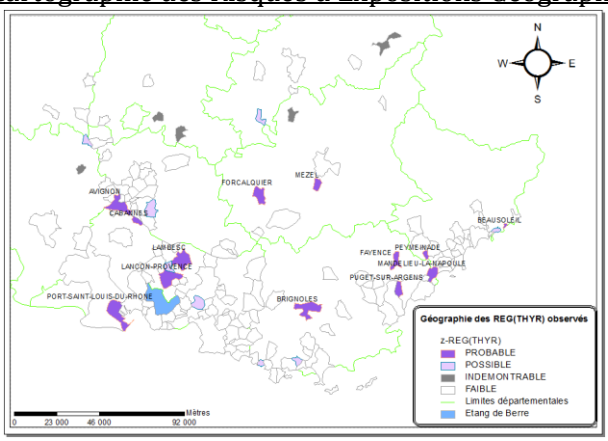


Figure 282 : Valeurs observées prises par $z^{REG.(THYR)}_{(U_k)}$

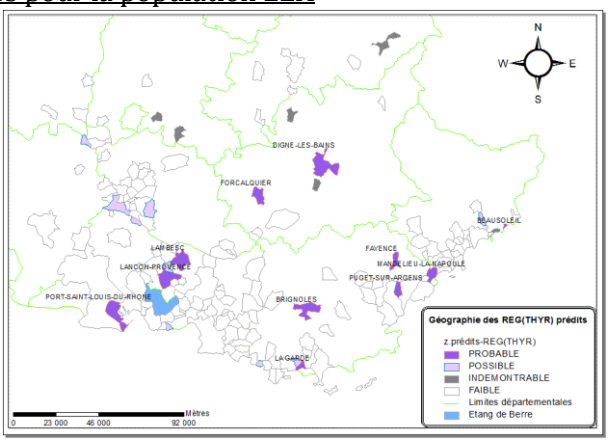
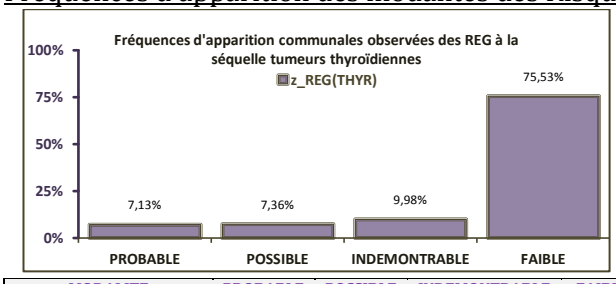


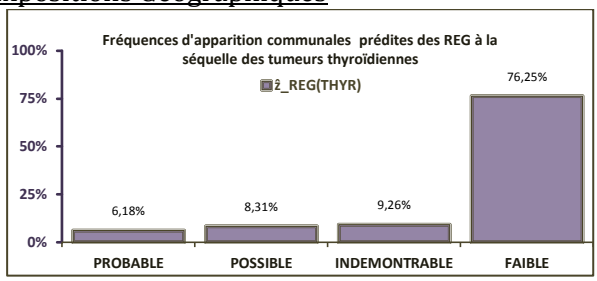
Figure 283 : Valeurs prédites prises par $\hat{z}^{REG.(THYR)}_{(U_k)}$

Fréquences d'apparition des modalités des Risques d'Expositions Géographiques



MODALITE	PROBABLE	POSSIBLE	INDEMONTRABLE	FAIBLE
Fréquences sur $z^{REG.(THYR)}_{(U_k)}$	7,13%	7,36%	9,98%	75,53%
Nombre de communes	30	31	42	318

Figure 284 : Fréquences spatiales estimées sur les $z^{REG.(THYR)}_{(U_k)}$



MODALITE	PROBABLE	POSSIBLE	INDEMONTRABLE	FAIBLE
Fréquences sur $\hat{z}^{REG.(THYR)}_{(U_k)}$	6,18%	8,31%	9,26%	76,25%
Nombre de communes	26	35	39	321

Figure 285 : Fréquences spatiales estimées sur les $\hat{z}^{REG.(THYR)}_{(U_k)}$

Cartographie, graphique et synthèse statistique des confusions géographiques des REG

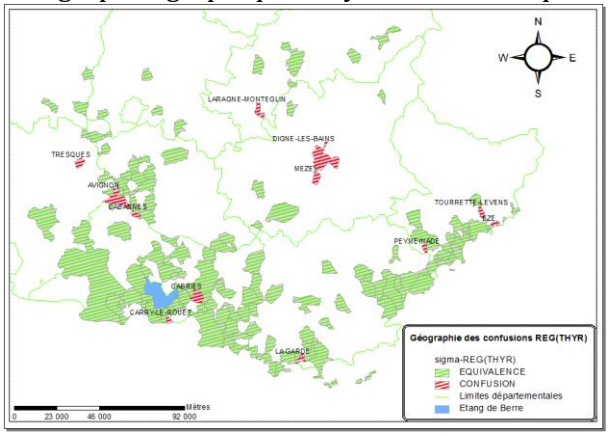
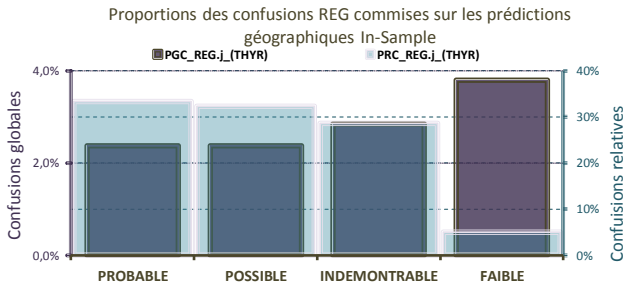


Figure 286 : Valeurs prises par $\varsigma^{REG(THYR)}_{(U_k)}$ dans les U_k de PACA et aux alentours



TUMEURS THYROÏDIENNES	PROBABLE	POSSIBLE	INDEMONTRABLE	FAIBLE	TOTAL
NTC_cj_géo_($\hat{z}^{REG.j}_{(U_k)}$)	10	10	12	16	48
PRC_rj_géo_($\hat{z}^{REG.j}_{(U_k)}$)	33,33%	23,26%	28,27%	5,03%	-
PGC_r.j_géo_($\hat{z}^{REG.j}_{(U_k)}$)	2,38%	2,38%	2,85%	3,80%	11,40%

Figure 287 : Synthèse et représentation graphique des $\varsigma^{REG(THYR)}_{(U_k)}$

SEQUELLES : TUMEURS SECONDAIRES

Risques d'Expositions Géographiques aux tumeurs secondaires - majeures

Les indicateurs $\hat{z}^{REG.(TUM2)}_{(U_k)}$ $\hat{z}^{REG.(TUM2)}_{(U_k)}$ et $\varsigma^{REG.(TUM2)}_{(U_k)}$ des illustrations cartographiques sont présentés pour les U_k, sises en région PACA et aux alentours, où des patients sont spatialisés.

Cartographie des Risques d'Expositions Géographiques pour la population LEA

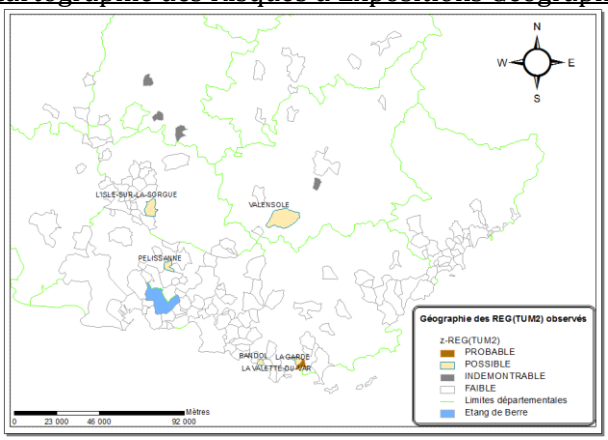


Figure 288 : Valeurs observées prises par $z^{REG.(TUM2)}_{(U_k)}$

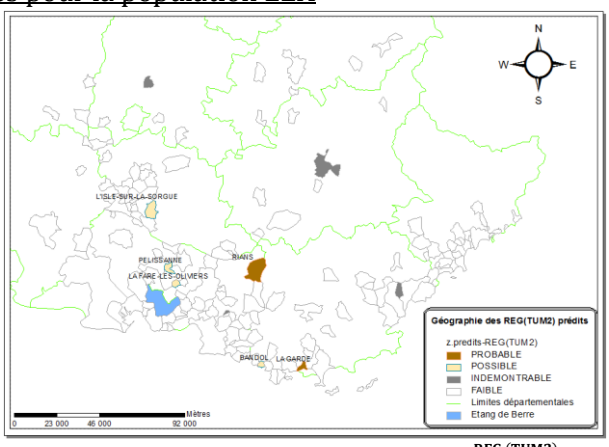
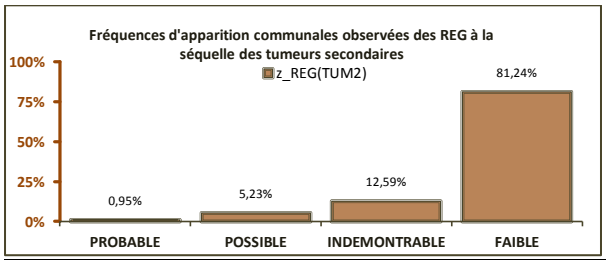


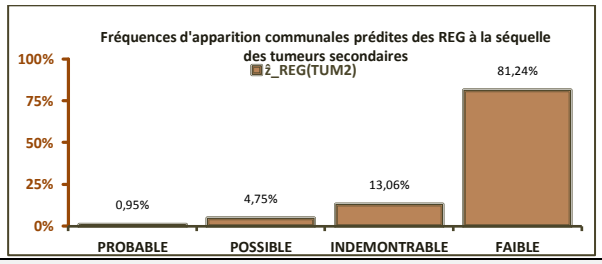
Figure 289 : Valeurs prédites prises par $\hat{z}^{REG.(TUM2)}_{(U_k)}$

Fréquences d'apparition des modalités des Risques d'Expositions Géographiques



MODALITE	PROBABLE	POSSIBLE	INDEMONTRABLE	FAIBLE
Fréquences sur $z^{REG.(TUM2)}_{(U_k)}$	0,95%	5,23%	12,59%	81,24%
Nombre de communes	4	22	53	342

Figure 290 : Fréquences spatiales estimées sur les $z^{REG.(TUM2)}_{(U_k)}$



MODALITE	PROBABLE	POSSIBLE	INDEMONTRABLE	FAIBLE
Fréquences sur $\hat{z}^{REG.(TUM2)}_{(U_k)}$	0,95%	4,75%	13,06%	81,24%
Nombre de communes	4	20	55	342

Figure 291 : Fréquences spatiales estimées sur les $\hat{z}^{REG.(TUM2)}_{(U_k)}$

Cartographie, graphique et synthèse statistique des confusions géographiques REG

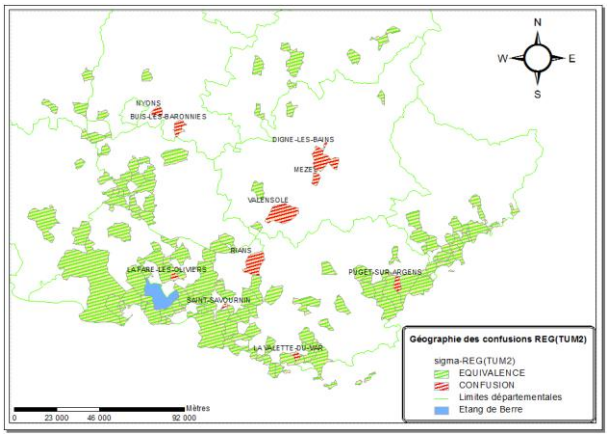
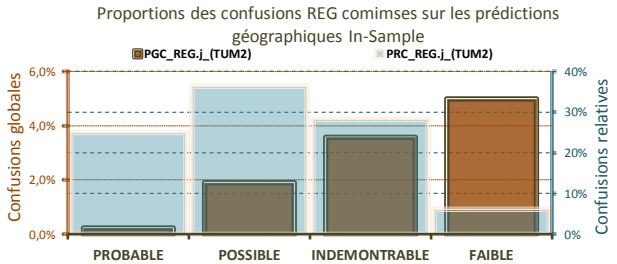


Figure 292 : Valeurs prises par $\varsigma^{REG(TUM2)}_{(U_k)}$ dans les U_k de PACA et aux alentours



TUMEURS SECONDAIRES	PROBABLE	POSSIBLE	INDEMONTRABLE	FAIBLE	TOTAL
NTC_cj_géo_$(\hat{z}^{REG.j}_{(U_k)})$	1	8	15	21	45
PRC_rj_géo_$(\hat{z}^{REG.j}_{(U_k)})$	25,0%	36,36%	28,3%	6,14%	-
PGC_r.j_géo_$(\hat{z}^{REG.j}_{(U_k)})$	0,24%	1,90%	3,56%	4,99%	10,69%

Figure 293 : Synthèse et représentation graphique des $\varsigma^{REG(TUM2)}_{(U_k)}$

ANALYSE ET REMARQUES

Descriptions numériques

CATA - la proportion globale, $PGC_{geo}^{r_j}(\cdot)$, de confusions sur $\hat{z}'^{REG(CATA)}_{(U_k)}$ s'élève à 18,76% soit 79 confusions. La proportion d'erreurs relatives sur les modalités, $PRC_{geo}^{r_j}(\cdot)$, est de : 33,33% sur PROBABLE soit 32 prédictions correctes sur 48 attendues, 50% sur POSSIBLE avec seulement 13 prédictions sur 26, 25,14% sur INDEMONTRABLE avec 41 prédictions correctes sur 55, 12,33% sur FAIBLE - i.e. seulement 256 prédictions correctes sur 292.

THYR - la $PGC_{geo}^{r_j}(\cdot)$ sur $\hat{z}'^{REG(THYR)}_{(U_k)}$ est de 11,40% avec 48 confusions. Les $PRC_{geo}^{r_j}(\cdot)$ sont de : 33,33% sur PROBABLE avec seulement 20 prédictions correctes sur 30, 32,26% sur POSSIBLE, 21 prédictions correctes sur 31, 28,57% sur INDEMONTRABLE avec tout de même 30 prédictions correctes sur 42, 5,03% FAIBLE - avec 302 prédictions correctes sur 318.

TUM2 - la $PGC_{geo}^{r_j}(\cdot)$ sur $\hat{z}'^{REG(TUM2)}_{(U_k)}$ s'élève à 10,70% soit 45 confusions. Les $PRC_{geo}^{r_j}(\cdot)$ sont de : 25,0% sur PROBABLE avec 3 prédictions correctes sur 4, 36,36% sur POSSIBLE soit 14 prédictions correctes sur 22, 28,30% sur INDEMONTRABLE avec seulement 38 prédictions correctes contre 53 attendues, 6,14% FAIBLE - avec près de 231 prédictions correctes sur 342

Analyse des confusions

CATA a le nombre de modalités PROBABLE et POSSIBLE le plus fort. Cette particularité vient du fait que la séquelle est précoce et que son incidence est élevée. Les confusions portent sur ces deux modalités. En revanche, INCERTAIN et FAIBLE sont assez bien prédites.

Les THYR ont un nombre de PROBABLE et POSSIBLE élevé à cause de leur incidence élevée sur la population spatialisée. Le cumul de ces modalités sur les observés et les prédits sont identiques et la majeure partie des confusions est faite avec FAIBLE.

Les TUM2 comptent seulement 4 PROBABLE, à cause de la latence et l'incidence rare de cette séquelle – raison pour laquelle cette modalité semble si bien prédite. La majorité des confusions s'opère entre FAIBLE et INDEMONTRABLE.

Qualité prédictive et remarques heuristiques

Les REG sont globalement bien prédits par les $\hat{z}_{(U_k)}^{REG.j}$. Les valeurs de $PGC_{geo}^{r_j}(\cdot)$ en sont la preuve :

Les REG : PROBABLE sont systématiquement sous-estimés. Ce constat s'illustre parfaitement sur les cartographies de $\hat{z}_{(U_k)}^{REG.j}$ et $z_{(U_k)}^{REG.j}$ présentées, pour toutes les séquelles, et était aussi nécessaire qu'évident puisque inhérent au processus d'estimation des $\hat{z}_{(U_k)}^{REG.j}$.

Les REG : POSSIBLE prédits sont généralement surestimés. Lorsque cette surestimation contrebalance la défaillance commise sur PROBABLE, alors les paramètres $\hat{\alpha}^j$ et surtout $\hat{\xi}_{géo}^j$ assurent parfaitement leur rôle et sont correctement spécifiés.

Les REG : INDEMONTRABLE sont très bien prédits, un temps sous-évalués, et curieusement de façon inversement proportionnelle à l'incertitude EpiGéoStat (chapitre2).

Les REG : FAIBLE sont, de fait, systématiquement sous-évalués et de façon plus ou moins forte selon la séquelle. Lorsque cette sous-estimation est effectuée au profit de la modalité INDEMONTRABLE cela n'a pas un impact fort sur l'interprétation épidémiologique des résultats. En revanche, lorsque les

confusions s'effectuent au profit de : POSSIBLE ou, *a fortiori, de* PROBABLE – ceci est préjudiciable pour la commune concernée puisque sa qualité en matière de santé environnementale* est engagée.

L'indicateur géographique du REG est conçu pour le management durable des espaces. En particulier les prédictions : $\hat{z}_{(U_k)}^{REG.j}$ qui caractérisent les Risques d'Expositions Géographiques morbides encourus par les populations locales, au regard des DES, FREC et FREPA auxquels elles sont assujetties.

Une mauvaise caractérisation des REG peut donc induire des craintes sociales et la mise en place de mesures politiques inadaptées aux besoins réels des populations. L'impact sanitaire des erreurs commises par les $\hat{z}_{(U_k)}^{REG.j}$ est développé en conclusion.

L'ampleur des confusions est liée à la nature du PM* d'intérêt. Il convient donc de hiérarchiser la qualité prédictive des $\hat{z}'^{REG.j}_{(U_k)}$. Pour cela le plus simple est de confondre simultanément, sur les histogrammes, les niveaux $PRC_{geo}^{r_j}(\cdot)$ et le caractère préjudiciable des confusions modales. Dans un premier temps il s'agit de comparer simultanément les niveaux de confusion de PROBABLE et POSSIBLE avec ceux de INDEMONTRABLE et FAIBLE qui sont systématiquement plus petits.
Seules, les TUM2 ont un niveau de $PRC_{geo}^{r_j}(\cdot)$ PROBABLE inférieur à INDEMONTRABLE. Les TUM2 sont donc les séquelles les mieux prédites. Dans un second temps il suffit de comparer les niveaux $PRC_{geo}^{r_j}(\cdot)$ PROBABLE avec POSSIBLE. Les THYR ont un niveau de PROBABLE supérieur à celui de POSSIBLE, alors que pour les CATA c'est l'inverse - d'ailleurs son niveau de PROBABLE est à peine supérieur à celui d'INDEMONTRABLE. La séquelle dont les REG sont, de loin, les moins bien prédits, est THYR.

REMARQUES GENERALES

L'algorithme MVG permet de caractériser les interactions statistiques entre la géographie de l'environnement, modélisée par les i.st.e* $x_{(U_k)}^l$, et celle d'états de santé, modélisée par les i.st.m* $z'^{j}_{Uk,c}$ et $z'^{j}_{Uk,q}$. En introduisant, en amont de MVG, un *boosting** (Han et Kamber, 2006) et en focalisant le processus d'apprentissage sur les patients dont les x_i^l décrivent avec le plus de certitude l'environnement géographique de leur lieu de vie - un second algorithme BVMG a été créé. Ce dernier a permis d'identifier des x_i^l explicatifs, par une approche individus-centrée*, plus complexe car directement menée sur les séquelles y_i^l, et par suite, de valider la pertinence des DES, des FREC et des FREPA spécifiés dans le cadre de l'approche géographique. MVG permet aussi de prédire, avec une grande acuité, la géographie des états de santé par les i.st.m* : $\hat{z}'^{j}_{Uk,c}$ et $\hat{z}'^{j}_{Uk,q}$. Mais ces indicateurs ne sont pas adaptés à la gestion des espaces.

Une stratégie de fusion permettant de les combiner en $\hat{z}_{Uk}^{REG.j}$ a été proposée. Cet indicateur caractérise les Risques d'Expositions Géographiques (REG) morbides au regard des caractéristiques environnementales des espaces. Les $\hat{z}_{Uk}^{REG.j}$ sont adaptés *au management durable des territoires*, par la réduction des risques d'expositions aux DES, FREC et FREPA à partir de mesures *individuelles*, i.e. médicales, ou *collectives*, i.e. politiques. La caractérisation des REG morbides est décrite selon quatre modalités : PROBABLE, POSSIBLE, FAIBLE, et INDEMONTRABLE.

L'interprétation des REG prédits est à considérer avec circonspection. En effet, si la stratégie d'estimation des $\hat{z}_{Uk}^{REG.j}$ est statistiquement fiable, les rares confusions s'avèrent très préjudiciables. En particulier lorsqu'un REG PROBABLE est prédit à la place d'un FAIBLE, les répercussions sociales, économiques ou sanitaires peuvent être lourdes de conséquences. Sur le plan médical, des traitements

préventifs aux effets indésirables, ou des changements d'habitudes inutiles, peuvent altérer la Qualité de Vie des patients. Sur le plan politique, des mesures visant à améliorer la qualité des tissus sanitaires*, ou à réguler les activités industrielles polluantes, peuvent engendrer des dépenses conséquentes avec des conséquences sur les autres secteurs socio-économiques et induire des phénomènes de paupérisation des populations locales – et de ce fait des risques de prédispositions géographiques aux phénomènes morbides (Mankiw, 1998).

Les algorithmes MVG et BMVG, ainsi que la stratégie de prédiction environnementale des REG, sont applicables et reproductibles à tous les états de santé étudiés en géographie. Cependant, malgré leur robustesse statistique, l'approche reste entachée d'incertitudes spatiotemporelles protéiformes. Le caractère *déterministe des interactions santé-environnement* ne peut donc pas atteindre *le niveau de rigueur requis en épidémiologie* (Hill, 1965). D'autant plus qu'il existe aussi un risque quant à la fiabilité et à l'interprétation des indicateurs géographiques utilisés pour identifier les DES, FREC et FREPA et caractériser les REG.

SYNTHESE DU CHAPITRE 4

L'objectif était de proposer une méthode adaptée à l'identification de Déterminants Environnementaux de Santé* (DES) et de Facteurs de Risques Environnementaux Contributifs* (FREC) ou de Facteurs de Risques Environnementaux Probablement Aggravants* (FREPA) liés à des états de santé particuliers et modélisés dans des *espaces géographiques* à partir de jeux de données multidimensionnelles.

L'analyse des *interactions combinées* entre des *Facteurs Environnementaux* (FE) multidimensionnels* et des états de santé est un problème complexe qui tend à se généraliser en géographie de la santé - un domaine où les méthodes utilisées sont indigentes pour y répondre. L'adaptation de Variable Selection Using RandomForest* (VSURF) à la dialectique géographique a engendré la création de l'algorithme MyVsurfGéo* (MVG).
Ses capacités explicatives et prédictives ont été validées par une approche individus-centrée*, du fait de l'impossibilité de travailler sur des échantillons de test. Cette approche épidémiologique de la problématique a augmenté la complexité des jeux de données (section.A). De fait, un processus boosting* - géographique - a été implémenté et a engendré un second algorithme nommé BoostMyVsurfGéo* (BMVG) (section.B).

MVG est un instrument d'analyse statistique adapté à la complexité spatiale des problématiques étudiées en géographie de la santé. Il permet : d'analyser les caractéristiques statistiques des jeux de données géographiques, de calibrer les Forêts Aléatoires* (FA) afin d'estimer de façon robuste les interactions santé environnement, de sélectionner un panel d'i.st.e* $x^{l}_{(U_k)}$ contenant une quantité d'informations suffisante pour expliquer la variabilité spatiale des PM* étudiés - modélisé par les i.st.m* : $z^{j}_{(U_k),c}$ et $z^{j}_{(U_k),q}$, et d'identifier un paquet de $x^{l}_{(U_k)}$ permettant de prédire les états de santé (section.B).

Tous les FE/FIM* mobilisés par *des* $x^{l}_{(U_k)}$ *explicatifs* ne sont pas éligibles au rang de DES. *Leurs scores randomForest* obtenus en régression* et l'analyse conjointe des similitudes entre les $x^{l}_{(U_k)}$ explicatifs ou prédictifs menée simultanément sur les i.st.m* $z^{j}_{(U_k),c}$ et $z^{j}_{(U_k),q}$, permettent d'identifier – à partir de *la grille de lecture MVG** - les DES, des FREC et des FREPA.

La géographie des états de santé étudiés peut être prédite, avec une grande acuité, par MVG à partir des i.st.m* : $\hat{z}'^{j}_{(U_k),c}$ et $\hat{z}'^{j}_{(U_k),q}$, et uniquement au regard des $x^{l}_{(U_k)}$ qui modélisent la géographie des DES, des FREC et des FREPA (section.B).
Cependant ces i.st.e* contiennent des incertitudes spatiotemporelles et sont difficilement interprétables. Une stratégie de fusion a été proposée afin d'augmenter leur intelligibilité en discrétisant les espaces par une typologie de Risques d'Expositions Géographiques (REG) aux phénomènes morbides. La robustesse de cet indicateur composite : $\hat{z}^{REG.j}_{(U_k)}$ a été testée et validée.
La géographie de la santé ne saurait se restreindre à des problématiques de modélisation. Elle a aussi pour devoir de caractériser les territoires, par des indicateurs géographiques adaptés à la gestion durable des espaces. C'est justement le rôle qu'assurent les $\hat{z}^{REG.j}_{(U_k)}$. Ils donnent aux praticiens de santé et aux politiques les moyens des outils d'aide à la mise en place de mesures préventives, individuelles ou collectives, afin de limiter les risques d'expositions aux FE/FIM* prédisposant aux phénomènes morbides, et en même temps d'évaluer, par le biais de scenarii, les coûts de ces mesures. Mais l'analyse des REG morbides prédits à partir de la géographie des DES, FREC et FREPA est à considérer avec circonspection. En effet, les erreurs de caractérisation des REG peuvent avoir des conséquences socio-économiques lourdes, notamment lorsqu'un REG PROBABLE est prédit pour un FAIBLE (section.C).

L'analyse conjointe des résultats de MVG et de BMVG sur les i.st.m* de types qualitatif et quantitatif, représentatifs de la géographie des CATA, des THYR et des TUM2 et sur les variables séquelles LEA, à partir des i.st.e* représentatifs de la géographie des FE/FIM* intégrés, a permis d'identifier des :

DES* systématiques, en l'occurrence les FIM* et en particulier, la géographie des traitements agressifs qui ont nécessité des irradiations corporelles totales ou des greffes, ainsi que le rôle explicatif du temps modélisé par la géographie des durées du suivi des patients. L'interprétation de ce dernier est ambivalente puisqu'elle suggère à la fois les effets secondaires, au long cours, des traitements et les effets combinés d'expositions à des Facteurs Environnementaux* (FE) délétères présents dans les lieux de vie des patients.

FREC mis en évidence de façon récurrente. Les FE-SAN* modélisés par les Distances Temporelles bruitées d'Accès aux plateaux techniques des établissements de santé et l'Accès Potentiel Localisé* aux professionnels libéraux. Les FE-PHY-CHIM spécifiés par les niveaux géographiques d'exposition à la radioactivité d'origine artificielle. Ainsi que, l'effet *direct des* paramètres météorologiques, ou *indirect* - comme principal vecteur de diffusion des radionucléides dans *les milieux environnementaux* ou *les milieux de contact*.

FREPA modélisant des prédispositions spatiales morbides contextuelles, induites par des FE-SOCIO.ECO. En l'occurrence, la géographie *de la défaveur sociale - i.e. des conjonctures de* pauvreté, des défaillances du système culturel, des catégories socioprofessionnelles défavorisées et de faibles revenus fiscaux. Ou encore, certains effets négatifs de l'urbanisation et du mode de développement économique comme l'insécurité, le stress d'origine sociale, ou des diffusions dans les compartiments environnementaux de substances toxiques (i.e. :HAP, MES, particules de diesel, métalloïdes, pesticides, Benzène, Formaldéhyde, Bitumes...) - liées à la spécialisation des espaces dans les domaines de l'industrie ou l'agriculture.

CONCLUSION DU CHAPITRE 4

La théorie des *FA couplée à la stratégie VSURF* et adaptée à la dialectique géographique par les algorithmes MVG et BMVG a permis d'identifier, avec une grande acuité, des DES, des FREC et des FREPA, permettant d'expliquer et prédire les états de santé étudiés, et par extension, de valider la seconde hypothèse fondamentale de cette thèse.

Cette dernière suppose, qu'en disposant de méthodes et d'outils, de modélisation et d'analyse, suffisamment puissants pour représenter la géographie des tous les Facteurs Environnementaux* (FE), médicaux et individuels potentiellement dangereux, il est possible - en dépit de nombreuses incertitudes induites par des processus aléatoires - d'expliquer, prédire et améliorer la santé environnementale* d'une population (chapitre.1).

La capacité de VSURF à caractériser les interactions entre les indicateurs spatiotemporels environnementaux* et morbides proposés montre que les *propositions heuristiques* émises, dans le cadre d'une approche *globale et interdisciplinaire des problématiques de santé,* sont *adaptées à la complexité des jeux de données* désormais utilisés *en géographie de la santé* (Morin, 1977).

Les mathématiques modernes en général, et plus spécifiquement la sphère des statistiques et des probabilités, offrent à la géographie des instruments manifestement pertinents*. En particulier, les Machine Learning* fondées sur des méthodes d'ensemble non paramétriques et randomisées constituent *un formidable moyen pour dégager de la gangue des données, le pur diamant de la véridique nature* (Benzécri, 1985).

Malgré cela, en géographie de la santé, cette *approche méthodologique* de la complexité est encore rare malgré les perspectives promises (Gatrell, 2005).

Attention pourtant à l'interprétation des résultats. Les indicateurs proposés ne sont pas forcément en adéquation avec la réalité géographique, ni avec les attentes sociales. De plus, les mathématiques sont subjectives et l'interprétation des DES* comme des *leviers d'accès à une bonne santé environnementale** peut s'avérer spécieuse…

CONCLUSION DE LA PARTIE II

Chapitre 3. Les méthodes de modélisation proposées sont basées sur le concept de *minimisation du biais conditionnel*.* Elles ont permis d'estimer des i.st.e* robustes $x^{l}_{(U_k)}$ représentant, à l'échelle des communes U_k, la géographie des FE/FIM* *pertinents* et curieux** intégrables.
Ces méthodes ont été élaborées de sorte que les $x^{l}_{(U_k)}$ caractérisent *l'environnement géographique des lieux de vie* avec des *distances a-spatialses morbides* harmonieuses, afin de pouvoir analyser rigoureusement leurs effets sur les états de santé étudiés (chapitre.3).

Les PM* d'intérêt – séquelles - sont modélisés par deux i.st.m. Le premier est quantitatif $z'^{j}_{(U_k),c}$, et s'attache à la variabilité spatiale des séquelles. Le second est qualitatif $z'^{j}_{(U_k),q}$ et renseigne sur la propension à développer la séquelle et sur la qualité spatiotemporelle des informations utilisées. La fusion de ces deux i.st.m* permet d'en générer un troisième $z^{REG.j}_{(U_k)}$ voué à caractériser les espaces par des Risques d'Expositions Géographiques (REG) morbides (chapitre.2).

Chapitre 4. La théorie des Forêts Aléatoires* (FA) couplée à la stratégie *VSURF* a été adaptée à la dialectique géographique, ce qui a conduit à la cristallisation de l'algorithme MyVSurfGéo (MVG).
Le processus heuristique permet d'identifier des Déterminants Environnementaux de Santé* (DES), des Facteurs de Risques Environnementaux Contributifs* (FREC) et des Facteurs de Risques Environnementaux Probablement Aggravants* (FREPA) des états de santé étudiés – séquelles. L'algorithme MVG explique et prédit, selon le contexte, les interactions entre les i.st.e* $x^{l}_{(U_k)}$ et les i.st.m* $z'^{j}_{(U_k),c}$, $z'^{j}_{(U_k),q}$. La robustesse des résultats a été validée statistiquement à partir de *prédictions géographiques*, puis par une *approche individus-centrée** – i.e. directement menée sur les variables épidémiologiques. Une procédure de *Boosting** a été implémentée en amont de MVG pour l'adapter à la complexité de l'analyse épidémiologique, ce qui a engendré une nouvelle méthode et l'algorithme : BootsMyVsurfGéo* (BMVG).
Corollairement, la caractérisation des espaces en fonction d'un REG peut être prédite par l'i.st.m* $\hat{z}^{REG.j}_{(U_k)}$ et uniquement à l'aide des i.st.e* $x^{l}_{(U_k)}$ modélisant la géographie des DES, FREC et FREPA *explicatifs* (chapitre.4).

Intérêt : l'interface SIG permet de ramener le raisonnement à l'espace, par le biais de d'instruments cartographiques, afin d'anticiper et de réduire les REG morbides, individuels ou collectifs, aux DES, aux FREC et aux FREPA. Ainsi, les politiques disposent d'outils opérationnels adaptés pour concevoir des mesures de santé publique, dont l'efficience est estimable sur les $x^{l}_{(U_k)}$.
Différents scenarii peuvent être modélisés afin de s'assurer que les coûts financiers engendrés aient bien les effets espérés en matière de réduction des REG, et par ailleurs, que les mesures prises soient en adéquation avec les attentes des populations locales. L'amélioration du *système de santé environnementale** menée sous l'égide du *contrôle social* doit permettre de tendre vers *le développement durable des territoires* (Salem, 1995).

En somme, l'analyse statistique multidimensionnelle des interactions santé environnement, menée conjointement par MVG et BMVG, permet d'identifier des leviers environnementaux opérationnels visant à réduire les REG morbides. Du moins, les résultats expérimentaux obtenus sur les états de santé étudiés – séquelles cataractes, tumeurs thyroïdiennes et tumeurs secondaires majeures - offrent en ce sens des perspectives prometteuses.

Cependant les *leviers environnementaux* mis en évidence sont entachés d'incertitudes, engendrées par des *biais heuristiques protéiformes,* qui constituent les limites de cette recherche. Ils sont déclinés dans la conclusion générale et des solutions pour y remédier sont aussi proposées. Dans la mesure où cette partie s'attache à des considérations méthodologiques, le biais associé convient d'être énoncé.

Limite induite par le biais de l'approche méthodologique retenue

La stratégie de sélection VSURF, sur laquelle sont fondées MVG et BMVG, est statistiquement fiable et parfaitement adaptée à l'identification géographique de FE/FIM* liés à des PM. De surcroit, *des résultats convaincants similaires ont été rapportés notamment dans le champ de la santé*. Cependant, il existe d'autres stratégies de sélection très puissantes et qui conduisent parfois à des conclusions complétement différentes, *et à l'heure actuelle, dans la pratique, il est impossible de distinguer laquelle s'approche le plus de la réalité* (Genuer, 2010).

En effet, des approches mathématiques différentes peuvent conduire à des solutions toutes autres pour un même problème donné. La complexité des théories mathématiques dissimule finalement une subjectivité aussi grande que celle que l'on confère systématiquement aux SHS. La spécification *a priori* des *coefficients experts* : ς_{ψ}^{j} illustre parfaitement cette remarque, et il en est de même pour les seuils VSURF : $\psi_{.}^{j}$, qui sont finalement très arbitraires lorsqu'on regarde attentivement la façon dont ils sont conçus (annexe 8).

En dépit de leur objectivité apparente, les méthodes mathématiques ne sont finalement que des points de vue particuliers, *qui ne consistent jamais qu'à découvrir un ordre et des symétries nouvelles, en copiant des symétries observées, puis à les imposer au monde qui nous entoure*. Or, les résultats obtenus sont intimement liés à la théorie – subjective et limitée - sur laquelle ces méthodes se fondent et, par conséquent, *ne doivent pas être trop hâtivement érigés en vérité* (Cartier, Villani et al., 2012).

CONCLUSION GENERALE

Eléments fondamentaux

Cette thèse propose une dialectique et un ensemble de méthodes pour modéliser, analyser et caractériser, dans des *espaces géographiques numériques*, les interactions spatiales entre l'état de santé d'une population et son environnement.

Ces méthodes sont appliquées à des séquelles : cataractes, tumeurs thyroïdiennes, tumeurs secondaires majeures, développées après le traitement d'une leucémie dans l'enfance – Cohorte LEA.

Chap.1. L'approche bibliographique a permis de construire une dialectique adaptée à la géographie de la santé. Les causes des maladies sont supposées multifactorielles et l'environnement géographique est examiné dans toute sa plénitude. Ce dernier est représenté par des *objets spatiaux* caractérisant des *coprésences* - ou *agrégations spatiotemporelles* - (Levy J., Lussault M., 2003) *de situations à risques ou des substances nocives combinées* (Leux et Guénel, 2010).

Ces expositions géographiques sont caractérisées par des Facteurs Environnementaux* (FE) *jugés pertinents** (Salem, 1995) à connotation sanitaire (FE-SAN), socio-économique (FE-SOCIO.ECO) et physicochimique (FE-PHY.CIM) - la géographie des Caractéristiques Individuelles et Médicales*(FIM) des populations est également prise en compte – sans quoi *l'analyse est biaisée* (Abramson J.H., Abramson Z.H., 1988). L'échelle des communes $\mathbf{U_k}$ a été retenue sur un compromis de considérations théoriques et phénoménologiques (i.e. liées à la granularité* des données disponibles et à l'intégration d'une dimension socio-politique).

Chap.2. La géographie des Phénomènes Morbides* (PM) est modélisée par deux i.st.m* qui prennent en compte la qualité spatiale EpiGéoStat des données et le temps d'exposition à l'environnement.

Le premier est quantitatif $z'^{j}_{(U_k),c}$, le second qualitatif $z'^{j}_{(U_k),q}$. Leur fusion caractérise les Risques d'Expositions Géographiques (REG) morbides par $z^{REG.j}_{(U_k)}$. Ce dernier est adapté à la gestion durable des territoires.

Chap.3. La géographie des FE/FIM* est représentée par des i.st.e* : $x^{l}_{(U_k)}$. Leur biais conditionnel* a été minimisé et leurs *distances a-spatiales morbides** ont été harmonisées, autant que faire se peut.

Les i.st.e* modélises la variabilité spatiotemporelle des expositions environnementales à des FE/FIM* *pertinents* et curieux**.

Chap.4. La sélection statistique des $x^{l}_{(U_k)}$ permettant d'expliquer et de prédire les i.st.e* : $z'^{j}_{(U_k),c}$, $z'^{j}_{(U_k),q}$. Elle a engendré deux algorithmes fondés sur la méthode VSURF (Genuer, Poggi et al., 2013).

Le premier, MVG*, identifie les effets des FE/FIM* sur la géographie des PM* étudiés. *Une grille de lecture* permet de les hiérarchiser en : Déterminants Environnementaux de Santé* (DES), Facteurs de Risques : Environnementaux Contributifs* (FREC) ou Potentiellement Aggravants* (FREPA). Et à partir des caractéristiques environnementales, de prédire les REG morbides qu'ils suggèrent par les i.stm $\hat{z}^{REG.j}_{(U_k)}$.

Les résultats obtenus ont été validés par le second algorithme : BMVG, dans le cadre d'une approche individus-centrée*.

Objectifs atteints et résultats obtenus

En géographie de la santé.

(i) Proposer des méthodes de modélisation et de sélection adaptées aux jeux de données géographiques contemporains.

(ii) Identifier les FE* géographiques qui conditionnent des PM* particuliers, grâce à une dialectique opérationnelle et reproductible à toutes les maladies.

(iii) Construire des indicateurs géographiques permettant de caractériser les REG morbides et de déterminer *des leviers environnementaux* permettant de les réduire.

En santé publique.
L'interface SIG permet de ramener la proposition heuristique (mathématique) à l'espace et d'offrir des moyens opérationnels représentatifs permettant d'analyser, d'anticiper et de réduire les REG morbides, individuels ou collectifs, liés aux DES, aux FREC et aux FREPA.

A l'échelle des individus. Les praticiens de santé peuvent proposer à leurs patients des solutions médicales ou comportementales préventives adaptées aux REG auxquels ils sont soumis du simple fait de leur situation géographique (Kenneth et Sander, 1998).

A l'échelle communautaire. Les politiques peuvent imaginer des mesures de santé publique visant à diminuer les expositions aux DES, aux FREC ou aux FREPA. Ils peuvent aussi estimer à partir de scenarii, l'efficience espérée de ces mesures sur les REG morbides, tout en s'assurant que les coûts engendrés soient en adéquation avec les besoins sanitaires et les attentes sociales des populations locales (Mankiw, 1998).

En épidémiologie : l'application aux données de la cohorte LEA a permis d'ouvrir des perspectives de recherche prometteuses. Les résultats obtenus corroborent *certains soupçons* déjà mis en exergue par les épidémiologistes et les géographes de la santé.

Projet LEA – résultats. La géographie des séquelles développées après le traitement d'une LA s'explique d'abord systématiquement par celle des Caractéristiques Individuelles et Médicales* (CIM) des patients (Michel, Auquier et al., 2007) – les DES. Et ensuite, par les effets combinés de l'accès géographique aux soins lié à des *tissus sanitaires** *défaillants* (Powell, 1995) et de l'exposition à la radioactivité environnementale et à des paramètres météorologiques particuliers (Leux et Guénel, 2010) - les FREC. Elle est potentiellement amplifiée par *des contextes territoriaux de défaveur sociale* (Rey, Jougla et al., 2009) – les FREPA.

Avant d'énoncer les limites et les solutions envisagées - pour y remédier, il convient de faire un résumé intelligible des contributions effectives et théoriques de cette thèse.

Les apports heuristiques en géographie de la santé des méthodes proposées sont synthétisés pour chacune des étapes du processus dialectique, dans la partie gauche du schéma synoptique 2.
La partie située à droite, de ce schéma, présente les retombées possibles de l'ensemble des contributions méthodologiques énoncées, dans le domaine de la santé. Ces apports théoriques devraient permettre d'améliorer la santé environnementale des sociétés humaines. Ils interviennent à quatre niveaux socio-spatiaux différents.

Schéma synoptique 294 : Les apports de la recherche

Les limites heuristiques

L'identification des DES, FREC et FREPA permet de réduire les REG morbides en jouant sur des *leviers environnementaux*. Mais encore faut-il que ces derniers soient en adéquation avec la réalité géographique et les attentes sociales qui pèsent sur les structures environnementales.

Limites théoriques

La modélisation, par définition, est une simplification de la réalité. La dialectique géographique est à la fois spatiale et temporelle ce qui, par nature, complexifie l'approche des problématiques de santé publique. La géographie *des états et des faits de santé* varie dans le temps, et comme *l'espace est rarement homogène*, ces phénomènes présentent des formes spatiales - *discontinues et anisotropes* (Salem, 1995). *Cette complexité qui caractérise les systèmes spatiaux* est constituée d'éléments qui interagissent entre eux et dont les formes *semblent a priori complètement chaotiques.* Les interactions spatiotemporelles déterministes et stochastiques entre ces éléments *s'emmêlent et contribuent tantôt à leur stabilité, tantôt à leur évolution*. La caractéristique fondamentale des systèmes complexes est l'émergence de régularités spatiales à des échelles macroscopiques (Pumain, 2004).

En dépit du fait que l'échelle d'investigation retenue et les méthodes de modélisation et de sélection proposées soient adaptées à l'exploration de la complexité spatiale et temporelle, la dialectique et les résultats présentés sont empreints d'incertitudes protéiformes. Une approche interdisciplinaire est nécessaire pour traiter cette fameuse complexité. Les *mathématiques* constituent un bon *angle d'attaque*, mais *ne sont jamais assez sollicitées par les géographes de la santé* (Gatrell, 2005), à l'instar de la *réflexion épidémiologique*, qui est souvent *introduite de façon partielle* alors qu'elle devrait intervenir depuis la spécification des objectifs jusqu'à l'interprétation des résultats (Morin, 1977).

Limites pragmatiques et perspectives de recherche

Afin de réduire les incertitudes inhérentes à la complexité spatiotemporelle, les regards des géographes, des mathématiciens et des épidémiologistes ont été croisés. Ils ont permis de mettre en évidence de nombreux biais heuristiques. Ces derniers sont déclinés et des solutions pour les éluder ou les atténuer sont proposées.

Biais d'interprétation.

L'identification des DES, FREC et FREPA pose le problème de la significativité et de l'interprétation des $x^{l}_{(U_k)}$. En effet, il est impossible de savoir si les indicateurs spatiotemporels sont bien représentatifs de la réalité géographique des phénomènes pour lesquels ils ont été proposés (Brook, Lohr et al., 1984).

Concernant les DES :* Les résultats obtenus confèrent une puissance explicative statistique particulièrement forte à l'i.st.e* : $x'^{\mathrm{DSUIVI}}_{(\mathrm{U_k})}$. Mais l'interprétation épidémiologique de ce FIM* est ambigüe. En effet, il modélise à la fois la géographie des effets secondaires au long cours des traitements, des séquelles mieux diagnostiquées - donc mieux traitées – et, des durées d'exposition élevées à un environnement nocif, sans pour autant qu'il soit possible de dissocier la part de chacun. Solution proposée : Supprimer cet i.st.e* dans l'approche individus-centrée*, et le remplacer par le temps d'exposition à l'environnement et par un autre i.st.e* intégrant conjointement la durée du suivi et la qualité des soins reçus, i.e. le nombre de fois que le patient a été vu, sur cette période. Cette proposition est inadaptée à l'approche géographique puisque le temps d'exposition à l'environnement est utilisé dans l'estimation de : $z'^{j}_{(\mathrm{U_k}),\mathrm{c}}$ et $z'^{j}_{(\mathrm{U_k}),\mathrm{q}}$ (Auquier, 2013b).

Concernant les FREC : une ambigüité analogue est soulevée pour l'i.st.e* $x^{\mathrm{NJAP}}_{(\mathrm{U_k})}$. S'agit-il d'une exposition géographique à des conditions climatologiques particulières ou une exposition indirecte à la radioactivité environnementale ? La pluie étant le principal vecteur de propagation des radionucléides dans l'environnement. Cet écueil concerne aussi tous les $x'^{\mathrm{l:SAN}}_{(\mathrm{U_k})}$ puisqu'ils sont redondants à tel point

qu'ils ne modélisent plus l'accès à un *item sanitaire* particulier mais la qualité globale des tissus sanitaires* territoriaux (chapitre.1) ; (chapitre.3). Solution proposée : Substituer ou intégrer des indicateurs modélisant différemment le FE* en question et étudier leurs capacités explicatives, i.e. introduire les niveaux géographiques pluviométriques pour valider l'effet de $x_{(U_k)}^{NJAP}$.

Concernant les FREPA : les i.st.e* composites modélisant des prédispositions géographiques morbides comme la *défaveur sociale* : $x_{(U_k)}^{FDep.}$, ou *a fortiori* $x_{(U_k)}^{URIN.}$ et $x_{(U_k)}^{PEST.}$ qui suggèrent, sans aucune distinction possible, à la fois l'exposition à des substances nocives et des effets économiques positifs et qui de fait, ne permettent pas d'identifier des *leviers environnementaux particuliers* (chapitre.1) ; (chapitre.3). Solution proposée : modéliser, dans la mesure du possible, tous les FE* qui les composent et les soumettre, indépendamment de tous les autres FE/FIM, à MVG et BMVG afin de déterminer ceux dont les interactions statistiques sont les plus fortes

Biais de surreprésentation.

La géographie de certains FE* est surreprésentée, ce qui a manifestement un impact sur leur éligibilité au rang de FREC ou FREPA. En l'occurrence, les expositions à la radioactivité sont modélisées par 12 i.st.e* dont 9 sont construits sur des mesures RNM. De plus, certains sont auto-corrélés puisqu'ils ont été obtenus par cokrigeage. Le même constat peut être étendu aux FE-SAN* représentés par 14 i.st.e, parmi lesquels 9 sont corrélés et 4 sont inconsistants (chapitre.1) (chapitre.3). Solution proposée : ces i.st.e* modélisent soit l'exposition globale à des radionucléides artificiels émis par les INB, ou à des tissus sanitaires* défaillants. Puisqu'ils sont corrélés, il peut être intéressant de les agréger par une méthode topologique, comme pour FDep., afin d'évaluer l'efficience statistique des mesures réglementaires ou d'aménagement.

Biais de non représentation.

L'analyse spatiale des interactions santé environnement nécessite de considérer l'environnement dans toute ses dimensions multiples. Or certains FE/FIM* pertinents* n'ont pas pu être modélisés faute de pouvoir accéder aux données. C'est le cas de la géographie des : facteurs génétiques, expositions urbaines aux RNI et aux Champs Electromagnétiques Extrêmement Basses Fréquences (CEM EBF), conduites à risques liées aux habitudes alimentaires régionalisées (chapitre.1), et bien d'autres encore… Solution proposée : lorsque l'accès aux données fait défaut, il y a deux possibilités. La première résulte du fait que la base SIG est en cours de construction et dans ce cas, la seule solution est l'attente. La seconde touche à des données mobilisables mais les partenariats sont trop longs à se mettre en place, ou alors les données ne sont pas fiables ou soumises au secret statistique et ne peuvent donc pas être utilisées.

Biais de représentation partielle.

Ce biais est analogue au précédent. La géographie des FE-SAN* n'a pu être modélisée que partiellement, puisque les indicateurs APL n'étaient disponibles que pour les ophtalmologues et les généralistes et que les DTA aux praticiens libéraux sont inconsistantes - excepté pour les radiologues. Aussi, sur le plan temporel les DTA, les APL ainsi que les AVR de l'Atlas Radon n'étaient disponibles qu'à une seule temporalité (chapitre.1) ; (chapitre.1). Solution proposée : s'agissant de l'obtention des APL et des DTA pour tous les items sanitaires* et à différentes dates, des conventions sont en cours d'élaboration avec les ARS. En revanche, pour les AVR Radon, l'IRSN ne semble pas vouloir participer à ce projet.

Biais de représentation grossière.

Certains FE-SOCIO.ECO* particulièrement pertinents* ont été intégrés de façon grossière. En l'occurrence, la qualité des politiques de durabilité et leurs répercussions morbides sont modélisées par : $x_{(U_k)}^{tx.MORT}$; $x_{(U_k)}^{txAccNAT}$; $x_{(U_k)}^{txAccPOP}$ $x_{(U_k)}^{tx.BAC}$. Mais *les niveaux géographiques d'espérance de vie* ou *les proportions d'individus vivant en-dessous du seuil de pauvreté* auraient été plus adaptés (chapitre.1). Solution proposée : Aucune, le problème est insoluble. Ces variables INSEE sont couvertes par le secret statistique à l'échelle des communes et sont inopérantes à l'échelle des régions à laquelle elles sont déclinées.

Biais des modélisations géographiques morbides.

Sur le sujet le Professeur Auquier est formel. *Les résultats obtenus ne sont pas utilisables car la réflexion et les objectifs épidémiologiques ne sont pas correctement spécifiés*. Les applications faites sur les données de la Cohorte LEA sont donc à considérer *comme des exercices de style* qui ont permis de montrer *l'intérêt, la robustesse et la pertinence des propositions heuristiques* de cette thèse. *Des perspectives de recherche prometteuses sont envisageables* - et même envisagées. Solution proposée : le seul moyen d'éluder le problème d'inconsistance* statistique, qui engendre trop d'incertitudes dans le cadre de l'approche géographique, *est de travailler sur le cumul temporel de l'ensemble des séquelles développées par les patients*. Perspective professionnelle : Un contrat postdoc est envisagé dans l'EA-3279

Biais de validation des résultats.

La validation des DES, des FREC et des FREPA identifiés est nécessairement statistique. Mais elle a été effectuée In-Sample faute de temps - le financement de cette thèse touchant à sa fin. La stratégie de validation *individus-centrée** est astucieuse et pertinente d'un point de vue épidémiologique. Cependant, elle a nécessité la confection de l'algorithme MVGV, dont la fiabilité et la puissance restent à prouver. Car si MVG peut être assimilée à la version aboutie VSURF – disponible actuellement en béta – BMVG est radicalement différente, donc incertaine… Solution proposée : valider les résultats en appliquant MVG et BMVG *Out-Of-Sample* (OOS), i.e. sur un autre échantillon de la population LEA, par exemple sur les patients inclus entre 2009 et 2011.

Biais méthodologique d'identification

A priori la stratégie de sélection VSURF, sur laquelle sont fondées MVG et BMVG, est *fiable* dans la pratique. Toutefois, d'autres méthodes de sélection, toutes aussi puissantes, peuvent conduire à des conclusions complétement différentes – à partir de jeux de données pourtant identiques. Au fond, les mathématiques sont *subjectives et les résultats obtenus ne doivent pas être érigés en vérité de façon précipitée*. Il ne s'agit jamais que de points de vue tous aussi intéressants les uns que les autres (chapitre.4) ; (Conclusion – Partie.II). Solution proposée : appliquer d'autres méthodes de sélection aux jeux de données utilisés. Puis comparer les résultats obtenus en analysant l'importance des variables explicatives qui se confondent et qui dissemblent. Ainsi des DES, des FREC et des FREPA plus consistants pourraient être identifiés.

Perspectives opérationnelles théoriques et Réflexion épistémologique.
Les solutions énoncées visent à débiaiser, à désupréposer, la dialectique en vue d'identifier des *leviers environnementaux opérationnels* permettant de réduire efficacement les REG morbides aux vrais DES, FREC ou aux FREPA. Et par conséquent, de garantir l'accès des populations locales à une bonne santé environnementale*.

Cependant l'approche des problématiques de santé publique est maculée par les incertitudes de la complexité géographique. En conséquence de quoi, *l'exposition aux risques, dans l'espace et le temps, est nécessairement exprimée sous la forme d'une relation possible entre le milieu et l'homme* (Bailly et Beguin, 2005). Quant à l'approche épidémiologique, individus-centrée*, elle supporte d'autres incertitudes (chapitre.4) mais reste plus rigoureuse et plus sûre. Toutefois, *des relations causales* sont difficiles – voire impossibles - à établir (Hill, 1965).
D'autant plus que les instruments de datamining* utilisés sont les mêmes. Et bien que par hypothèse, la puissance statistique soit garantie, au fond rien n'est certain – *les Forêts de la réalité sont Aléatoires* (chapitre.4). Par conséquent l'identification des *vrais* DES, FREC et FREPA est impossible. Heureusement, *la recherche de la vérité est à la fois difficile et facile, [car si] nul ne peut l'atteindre absolument, [nul ne peut] la manquer tout à fait* (Aristote, 1862)

La vérité est inaccessible mais *la quantité d'absolu qu'elle contient peut néanmoins être mise au service de la vie* (Nietzsche, 1972). Augmenter cette quantité d'absolu est le devoir du scientifique. *Il s'agit de mieux connaître pour mieux vivre.* Que tout *soit incertain* n'est pas une raison suffisante pour renoncer, il faut *progresser désespérément* (Comte-Sponville, 2000).

La complexité n'est pas, non plus une difficulté suffisante pour cesser de chercher, non pas tant *la vérité,* mais une vérité suffisamment proche de la réalité pour améliorer la santé environnementale* – et dont les *résonnances sociales* augmentent le niveau global de Qualité de Vie des populations (Wallace R., Wallace D. et al., 1999). En effet, *une population en bonne santé* est une population *plus active*, *plus impliquée dans les activités économiques, sociales, sanitaires et environnementales* – créant ainsi des dynamiques *propices au développement durable* (Dumont. Gérard-François, 2004).

En somme, *la santé environnementale* des populations* ne doit *pas être appréhendée par les sociétés humaines comme un coût mais comme un investissement* (Bailly, 2000). La conception de stratégies opérationnelles vouées à l'émergence de territoires durables est l'objet de la *nouvelle géographie*. Celle-ci a pour but d'introduire des méthodes quantitatives et rigoureuses d'analyse. *La médico-métrie géographique* a pour objet de construire des indicateurs spatiaux qui, interfacés dans un SIG, permettent une approche spatiale interdisciplinaire et l'intégration de dimensions : épidémiologique, économique, sanitaire, géographique (i.e. en aménagement), sociale, psychologique… en vue de planifier des politiques de santé publique efficaces (Bailly, 1981).
L'efficience des mesures de santé environnementale* relève donc d'un art de gouverner particulier qui doit intégrer une démarche réflexive associant géographie, santé, et territoire - i.e. qui ne saurait se passer de la compréhension des processus socio-spatiaux. Et, il appartient au géographe de plaider pour injecter de la conscience géographique en santé publique (Amat-Roze, 2011).

Afin de réduire les incertitudes liées à la complexité d'une approche spatiotemporelle, l'interdisciplinarité est nécessaire. Dans cette thèse le regard du géographe a été croisé avec celui du statisticien et de l'épidémiologiste. Mais d'autres points de vue doivent être intégrés.
En l'occurrence, celui des économistes qui a pour but, à partir d'un compromis bénéfice/risque, de rationaliser le coût d'activation *des leviers environnementaux*. Il s'agit de s'assurer que les gains relatifs de richesses (économiques ou pas) soient supérieurs aux dépenses. Les mesures *somptuaires sont associées à des risques élevés de paupérisation* – affectant les populations les plus vulnérables – et créant

des *contextes de prédispositions géographiques morbides* et des *effets domino négatifs protéiformes* difficilement estimables (Mankiw, 1998).

Un point de vue tout autre, celui des sociologues, présente aussi un intérêt majeur en matière de rationalisation des politiques de santé publique. Il consiste à s'assurer que les mesures imaginées sont adaptées non seulement aux besoins, i.e. aux états de santé, mais aussi aux attentes des populations, i.e. à leur désir d'améliorer certaines structures environnementales plutôt que d'autres. Les attentes touchent à la notion de valeur que la population octroie *au fait de santé* – le levier environnemental - sur lequel on s'apprête à jouer. La santé doit être intégrée transversalement dans toutes les thématiques territoriales de sorte que le *contrôle politique* de la santé environnementale* reste sous le joug *du contrôle social,* pour garantir une dynamique territoriale de *durabilité* (Salem, 1995).

En d'autres termes, la réduction des REG menée sur un DES* peut s'avérer stratégiquement moins rationnelle que celle menée sur un FREC ou un FREPA. La logique interdisciplinaire permet d'optimiser les actions, individuelles ou collectives, visant à améliorer la santé environnementale* des populations dans des espaces donnés. En somme, les mesures politiques doivent être en adéquation avec les besoins (i.e. les états de santé observés) et les attentes (i.e. la façon dont les populations locales perçoivent la Qualité de Vie environnementale), pour garantir - *ou plutôt dirais-je, pour augmenter les chances de parvenir à* – une gestion durable des territoires.

Quels que soient les progrès des connaissances humaines, il y aura toujours place pour l'ignorance et par suite pour le hasard et la probabilité (Borel, 1928).

BIBLIOGRAPHIE

Abramson J.H, Abramson Z.H. (1988). Making Sense of Data: A Self-Instruction Manual on the Interpretation of Epidemiological Data. Oxford: Oxford University Press.

ADEME. (2012). *Polluants de l'air extérieur | BULDAIR*. Adresse Internet: www.ademe.fr, Consulté le 19 juin 2012.

AECB. (1991). *Childhood leukemia around Canadian nuclear facilities - Phase II Final report.* Ottawa, Canada: Ontario Cancer Traitment and Reaserch Foundation, Ottawa, Canada.

Afsset. (2009a). *Cancer et Environnement.* Paris: Avis de l'Agence française de sécurité sanitaire de l'environnement et du travail (Afsset). *www.afsset.fr*. Maisons-Alfort: afssef.

Afsset. (2009b). *Propositions de Valeurs Guides de qualité d'Air Intérieur.* Paris: ANSES.

Ahola K., Honkonen T., Kivimäki M., Virtanen M., Isometsä E, Aromaa A., Lönnqvist J. (2006). Contribution of burnout to the association between job strain and depression: the health 2000 study. *Journal of Occup Environ Med, 48*, pp. 1023-1030.

Akerstedt T. (2006). Psychosocial stress and impaired sleep. *Scandinavian Journal of work, Environment & Health*, (32), 493-501.

Amat-Roze J-M (2011). La santé, une construction interdisciplinaire. L'exemple du dialogue Géographie-Santé-Territoire. ARSI, Recherche en soins infirmiers, 3(106), pp. 5-15.

ANFR. (2012). *ANFR Agence Nationale des Fréquences*. Adresse Internet: http://www.cartoradio.fr/netenmap.php?cmd=zoomfull. Consulté le 15 mai 2012.

ANS, IRSN. (2013). Réseau National de Mesures de la radioactivité de l'environnement (RNM). Adresse Internet: http://www.mesure-radioactivite.fr/public/spip.php?page=carte. Consulté le 04 mai 2013.

Anses. (2012). *Agence nationale de sécurité sanitaire, de l'alimentation, de l'environnement et du travail*. Adresse Internet www.anses.fr: http://www.anses.fr/. Consulté le 05 juillet 2012.

ArcGis: Analysis toolbox. (2013). *ArcGIS Ressource Center*. Adresse Internet: http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/An_overview_of_the_Analysis_toolbox/000800000002000000/. Consulté en 2013.

ArcGIS: Iterator. (2013). *An overview of the Iterator toolset.* Adresse Internet: http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/An_overview_of_the_Iterator_toolset/00400000000n000000/. Consulté en 2013.

ArcGIS: ModelBuilder. (2013). *Desktop Help 10.0 - What is ModelBuilder?* Adresse Internet: http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/What_is_ModelBuilder. Consulté en 2013.

Aristote (1862). *Physique d'Aristote, Livre V.* Paris: GF Flammarion.

Arnaud M., Emery X. (2000). Estimation et interpolation spatiale. *Hermes Science*, 20. Cachan: Lavoisier.

ARS. (2012). Agences Régionales de Santé (ARS), Adresse Internet: *C@rtoSanté*. Site:Carto.ars.sante.fr/: Carto.ars.sante.fr/. Consulté en 2012.

ASN. (2010). Autorité de Sûreté Nucléaire (ASN). La surveillance de la radioactivité de l'environnement. *Contrôle : La revue de l'ASN*,188, 156p.

ASN. (2011). Autorité de Sûreté Nucléaire (ASN). *Décision n° 2011-DC-0204 de l'Autorité de sûreté nucléaire du 4 janvier 2011 établissant la liste des installations nucléaires de base au 31 décembre*

2010. Adresse Internet: http://www.asn.fr/index.php/Les-actions-de-l-ASN/La-reglementation/Bulletin-officiel-de-l-ASN/Décisions-de-l-ASN/Décision-n-2011-DC-0204-de-l-ASN-du-4-janvier-2011. Consulté le 23 février 2013.

ASN. (2013). Autorité de Sûreté Nucléaire (ASN). Adresse Internet: www.asn.fr: http://www.asn.fr/. Consulté en 2012-2013

Auquier P. (2010) Entretien au sujet des données de la Cohorte LEA. Marseille: Laboratoire de Santé Publique de la Timone, le 10 novembre 2010

Auquier P. (2012). Entretien au sujet *la qualité de vie*. Marseille, Bouches du Rhône, France: Laboratoire de Santé publique de la Timone. Marseille: Laboratoire de Santé Publique de la Timone, le 04 mai 2012.

Auquier P., Michel G. (2012) Mise au point d'une stratégie de pondération épidémiologique adaptée à la spatialisation de la base LEA. *Réunion d''avancement de la thèse* . Marseille: Laboratoire de Santé Publique de la Timone, le 07 octobre 2012.

Auquier P. (2013a) Entretien au sujet de l'accès aux soins et des protocoles de traitement prescrits. Marseille: Laboratoire de Santé Publique de la Timone, le 03 janvier 2013

Auquier, P. (2013b). Entretien au sujet de l'interprétation épidémiologique de la géographie des durées du suivi médical. Marseille: Laboratoire de Santé Publique de la Timone, le 15 août 2013

Baillargeon S. (2005). *Le krigeage : revue de la théorie et application à l'interpolation spatiale de données de précipitations.* Québec: Université de Laval.

Bailly A. (1981). *La géographie du bien-être* (éd. Presses Universitaires de France - PUF). Paris: Broché.

Bailly A. (2000). *Développement Social Durable Des Villes - Principes Et Pratiques.* Paris: Economica.

Bailly A., Beguin H. (2005). *Introduction à la géographie humaine* (éd. 8). Paris: Armand Colin.

Baraud Y., Giraud C., Huet S. (2009). *Annals of Statistics : Guaussian model selection with an unknown variance.* Nice: Peter Bühlmann.

Barlet M., Coldefy M., Collin C., Lucas-Gabrielli V. (2012). L'accessibilité Potentielle Localisée (APL) une nouvelle mesure de l'accessibilité aux médecins généralistes libéraux. *Question d'économie de la Santé*. Paris: Institut de recherche et documentation en économie de la santé (Irdes).

Barnett S., Roderick P., Martin D., Diamond I., Wrigley H. (2002). Interrelations between three proxies of health care need at the small area level: an urban/rural comparison. *Journal of Epidemiol Community Health, 56*(10), pp. 754-761.

Baysson H., Billon S., Catelinois O., Jean-Gambard P., Laurier D., Rogel A., Tirmarche M. (2004). Exposition de la population française à la radioactivité environnementale d'origine naturelle. *Environnement, Risques et Santé, 3*(6).

Ben Ishak A., Ghattas B. (2005). An efficient method for variable selection using svm-based criteria. Marseille: *Pré-publication de l'institut de Mathématiques de Luminy*.

Benach J., Yasui Y. (1999). Geographical patterns of excess mortality in Spain explained by two indices of deprivation. *Epidemiol Community Health,* 53, pp. 423-431.

Benzécri J-P. (1985). *Analyse des données* (éd. 3e). Paris: Economica.

Bernard P-M., Lapointe C. (2003). *Mesures statistiques en épidémiologie.* Quebec: Presses de l'Université du Québec.

Bernard S., Heutte L., Adam S. (2008). Etude de l'influence des paramètres sur les performances des Forêts Aléatoires. *Acte du dixième Colloque International Francophone sur l'Ecrit et le Document*. Rouen, France: Université de Rouen, LITIS EA 4108.

Bernoulli J. (1713). *Ars conjectandi, opus posthumum, Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilae reticularis.* Bâle: Thurneysen Brothers.

Biau G, Devroye L, Lugosi. G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research 9* : pp. 2039-2057.

Blanpain N., Chardon O. (2008). *Projections de population à l'horizon 2060.* INSEE, division: Enquêtes et études démographiques. Paris: INSEE.

Bloch L., Ricordeau P. (1996). *La régulation du système de santé en France 11*. Paris: Revue Française d'Economie.

Blondeau P. (2009). Cataracte. Adresse Internet http://www.passeportsante.net/fr/Maux/Problemes/Fiche.aspx?doc=cataracte_pm. Consulté le 01 Avril 2011.

Borel E. (1928). *Le hasard.* Paris: Alcan.

Borel E. (1967). *Probabilité et la vie.* Paris: Presses Universitaires de France.

Brauer V.F., Below H., Kramer A., Furthrer D., Paschke R. (2006). The role of thiocyanate in etiology of goiter in an industrial metropolitan area. *Eur J Endocrinol, 154*, pp. 229-235.

Braverman LE., He X., Pino S., Cross M., Magnani B., Lamm SH., and al. (2005). The effect of perchlorate, thiochlorate, and nitrate on thyroide function in workers exposed to perchlorate long-term. *J Clin Endocrinol Metab, 90*, pp. 700-706.

Breiman L. (1996). Bagging predictors. *Machine Learning*. New-York, U.S.A: 123 - 140.

Breiman L. (2001). Random Forests. *Machine Learning, 45*(1), Bristol, Royaume-Uni: P.A. Flach.

Breiman L. (2004). Manual - Setting Up, Using, And Understanding Random Forests V4.0. 15: ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using_random_forests_v4.0.pdf.

Breiman L., Cutler A. (2005). Random Forests, Berkeley. Available from : http ://www.stat.berkeley.edu/users/breiman/RandomForests/.

Breiman L., Friedman J., Olshen R. and Stone C. (1984). Classification And Regression Trees. *Belmont CA*. New-York, U.S.A: Wadsworth Center.

Brook RH., Lohr KN., Chassin M., Kosecoff J., Fink A., Solomon D. (1984). Geographic variations in the use of services: do they have any clinical significance? *Health Affairs, 3*, pp. 64-73.

Brown P. (1997). Popular epidemiology revisited. *Current Sociology, 45*(3), pp. 137-156.

Brucker-Davis F. (1998). Effects of environmental synthetic chemical on thyroïd function. *Thyroid, 1*, 827-856.

Brunet R. (1968). *Les phénomènes de discontinuité en géographie.* Paris: CNRS, Mémoire et Documents du Centre de Recherche et Documentation cartographiques et géographiques.

Brunet R, Ferras R, et Théry H. (2009). *Les mots de la géographie, Dictionnaire critique* (éd. 3ème, 1er ed: 1995). (L. D. Française, Éd.) Montpellier-Paris: Reclus.

Carretier J., Luporsi E. (2011). *CAV 2011: Cancer et environnement etat actuel des connaissances. Adresse Internet:* http://www.canal-u.tv/producteurs/canal_u_medecine/dossier_programmes/cancerologie/colloque_et_evenement/centre_alexis_vautrin/cancer_et_environnement/cav_2011_cancer_et_environnement_etat_actuel_des_connaissances. Consulté le 02 janvier 2012.

Carstairs V, Morris R. (1989). Deprivation: explaining differences in mortality between Scotland and England and Wales. *Bmj, 299*(6704), pp. 886-889.

Cartier P., Dhombres J., Heinzmann G., Villani C. (2012). *Mathématiques en liberté: Liberté, réalité, responsaibilité.* (L. v. brûle, Éd.) Broché.

Caudeville J. (2011). *Développement d'une plateforme intégrée pour la cartographie de l'exposition des populations aux substances chimiques ; Construction d'indicateurs spatialisés en vue d'identifier les inégalités environnementales à l'échelle régionale.* Thèse soutenur à Compiègne: Université de technologie Compiègne.

Caudeville J., Boudet C., Denys S., Bonnard R., Govaerty G., Cicolella A. (2012). Caractérisation des inégalités environnementales en Picardie fondée sur l'utilisation couplée d'un modèle multimédia et d'un système d'information géographique. (J. Libbey, Éd.) *Environnement Risques Santé, 10*(6), pp. 485-494.

Caudeville J., Bonnard R., Boudet C., Denys S., Govaert G., Cicolella A. (2012). Developement of spatial stochastic multimedia exposure model to assess population exposire at a region scale. (Elsevier, Éd.) *Science of the Total Environment, 432*, pp. 297-308.

CGDD, SOeS. (2009). CORINE Land Cover France. Commissariat Général au Développement Durable, *Guide d'utilisation*. Paris: Document Technique. pp. 1-22

Chaix B., Merlo J., Chauvin P. (2005). Comparaison of a spatial approach with the multilevel approach for investigating place effects on health: the example of healthcare utilisation in France. *Journal of Epidemiology and Community Health, 59*(6), pp. 517-526.

Charre J. (1995). *Statistique et territoire.* Statistique et territoire. GIP-RECLUS.

Chasles V., Fervers B. (2011). Expositions environnementales et cancer : risques perçus, risques réels. *Espace population et société*,1, pp. 125-136.

Chernick M. R. (1999). *Boostrap methods: a practitioner's guide.* New-York: Wiley.

Cleveland W., Devlin S. (1988). Locally weighted regression : An approach to regression analysis by locating. *Journal of the American Statistical Association*, *83*(403): pp. 596-610.

Coldefy M., Com-Ruelle L., Lucas-Gabrielli V. (2011). Distance et temps d'accès aux soins en France Métropolitaine. *Questions d'économie de la Santé, 164*(8),. Paris: IRDES.

Colle C., Adam C., Garnier-Laplace J., Roussel-Debet S., Beaugelin-Seiller K., Germain P. (2005). *Césium 137 et environnement.* Service d'étude du comportement des radionucléides dans les écosystèmes. Paris: IRSN.

Commission des Communautés Européennes (2003). *Stratégie Européenne en matière d'environnement et de Santé.* Bruxelles: Commission des Communautés Européennes.

Commission des Communautés Européennes (2007). *Ensemble pour la santé: une approche stratégique pour l'EU 2008-2013.* Bruxelles: Commission des Communautés Européennes.

Comte-Sponville A. (2000). *Présentation de la philosophie.* Paris: Albin Michel.

Couet C. (2006). La mobilité résidentielle des adultes : existe-t-il des "parcours type" ? Dans *portrait social*. Paris: INSEE. pp. 159-179.

Cressie N. A. C. (1993). Statistics for spatial data. *Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics*. New York: John Wiley & Sons Inc. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.

Cressie N. A. C. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology 17*, pp. 653-702.

Dauphiné A.., Voiron-Canicio C. (1988). *Variogrammes et structures spatiales.* Montpellier: Reclus.

Davis J-C. (1986). *Statistis and Data Analysis in Geology 2nd ed.* New-York: Jhon Wiley & Sons.

Dejour-Salamancad D., Gomes-Do-Espirito-Santo M.E., Chappert J-L., Garcia S., Creteur X., Isnard H. (2005). Investigation d'un signalement d'agrégats de cancers de l'enfant à Saint-Cyr-l'École. *Bulletin Épidémiologique Hebdomadaire, 49*(50).

Deluzarche C. (2012). *Les aliments les plus consommés par départements*. Adresse Internet: http://www.journaldunet.com/economie/distribution/consommation-alimentaire-en-france.shtml. Consulté le 01 juin 2012.

DGS / Bureau de la qualité des eaux. (2008). *Bilan de la qulité de l'eau au robinet du consommateur vis-à-vis des pesticides.* Paris: Direction Générale de la Santé.

DILA. (2002). Secrétariat général du gouvernement et la Direction de l'information légale et Administrative (DILA). Légifrance : Le service public de l'accès au droit. Adresse Internet : http://www.legifrance.gouv.fr/. Consulté le 17 mai 2012.

Draper G et al. (2005). Childhood cancer in relation to distance from hight voltage power lines in England and Wales: : a case-control study. *Binj*, 330 : 1290.

DREES. (2012). *Les distances d'accès aux soins en France métropolitaine au 1er janvier 2007*. Adresse Internet: http://www.drees.sante.gouv.fr/les-distances-d-acces-aux-soins-en-france-metropolitaine-au-1er-janvier-2007,9026.html. Consulté en 2012-2013.

DREES. (2013). *L'accessibilité Potentielle Localisée 2010*. Adresse Internet: http://www.google.fr/#gs_rn=18&gs_ri=psy-ab&cp=16&gs_id=1q&xhr=t&q=ressources+textuelles&es_nrs=true&pf=p&output=search&sclient=psy-ab&oq=ressources+tecxt&gs_l=&pbx=1&bav=on.2,or.r_qf.&bvm=bv.48572450,d.d2k&fp=d78fb448b3696df&biw=1600&bih=775, Consulté le 06 juin 2013.

Dubois D., Prade H. (1994). La fusion d'informations imprécises. *Traitement du signal, 11*(6), pp. 447-458.

Dubois D., Prade H. (2004). On the use of aggregation operations in information fusion processes. (ELSEVIER, Éd.) *Fuzzy Sets and Systems*, 142, pp. 144-161.

Dumond G-F. (2004). *Les Populations du monde (deuxième édition).* Paris: Armand Colin.

Durand-Dastès F., Mutin G. (1995). *Géographie Universelle.* Paris: Reclus.

Efron B., Tibshirani R. (1997). Improvements on cross-validation : The. 632+ bootstrap method. *Journal of the American Statistical Association*, 438(92), pp. 548-560.

Elliott P., Briggs D., Morris S, et al. (2001). Risk of adverse birth outcomes in populations living near landfill sites. *Br Med J., 323*, pp. 363-368.

ESRI. (2013). ArcGis Ressource Center: *Desktop help 10.0. Adresse Internet:* http://help.arcgis.com/en/arcgisdesktop/10.0/help/. Consulté entre 2011-2013,

Fisher R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2), pp. 179-188.

Fotheringham A., Brunsdon C., Charlton M. (2002). *Geographically Weighted Regression : the analysis of spatially varying relationships.* Chichester: John Wiley & Sons Ltd.

Fromageot, A. Coppieters Y., Parent F., Lagasse R. (2005). Epidémiologie et Géographie: une interdisciplinarité à développer pour l'analyse des relations entre santé et environnement. *Environnement Risques Santé, 4*(6), pp. 395-403.

Furtos J. (2007). Les effets cliniques de la souffrance psychologique d'origine sociale. (S. et Scociété, Éd.) *Mental'idées*, 11, pp. 24-33.

Gaitan E. (1983). Endemic goiter in western Colombia. *Ecol Dis, 2*, pp. 295-3008.

Garry V.F., Danzl T.J., Tarone J., Griffith J., Cervenka J., Krueger L., and al. (1992). Chromosome rearrangements in fumigant appliers: possible relationship to non-hodgkin's lymphoma risk. *Cancer Epidemiol Biomarkers Prev, 1*, pp. 287-291.

Gatrell A.. (2005). Complexity theory and geographies of health: a critical assessment. *Social Science & Medicine*, 60, pp. 2661-2671.

Gaudart J. (2007, Novembre 20). Analyse spatio-temporelle et modélisation des épidémies : application au paludisme à P. falciparum. *Thèse : Ecole doctorale mathématique et Informatique, E.D. 184*. Marseille, Bouches-du-Rhône, France: Université de la Méditéranée, faculté d'Aix-Marseille.

Gay J-R., Korre A. (2006). A spatially-evaluated methodology for assessing risk to a population from contaminated land. *Environmental Pollution, 142*, pp. 227-234.

Gehring U., Casas M., Brunekreef B., Bergström A., Peter Bonde J., Botton J., Chévrier C., Cordier S., Heinrich J., Hohmann C., Keil T., Sunyer J., Toft G., Wickman M., Vrijheid M., Nieuwenhuijsen M.. (2013). Environmental exposure assessment in European birth cohorts: results from the ENRIECO project. *Environmental Health, 12*(8), pp. 1-14.

Genuer R. (2013). Validation de MyVsurfGéo sur des jeux de données jouées et sur les données iris. *Entretiens téléphoniques et échanges de mails* du 11 juillet 2013

Genuer R., Poggi J-M., Tuleau-Malot C. (2013). *VSURF: Variable Selection Using Random Forests.* Adresse Internet: cran.r-project.org: http://cran.r-project.org/web/packages/VSURF/index.html. Consulté le 09 juillet 2013

Genuer R. (2012). Entretien au sujet des Forêts Aléatoires et de la procédure de sélection de variables mise au point par Robin Genuer dans sa thèse, le 01 octobre 2012.

Genuer R. (2010). Forêts aléatoires : aspects théoriques, sélection de variables et applications. *Thèse pour le grade de Docteur en sciences*. Paris, France: Université Paris XI Orsay.

Gérard J. (2012). *Un homme abattu dans les quartiers nord de Marseille*. Consulté le 28 janvier 2013, Adresse Internet: http://www.lemonde.fr/societe/article/2012/11/23/un-homme-abattu-dans-le-13e-arrondissement-de-marseille.html. Consulté le 23 novembre 2012

Geurts P., Ernst D., Wehenkel L. (2006). Extremely randomized trees. *Machine Learnin,* 63(1), pp.3-42.

Ghanbari-Niaki A., Saghebjoo M., Lamir A.R., Fathi R., Kraemer R.R. (2009). Acute circuit-resistance excercise increases expression of lymphocyte agouti-related protein in young women. (T. R. Journal, Éd.) *Experimental Biology and Medicine, 235*(3), pp. 326-334.

Ghattas B., Ben Ishak A. (2008). Sélection de variables pour la classification binaire en grande dimension : comparaisons et application aux données de Biopuces. *Journal de la Société Française de Statistique, tome 149, n°3*. Paris, France: SFdS-Institut.

Godin C. (2007). *Le comptoir philosophique.* Paris: First.

Godin I., Kittel F., Coppieters Y. Siegrist J. (2005). A prospective study of cumulative job stress in relation to mental health. *BMC Public Health,* (15), 67p.

Goovaerts P. (2006). Geostatistical analysis of disease data: accounting for spatial support and population density in the isoplelth mapping of cancer mortality risk area-to-point Poisson kriging. *Int J Health Geogr, 5*(52), p.53.

Goovaerts P. (1997). *Geostatistics for Natural Ressources Evaluation.* New York: Applied Geostatistics Series. Oxford University Press.

Grataloup C. (1996). *Lieux d'histoire, essai de géohistoire systématique.* Montpellier: GIP-RECLUS.

Gratton Y. (2002). Le krigeage : la méthode optimale d'interpolation spatiale. 4.

Grawtiz M. (2000). *Méthodes des sciences sociales* (éd. 11e). (Dalloz, Éd.) Paris: Broché.

GRNC. (1999). *Inventaire des rejets radioactifs des installations nucléaires.* Groupe Radioécologique Nord Cotentin. Fontenay aux Roses: IRSN.

Groupe CHADULE. (1997). *Initiation aux pratiques statistiques en géographie.* Paris: Armand Colin / Masson.

Haddad S. (1992). *Utilisation des services de santé en pays en développement.* Institut d'analyse des systèmes biologiques et socio-économiques. Lyon: Université de Lyon Claude Bernard.

Hamilton J-D. (1994). *Time Series Analysis.* Princeton: Princeton University Press.

Han J., Kamber M. (2006). *Data Mining : Concepts and Techniques.* (Elsevier, Éd.) University of Illinois at Urbana-Champaign: Morgan Kaufmann.

Harrouin J., Aligon A., Ruchon F., Broïdo M. (2012). *Eco-Santé France*, Adresse Internet: www.ecosante.fr: http://www.ecosante.fr/index2.php. Consulté le 17 mai 2012.

HAS, INCa. (2010). *La prise en charge du cancer*. Adresse Internet: www.e-cancer.fr: http://www.e-cancer.fr/les-cancers/cancers-de-la-thyroide. Consulté le 07 avril 2012.

Hastie T-J., Tibshirani R-J. (1990). Generalized additive models. Dans v. 43, *Monographs on Statistics and Applied Probability.* London: Chapman and Hall Ltd.

Hawley JK. (1985). Assessement of health risk from exposure to the contaminated soil. *Risk Anal*, *5*, pp. 283-302.

Head J., Standsfeld S-A., Siegrist J. (2004). The psychosocial work environment and alcohol dependence: a prospective study. *Occup Environ Med*, *61*, pp. 219-224.

Henry-Amar. M. (1999). *Tumeurs et leucémies induites*. Consulté le 07 avril 2011, Adresse Internet: www.medespace.com : http://www.medespace.com/cancero/doc/tumleucind.html

Hill A-B. (1965). The environment and the disease: association or causation. *Proceeding of the Royal Society of Medecin,* 58, pp. 796-798.

Hubert J-P. (2009). *Dans les grandes agglomérations, la mobilité quotidienne des habitants diminue, et elle augmente ailleurs.* Université de Paris-Est, Inrets-DEST et division conditions de vie des ménages. Paris: INSEE.

Hutchinson M-F. (1991). The application of thin plate smoothing splines to continent-wide data assimilation. Dans J. Jasper, *Data Assimilation Systems, Bureau of Meteorology Research Report No. 27*, pp. 104-113. Melbourne: Bureau of Meteorology Research Report.

IARC. (2007). *Attribuable causes of cancer in France in the Year 2000*. éd. World Health Organization,3, Lyon: International Agency for Research on Cancer (IARC).

IARC. (2008). *World Cancer Report.* (W. H. /WHO, Éd.). Adresse Interenet: International Agency for Research on Cancer (IARC): www.irac.fr/fr/Media-Centre/IARC-News/World-Cancer-Report-2008.

IARC. (2012). Review of Human Carcinogens Radiation. *Monographs on the Evaluation of Carcinogenic Risks to Humans, 100D*, 341.

IARC. (2013). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. Adresse Internet: http://www.iarc.fr/en/publications/list/monographs/index.php. Consulté en 2012-2013.

Ifss. (2012). *Exposition aux UV*. Adresse Internet: http://www.ifss.fr/fr/maitrise-desuv/. Consulté en 2011-2012

IGN (2004). Institut Géographique National: *Géofla : Descriptif de livraison, Format Shape file.* Paris: I.G.N.

INERIS. (2012). *Institut National de l'Environnement Industriel et des Risques*. Adresse Internet: http://www.ineris.fr/. Consulté le 2012-2013.

INHETSJ. (2012). *Présentation de l'ONDRP.* Adresse Internet: www.inhesj.fr: http://www.inhesj.fr/?q=content/presentation-de-londrp. Consulté en 2012-2013

INSEE. (2010). *Revenu fiscaux localisés des ménages (RFL).* Paris: INSEE.

INSEE. (2012a). *Définition et méthodes - Exploitations agricoles*. Adresse Internet: www.insee.fr: http://www.insee.fr/fr/methodes/default.asp?page=definitions/exploitation-agricole.htm. Consulté en 2013.

INSEE. (2012b). *Insee - Code Officiel Géographique*. Adresse Internet: www.insee.fr: http://www.insee.fr/fr/methodes/nomenclatures/cog/default.asp. Consulté en 2012

INSEE. (2012c). *www.insee.fr*. Adresse Internet: www.insee.fr: http://www.insee.fr. Consulté le 2012-2013.

INSEE. (2013). *Le secret statistique et la protection des données*. Adresse Internet: www.insee.fr: http://www.insee.fr/fr/insee-statistique-publique/default.asp?page=statistique-publique/secret-statistique.htm#p1. Consulté le 30 janvier 2013

Inserm - expertise collective. (2005). *Cancer Approche Environnementale et Méthodes.* Paris: JOUVE.

Inserm - Expertise collective. (2011). Stress au travail et santé : Situation chez les indépendants. Synthèse et recommandations. Dans I. n.-(. collective), *Synthèse et recommandations d'expertise collective (fascicule.)* Paris: Inserm. 77p.

Institute for Statistics and Mathematics. (1997). *The Comprehensive R Archive Network*. Consulté en 2010-2013, Adresse Internet: www.cran.r-project.org: http://cran.r-project.org/index.html

Irdes. (2012). *Institut de recherche et de documentation en économie de la santé (Irdes)*. Adresse Internet: www.irdes. Consulté en 2012-2013

IRSN. (2001). *Que faut-il savoir sur le Radon*. Adresse Internet: www.irsn.fr: http://www.irsn.fr/FR/base_de_connaissances/Environnement/radioactivite-environnement/radon/Pages/. Consulté le 30 mai 2012.

IRSN. (2004). *Rayonnements ionisants et santé.* Adresse Internet: http://www.irsn.fr/FR/base_de_connaissances/librairie/Documents/publications_pour_les_professionnels/IRSN_ColPro_rayonnements_ionisants_sante.pdf.

IRSN. (2009). *Bilan de l'état radiologique de l'environnement français.* Nanterre: Institut de Radioprotection et de Sûreté Nucléaire.

IRSN. (2012a). *Portail de la mesure de la radioactivité dans l'environnement*. Adresse Internet: www.irsn.fr: http://sws.irsn.fr/sws/mesure/index. Consulté en 2012-2013

IRSN. (2012b). *Présentation de l'IRSEN*. Adresse Internet: www.irsn.fr: http://www.irsn.fr/FR/IRSN/presentation/Pages/Presentation.aspx. Consulté le 07 février 2012.

IRSN, INERIS. (2008). *Le radon, synthèse des connaissances et des premières investigations en environnement minier.* Paris: IRSN/DEI/SARG/INERIS-DRS.

Jacob S., Bertrand A., Bernier M-O. (2010). Occupational Cataracts and Lens Opacities in interventional Cardiology: the O'CLOC study. (B.P. Health, Éd.) *EUROSAFE*, 5.

Jacquez G-M., Goovaerts P., Rogerson P. (2005). Space-Times intelligence systems: Technology applications and methods. *Journal of Geographycal Systems, 7*(1), pp. 1-5.

Jeauneau E. (1967). *Nani gigantum humeris insidentes: éssai d'interprétation de Bernard de Chartres.* Paris: Vivarium.

Johnson J-R, Myers D-K Jackson J-S, Dunford D-W. Gragtmans N-J., Wyatt H-M., Jones A-R., Recy D-H. (1995). Relative Biological Effectiveness of tritium for Induction of Myeloïd Leukemia in CBA/H Mice. *Rad Res, 144*, pp. 82-89.

Johnston Fay H., Henderson Sarah B., Chen Yang, Randerson James T., Marlier Miriam, De Fries Ruth S., Kinney Patrick, Bowman David M-J.S., Brauer Michael. (2012). Estimated global mortality attribuable to smoke from landscape fires. *Environement Health Perspectives, 120*(5), pp. 695-701.

Journel A-G., Huijbergts C-J. (1978). *Mining Geoststistics.* (Elsevier, Éd.) Maryland Heights, USA: Academic Press.

Kamiguchi Y, Tateno H. Mikamo K. (1990). Dose-response relationship for the induction of structural chromosomal aberration in human spermatozoa after in vitro exposure to tritium ß-rays. *Mut Res,*228, pp. 125-131.

Kenneth J. R., Sander G. (1998). *Modern Epidemiology* (éd. 2). New-York: Lippincott-Raven.

Kleijnen J.P.C. (2011). Simulation, optimization via Boostrapped Kriging. *Disvussion Paper*. (T. University, Éd.) Tilburg, Netherlands: Tilburg University.

Klein K. (1989). *The politics of the NHS* (éd. 2nd). Harlow, Longman.

Krewski D. (2009). Evaluating the Effects of Ambient Air Pollution on Life Expectancy. *New England Journal of Medicine, 360*(4), pp. 413-415.

Bellin L., Morin A.-C., Perrel C., Pfister C. (2011). *Le découpage en unités urbaines de 2010.* Paris: INSEE.

La Poste. (2012). *La Poste SNA - Acceuil SNA.*, Adresse Internet: www.laposte.fr: http://www.laposte.fr/sna/rubrique.php3?id_rubrique=59. Consulté en 2012

Lahousse P., Piédanna V. (1998). *L'outil statistique en géographie, tome I: Les distribution à une dimension.* Paris: Armand Colin.

Laplace P-S. (1820). *Théorie Analytique des Probabilités* (éd. 3e: 2009). Paris: Mme Ve COURCIER.

Leduc F. (2011). Entretien au sujet des bases INSEE de données du sous-thème "personnels et équipements de santé". Le 5 juin 2011.

Leplège A., Ecosse E., Verdier A., V-Perneger T. (1998). *The French SF-36 Health Survey: Translation, Cultural Adaptation and Preliminary Psychometric Evaluation.* France: Elsevier Science Inc.

Lesur A. (2011). *CAV 2011- Activité physique et cancer.* Adresse Internet: http://www.canal-u.tv/video/canal_u_medecine/. Consulté le 06 mai 2012

Leux C., Guénel P. (2010). Risk factors of thyroid tumors: role of environmental and occupational exposures to chemical pollutants. *épidémiologie Santé Publique, 58*(5), pp. 359-367.

Levy J., Lussault M. (2003). *Dictionnaire de la géographie et de l'espace des sociétés* (éd. 2003). Paris: Belin.

Liaw A. (2013). *Package randomForest: Breiman and Cutler's random forests for classification and regression.* (CRAN, Éd.) Récupéré Adresse Internet: cran.r-project.org: http://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Liaw A., Wiener M. (2006). Classification and Regression with randomForest. *R News*(ISSN 1609-3631), pp. 18-22.

Ligges U. (2013). *Package 'tuneR' : Analysis of music and speech.* Vienne: CRAN.

Litva A., Eyles J. (1995). Coming out: exposing social theory in medical geography. *Health and Place*, 5-14.

Loader C. (1999). Local regression and likelihood. Statistics and Computing. *Springer-Verlag*, New York.

Lucas-Gabrielli V. (2012). Entretien téléphonique au sujet des DTA. Le 5 mai 2012.

Lütkepohl H. (1991). *Introduction to multiple time series analysis* (éd. 2nd). (U. d. Michigan, Éd.) Michigan: Springer-Verlag.

Mafijul I.M., Alam M., Tariquzaman Md., Alamgir K.M., Pervin R., Begum M., Khan Md.M.H. (2013). Predictors of the number of under-five malnourished children in Bangladesh: application of the generalized poisson regression model. *BMC Public Health, 13*(11), pp. 1-8.

Mandin C. (2004). Exposition de la population fraçaise au bruit de fond du formaldéhyde et risques sanitaires associés. Ministère de l'Ecologie et du Développement Durable. Paris: INERIS.

Mankiw G. N. (1998). *Principes de l'économie* (éd. 2nd). (Economica, Éd., & M. Taylor, Trad.) Broché.

Marcotte D. (2008). *Cours GML6402: Géostatistiques, Marcotte Home Page*., Adresse Internet: www.geo.polymtl.ca: http://geo.polymtl.ca/~marcotte/.

Matheron G. (1962). Traité de géostatistique appliquée, Tome I. *Mémoires du Bureau de Recherches Géologiques et Minières, No.14.* Paris: Editions Technip.

Matheron G. (1963). Traité de géostatistique appliquée, II : Le Krigeage. *Mémoires du Bureau de Recherches Géologiques et Minières, No.24*. Paris: Editions B. R. G. M.

Matheron G. (1965). *Les variables régionalisées et leur estimation.* Paris: Masson.

Matheron G. (1989). *Estimating and choosing.* Berlin: Springer.

McCullagh P., Nelder J. (1989). *Generalized Linear Models.* Boca Raton: Chapman & Hall / CRC.

McGuirk et Porell. (1984). Spatial Patterns of Hospital Utilization: the Impact of Distance and Time. *Inquiry 21*, pp. 84-95.

McKone T.E., McLeod M. (2003). Tracking multiple patchways of human multiple exposur to persistent multimedia polluants: regional, continental and global-scale models. *Annu Rev Environ Ressour, 5*, pp. 463-492.

Melchior M., Caspi A., Milne BJ., Danese A., Poulton R., Moffit T-E. (2007). Work stress precipitates depression and anxiety in young, working women and men. *Psychol Med*, 37, pp. 1119-1129.

Mercat-Rommens C., Chojnacki E., Baudrit C. (2008). Représentation et propagation de la connaissance imprécise : ce que les théories de l'incertain peuvent apporter aux sciences environnementales. Dans D. F. Paul Allard, *Incertitude et Environnement : La fin des certitudes scientifiques*. Aix-En-Provence: EDISUD. pp. 179-192

Merleau-Ponty M. (1945). *La Phénoménologie de la perception.* Paris: Gallimard.

Météo-France (2010). Note sur les produits Météo-France. Paris.

Météo-France (2011). *Accueil Prévisions météo de Météo-France*. Adresse Internet: www.meteofrance.com: http://entreprise.meteofrance.com/. Consulté le 15 férvrier 2012.

Michel G., Bordigoni P., Simeoni M-C., Curtillet C., Hoxha S., Robitail S., Thuret I., Pall-Kondolff P., Chambost H., Orbicini D., Auquier P. (2007). Health status and quality of life in long-term survivors of childhood leukaemia: the impact of haematopoietic stem cell transplantation. *Bone Marrow Transplantation, 40*(9), pp. 897-904.

Microsoft. (2013). *Aide et Support Microsoft*. Adresse Internet: support.microsoft.com: http://support.microsoft.com/. Consulté entre 2011-2013

Kyung M.K, Yul J.C., Kim S.K, Yoo T.K, Kim D.W. (2010). Effects of radiation emitted by WCDMA mobile phone on electromagnetic hypersensitive subjects. *Environmental Health, 11*(69), pp. 2-8.

Ministère de la Santé et de la Protection sociale, Ministère de l'Écologie et du Développement durable, Ministère de l'Emploi du Travail et de la Cohésion sociale, Ministère délégué à la Recherche. (2004). *Plan Santé Environnement 2004-2008.* Paris: Ministère de la Santé, de l'Environnement et du Travail.

Ministère de l'Ecologie de l'Energie et du Développement Durable, Ministère de la Santé et des Sports, Ministère de l'Enseignement Supérieur et de la Recherche. (2009). *Plan Santé Environnement 2 2009-2013.* Paris: Direction générale de la santé.

Ministère de l'Ecologie, du Développement durable et de l'Energie. (2013). *Les engagements du Grenelle de l'environnement*. Adresse Internet: http://www.developpement-durable.gouv.fr/-Les-engagements-du-Grenelle-de-l-.html. Consulté le 19 septembre 2013,

Ministère de l'Enseignement Supérieur et de la Recherche, Ministère de la Santé et des Sports. (2009). *Pan Cancer 2009-2013.* Boulogne-Billancourt: Institut National du Cancer (INCa).

Ministère des affaires sociales et de la santé. (2012). *Ministère en charge de la santé*. Consulté en 2012, Adresse Internet: www.sante.gouv.fr.

Morin E. (1977). *La méthode.* Paris: Seuil.

Myers Donald E. (1994). Spatial interpolation : an overview. *Geoderma*, n°62, pp. 17-28.

Nielsen. (2012). *Nielsen | France*. Adresse Internet: http://www.nielsen.com/fr/fr.html. Consulté en 2012.

Nietzsche F. (1972). *Ainsi parlait Zarathoustra* (éd. 6ème). (L. L. Poche, Éd.) Paris: Les Classiques de Poche.

Olshansky S.J. et al. (2005). A Potential Decline in Life Expectancy in the United States in the 21st Century. *New England Journal of Medicine, 352*(11), pp. 1138-1145.

OMS. (2004). *OMS/EUROPE | Cinquième Conférence ministérielle sur l'environnement et la santé*., Adresse http://www.euro.who.int/fr/what-we-do/event/fifth-ministerial-conference-on-environment-and-health/past-conferences/fourth-ministerial-conference-on-environment-and-health,-budapest,-hungary,-2004. Consulté le 19 septembre 2013.

OMS. (2010). *OMS/EUROPE | Cinquième Conférence ministérielle sur l'environnement et la santé*., Adresse Internet: http://www.euro.who.int/fr/what-we-do/event/fifth-ministerial-conference-on-environment-and-health. Consulté le 19 septembre 2013.

OMS., CC. (1994). *Deuxième Conférence européenne sur l'environnement et la santé Helsinki 1994.* Organisation Mondiale de la Santé (OMS) Bureau régional de l'Europe. Scherfigsvej: Département Environnement et santé Bureau régional de l'OMS en Europe.

ONDRP. (2011). *La criminalité en France. Synthèse du rapport de l'Observatoire national de la délinquance et des réponses pénales.* PARIS: Ecole Militaire.

ONDRP. (2012). *CartoCrime.net*. Adresse Internet: http://www.cartocrime.net/Cartocrime2/index.jsf. Consulté en décembre 2012.

Overmars K.P., Verburg P.H., Bakker M.M., Staritsky I., Hellemann F. (2008). Translating land use change to landscape change for a scenario study in Europe. *Revue Internationale de Géomatique, 18*(3), pp. 327-344.

Pavuk M., Certhan J-R., Lynch C.F., Schecter A., Petrick J., Chovancova J., (2004). Environemental exposure to PCB and cancer incidence in estern Slovenia. *Chemosphere, 54*, pp. 1509-1520.

Peguy C-P. (1996). *L'horizontal et le vertical* (éd. 176). Montpellier: RECLUS.

Penchansky R. et Thomas J.W. (1981). The Concept of Access. *Medical Care*, 19(2), pp. 127-140.

Pennington DW., Margni M., Amman C., Jolliet O. (2009). Multimedia fate and human intake modeling: spatial versus nonspatial insights for chemical emissions in Western Europe. *Environ Sci Technol, 39*(11), pp. 19-128.

Picheral H. (2001). *Dictionnaire raisonné de géographie de la santé, Atelier Géographie de la santé.* Montpellier: Presse universitaire.

Picheral H. (1996). *Mots et concepts de la géographie de la santé* (éd. 2nd). Montpellier: GEOS : Atelier Géographie de la santé.

Picheral H. (1989). *Géographie de la santé; Premier cours européen de géographie de la santé.* Paris: Université P. & M. CURIE.

Pierrard O. (2013). Entretien au sujet de l'accessibilité et de la constitution des bases de données. *Chef du laboratoire des Effluents*. Paris.

Pison G. (2005). France, 2004 : l'espérance de vie franchit le seuil de 80 ans. *Population & Sociétés*(410).

Pistocchi A. Sarigiannis D.A., Vizcaino P. (2010). Spatially explicit multimedia fate models for polluants in Europe: state of the art and perspectives. *Total Environ, 408*(38), pp.,17-30.

Pope C.A., Dockery D.W. (2006). Health effects of fine particulate air pollution : lines that connect. *Journal Air Waste Management Association, 56*(6), pp.709-742.

Powell M. (1995). On the outside looking in: medical geography, medical geographers and access to health care. *Health and Place*, 41-50.

Premium Consultants (2008). *VBA pour Excel 2007.* Paris: Micro Application.

Pumain D. (2004). Système. *Hypergéo*, 4.

Pumain D. et Saint-Julien T. (1997). *L'analyse spatiale.* Paris: Masson & Armand Colin.

Quételet A. (1969). *La physique sociale: ou essai sur le développement des facultés de l'homme.* Bruxelles: Académie Royale de Belgique.

Reilly C. (1991). Metal contamination of food. (Elsevier, Éd.) *Applied Science*(2nd ed.), 235p.

Rémy E., Handschumacher P., Cinqualbre J. (2011). Les disparités spatiales du recours à un service médical hautement spécialisé : le cas de la transplantation hépatique au CHRU de Strasbourg ». *Espace Population Sociétés*, 79-96.

Rey G., Jougla E., Fouillet A. and Hénon D. (2009). Ecological association between a privation index and mortality in France over the period 1997-2001: variations with spatial scale, degree of urbanicity, age, gender and cause of death. *BMC Public Health*, 1-12.

Rican S., Salem G., Vaillant Z., Jougla E. (2009). *Dynamiques sanitaires des villes françaises.* Paris: Data-Documentation Française.

Ripley B.D. (1981). *Spatial statistics.* New York: Wiley Series in Probability and Statistics.

RNM. (2010). *Réseau National de Mesures, Les acteurs du Réseau national*. Adresse Internet: http://www.mesure-radioactivite.fr/public/s-acteurs.html. Consulté en 2012-2013.

Royston P. (1995). Algorithm AS R94 (SWILK sub routine). *Applied Statistics, 44*(4).

Salem G. (1995). Géographie de la santé, santé de la géographie. *Espace, Population, Société*, pp. 25-30.

Salem G., Rican S., Kurzinger M-L. (2006). *Atlas de la santé en France. Volume 2: Comportement et Maladies.* Paris.

Salem G., In. : Lévy J., Lussault M. (2003). *Dictionnaire de la géographie et de l'espace des sociétés.* Paris: Belin.

Sanders L. (1992). *Sytème Villes et Synergétique.* Paris: Broché.

Saporta G. (2006). *Probabilité, analyse des données et Statistique.* Paris: Technip.

SFO. (2013). Les cataractes de l'adulte. Société française d'Ophtalmologie (SFO). Adresse Internet: http://www.sfo.asso.fr/. Consulté le 15 février 2013.

Sobol H. (2004). *Prédispositions génétiques au cancer*. Adresse Internet: http://college-genetique.igh.cnrs.fr/Enseignement/genformclin/gencancer.html. Consulté le 12 juin 2012.

Somol P., Pudil J., Novovicavà J., Paclk P. (1999). Adaptive floating search methods in feature selection. *Pattern Recognition Letters, 20*, pp. 1157-1163.

Stein R. A. (2012). Epigenetics and environmental exposures. *J Epidemiol Community Health*, *66*, pp. 8-13.

Takada Y., Urano T., Ihara H., Takada A. (1995). Changes in the central and peripheral serotonergic system in rats exposed to water-immersion restrained stress and nicotine serotonergic system in rats exposed to water-immersion restrained stress and nicotine. *Neurosci Res*, 23, pp. 305-311.

Talen E., Anselin L. (1998). Assessing spatial equity: an evaluation of measures of accessibility to public playgrounds. *Environment and Planning A, 30*(4), pp. 595-613.

Therneau T., Atkinson B., Ripley B. (2013). *Recursive partitioning and regression trees.* (CRAN, Éd.) Adresse Internet: http://cran.r-project.org/web/packages/rpart/rpart.pdf.

Tillaut H. (2005). Recensement des agrégats de pathologies non infectieuses, en France, 1997-2002. *Bulletin Epidémiologique Hebdomadaire, 50*(49).

Tissot C., Cuq F. (2004). Apport des SIG pour la modélisation spatio-temporelle d'activités humaines. *Revue Internationale de Géomatique, 14*(1), pp.83-96.

Tobler W. R.. (1979). Lattice Tuning. *Geographical Analysis, 11*(1), pp.36-44.

Tobler W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*. Worcester: Clark University. pp. 234-240.

Townsend P. (1987). Deprivation. *Int Soc Pol, 16*(2), pp.125-148.

Tuleau-Malot C. (2005). Sélection de variables pour la discrimination en grande dimension et classification de données fonctionnelles. *Thèse présentée pour obtenir le grade de Docteur en sciences*. Paris, France: Université Paris XI Orsay.

Tuleau-Malot C. (2011). Entretien informel sur le paramètre mtry à Nice au Laboratoire Dieudonné, UNSA. Le 05 juillet 2011.

Twellar M., Winants Y., Houkes I. (2008). How healthy are Dutch general practitioners? Self-reported (mental) health among Dutch general practitioners. *Eur J Gen Pract, 14*, pp.4-9.

Unité Cancer et Environnement. (2012). Cancer et environnement : portail officiel sur les facteurs environnementaux liés au cancer, le cancer professionnel, alimentation et cancer. Adresse Internet : http://www.cancer-environnement.fr/. Consulté en 2011-2012.

Valérie J. (2012). Entretien sur les mesures climatologiques Météo-France mobilisables. Le 17 juin 2012.

Valetich S. (2012). Entretien téléphonique au sujet des zones géographiques définies par les codes postaux. Paris: La Poste. Le 17 août 2012.

Vallin J. (2001). *La population française.* Edition: La Découverte. 128p.

Vapnik V. (1995). The Nature of Statistical Learning Theory. *Springer Verlag*, New York.

Voiron-Canicio C. (1995). *Analyse spatiale et analyse d'images.* Montpellier: RECLUS.

Voiron-Canicio C. (2006). L'espace dans la modélisation des interactions nature-société. *Colloque Interactions Nature-Société, analyse et modèles*, 6p.: La Baule.

Voltaire (1490). *Micromégas* (éd. ed: 2000). (L. L. Poche, Éd.) Paris: Le Livre de Poche.

Wackernagel H. (1985). *L'inférence d'un Modèle Linéaire en Géostatistique Multivariable, Thèse de Doctorat.* Fontainebleau: Ecole des Mines de Paris.

Wackernagel H. (1993). Cours de Géostatistique Multivariable. *C-146, 4ème Edition*. Ecole des Mines de Paris: Centre de Géostatistique.

Wackernagel H. (2003). *Multivariate Geostatistics: An Introduction with Application* (éd. 3rd). Berlin: Springer.

Wahba G. (1990). Spline models for observational data. *CBMS-NSF Regional Conference Series in Applied Mathematics,* 59. Philadelphia: Society for Industrial and Applied Mathematics (SIAM).

Wallace R., Wallace D., Ullmann J.E., Andrews H. (1999). Deindustrialization, inner-city decay, and the hierarchical diffusion of AIDS in the USA: how neoliberal and cold war policies magnified the ecological niche for emerging infections and created a national security crisis. *Environment and Planning, 31*, pp. 113-139.

Wong O. (1995). Risk of acute myeloid leukaemia and multiple myeloma in workers exposed to benzene. *Occup Environ Med, 52*, pp. 380-384.

World Cancer Research Fund & American Institute for Cancer Research. (2007). *Food, Nutrition, Physical Activity and the Prevention of Cancer: a Global erspective.* Washington: DC, AICR.

World Health Organization. (2009). Environmental Health Criteria Series. Dans *Principles for Modeling Dose-Response for the Risk Assessment of Chemicals*. Geneva: WHO Press. 124p.

Zablobska L.B., Ashmore J.P., Howe G.R. (2004). Analysis of mortality among Canandian nuclear power industry workers after chronic low-dose exposure to ionizing radiation. *Radiat Res,* 161, pp. 633-641.

Zaidi S.S., Bhatnargar V.K., Gandhi S.J., Shah M.P., Kulkarni P.K., Saiyed H.N. (2000). Assessment of thyroid function in pesticide formulations. *Hum Exp Toxicol, 19*, pp. 497-501.

Zeitouni K. (2006). *Analyse et extraction de connaissances des bases de données spatiotemporelles ; Habilitation à Diriger des Recherches.* Paris: Université de Verseilles Saint-Quentin-en-Yvelines.

Zorman M. (2001). *CogniSciences - Précarisation et apprentissages scolaires.* Adresse Internet: http://www.cognisciences.com/IMG/Precarisation_Inserm_Zorman.pdf

ANNEXES

Annexes déclinées par partie et par chapitre.

TABLES DES ANNEXES

ANNEXE 1 : VARIABLES METEO-FRANCE

Les données commandées sont des mesures temporelles géo-localisées : $m^{l:Météo-France}_{(s_g),t}$ qui décrivent trois paramètres météos :

- **Les moyennes des cumuls journaliers du rayonnement global : $\mathbf{m^{RAY}_{(s_g),t}}$** - unité : joules/cm^2, nombre de sites renseignés : 27, nombre de données manquantes : 0.
- **Les nombres de jours annuels avec une précipitation supérieure à 1mm : $\mathbf{m^{NJAP}_{(s_g),t}}$** - unité : jours, nombre de sites renseignés : 54, nombre de données manquantes : 0.
- **Les moyennes annuelles des températures quotidiennes : $\mathbf{m^{TEMP}_{(s_g),t}}$** - unité : degrés Celsius, nombre de sites renseignés : 54, nombre de données manquantes : 0.

La période recouverte s'étend de 1980 à 2010. Les lieux où sont situées les stations météo, leur localisation ainsi que les paramètres mesurés sur chacune d'elles sont listés ci-après.

Numéro	Nom	Coordonnées		Lambert II étendu		Altitude	RAY	NJAP	TEMP
2320001	ST QUENTIN	Latitude	49°49'06"N	Lambert Y (hm)	25361	98 mètres	OUI	OUI	OUI
		Longitude	3°12'18"E	Lambert X (hm)	6626				
3060001	VICHY-CHARMEIL	Latitude	46°10'00"N	Lambert Y (hm)	21302	249 mètres	OUI	OUI	OUI
		Longitude	3°23'54"E	Lambert X (hm)	6820				
5046001	EMBRUN	Latitude	44°33'54"N	Lambert Y (hm)	19604	871 mètres	OUI	OUI	OUI
		Longitude	6°30'06"E	Lambert X (hm)	9309				
6088001	NICE	Latitude	43°38'54"N	Lambert Y (hm)	18619	2 mètres	OUI	OUI	OUI
		Longitude	7°12'30"E	Lambert X (hm)	9933				
8105005	CHARLEVILLE-MEZ	Latitude	49°46'54"N	Lambert Y (hm)	25342	147 mètres	OUI	OUI	OUI
		Longitude	4°38'30"E	Lambert X (hm)	7663				
10030001	TROYES-BARBEREY	Latitude	48°19'24"N	Lambert Y (hm)	23709	112 mètres	OUI	OUI	OUI
		Longitude	4°01'12"E	Lambert X (hm)	7248				
11069001	CARCASSONNE	Latitude	43°12'54"N	Lambert Y (hm)	18014	128 mètres	OUI	OUI	OUI
		Longitude	2°17'42"E	Lambert X (hm)	5966				
12145001	MILLAU	Latitude	44°07'06"N	Lambert Y (hm)	19021	714 mètres	OUI	OUI	OUI
		Longitude	3°01'06"E	Lambert X (hm)	6547				
13054001	MARIGNANE	Latitude	43°26'12"N	Lambert Y (hm)	18304	9 mètres	OUI	OUI	OUI
		Longitude	5°12'54"E	Lambert X (hm)	8334				
14137001	CAEN-CARPIQUET	Latitude	49°10'48"N	Lambert Y (hm)	24683	67 mètres	OUI	OUI	OUI
		Longitude	0°27'18"W	Lambert X (hm)	3962				
15014004	AURILLAC	Latitude	44°53'54"N	Lambert Y (hm)	19887	639 mètres	OUI	OUI	OUI
		Longitude	2°25'12"E	Lambert X (hm)	6066				
18033001	BOURGES	Latitude	47°03'30"N	Lambert Y (hm)	22288	161 mètres	OUI	OUI	OUI
		Longitude	2°21'30"E	Lambert X (hm)	6017				
19031008	BRIVE	Latitude	45°08'48"N	Lambert Y (hm)	20168	112 mètres	OUI	OUI	OUI
		Longitude	1°28'24"E	Lambert X (hm)	5321				
20004002	AJACCIO	Latitude	41°55'00"N	Lambert Y (hm)	16790	5 mètres	OUI	OUI	OUI
		Longitude	8°47'30"E	Lambert X (hm)	11368				
20148001	BASTIA	Latitude	42°32'24"N	Lambert Y (hm)	17531	10 mètres	OUI	OUI	OUI
		Longitude	9°29'06"E	Lambert X (hm)	11879				

21473001	DIJON-LONGVIC	Latitude	47°16'00"N	Lambert Y (hm)	22556	219 mètres	OUI	OUI	OUI
		Longitude	5°05'18"E	Lambert X (hm)	8081				
22372001	ST BRIEUC	Latitude	48°32'06"N	Lambert Y (hm)	24057	135 mètres	OUI	OUI	OUI
		Longitude	2°51'06"W	Lambert X (hm)	2168				
25056001	BESANCON	Latitude	47°14'54"N	Lambert Y (hm)	22563	307 mètres	OUI	OUI	OUI
		Longitude	5°59'18"E	Lambert X (hm)	8763				
26198001	MONTELIMAR	Latitude	44°34'48"N	Lambert Y (hm)	19563	73 mètres	OUI	OUI	OUI
		Longitude	4°43'54"E	Lambert X (hm)	7904				
28070001	CHARTRES	Latitude	48°27'36"N	Lambert Y (hm)	23850	155 mètres	OUI	OUI	OUI
		Longitude	1°30'00"E	Lambert X (hm)	5381				
29075001	BREST-GUIPAVAS	Latitude	48°26'36"N	Lambert Y (hm)	24042	94 mètres	OUI	OUI	OUI
		Longitude	4°24'42"W	Lambert X (hm)	1011				
31069001	TOULOUSE-BLAGNAC	Latitude	43°37'12"N	Lambert Y (hm)	18470	151 mètres	OUI	OUI	OUI
		Longitude	1°22'42"E	Lambert X (hm)	5225				
32013005	AUCH	Latitude	43°41'18"N	Lambert Y (hm)	18557	122 mètres	OUI	OUI	OUI
		Longitude	0°36'00"E	Lambert X (hm)	4599				
33281001	BORDEAUX-MERIGNAC	Latitude	44°49'48"N	Lambert Y (hm)	19857	47 mètres	OUI	OUI	OUI
		Longitude	0°41'24"W	Lambert X (hm)	3604				
34154001	MONTPELLIER	Latitude	43°34'36"N	Lambert Y (hm)	18430	2 mètres	OUI	OUI	OUI
		Longitude	3°57'42"E	Lambert X (hm)	7315				
35281001	RENNES-ST JACQUES	Latitude	48°04'06"N	Lambert Y (hm)	23489	36 mètres	OUI	OUI	OUI
		Longitude	1°44'00"W	Lambert X (hm)	2967				
36063001	CHATEAUROUX DEOLS	Latitude	46°52'06"N	Lambert Y (hm)	22079	158 mètres	OUI	OUI	OUI
		Longitude	1°44'24"E	Lambert X (hm)	5546				
37179001	TOURS	Latitude	47°26'36"N	Lambert Y (hm)	22729	108 mètres	-	OUI	OUI
		Longitude	0°43'36"E	Lambert X (hm)	4786				
38384001	GRENOBLE-ST GEOIRS	Latitude	45°21'48"N	Lambert Y (hm)	20448	384 mètres	-	OUI	OUI
		Longitude	5°18'48"E	Lambert X (hm)	8332				
44020001	NANTES-BOUGUENAIS	Latitude	47°09'00"N	Lambert Y (hm)	22464	26 mètres	-	OUI	OUI
		Longitude	1°36'30"W	Lambert X (hm)	3009				
45055001	ORLEANS	Latitude	47°59'00"N	Lambert Y (hm)	23317	125 mètres	-	OUI	OUI
		Longitude	1°46'36"E	Lambert X (hm)	5582				
47091001	AGEN	Latitude	44°10'18"N	Lambert Y (hm)	19094	58 mètres	-	OUI	OUI
		Longitude	0°35'36"E	Lambert X (hm)	4605				
49020001	BEAUCOUZE	Latitude	47°28'42"N	Lambert Y (hm)	22797	50 mètres	-	OUI	OUI
		Longitude	0°36'48"W	Lambert X (hm)	3776				
51183001	REIMS-COURCY	Latitude	49°18'18"N	Lambert Y (hm)	24801	91 mètres	-	OUI	OUI
		Longitude	4°03'00"E	Lambert X (hm)	7247				
54526001	NANCY-ESSEY	Latitude	48°41'12"N	Lambert Y (hm)	24170	212 mètres	-	OUI	OUI
		Longitude	6°13'12"E	Lambert X (hm)	8860				
60639001	BEAUVAIS-TILLE	Latitude	49°26'42"N	Lambert Y (hm)	24944	89 mètres	-	OUI	OUI
		Longitude	2°07'36"E	Lambert X (hm)	5848				
61001001	ALENCON	Latitude	48°26'42"N	Lambert Y (hm)	23853	143 mètres	-	OUI	OUI
		Longitude	0°06'36"E	Lambert X (hm)	4352				
63113001	CLERMONT-FD	Latitude	45°47'12"N	Lambert Y (hm)	20877	331 mètres	-	OUI	OUI
		Longitude	3°08'54"E	Lambert X (hm)	6632				

64549001	PAU-UZEIN	Latitude	43°23'06"N	Lambert Y (hm)	18242	183 mètres	-	OUI	OUI
		Longitude	0°24'54"W	Lambert X (hm)	3766				
65344001	TARBES-OSSUN	Latitude	43°11'12"N	Lambert Y (hm)	18012	360 mètres	-	OUI	OUI
		Longitude	0°00'00"E	Lambert X (hm)	4097				
66136001	PERPIGNAN	Latitude	42°44'12"N	Lambert Y (hm)	17483	42 mètres	-	OUI	OUI
		Longitude	2°52'18"E	Lambert X (hm)	6440				
67124001	STRASBOURG-ENTZHEIM	Latitude	48°32'54"N	Lambert Y (hm)	24077	150 mètres	-	OUI	OUI
		Longitude	7°38'24"E	Lambert X (hm)	9914				
69029001	LYON-BRON	Latitude	45°43'30"N	Lambert Y (hm)	20840	197 mètres	-	OUI	OUI
		Longitude	4°56'12"E	Lambert X (hm)	8024				
71105001	MACON	Latitude	46°17'48"N	Lambert Y (hm)	21470	216 mètres	-	OUI	OUI
		Longitude	4°47'54"E	Lambert X (hm)	7896				
72181001	LE MANS	Latitude	47°56'42"N	Lambert Y (hm)	23296	48 mètres	-	OUI	OUI
		Longitude	0°11'36"E	Lambert X (hm)	4399				
75114001	PARIS-MONTSOURIS	Latitude	48°49'18"N	Lambert Y (hm)	24248	75 mètres	-	OUI	OUI
		Longitude	2°20'12"E	Lambert X (hm)	6000				
77306001	MELUN	Latitude	48°36'36"N	Lambert Y (hm)	24013	91 mètres	-	OUI	OUI
		Longitude	2°40'42"E	Lambert X (hm)	6253				
78621001	TRAPPES	Latitude	48°46'24"N	Lambert Y (hm)	24196	167 mètres	-	OUI	OUI
		Longitude	2°00'30"E	Lambert X (hm)	5759				
79191005	NIORT	Latitude	46°18'54"N	Lambert Y (hm)	21498	57 mètres	-	OUI	OUI
		Longitude	0°24'00"W	Lambert X (hm)	3892				
82121002	MONTAUBAN	Latitude	44°01'36"N	Lambert Y (hm)	18923	106 mètres	-	OUI	OUI
		Longitude	1°22'36"E	Lambert X (hm)	5229				
85191003	LA ROCHE SUR YON	Latitude	46°42'00"N	Lambert Y (hm)	21956	90 mètres	-	OUI	OUI
		Longitude	1°22'42"W	Lambert X (hm)	3159				
87085006	LIMOGES-BELLEGARDE	Latitude	45°51'36"N	Lambert Y (hm)	20963	402 mètres	-	OUI	OUI
		Longitude	1°10'30"E	Lambert X (hm)	5097				
89346001	AUXERRE	Latitude	47°48'00"N	Lambert Y (hm)	23120	207 mètres	-	OUI	OUI
		Longitude	3°32'42"E	Lambert X (hm)	6905				
91027002	ORLY	Latitude	48°43'00"N	Lambert Y (hm)	24131	89 mètres	-	OUI	OUI
		Longitude	2°23'00"E	Lambert X (hm)	6035				
						TOTAL	**27**	**54**	**54**

Tableau 54 : Numéros, lieux, coordonnées géographiques, altitudes et paramètres météo mesurés par les stations mises à disposition par Météo-France.

ANNEXE 2 : VARIABLES CORINE LAND COVER

Nomenclature et charte graphique

La nomenclature est hiérarchisée en 3 niveaux. Le niveau 1 comprend 5 postes, le niveau 2 en compte 15 et le troisième niveau décrit 44 postes.

La nomenclature déclinée ci-contre représente les postes de niveau 3.
Chacun d'eux est numéroté, le nombre de chiffres dépend du niveau d'investigation choisi. *Par exemple. 221 pour 2.. : territoires agricoles, .2. : cultures permanentes, ...1 : vignobles).*

En dépit du désir de « cohérence spatiale » de la base CORINE Land Cover (CLC), certains regroupements de zones très hétéroclites ont dû être effectués pour satisfaire la contrainte des unités spatiales minimales de 25 hectares.

La documentation technique de la base de données conseille de prendre des précautions particulières quant à l'utilisation des variables : {242= « *systèmes parcellaires et culturaux complexes* »}, {243= « *territoires occupés par l'agriculture, avec présence de végétation naturelle importante* »} et {324 = « *forêts et végétation arbustive en mutation* »}.

Un descriptif précis des zones incluses dans chacune des variables de la nomenclature est disponible dans la documentation technique de la BD (CGDD, SOeS, 2009).

code niv3	*libellé*	*couleur*
111	Tissu urbain continu	
112	Tissu urbain discontinu	
121	Zones industrielles et commerciales	
122	Réseaux routier et ferroviaire et espaces associés	
123	Zones portuaires	
124	Aéroports	
131	Extraction de matériaux	
132	Décharges	
133	Chantiers	
141	Espaces verts urbains	
142	Equipements sportifs et de loisirs	
211	Terres arables hors périmètres d'irrigation	
212	Périmètres irrigués en permanence	
213	Rizières	
221	Vignobles	
222	Vergers et petits fruits	
223	Oliveraies	
231	Prairies	
241	Cultures annuelles associées aux cultures permanentes	
242	Systèmes culturaux et parcellaires complexes	
243	Surfaces essentiellement agricoles, interrompues par des espaces naturels importants	
244	Territoires agro-forestiers	
311	Forêts de feuillus	
312	Forêts de conifères	
313	Forêts mélangées	
321	Pelouses et pâturages naturels	
322	Landes et broussailles	
323	Végétation sclérophylle	
324	Forêt et végétation arbustive en mutation	
331	Plages, dunes et sable	
332	Roches nues	
333	Végétation clairsemée	
334	Zones incendiées	
335	Glaciers et neiges éternelles	
411	Marais intérieurs	
412	Tourbières	
421	Marais maritimes	
422	Marais salants	
423	Zones intertidales	
511	Cours et voies d'eau	
512	Plans d'eau	
521	Lagunes littorales	
522	Estuaires	
523	Mers et océans	

Tableau 55 : Type de surface et code couleur associés aux postes de niveau 3 de la nomenclature CORINLE Land Cover.

ANNEXE 3 : COMPLEMENTS THEORIQUES SUR L'ACP

Le cadre mathématique de l'Analyse en Composante Principale (ACP) est le suivant : on observe p variables qui constituent les coordonnées de n individus. L'ACP est la mère de la plupart des méthodes d'analyse en statistiques descriptives multidimensionnelles.

Principe mathématique :

Il s'agit d'une phase indispensable dans le processus de dépouillement d'observations. Le jeu de données et ses caractéristiques intrinsèques sont représentés par la matrice X :

$$\mathrm{X} = \begin{pmatrix} x_1^1 & \cdots & x_1^p) \\ \vdots & \ddots & \vdots \\ x_n^1 & \cdots & x_n^p \end{pmatrix}$$

Dans l'ACP on considère parfois un ensemble de données actives et de données inactives. Les éléments dits *illustratifs* peuvent aussi bien appartenir à l'espace des variables qu'à l'espace des individus. Si les données sont recueillies de façon aléatoire, tous les individus ont la même importance :

$$\mathrm{D} = \frac{1}{n} \cdot \mathbb{I}_n = \frac{1}{n} \cdot \begin{pmatrix} 1 & \ddots & 0 \\ \ddots & 1 & \ddots \\ 0 & \ddots & 1 \end{pmatrix}$$

Il s'agit du cas général mais il n'en est pas toujours ainsi. Dans certaines études il convient de travailler avec des poids : p_i différents d'un individu à l'autre. Dans ce cas : $\mathrm{D} = p_i \cdot \mathbb{I}_n$, et une contrainte de non biais est imposée tel que : $\mathrm{trace(D)} = \sum_{\forall i=j}(\mathrm{d}_i^j) = \sum_{\forall i=1,\dots,n}(p_i) = 1$.
On appelle centre de gravité le vecteur des moyennes empiriques : $g = \mathrm{X}^t \mathrm{D} \mathcal{V}_{(1)}$ où $\mathcal{V}_{(1)}$ est un vecteur de $\mathbb{R}^n$ constitué uniquement de : 1.
On procède ensuite généralement à la transformation suivante : $\mathrm{Y} = \mathrm{X} - \mathcal{V}_{(1)} g^t$: le tableau des données centrées. Puis on estime la matrice des *variances-covariances* : $\mathrm{V} = \mathrm{X}^t \mathrm{DX} - gg^t = \mathrm{Y}^t \mathrm{DY}$. On note $\mathrm{D}_{1/\hat{\sigma}}$ la diagonale de l'inverse des racines de V, i.e. la matrice diagonale des estimateurs de l'écart-type

$$\mathrm{D}_{1/\hat{\sigma}} = \begin{pmatrix} \frac{1}{\hat{\sigma}_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\hat{\sigma}_p} \end{pmatrix}$$

Bien sûr $\mathrm{D}_{1/\hat{\sigma}^2}$ est la matrice diagonale des inverses des variances d'estimation biaisées. Il est désormais possible de travailler avec le tableau des données centrées réduites : $\mathrm{Z} = \mathrm{YD}_{1/\hat{\sigma}}$:

$$z_i^j = \frac{x_i^j - \bar{x}_n^j}{\hat{\sigma}_j}, \quad \forall\, i = \{1, \dots, n\}, \forall\, j = \{1, \dots, p\}$$

La matrice regroupant les coefficients de corrélation linéaires est de la forme :

$$\mathrm{R} = \mathrm{D}_{1/\hat{\sigma}} \mathrm{V} \mathrm{D}_{1/\hat{\sigma}} = \mathrm{Z}^{\mathrm{t}} \mathrm{DZ} = \begin{pmatrix} 1 & \cdots & r_{1,p} \\ \vdots & \ddots & \vdots \\ r_{p,1} & \cdots & 1 \end{pmatrix}$$

Chaque point est défini par p coorodonnées, il est projeté sur F l'espace des individus qui est muni d'une structure euclidienne permettant de définir des distances entre les individus. En conséquence de quoi l'ACP fait intervenir la notion de métrique de façon à définir des distances statistiques cohérentes entre les individus.

A l'accoutumée on suppose que les axes de F sont orthogonaux. Les distances qui séparent les individus sont définies comme des formes quadratiques : $d^2(e_i; e_j)$. L'espace des individus est donc muni d'un produit scalaire : $\langle e_i | e_j \rangle = e_i^t \mathrm{M} e_j$ avec : $e_i = \begin{pmatrix} e_i^1 & \cdots & e_i^p \end{pmatrix}$ et d'une métrique : M. La plus utilisée est $\mathrm{M} = \mathrm{D}_{1/\hat{\sigma}^2}$. En choisissant cette métrique chaque coordonnée est divisée par l'estimateur de la variance, ce qui permet de s'affranchir d'un système unitaire disparate. De fait, l'inertie totale du nuage initial ne dépend plus des valeurs de : X, puisque :

$$I_g = \mathrm{trace}(\mathrm{MV}) = \mathrm{trace}\left(\mathrm{D}_{1/\hat{\sigma}^2}\mathrm{V}\right) = \mathrm{trace}\left(\mathrm{D}_{1/\hat{\sigma}}\mathrm{V}\mathrm{D}_{1/\hat{\sigma}}\right) = \mathrm{trace}(\mathrm{R}) = p$$

L'espace des variables $x_.^j$ est supposé connu, pour chacun des n-termes. Afin d'étudier la proximité entre les variables, cet espace est aussi muni d'une métrique, i.e. une matrice d'ordre n symétrique et définie positive. En l'occurrence, pas d'ambigüité possible, la métrique $\mathrm{M} = \mathrm{D} \in \mathcal{M}_{n\times n}(\mathbb{R})$. En effet, le produit scalaire entre : $x_.^j$ et $x_.^k$ n'est autre que : $\langle x_.^j | x_.^k \rangle = cov(x_.^j; x_.^k) = \sum_{i=1}^n p_i \cdot x_i^j \cdot x_i^k - (\bar{x}_.^j \cdot \bar{x}_.^k)$. La « longueur » entre les variables (i.e. leur norme) est définie par $\left\|x_.^j\right\|_{\mathrm{D}}^2 = \hat{\sigma}_j^2$: l'angle entre deux variables centrées est donnée par la relation :

$$cos(\theta_{jk}) = \frac{\langle x_.^j | x_.^k \rangle}{\|x_.^j\| \cdot \|x_.^j\|} = r_{j.k}$$

A chaque variable $x_.^j$ on peut associer un axe de l'espace des individus F et un vecteur de l'espace des variables E. On peut aussi calculer d'autres variables que : $x_.^1, \ldots, x_.^P$ par combinaison linaire. L'idée est de projeter les individus sur de nouveaux axes : Δ qui forment de nouvelles composantes $c_.^j$ Cette nouvelle variable c est associée aux entités mathématiques suivantes : un axe Δ de F de vecteur unitaire a ; un vecteur c de E espace de variables qui est une forme linéaire de u appelée facteur.

Le principe de l'ACP est d'obtenir une représentation simplifiée du nuage de points dans un sous-espace de dimension plus faible $k \leq \{p \in \mathbb{N}^+\}$.

Les éléments principaux de l'ACP

Projection des individus dans un sous-espace : F_k
L'idée est de déterminer une droite maximisant l'inertie du nuage de points projeté. On appelle :
Axes principaux : a un vecteur porté par la droite Δ ; Et dont le projeteur M-orthogonal sur cette droite, tel que

$$\mathrm{VM}a = \lambda a$$

Comme M est régulière le sous-espace vectoriel F_k de dimension k est engendré par les k axes de a qui sont les vecteurs propres de VM associées aux k plus grandes valeurs propres λ_k. Dans la pratique le calcul de a n'a aucun intérêt.

Facteurs principaux :
A l'axe a on associe à la forme linéaire de u un élément de $\mathbb{R}^p$ (dual de l'espace des individus) qui définit une combinaison linéaire des variables $\begin{pmatrix} x_.^1 & \cdots & x_.^p \end{pmatrix}$. L'axe principal a est M-normé, donc les facteurs principaux : u sont M^{-1}-normés. On peut donc écrire $u = \mathrm{M}a$, en conséquence de quoi :

$$\mathrm{MV}u = \lambda u$$

Comme M est régulière le sous-espace vectoriel E_k de dimension k est engendré par les k facteurs principaux qui sont les vecteurs propres de MV associés aux k plus grandes valeurs propres λ_k.
Les u facteurs principaux permettent de définir les composantes principales. Rappel de topologie : lorsque $\mathrm{MV} \in \mathcal{M}_{\mathrm{p}\times\mathrm{p}}(\mathbb{R})$ le nombre de vecteurs propres maximal est de $\mathrm{k} = \mathrm{p}$

Composantes principales :
Les variables $c_.^j$ sont des éléments de $\mathbb{R}^n$ définis par les facteurs principaux u_i^j ; $c_.^j$ est un vecteur renfermant les coordonnées des individus projetés sur l'axe unitaire : $a_.^j$

$$c_.^j = \mathrm{X}u_.^j$$

La variance de la composante $c_{.}^{j}$ vaut :

$$\mathbb{V}(c_{.}^{j}) = \lambda_j$$

Comme on travaille généralement avec le tableau des variables centrées réduites on peut aussi utiliser – pour s'affranchir des biais inhérents à un système unitaire disparate – la transformation :

$$c = \mathrm{Z}u$$

Qualité de l'ACP

Il existe de nombreux critères dans la littérature pour juger de la qualité d'une ACP. Le plus classique est le Pourcentage Expliqué d'Inertie Projetée (PEIP) sur les k premières composantes :

$$\mathrm{PEIP.k} = \frac{\sum_{j=1}^{\forall k \leq p} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

Le second critère ubiquitaire est l'analyse des corrélations : variables-facteurs. Il s'agit de la méthode la plus naturelle pour donner une signification aux composantes principales c. L'idée est de relier les $c_{.}^{j}$ aux variables $x_{.}^{j}$ en calculant les coefficients de corrélation linéaire : $r(c\,; x_{.}^{j})$. Lorsqu'on travaille avec des variables centrées réduites, on a $\mathrm{R}u = \lambda\mathrm{u}$ dont on déduit aisément :

$$r(c_{.}^{j}; x_{.}^{j}) = \sqrt{\lambda} u_{.}^{j}, \quad \forall\, j = \{1, \dots, \{k \leq p\}\}$$

On synthétise usuellement l'information obtenue dans un graphique appelé : *le cercle des corrélations*. A l'intérieur de celui-ci chaque variable $x_{.}^{j}$ est repérée par un point d'abscisse $r(c_{.}^{1}; x_{.}^{j})$ et d'ordonnée $r(c_{.}^{2}; x_{.}^{j})$. On parlera de corrélation positive avec la composante : 1 lorsque les variables sont situées à proximité de $c_{.}^{1}$ et dans la partie droite du cercle. A l'inverse on parlera de corrélation négative avec la composante : 1 pour les variables situées à proximité de $c_{.}^{1}$ mais dans la partie gauche du cercle (Saporta, 2006).

ANNEXE 4 : COMPLEMENTS THEORIQUES SUR L'INTERPOLATION SPATIALE

L'interpolation spatiale permet d'estimer, sur l'intégralité de la zone d'investigation, la variabilité spatiale de phénomènes d'intérêt dont la géographie est décrite de façon imprécise ou fragmentaire. Il s'agit de prédire en tous points les valeurs inconnues - notées $m^{l}_{(s_o)}$ - à partir d'un nombre restreint de points connus – en l'occurrence les. $m^{l}_{(s_g)}$. La qualité d'une interpolation spatiale dépend avant tout du choix de la méthode d'estimation (Myers Donald, 1994).

Chaque méthode est composée d'un ensemble de techniques qui influencent les résultats obtenus. Chacune de ces techniques engendre un prédicteur spatial particulier. Le choix du prédicteur est conditionné par sa capacité à optimiser *l'effet information* i.e. par la précision et la nature des mesures, ainsi que par sa propension à maximiser *l'effet de support, i.e.* par la quantité et la répartition spatiale des réalisations disponibles (Marcotte, 2008).
Le choix de la méthode d'interpolation spatiale repose essentiellement sur des considérations théoriques. En contrepartie, celui de la technique utilisée se base sur des considérations pratiques, spécifiées conditionnellement aux données inputs considérées, i.e. les $m^{l}_{(s_g)}$ i.e. des mesures spatiales géo-localisées. L'interpolation spatiale se fonde sur le concept de variables régionalisées et, toute mesure géo-localisée peut être interprétée comme la réalisation d'une variable régionalisée.

La géostatisque s'impose sur toutes les autres méthodes d'interpolation spatiale. Les techniques de krigeage ordinaire (KO) et de cokrigeage ordinaire font partie des plus classiques. Elles permettent d'estimer en tous points les valeurs inconnues de $m^{l}_{(s_o)}$ par $\widehat{m}^{l}_{(s_o)}$. Le choix de l'une ou de l'autre dépend de la structure et de la densité spatiale des données d'échantillonnage $m^{l}_{(s_g)}$. Lorsque la *densité spatiale d'échantillonnage est défaillante,* la variance des interpolateurs spatiaux peut être améliorée par l'introduction de variables auxiliaires : $m^{k}_{(s_g)}$ (Wackernagel, 2003).

Les variables régionalisées

Il existe une grande quantité de méthodes d'interpolation spatiale. Elles sont issues des théories de l'autocorrélation spatiale. Les variables régionalisées (v.r.) en constituent la base. Les théories de l'autocorrélation imposent l'utilisation d'un vocabulaire et de notations vernaculaires spécifiques. Ces derniers ont déjà été utilisés à demi-mot dans le paragraphe précédent. Pour que ce qui suit soit plus limpide il convient de rappeler quelques notions fondamentales.

Dans la théorie, la zone d'investigation est appelée *champ*. Il est noté $D(\omega)$ et ω est un évènement temporel ou épisodique, i.e. une période temporelle. Les processus sont décrits par des mesures géo-localisées. Toutes les variables régionalisées peuvent être modélisées par une fonction aléatoire, notée à l'accoutumée $\{Z_{(s)}, s \in D(\omega)\}$. Lorsqu'on modélise $Z_{(s,\omega)}$ par une fonction aléatoire à accroissements stationnaires, i.e. une Fonction Aléatoire Intrinsèque (FAI), on peut l'estimer en tous points du champ, conditionnellement à des données spatiales d'échantillonnage notées : $z_{(s_i)}$, à partir des opérateurs spatiaux, où $s_i \; \forall \; i = \{1, \dots, n\}$ sont des sites, i.e. des points précis de l'espace spécifiés à partir de leurs coordonnées géographiques : $s_i = (x.\text{géo}_i;\ y.\text{géo}_i)$ (Matheron, 1965).

L'interpolation spatiale : généralités et application

L'intérêt de l'interpolation spatiale est d'évaluer en tous points de l'espace les valeurs inconnues des diverses mesures spatiales notées $\{m^{l}_{(s_o)} = ?\}$. Les méthodes classiques sont déclinées dans le paragraphe suivant. Une attention particulière est portée au krigeage ordinaire (KO) et au cokrigeage ordinaire (CKO) puisque ce sont celles qui sont utilisées dans le cadre de cette thèse.
Avant d'utiliser une méthode d'interpolation spatiale particulière il convient d'évaluer sa robustesse. Pour ce faire, le plus simple est de mettre en perspective ses caractéristiques avec celles des autres méthodes. L'intérêt est de montrer la suprématie des géostatistiques pour reconstituer la majeure partie des phénomènes étudiés en géographie. Chaque méthode, même la plus robuste, assigne au respect de certaines hypothèses.
En l'occurrence le KO et le CKO font allégeance à l'hypothèse intrinsèque. L'idée de ce postulat géostatistique ressemble à s'y méprendre à la première loi de la géographie selon laquelle *tout interagit avec tout mais deux objets proches ont plus de chance de le faire que deux objets éloignés* (Tobler, 1970). En somme, l'hypothèse d'accroissements spatiaux stationnaires suppose que la valeur inconnue $m^{l}_{(s_o)}$ sise en s_o ressemble aux valeurs connues de $m^{l}_{(s_g)}$ les plus proches. Par conséquent, $m^{l}_{(s_o)}$ peut être estimée de façon précise conditionnellement aux $m^{l}_{(s_g)}$ disponibles dès lors que les interactions spatiales entre les données d'échantillonnage sont prises en compte.

En interprétant les mesures agrégées dans le temps : $m^{l}_{(s_g)}$ comme les réalisation d'une v.r., un cadre de travail est fixé ce qui permet d'utiliser certains outils géostatistiques. L'hypothèse de la stationnarité spatiale intrinsèque est supposée et peut-être vérifiée à l'aide des variogrammes. Ce sont des outils mathématiques permettent d'explorer la structure des phénomènes spatiaux à différentes échelles. En cas d'accroissements stationnaires, le variogramme est ajusté à un modèle de tendance permettant de résoudre le système de krigeage. Les opérateurs du krigeage ordinaire (KO) ou de son extension multivariable, le Cokrigeage ordinaire (CKO), permettent d'interpoler les valeurs inconnues du champ et ainsi atteindre l'objectif qui est fixé (Matheron, 1962). En somme, ici la FAI est notée : $M^{l}_{(s)}$ et ses réalisations fragmentaires sont accessibles par le biais de valeurs géo-localisées, notées les $m^{l}_{(s_g)}, \forall\, g \in \{1, \dots, n^{l}_{s}\}$.

Revue des méthodes d'interpolation spatiale et prééminence de la géostatistique

Il existe deux grandes familles de méthodes d'interpolation spatiale. Les méthodes déterministes et les méthodes stochastiques fondées sur les probabilités qui incorporent l'idée du hasard.
Les méthodes déterministes les plus connues sont les méthodes barycentriques (Arnaud et Emery, 2000), aussi appelées moyennes mobiles (Ripley, 1981) ou encore approximation de Kernel (Myers, D., 1994). Elles présentent l'avantage de prendre en compte le voisinage et la distance entre les observations pour estimer les poids des modèles. La technique la plus connue étant l'inverse de la distance à la puissance "d" (Baillargeon, 2005). En contrepartie elles présentent l'inconvénient d'être grossières, raison pour laquelle on préfèrera utiliser certaines méthodes de partitionnement de l'espace (Arnaud et Emery, 2000) ; (Ripley, 1981), qui forment en fait un sous-ensemble de cette première catégorie. Les polygones de Thiessen, le diagramme de Voronoï et la mosaïque de Dirichlet sont les techniques les plus triviales. A l'aune des méthodes barycentriques, elles permettent de spécifier les critères de voisinage en partitionnant le champ. Mais cette spécification est limitée à la seule répartition spatiale des données d'échantillonnage. Elles présentent néanmoins souvent l'avantage d'être exactes, i.e. que les prédictions correspondent aux valeurs connues. Toutefois les couches de résultat sont généralement empreintes de changements abrupts et de surcroit, l'interpolation est parfois limitée à l'enveloppe convexe des sites connus. Les splines forment la troisième méthode déterministe listée. Les plus connues sont les *thin plate spline* (Wahba, 1990). Il s'agit d'ailleurs de la seule méthode déterministe d'interpolation spatiale susceptible de concurrencer certaines techniques de krigeage. Les résultats sont des couches de surface ajustées par minimisation d'un critère de flexion (Hastie et Tibshirani, 1990). De fait la structure spatiale d'échantillonnage est nécessairement prise en compte. En outre, et contrairement aux méthodes barycentriques, certaines splines permettent d'effectuer des

interpolations multi-variables, e.g. les *partial thin splines* (Hutchinson, 1991). Pour autant, comme toutes les méthodes déterministes, les splines ne permettent pas d'estimer l'erreur commise sur les prédictions.

Du côté des méthodes stochastiques les plus courantes, il y a les *régressions* et les méthodes de *krigeage*. Les méthodes stochastiques sont composées d'une fonction déterministe et d'une fonction aléatoire. Comme les splines, les régressions globales ajustent des couches de surface aux valeurs régionalisées. Ce type d'interpolation est aussi appelé *surface de tendance* (Ripley, 1981). S'agissant d'une méthode stochastique, l'erreur d'estimation peut être évaluée. Comme tous les modèles de régression, des variables auxiliaires peuvent être incorporées. Cependant l'estimation des poids du modèle, par le critère des moindres carrés ordinaires, suppose que les erreurs soient indépendantes et identiquement distribuées, ce qui n'est presque jamais le cas des variables régionalisées. Les régressions spatiales globales sont du reste approximatives. En outre, le même poids est attribué à chacune des valeurs connues. A cause de ces faiblesses, des techniques de *régression locales* ont été mises au point (Cleveland et Devlin, 1988) - connues aussi sous le nom *de régressions pondérées géographiquement* (Fotheringham, Brunsdon et al., 2002). Désormais les paramètres des modèles sont estimés par le critère des moindres carrés pondérés. Plusieurs fonctions de poids sont proposées dans la littérature. Ces dernières permettent de prendre en compte la distance entre les observations et de définir une taille de voisinage (Loader, 1999). Mais ces caractéristiques restent néanmoins soumises à la subjectivité d'un choix arbitraire d'une fonction de pondération.

L'ultime méthode d'interpolation stochastique listée est le krigeage. Elle semble être la plus adaptée aux jeux de données spatialisées (Baillargeon, 2005). En effet, à l'instar des méthodes barycentriques et des régressions, l'opérateur krigeage permet de choisir entre une interpolation globale ou locale et comme il s'agit d'une méthode stochastique, l'erreur de prévision peut être estimée en tous points (Cressie, 1993). Tout comme les splines, des extensions multi-variables existent, c'est le cas du krigeage avec dérive externe (Goovaerts, 1997) et *a fortiori* du cokrigeage (Wackernagel Hans, 1993). Par contre, le krigeage ne permet pas d'ajuster des couches de surface aux données régionalisées. L'estimateur krigeage est qualifié de BLUP : best linear unbiased predictor (Matheron, 1963). Le krigeage présente les mêmes avantages que les régressions géographiques puisqu'il prend en considération la distance entre les observations et le voisinage. Mais contrairement aux régressions géographiques, le krigeage permet de spécifier objectivement la structure spatiale du phénomène à partir d'une contrepartie empirique : le variogramme, en supposant les erreurs d'estimation spatialement interdépendantes. On appelle cela l'hypothèse d'accroissement stationnaire. L'analyse variographique permet de valider cette hypothèse et par la même occasion, de spécifier un modèle de fonction aléatoire régionalisée et une taille du voisinage. En conséquence de quoi, le krigeage affirme sa suprématie parmi toutes les autres méthodes d'interpolation spatiale. Enfin le krigeage ou le cokrigeage englobent plusieurs techniques du type : simple, ordinaire, universel, disjonctif... Chacune d'elles a ses spécificités. Dans le cadre de l'hypothèse intrinsèque l'utilisation de la technique dite ordinaire est préconisée (Matheron, 1965).

Les Variogrammes : structure spatiale, anisotropie et ajustement

Le variogramme est l'outil adapté au krigeage et au cokrigeage. Typiquement le variogramme est noté : $\gamma(h)$. C'est une fonction mathématique utilisée pour résumer l'information contenue par les variables régionalisées. Il est partiellement continu dans le champ lorsque le phénomène d'intérêt est spatialement auto-corrélé, c'est-à-dire lorsque la variabilité des réalisations de la v.r. est en partie déterminée par la notion de distance (Matheron, 1963). Le variogramme permet d'explorer, à différentes échelles, la structure spatiale des réalisations d'une v.r. C'est un outil de mesure qui s'attache aux dissemblances. Il permet de caractériser la stationnarité d'un phénomène. *A priori* on suppose que les accroissements des $m^{l}_{(s)}, \forall\, s \subseteq \mathcal{D}(\omega.)$ sont à peu près stationnaires. La variographie permet de vérifier cette hypothèse et d'en mesurer l'ampleur.

Le KO et le CKO font allégeance aux attentes de l'hypothèse intrinsèque. Cette dernière implique que la moyenne des accroissements spatiaux est nulle, quelle que soit la valeur de h, à l'intérieur d'un certain voisinage et que sa variance existe et qu'elle est constante à des échelles locales (Journel et Huijbergts, 1978) – en d'autres termes :

$$\begin{cases} \mathbb{E}\left[M^{l}_{(s_g)} - M^{l}_{(s_g+h)}\right] = 0 \\ \mathbb{V}\left[M^{l}_{(s_g)} - M^{l}_{(s_g+h)}\right] = \{2\gamma(h) < +\infty\} \end{cases}$$

Ainsi les valeurs situées en $\{s_i \overset{\text{def}}{=} s_g\}$ ressemblent globalement à celles sises en $\{s_j \overset{\text{def}}{=} s_g + h\}$ lorsque h est "petit", et en sont statistiquement différentes dès lors que la valeur de h est "grande". Ainsi, les jeux de données spatiaux ont des structures très particulières. Les mesures proches se ressemblent. Les plus éloignées dissemblent. La présence d'*extrema* vient généralement biaiser l'estimation de la moyenne lorsqu'elle est calculée sur l'intégralité des données d'échantillonnage. Dans un cadre univarié le variogramme est le moyen le plus adapté pour explorer la structure des v.r. : $m^{l}_{(s_g)}$. La force du variogramme est de s'affranchir, dans sa formulation mathématique, de l'estimation globale de la moyenne (Marcotte, 2008). Dans la pratique, le semi-variogramme omnidirectionnel s'estime de la façon suivante :

$$\hat{\gamma}^{l}(h) = \frac{1}{2 \cdot N(h)} \cdot \sum_{i=1}^{n^{l}_{s}} \sum_{j>i} \left(m^{l}_{(s_i)} - m^{l}_{(s_j)}\right)^2 \cdot \left(\mathbb{1}_{(s_i,\, s_j) \in D(\omega.)}\right) \cdot \left(\mathbb{1}_{(h_{gi} \leq \{h_{crit}\})}\right)$$

Cette fonction empirique représente la différence au carré de toutes les combinaisons linéaires des réalisations disponibles de $m^{l}_{(s_g)}$, i.e. pour tous les sites s_i et s_j du champ $D(\omega)$, avec $h_{ij} = \|s_g - s_i\|$ une distance euclidienne et N(h) le nombre de couples de points espacés d'une distance inférieure ou égale à "h". Le semi variogramme est estimé au maximum pour des valeurs de h inférieures à la moitié de la distance maximale d'échantillonnage. Cependant cette valeur est généralement réduite à $h_{crit.} \leq \frac{1}{3} \max_{\forall j \neq i}\{h_{ij}\}$ (Arnaud et Emery, 2000).

Il s'agit d'une règle empirique certes, mais raisonner sur de trop grandes distances revient à étudier des dissemblances entre $m^{l}_{(s_i)}$ et $m^{l}_{(s_j)}$ pour des sites qui se situent aux antipodes l'un de l'autre.

Lorsque le semi-variogramme estimé sur $m^{l}_{(s_g)}$ dénote la présence d'une stationnarité intrinsèque, il est ajusté à un modèle théorique. Conséquemment un KO peut être envisagé. Il existe un grand nombre de fonctions mathématiques admissibles décrites dans la littérature à cet effet. Les plus classiques sont présentées ci-après. Le choix du modèle d'ajustement se fait *a priori* selon les caractéristiques de forme et d'échelle qui leur sont propres. Le choix et l'ajustement du semi-variogramme constituent la partie la plus délicate du krigeage, *c'est un art plutôt qu'une science* (Gratton, 2002). Mais il n'en demeure pas

moins : *qu'il existe un lien étroit entre la variable étudiée et le type de phénomène que l'on est susceptible de rencontrer* (Marcotte, 2008).

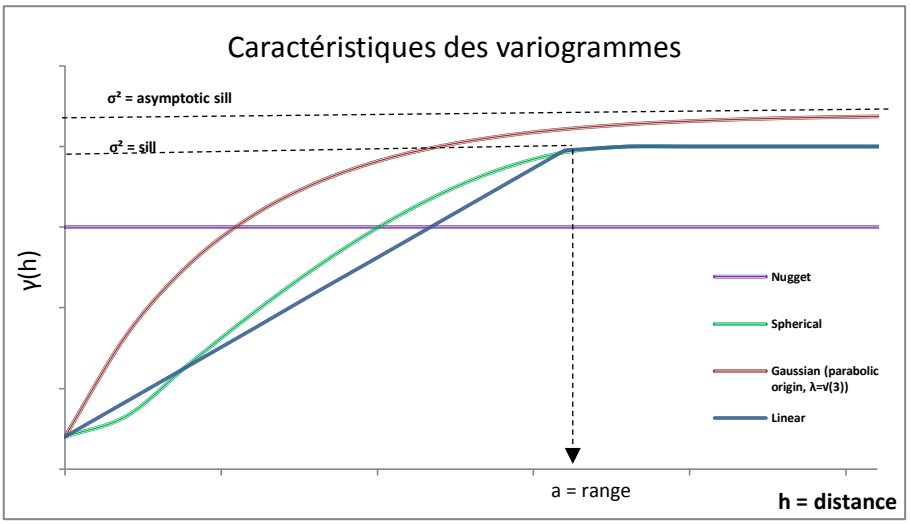


Figure 295 : Représentation caractéristique des modèles de variogrammes couramment utilisés

On appelle *effet pépite* $\{\sigma^2_{o,\mathrm{l}} = \text{"nugget"}\}$ la valeur : $\lim_{h\to 0^+}\{\hat{\gamma}^{\mathrm{l}}(h)\}$. Il est associé à un bruit blanc faible qui intervient à l'origine du variogramme. Cette variabilité est généralement toujours présente. Elle est due soit à l'imprécision des mesures, soit à des perturbations aléatoires de la v.r. à micro-échelle. Parfois l'effet pépite est amplifié par l'éloignement des valeurs spatialisées notamment dans les jeux de données qui ne sont pas uniformément reparties dans le champ ou dans ceux dont la densité d'échantillonnage est faible. (Cressie, 1993).

On appelle « palier » du variogramme $\{\sigma^2_{a,\mathrm{l}} = \text{"sill"}\}$ la valeur limite à laquelle la variance se stabilise $\{(\hat{\gamma}^{\mathrm{l}}(h) - \hat{\sigma}^2_{o,\mathrm{l}}) \approx \mathrm{c}^{\mathrm{l}} \in \mathbb{R}^+_*\}$. Ce seuil dénote une indépendance spatiale de la v.r. : $\mathrm{m}^{\mathrm{l}}_{(s_g)}$. On distingue deux types de variogrammes : ceux pourvus d'un palier fixe, c'est le cas par exemple du modèle linéaire avec palier et du modèle sphérique, et ceux ayant une portée asymptotique, i.e. ceux dont le palier n'est jamais vraiment atteint. Le modèle gaussien fait partie de cette seconde catégorie. *Il décrit des variables ayant une grande continuité de forme, dont la topographie fait partie* (Marcotte, 2008). Il est aussi adapté à la modélisation de nombreux phénomènes météorologiques ou, de la radioactivité environnementale liée à l'activité volumiques de radionucléides présents dans les milieux eau, air, sol.
Dernière caractéristique importante : la « portée » : $\{a^{\mathrm{l}} = \text{»range »}\}$ du variogramme, correspond à la distance à laquelle ledit palier est atteint. Cette mesure caractérise l'intensité des forces d'inertie qui assurent la cohésion spatiale du phénomène jusqu'à un certain voisinage. La portée empirique s'estime de la façon suivante :

$$\hat{a}_{\mathrm{l}} = \underset{\forall\, h \leq \mathrm{h}_{\mathrm{crit}}}{\operatorname{argmin}} \{(\hat{\gamma}^{l}(h) - \hat{\sigma}^2_{o,\mathrm{l}}) = \hat{\sigma}^2_{a,\mathrm{l}}\}$$

Pour les variogrammes à paliers asymptotiques, on concèdera que la portée est la distance pour laquelle le variogramme atteint 95% de sa valeur limite (Baillargeon, 2005).
La formulation mathématique du semi-variogramme $\hat{\gamma}^{\mathrm{l}}(\mathrm{h})$ estimé énoncée précédemment est celle du variogramme omnidirectionnel. Or, si les phénomènes spatiaux prennent des valeurs similaires lorsqu'elles sont proches l'une de l'autre, leur structure est rarement isotrope. Bien que les causes d'une anisotropie soient généralement ineffables car induites par une complexité spatiale multifactorielle, les variogrammes directionnels permettent néanmoins de les déceler (Wackernagel Hans, 1993).
L'irrégularité de la répartition des données d'échantillonnage, les singularités géométriques du champ, la structure spatiale du phénomène, ainsi que les données extrêmes, viennent bruiter le variogramme empirique. Ces caractéristiques constituent, entre autres, des sources d'anisotropie qui complexifient

l'ajustement à une courbe de tendance. On distingue communément deux types d'anisotropies. La première est géométrique. Elle se caractérise par des pépites et des paliers identiques mais des portées différentes selon les directions d'exploration du champ. Elle présente l'avantage d'être facilement corrigeable. La seconde est l'anisotropie zonale. Plus grave et plus difficile à corriger car les valeurs des portées mais aussi celles des paliers et des pépites varient selon la direction d'analyse (Wackernagel Hans, 1993).
Pour construire des variogrammes directionnels on utilise des fenêtres angulaires. Elles ne prennent en compte que les paires de points se trouvant dans une direction : $\theta_k \in [\![0°; 180°]\!]$ (avec le Nord géographique) et pour une tolérance d'ouverture donnée : $\varphi \in [\![30°; 45°]\!]$ (ESRI, 2013).

Les variogrammes directionnels s'estiment de la façon suivante :

$$\hat{\gamma}^{l}(h, \theta_k) = \hat{\gamma}(h_{ij}) \cdot \mathbb{1}_{\{\theta_{ij}\} \in \{\theta_k \pm \frac{\varphi}{2}\}} \; ; \quad \forall\, j \neq i \in \{1, \dots, n^{l}_{s.}\}$$

Lorsqu'une anisotropie géométrique est décelée, des modèles directionnels sont ajustés et seules les valeurs de $\hat{a}_{\theta,l}$ varient. L'anisotropie se corrige par le biais d'une ellipse d'iso-valeur. L'axe majeur de l'ellipse est celui de la plus grande portée d'échantillonnage :

$$\hat{a}^{l}_{\mathscr{g}} = \max_{\forall \theta_k \in [\![0°;180°]\!]} \left\{ \underset{\forall h \leq h_{(crit;\theta_k)}}{\operatorname{argmin}} \left\{ \hat{\gamma}^{l}(h, \theta_k) \approx \sigma^2_{a,l} \right\} \right\}$$

Une fois obtenu on en déduit $\hat{\theta}^{l}_{\mathscr{g}}$ et corollairement $\hat{\theta}^{l}_{\mathscr{p}} = \left|90 + \hat{\theta}^{l}_{\mathscr{g}}\right|$. L'axe mineur de l'ellipse est évalué comme étant la valeur de la plus grande portée orthogonale notée : $\hat{a}^{l}_{\mathscr{p}}$. La correction d'une anisotropie géométrique revient donc à construire un variogramme omnidirectionnel en introduisant une matrice de dilatation ou de rotation – notées : D (Marcotte, 2008).

En pratique, lorsque la superposition des variogrammes directionnels présente une allure de *rose des sables* cela dénote la présence d'une anisotropie géométrique (Gratton, 2002). On appelle facteur d'anisotropie le rapport $\left\{F^{l}_{a} = \frac{\hat{a}^{l}_{\mathscr{g}}}{\hat{a}^{l}_{\mathscr{p}}}\right\}$. Dans la pratique une correction d'anisotropie n'a de sens que s'il est au moins égal à 1.5 et si le nombre de données est supérieur à 50 (Marcotte, 2008).
Le variogramme empirique sert de support au modèle théorique. Parmi le large panel de fonctions admissibles, la combinaison d'un modèle gaussien et d'une pépite a souvent été utilisée comme ajustement aux $\hat{\gamma}^{l}(h)$. Son équation caractéristique est donnée par :

$$\tilde{\gamma}^{l}(h) = \begin{cases} \hat{\sigma}^2_{o,l} & h = 0 \\ \hat{\sigma}^2_{o,l} + \hat{\sigma}^2_{a,l} \cdot \left(1 - e^{-\left(\frac{h}{\lambda \cdot a}\right)^2}\right) & 0 < h \leq h_{crit} \end{cases}$$

Il existe toutefois un grand nombre d'autres modèles admissibles – un peu plus d'une dizaine sont disponibles dans l'extension Géostatistical Analyst d'ArcGis.10 – une trentaine sont couramment utilisés et particulièrement cités dans la littérature (Marcotte, 2008). Pour simplifier les écritures on notera : $\tilde{\gamma}^{l}(h) \overset{\text{def}}{=} D \cdot \tilde{\gamma}^{l}\left(h \cdot \tilde{\gamma}^{l}(h; \theta_k)\right)$, ce qui implique que le modèle de variogramme spécifié à fait l'objet d'une correction lorsqu'une anisotropie a été mise en évidence, par les variogrammes directionnels.

S'agissant du calibrage des paramètres du modèle, il résulte d'une minimisation de la variance effectuée conditionnellement aux données disponibles :

$$\{\hat{\vartheta}^{\mathrm{l}} = \hat{a}^{\mathrm{l}}, \hat{\sigma}^2_{a,\mathrm{l}}, \hat{\sigma}^2_{o,\mathrm{l}}, \lambda_i\} = \underset{\forall\, \vartheta^{\mathrm{l}} \in \mathbb{R}_+^{|\vartheta^{\mathrm{l}}|}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n_s} \sum_{j \neq i} \mathrm{w}_i^{\mathrm{l}} \cdot \left(\hat{\gamma}^{\mathrm{l}}(\mathrm{h_{ij}}) - \tilde{\gamma}^{\mathrm{l}}(\mathrm{h_{ij}};\, \vartheta) \right)^2 \right\} \;\; \forall\, h_{ij} \leq h_{crit}$$

Comme les réalisations de $\mathrm{M}^{\mathrm{l}}_{(s)}$ sont à valeurs réelles, cela revient à minimiser une fonction quadratique. Pour ce faire on utilise le critère des moindres carrés pondérés (mcp), dont les poids de l'équation $\mathrm{w}_i^{\mathrm{l}}$ sont ceux de Cressie (Cressie, 1985).

Voilà succinctement présentées les caractéristiques de base de la modélisation du variogramme expérimental. Il en existe d'autres, telles que : le contrôle de la tolérance d'ouverture de la fenêtre, l'agrégation ensembliste liminaire, dite : « binning », le *pas d'agrégation* qui permettent de grouper certains quadrats par classes de distance, les structures *gigognes,* les dérives... Ces éléments sont d'une importance majeure, mais ils ne seront pourtant pas détaillés. Il est simplement précisé qu'ils permettent de mieux interpréter les semi variogrammes et d'obtenir des ajustements plus robustes (ESRI, 2013).

Le krigeage ordinaire : théorie et recommandations pratiques

Le krigeage est une méthode d'interpolation géostatistique qui a été développée par Georges Matheron à l'Ecole des Mines de Paris. Historiquement, elle découle des travaux d'un ingénieur minier sud-africain nommé Krige. L'idée est de modéliser des phénomènes continus dans l'espace et susceptibles d'être décrits par des variables régionalisées (Matheron, 1965). Le krigeage est une sorte de régression spatiale, dont les valeurs de la composante déterministe ${\varphi^{\mathrm{l}}}_{(\cdot)}$ fluctueraient selon deux composantes aléatoires distinctes. La première, ${\delta^{\mathrm{l}}}_{(\cdot)}$ représente des variations spatialement corrélées, à moyenne et à micro échelles, qui peuvent être approchées par le variogramme. Et la seconde ${\varepsilon^{\mathrm{l}}}_{(\cdot)}$ est un terme d'erreur du type bruit blanc faible - purement aléatoire. De fait l'estimateur peut s'écrire :

$$\mathrm{M}^{\mathrm{l}}{}_{(s)} = {\varphi^{\mathrm{l}}}_{(s)} + {\delta^{\mathrm{l}}}_{(s)} + {\varepsilon^{\mathrm{l}}}_{(s)} \qquad \forall\, s \in D(\omega)$$

Comme le krigeage est une méthode stochastique, l'erreur d'estimation commise sur chaque prédiction est statistiquement estimable (Baillargeon, 2005). Le postulat de stationnarité intrinsèque s'adapte à un grand nombre de phénomènes spatiaux. Le variogramme permet de s'affranchir de l'estimation globale de la moyenne puisque celle-ci est constamment ré-estimée localement (Marcotte, 2008). Il suffit donc que la moyenne et la variance soient constantes à l'intérieur un certain voisinage. L'opérateur krigeage tient compte de la structure spatiale des réalisations. L'estimateur est linéaire, sans biais et de variance minimale. Il est aussi qualifié d'exact, i.e. que les valeurs connues correspondent exactement à celles prédites.

Le modèle de krigeage se construit conditionnellement aux données disponibles : $\mathrm{m}^{\mathrm{l}}_{(\mathrm{s})} = \left(\mathrm{m}^{\mathrm{l}}_{(\mathrm{s}_1)}, \ldots, \mathrm{m}^{\mathrm{l}}_{(\mathrm{s}_n)}\right)$. L'estimation de $\mathrm{m}^{\mathrm{l}}_{(\mathrm{s}_o)}$ par $\widehat{\mathrm{m}}^{\mathrm{l}}_{(\mathrm{s}_o)}$ nécessite donc la sélection de données d'échantillonnage se trouvant dans un certain « voisinage de krigeage » spécifique. Autrement dit les $\mathrm{m}^{\mathrm{l}}_{(\mathrm{s}_g)}$ prises en compte sont uniquement celles définies par les indices :

$$\left\{\mathrm{i} \in \mathrm{V}^{\mathrm{l}}(\mathrm{s_o})\right\} \stackrel{\text{def}}{=} \left\{[\mathrm{i}] = [1], \ldots, \left\{\left[\mathrm{n}^{\mathrm{l}}_o\right] \leq \mathrm{n}^{\mathrm{l}}_s\right\}\right\}$$

L'estimateur du krigeage ordinaire est de la forme :

$$\mathrm{M}^{\mathrm{l}}_{(\mathrm{s}_o)} = \sum_{\forall i \,\in \mathrm{V}^{\mathrm{l}}(\mathrm{s_o})} \lambda_{\mathrm{i}} \cdot \mathrm{M}^{\mathrm{l}}_{(\mathrm{s}_i)}$$

La contrainte de non biais impose d'un point de vue théorique que :

$$\mathbb{E}\left[\widehat{\mathrm{M}}^{\mathrm{l}}_{(\mathrm{s}_o)} - \mathrm{M}^{\mathrm{l}}_{(\mathrm{s}_o)}\right] = 0$$

De fait, une première condition s'impose sur les poids du modèle : $\sum_{i \,\in \mathrm{V}(s_o)} \lambda_{\mathrm{i}} = 1$. Par ailleurs, le modèle de K.O. garantit une variance minimale, c'est-à-dire que l'estimation des poids du modèle est soumise à une seconde contrainte :

$$\hat{\lambda} = \underset{\forall \lambda \in \mathbb{R}^{n_o^l}}{\operatorname{argmin}}\left\{\mathbb{V}\left[\widehat{M}^l_{(s_o)} - M^l_{(s_o)}\right]\right\}$$

En développant on obtient : $\mathbb{V}\left[\widehat{M}^l_{(s_o)} - M^l_{(s_o)}\right] = -\lambda^t \Gamma_{ij} \lambda + 2\lambda \gamma_{go}$. Avec λ et γ_{go} des vecteurs de dimension $\left([n^l_o] \times 1\right)$; Et Γ_{ij} une matrice de dimension : $\left([n^l_o] \times [n^l_o]\right)$. Dans la mesure où les réalisations de $M^l_{(s_.)}$ sont à valeurs réelles, il s'agit de minimiser une fonction convexe. Par conséquent des solutions existent nécessairement. Il s'agit donc d'un problème d'optimisation qui, typiquement, se résout en introduisant un lagrangien, que sera noté: μ. La fonction à minimiser devient : $L(\lambda; \mu) = -\lambda^t \Gamma_{gi} \lambda + 2\lambda^t \gamma_{go} + 2\mu\left(\lambda \mathbb{I}_{n^l_o} - 1\right)$; avec $\mathbb{I}_{n^l_o} = (1, \dots, 1)^t$ et $\dim\left(\mathbb{I}_{n^l_o}\right) = n^l_o$. De telle sorte qu'il ne reste plus qu'à annuler les dérivées partielles $\frac{L(\lambda;\mu)}{\partial \lambda} = 0$. En incluant ensuite le *lagrangien* dans le vecteur des poids λ en vue de l'estimer, et symétriquement $\mathbb{I}_{n_o}$ dans la matrice Γ_{gi} ainsi qu'un zéro en bout de diagonale : $\left\{\Gamma_{(\boldsymbol{n_o^l+1})(\boldsymbol{n_o^l+1})} = 0\right\}$ pour respecter les contraintes énoncées. Le système de krigeage s'écrit sous forme matricielle :

$$\begin{pmatrix} \tilde{\gamma}^l(h_{11}) & \dots & \tilde{\gamma}^l(h_{1j}) & \dots & \tilde{\gamma}^l\left(h_{1n^l_o}\right) & 1 \\ \vdots & \ddots & & & \vdots & \vdots \\ \tilde{\gamma}^l(h_{i1}) & & \tilde{\gamma}^l(h_{ij}) & & \tilde{\gamma}^l\left(h_{in^l_o}\right) & 1 \\ \vdots & & \vdots & \ddots & \vdots & \vdots \\ \tilde{\gamma}^l\left(h_{n^l_o 1}\right) & \dots & \tilde{\gamma}^l\left(h_{n^l_o h}\right) & \dots & \tilde{\gamma}^l\left(h_{n^l_o n^l_o}\right) & 1 \\ 1 & \dots & 1 & \dots & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_i \\ \vdots \\ \lambda_{n^l_o} \\ \mu \end{pmatrix} = \begin{pmatrix} \tilde{\gamma}^l(h_{1o}) \\ \vdots \\ \tilde{\gamma}^l(h_{io}) \\ \vdots \\ \tilde{\gamma}^l\left(h_{n^l_o o}\right) \\ 1 \end{pmatrix}$$

Comme $\Gamma_{\mathbf{ij},\mathbf{1}}$ est supposée symétriquement positive on en déduit : $\hat{\lambda}_\mu = \Gamma^{-1}_{hg,1} \gamma^l_{go,1}$ et en conséquence de quoi, chaque site : s_o, de valeur inconnue est estimé par :

$$\hat{m}^l_{(s_o)} = \sum_{\forall i \in V^l(s_o)} \hat{\lambda}_i \cdot m^l_{(s_i)}$$

Quant à la variance de krigeage elle est donnée par la relation :

$$\hat{\sigma}^2_{\hat{m}^l_{(s_o)}} = \sum_{\forall i \in V^l(s_o)} \hat{\lambda}_i \cdot \tilde{\gamma}^l(h_{io}) + \hat{\mu}$$

A supposer que les erreurs d'estimation soient normalement distribuées (Davis, 1986), alors :

$$\mathbb{P}\left(m^l_{(s_o)} \in \left\{\hat{m}^l_{(s_o)} \mp 2 \cdot \hat{\sigma}^2_{\hat{m}^l_{(s_o)}}\right\}\right) \leq 95\%$$

Lorsque l'hypothèse de stationnarité intrinsèque est respectée, en dépit de sa simplicité, le KO demeure la méthode de prédilection en interpolation spatiale.

Dans la pratique : l'analyse variographique permet de conjecturer une stationnarité d'ordre deux et de valider l'utilisation du KO. La résolution du système de krigeage est parfaitement automatique. Seul le choix de sélection du voisinage reste, en partie, à la discrétion du modélisateur. Les caractéristiques de la fenêtre glissante $V^l(s_o)$ sont spécifiées conditionnellement à celles du variogramme. Une forme elliptique lui est conférée lorsqu'une anisotropie géométrique est mise en exergue et que la densité spatiale d'échantillonnage est suffisante. Elle est orientée dans la direction de la plus grande portée : $\hat{\theta}^l_{\mathcal{g}}$. Les axes de $V^l(s_o)$ sont dimensionnés à partir des estimations de: $\hat{a}^l_{\mathcal{g}}$ et $\hat{a}^l_{\mathcal{p}}$. La fenêtre peut être divisée en secteurs de manière à rendre le plus uniforme possible la sélection du voisinage de krigeage. Le nombre de secteurs dépend de la densité et de la répartition spatiale des données d'échantillonnage. Afin de procéder à une interpolation spatiale de qualité, il s'agit de faire en sorte que le nombre de $m^l_{(s_g)}$ inclus dans $V^l(s_o)$ soit au moins égal à 10 mais n'excède pas 20. A chaque fois, les points situés à l'extérieur de la zone de recherche, comme ceux en excès mais néanmoins à l'intérieur, sont rejetés. Les autres sont pondérés proportionnellement à la distance avec le site à interpoler. C'est ce qu'on appelle

l'effet d'écran. Enfin, pour les zones du champ sous-échantillonnées il est conseillé de spécifier une tolérance de façon à inclure les points, les plus proches, mais situés en dehors de $V^{l}(s_o)$ dans la direction des secteurs défaillants, afin de respecter la règle empirique des *au moins 8 points* (Marcotte, 2008).
Comme l'opérateur KO est ponctuel, les résultats obtenus sont positionnés sur des grilles régulières spatialisées. Le krigeage ne permet pas d'obtenir une interpolation continue dans l'espace géographique. Pourtant la plupart des logiciels renvoie des couches de surface, type images raster ou TIN. Ce type d'output fait intervenir une procédure de lissage. Les plus courantes sont : la triangulation de Delaunay ou le lissage polynomial par fonction splines, par défaut sous ArcGis (ESRI, 2013).
Parfois la répartition ou/et la densité d'échantillonnage des réalisations d'une v.r. ne permettent pas d'interpoler dans la zone d'investigation, de façon robuste, les réalisations d'une v.r. par le biais d'un KO. Une pierre de touche pour pallier cette défaillance est de se tourner vers une méthode géostatistique multivariable.

Le cokrigeage ordinaire : théorie et recommandations pratiques

Le cokrigeage est l'extension géostatistique multivariable du krigeage. Il inclut les différentes variations géostatistiques de type : simple, ordinaire, universel, disjonctif, d'indicatrice et probabiliste. Ici il est question du Cokrigeage ordinaire (CKO).

Le CKO est une méthode d'interpolation spatiale permettant de modéliser une FAI. $M^{l}_{(s_g)}$ en tenant compte de ses propriétés d'autocorrélation intrinsèque, et parallèlement, de l'information de FAI auxiliaires $M^{k}_{(s_g)}$, $\forall\, k \in \mathbb{N}^{*}_{+}$ et en particulier de leurs caractéristiques d'autocorrélations *simples* et *croisées*. L'introduction des $M^{k}_{(s_g)}$ a pour dessein l'amélioration des prédictions spatiales sur les sites inconnus s_o du champ lorsque la variable d'intérêt est sous échantillonnée. Le krigeage est un modèle statistiquement plus frustre que le cokrigeage, i.e. statistiquement meilleur en terme de variance puisqu'il utilise une seule variable explicative et qu'il nécessite l'estimation d'un minimum de paramètres.
Le cokrigeage permet d'interpoler les valeurs inconnues d'une v.r. en l'assimilant à une FAI. Il possède les mêmes propriétés que celles du KO, i.e. qu'il est sans biais, de variance minimale et les estimés sont à valeurs réelles dès lors que $M^{l}_{(s)} \in \mathbb{R}$. Ces caractéristiques sont supposées vraies, du moins à des portées circonscrites par la taille du voisinage de krigeage défini indépendamment pour chaque FAI.

En théorie, l'estimateur du cokrigeage ne fournit jamais des prédictions de moins bonne qualité que celles du krigeage. En effet, si les corrélations spatiales croisées sont nulles alors l'opérateur CKO donne des prédictions identiques à celle du KO (Wackernagel, 2003).
Cependant, dans la pratique les choses sont plus complexes. Pour chaque site interpolé, i.e. $\forall\, s_o \in \mathcal{D}(\omega)$, l'estimateur du CKO, à l'aune de celui du KO, permet d'utiliser l'information contenue dans des v.r. auxiliaires. Mais l'introduction des variables auxiliaires induit de la variabilité et nécessite l'estimation d'un nombre de paramètres plus grand, qui est directement proportionnel au nombre de v.r. auxiliaires considérées. En somme, le CKO réduit le biais d'une estimation qui serait effectuée par KO mais en contrepartie augmente son erreur en terme de variance tout en préservant un gain dans ce compromis (Marcotte, 2008).
A supposer qu'une v.r. soit modélisée par une FAI : $M^{l}_{(s)}$ en utilisant l'opérateur du CKO dans un cadre bivarié, alors le cadre géostatistique est le suivant : la FAI d'intérêt est modélisée par : $M^{l}_{(s)} = m^{l} + \varepsilon^{l}(s)$, la FAI auxiliaire par $M^{k}_{(s)} = m^{k} + \varepsilon^{k}(s)$, avec $\{m^{l}, m^{k}\} \in \mathbb{R}^2$ des constantes inconnues non nulles, et $\{\varepsilon^{l}(s), \varepsilon^{k}(s)\} \sim \mathcal{BBF}\left(0; \{\sigma^2_{\{l,k\}} < +\infty\}\right)$ des bruits blancs faibles (BBF) de moyenne nulle et de variance inconnue mais constante ; ce qui implique un espace d'Hilbert, qui assure les propriétés statistiques suivantes :

$$\mathbb{E}\left(M^{l}_{(s)}, M^{k}_{(s)}\right) = \mathbb{E}\left(\mu^{l} + \varepsilon^{l}(s)\right), \mathbb{E}\left(\mu^{k} + \varepsilon^{k}(s)\right) = \left(m^{l}, m^{k}\right)$$
$$\mathbb{V}\left(M^{l}_{(s)}, M^{k}_{(s)}\right) = \mathbb{V}\left(\mu^{l} + \varepsilon^{l}(s)\right), \mathbb{V}\left(\mu^{k} + \varepsilon^{k}(s)\right) = \left(\sigma^{l}, \sigma^{k}\right)$$

L'opérateur CKO converge vers la valeur vraie de $M^{l}_{(s)}$, en utilisant $M^{k}_{(s)}$ et en incorporant l'idée du hasard, par le biais de fluctuations aléatoires. La partie de la variance estimable par le biais des autocorrélations et des corrélations croisées est incorporée ici dans les parties déterministes m^{l} et m^{k} afin de simplifier les notations. Toutefois, on remarque que le CKO introduit au moins deux erreurs aléatoires (dans le cas bi-varié) – qui, dans le pire des cas, s'additionnent - alors qu'en KO il n'y en a qu'une.

Par conséquent, l'utilisation CKO n'a de sens que lorsque la FAI d'intérêt : $M^{l}_{(s)}$ est *défaillante, i.e.* que l'opérateur du KO ne permet pas fournir une interpolation spatiale consistante sur l'intégralité de $\mathcal{D}(\omega)$ et qu'il existe des FAI auxiliaires $M^{k}_{(s)}$ stationnaires et grevées de corrélations simples et croisées avec $M^{l}_{(s)}$. Dans ce cas le CKO fournit un meilleur estimateur que le KO. La figure subséquente illustre l'idée des corrélations croisées entre $M^{l}_{(s)}$ et $M^{k}_{(s)}$ sur une des trois dimensions de l'espace géographique.

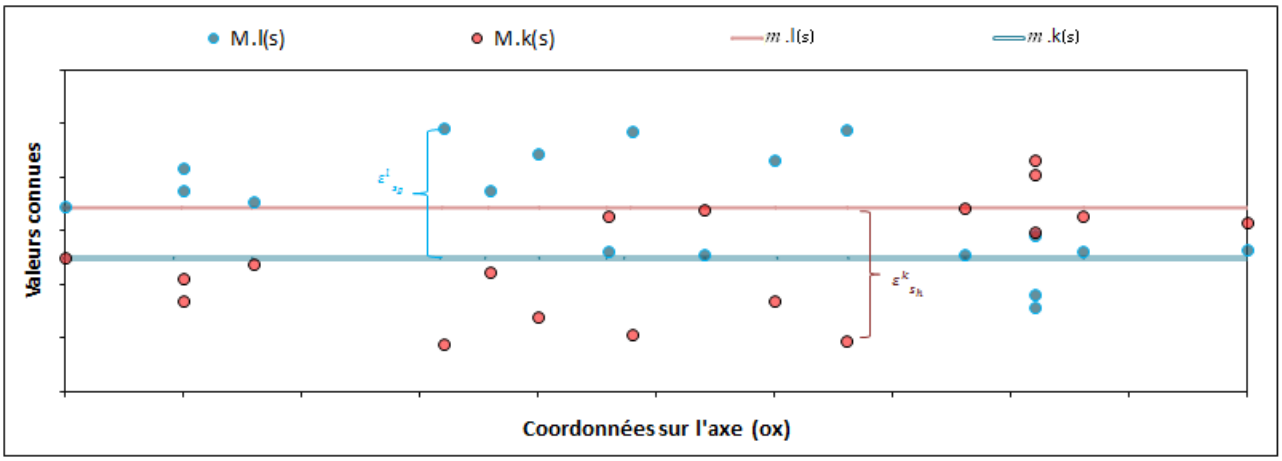


Figure 296 : Principe de l'analyse unidimensionnelle des corrélations croisées

On dira que $M^{l}_{(s)}$ et $M^{k}_{(s)}$ entretiennent des corrélations croisées négatives lorsque $M^{l}_{(s)}$ est en-dessous de m^{l} et que $M^{k}_{(s)}$ est au-dessus de m^{k} ; $M^{l}_{(s)}$ et $M^{k}_{(s)}$ sont corrélées positivement dans le cas contraire. L'utilisation judicieuse des corrélations spatiales croisées est la clé de voute du CKO. A l'inverse, lorsque $M^{l}_{(s)} \approx m^{l}$, que $M^{k}_{(s)} \approx m^{k}$ et une fois que les FAI sont centrées réduites, si elles ne s'entrecroisent pas alors il n'y a pas de corrélations croisées entre $M^{l}_{(s)}$ et $M^{k}_{(s)}$ – du moins sur l'axe considéré - et dans ce cas, le CKO ne présente aucun intérêt. Au contraire. Dans l'illustration graphique précédente les réalisations connues de : $M^{l}_{(s)}$ et $M^{k}_{(s)}$ sont mesurées sur des sites s_g identiques. Mais cette condition n'est pas nécessaire. Il s'agit même d'un cas particulier (ESRI, 2013).

A l'instar du KO le CKO utilise des semi-variogrammes $\gamma^{l}(h), \gamma^{k}(h)$ ou des fonctions de covariance $\mathbb{C}^{l}(h), \mathbb{C}^{k}(h)$ pour modéliser les autocorrélations intrinsèques de chaque variable. Dans ce cas, les champs de vecteurs sont définis dans $\mathbb{R}^{1}$. Ensuite pour modéliser les corrélations croisées on utilise des variogrammes-croisés $\gamma^{l,k}(h)$ ou des covariances-croisées $\mathbb{C}^{l,k}(h)$, dont les champs vectoriels sont définis dans $\mathbb{R}^{2}$. L'ajustement à des modèles théoriques passe par l'estimation de nombreux paramètres à partir de leurs contreparties empiriques, i.e. que les modèles sont estimés conditionnellement aux mesures d'échantillonnage disponibles : $m^{l}_{(s)}, m^{k}_{(s)} \forall\, k \in \mathbb{N}^{*}_{+}$. L'optimisation de l'ajustement nécessite parfois des transformations mathématiques des réalisations fragmentaires des FAI ou encore la correction d'éventuelles dérives spatiotemporelles, e.g. lorsque les $m^{l}_{\cdot}$ et $\varepsilon^{l}(s)$ (et/ou $m^{k}_{\cdot}$ et $\varepsilon^{k}(s)$) ou que les termes d'erreurs $\left(\varepsilon^{l}(s), \varepsilon^{k}(s)\right)$ sont corrélés (Matheron, 1989).

Les covariances permettent d'analyser les autocorrélations spatiales *simples ou croisées* des FAI. Par définition, plus les réalisations spatiales des FAI sont proches, plus elles se ressemblent. A l'inverse,

plus leurs réalisations sont éloignées, moins elles sont auto-corrélées, donc plus elles dissemblent. En pratique, l'estimation des autocorrélations croisées $\mathbb{C}^{l,k}(h)$ s'effectue à partir des réalisations fragmentaires des FAI : $m^{l}_{(s_i)}$ et $m^{l}_{(s_i)}$ par le biais des corrélogrammes croisés empiriques : $\hat{\mathbb{C}}^{l,k}(h)$. Généralement, les autocorrélations croisées maximales sont observées à l'origine ou à son voisinage, i.e. pour $\hat{\mathbb{C}}^{l,k}(h \approx 0)$. Ensuite elles décroissent proportionnellement à la distance si tant est que : $\{h = (s_i - s_j) = \text{"petit"}\}$. Au-delà d'une certaine distance, i.e. $a = \{h = \text{"grand"}\}$ elles sont annihilées, à tel point les corrélations croisées empiriques fluctuent aléatoirement autour d'une valeur nulle, i.e. $\hat{\mathbb{C}}^{l,k}(h = \text{"grand"}) \approx 0$. Comme pour le variogramme on appelle *portée* la distance minimale à laquelle le corrélogramme devient nul en moyenne, i.e. à laquelle $\underline{\mathbb{C}^{l}_{a}}(\{h \geq a\}) \approx 0$; qui en pratique s'estime conditionnellement au covariogramme croisé empirique, de sorte que :

$$\hat{a} = \underset{\forall h \in \mathbb{R}}{\operatorname{argmin}} \left\{ \left| \hat{\mathbb{C}}^{l,k}(h) \right| \leq \left\{ \varepsilon \longrightarrow 0^{+} \right\} \right\}$$

En somme, la covariance décroit avec la distance : h et de fait, permet de caractériser les similarités spatiales.

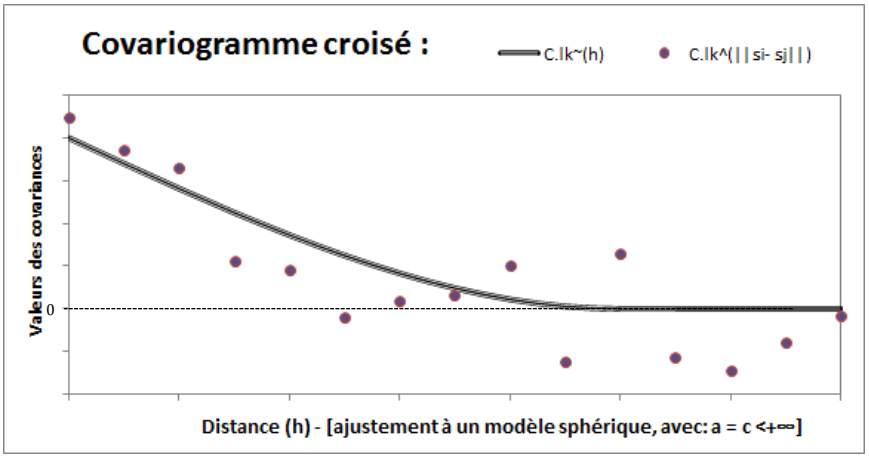


Figure 297 : Schéma de principe d'un covariogramme croisé

La fonction de covariance croisée $\mathbb{C}^{l,k}(h)$ est définie, dans le cadre de l'hypothèse de stationnarité d'ordre deux conjointe, par un ensemble de N FAI : $M^{k}_{(s)}$ admettant les propriétés mathématiques suivantes :

$$\begin{cases} \mathbb{E}\left[M^{k}_{(s)}\right] = m^{k} \in \mathbb{R}^{n_s^l} \\ \mathbb{E}\left[\left(M^{k}_{(s)} - M^{k}\right) \cdot \left(M^{l}_{(s+h)} - M^{l}\right)\right] = \mathbb{C}^{l,k}(h) \end{cases}, \forall\, s \in \mathcal{D}(\omega), l \neq k \in \{1, \dots, N\}$$

En l'occurrence : $M^{l}_{(s)}$ est la FAI d'intérêt tel : $\{M^{l}_{(s)} \supset \cup_{k=1}^{N}(M^{k}_{(s)})\}$; La moyenne de chaque variable $M^{k}_{(s)}$ est supposée invariante en tout point de $\mathcal{D}(\omega)$, du moins à l'intérieur d'un certain voisinage. La covariance d'une paire de réalisations des $M^{k}_{(s)}$ ne dépend que du vecteur différence, i.e. que $\mathbb{C}^{l,k}(s_i, s_j) = \mathbb{C}^{l,k}(s_i - s_j)$. La fonction de covariance multi-variable est définie par

$$\mathbb{C}^{l,k}(h) = \sum_{l=1}^{N}\sum_{k=1}^{N}\sum_{i=1}^{n_i}\sum_{j=1}^{n_j} \lambda_i^l \cdot \lambda_j^k \cdot \mathbb{C}^{l,k}(s_i - s_j)$$

Où λ_i^l et λ_j^k sont les $n_i + 1$ pondérateurs de la fonction de covariance croisée qui n'est pas *a priori* une fonction paire. Cependant l'égalité suivante est toujours valable : $\mathbb{C}^{l,k}(h) = \mathbb{C}^{k,l}(-h)$; Le maximum de la fonction de covariance croisée est atteint à $\mathbb{C}^{l,k}(\{h = 0\})$ ou au voisinage de l'origine et résulte, le plus souvent, d'un *effet retard spatiotemporel* (Wackernagel Hans, 1993).

Le variogramme croisé $\gamma^{l,k}(h)$ *fait son apparition dans le cadre de l'hypothèse de stationnarité intrinsèque conjointe* de N FAI. Sous cette condition on peut écrire :

$$\begin{cases} \mathbb{E}\left[\mathrm{M}^{\mathrm{k}}_{(\mathrm{s}_i)} - \mathrm{M}^{\mathrm{k}}_{(\mathrm{s}_j)}\right] = 0 \\ \mathbb{COV}\left[\left(\mathrm{M}^{\mathrm{l}}_{(\mathrm{s}_i)} - \mathrm{M}^{\mathrm{l}}_{(\mathrm{s}_j)}\right); \left(\mathrm{M}^{\mathrm{k}}_{(\mathrm{s}_i)} - \mathrm{M}^{\mathrm{k}}_{(\mathrm{s}_j)}\right)\right] = 2 \cdot \gamma^{\mathrm{l,k}}\left(\left\|\mathrm{s_i} - \mathrm{s_j}\right\|\right) \end{cases} \quad \forall\, \mathrm{s}_i, \mathrm{s}_j \in \mathcal{D}(\omega), l, k \in \{1, \dots, N\}$$

Le variogramme croisé est par définition une fonction paire

$$\gamma^{\mathrm{l,k}}(\mathrm{h}) = \frac{1}{2} \cdot \mathbb{E}\left[\left(\mathrm{M}^{\mathrm{l}}_{(s)} - \mathrm{M}^{\mathrm{l}}_{(\mathrm{s+h})}\right) \cdot \left(\mathrm{M}^{\mathrm{k}}_{(\mathrm{s})} - \mathrm{M}^{\mathrm{k}}_{(\mathrm{s+h})}\right)\right], \quad \text{avec: } \mathrm{h} = \left(\mathrm{s_i} - \mathrm{s_j}\right)$$

La relation suivante permettant de lier le variogramme croisé et le covariogramme croisé, dans le cadre de l'hypothèse de stationnarité intrinsèque conjointe, est donnée par :

$$\gamma^{\mathrm{l,k}}(\mathrm{h}) = \mathbb{C}^{\mathrm{l,k}}(\{\mathrm{h} = 0\}) - \frac{1}{2} \cdot \left(\mathbb{C}^{\mathrm{l,k}}(\mathrm{h}) + \mathbb{C}^{\mathrm{l,k}}(-\mathrm{h})\right)$$

Par décomposition de la fonction de covariance croisée on s'aperçoit que le variogramme croisé n'incorpore que les termes pairs de celle-ci. Moralité, une partie de l'information est perdue (i.e. celle permettant de mesurer le degré d'asymétrie des covariances croisées que certains processus peuvent dessiner en fonction de l'orientation d'analyse). Cette spécificité géostatistique dans le cadre multi-variable suggère le gros handicap du variogramme croisé. A l'aune du KO, en CKO, la prégnance du variogramme sur le diagramme de covariance est très discutable (Wackernagel Hans, 1993).

La matrice $\mathbb{C}^{\mathrm{l,k}}(\mathrm{h})$ de FAI conjointement stationnaire est définie semi-positive hermitienne. La généralisation multivariable et multidimentionnelle du théorème de Bochner assure que les covariances croisées $\mathbb{C}^{\mathrm{l,k}}(\mathrm{h})$ – et bien sûr les covariances simples $\mathbb{C}^{\mathrm{k,k}}(\mathrm{h})$ – sont admissibles lorsque la matrice des fonctions de covariance admet une densité spectrale $\mathrm{f}^{\mathrm{l,k}}(\mathrm{w})$. En se limitant à des fonctions de covariance intégrables en valeur absolue, i.e. lorsque $\mathrm{M}^{\mathrm{k}}_{(\mathrm{s})} \in \mathbb{R}^{\mathrm{N}}$ et pour lesquelles une densité spectrale existe et peut être inversée par une transformée de Fourrier, on a alors la relation :

$$\mathrm{f}^{\mathrm{l,k}}(\mathrm{w}) = \frac{1}{(2\pi)^n} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} e^{-i\mathrm{w}^{\mathrm{T}}h} \mathbb{C}^{\mathrm{l,k}}(\mathrm{h})\, d\mathrm{h}$$

Lorsque pour toutes les fréquences $\mathrm{w}\,[2\pi]$, la matrice des densités spectrales des fonctions de covariance est définie semi-positive, i.e. que ses valeurs propres sont positives, cela implique la relation de Cauchy-Schwarz : $\left|\mathrm{f}^{\mathrm{l,k}}(\mathrm{w})\right|^2 \leq \mathrm{f}^{\mathrm{k,k}}(\mathrm{w}) \cdot \mathrm{f}^{\mathrm{l,l}}(\mathrm{w})$ et par conséquent que le modèle CKO peut être spécifié conditionnellement aux réalisations fragmentaires de $\mathrm{M}^{\mathrm{l}}_{(s)}$ et des $\mathrm{M}^{\mathrm{k}}_{(s)}$ (Wackernagel Hans, 1993).

Ensuite il convient d'ajuster chaque FAI à un modèle de variogramme à l'instar de ce que l'on fait pour un KO, i ;e que $\tilde{\gamma}^{\mathrm{l}}(\mathrm{h})$ est spécifié à partir de sa contrepartie empirique $\hat{\gamma}^{\mathrm{l}}(\mathrm{h})$ estimée conditionnellement aux $\mathrm{m}^{\mathrm{l}}_{(\mathrm{s}_i)}$ et pour les variables auxiliaires : les $\tilde{\gamma}^{\mathrm{k}}(\mathrm{h})$ sont ajustés aux $\hat{\gamma}^{\mathrm{k}}(\mathrm{h})$ estimés conditionnellement aux $\mathrm{m}^{\mathrm{k}}_{(\mathrm{s}_i)}, k = \{1, \dots, (N-1)\}$. Ensuite il faut spécifier des modèles mathématiques admissibles $\tilde{\mathbb{C}}^{\mathrm{l,k}}(\mathrm{h})$ conditionnellement aux $\hat{\mathbb{C}}^{\mathrm{l,k}}(\mathrm{h}),\ \forall l, k \in \{1, \dots, N\}$ (Wackernagel Hans, 1993).

A l'instar du KO les variogrammes directionnels $\hat{\gamma}^{\mathrm{l}}(\mathrm{h};\theta)$, $\hat{\gamma}^{\mathrm{k}}(\mathrm{h};\theta)$ ou les corrélogrammes croisés directionnels $\hat{\mathbb{C}}^{\mathrm{l,k}}(\mathrm{h};\ \theta)$ permettent de mettre en évidence d'éventuelles anisotropies, et si besoin est, de les corriger. Afin de ne pas alourdir les écritures cette spécificité est supposée et systématiquement intégrée. La mise en œuvre d'un CKO, dès lors que $\{\mathrm{N} > 2\}$, est ô combien plus complexe que celle d'un KO, à tel point que dans le logiciel ArcGis.10 seul le *Modèle Corégionalisation Linéaire de CKO* est implémenté (ESRI, 2012).

En d'autres termes, l'utilisation du logiciel ArcGis.10 est limitée en matière de modélisation géostatistique multivariable à cause de *la contrainte d'admissibilité*. En effet, les algorithmes implémentés acculent à utiliser des modèles « simples » de CKO. Et donc par extension à faire allégeance à des hypothèses très restrictives.

Le Modèle de CKO à N v.r. dans le cadre de la *Corégionalisation Linéaire* implique l'ajustement de P modèles élémentaires et suppose que les $\left\{n_{\mathbb{C}} = \binom{\mathrm{N}+1}{2}\right\}$ variogrammes simples : $\left\{\gamma^{\mathrm{k}}(\mathrm{h}) \cup \gamma^{\mathrm{l}}(\mathrm{h})\right\}$ et *covariogrammes croisés* $\mathbb{C}^{\mathrm{l,k}}(\mathrm{h})$ peuvent être modélisés comme la somme d'une seule combinaison

linéaire de modèles mathématiques *admissibles*. Ensuite on consent à une hypothèse encore plus forte, à savoir que *chaque covariance est symétrique en* h. Sous ces deux postulats grevés de l'hypothèse de stationnarité intrinsèque conjointe, tous les modèles de variogrammes peuvent être ramenés à des modèles de covariances *simples* grâce à l'égalité suivante : $\gamma^{k}(h) = \left(\mathbb{C}^{k}(0) - \mathbb{C}^{k}_{pair}(h)\right)$ - sous contrainte d'utilisation de fonctions mathématiques élémentaires admissibles ajustables aux variogrammes et covariogrammes empiriques on peut écrire :

$$\left|\mathbb{C}^{k',k}(h)\right| = B_1\mathbb{C}_1(h) + B_2\mathbb{C}_2(h) + \cdots + B_P\mathbb{C}_P(h), \qquad \forall k', k = \{1, , \ldots l, \ldots, N\}$$

Où les $B_1, \ldots, B_P$ sont des matrices de : $\dim(B_{\_}) = (N \times N)$ définies symétriquement positives afin d'assurer l'admissibilité du modèle général. En contrepartie, la dérivation des fonctions de covariance est très difficile à réaliser sur le plan mathématique. En toute rigueur, dès lors que $\{N > 2\}$ et que l'on utilise des modèles de variogrammes simples ou de covariogrammes croisés différents – même s'ils sont admissibles individuellement - il est nécessaire de vérifier la validité du modèle global dans le domaine spectral. Quelques algorithmes existent mais ils sont rarement utilisés à cause de leur instabilité (Marcotte, 2008).
Il convient donc d'insister sur le fait que l'hypothèse des covariances symétriques suppose que $\{\mathbb{C}^{l,k}(h) = \mathbb{C}^{k,l}(-h)\}$, autrement dit, que dans la direction de h, les réalisations des variables auxiliaires précèdent ou suivent toujours, en moyenne, celles de la variable d'intérêt. Or, dans la pratique ce n'est pas forcément évident, du moins sur l'intégralité le champ. Les écueils de cette hypothèse s'apprécient d'ailleurs visuellement sur les diagrammes des covariances expérimentales.

Ainsi, les stratégies de modélisation des FAI diffèrent entre le KO et le CKO et dans la pratique rien ne garantit que la complexité du CKO apporte un gain significatif, ni même qu'il ne détériore pas les prédictions d'un KO. La structure spatiale des $m^{k}_{(s)}, k = \{1, \ldots, N\}$ permet néanmoins de choisir l'opérateur géostatistique le plus adapté vis-à-vis de la répartition spatiale des données disponibles. Selon les caractéristiques des $m^{k}_{(s)}$ on distingue trois cas : *l'isotopie* - tous les points d'échantillonnage sont identiques pour la variable d'intérêt et les variables auxiliaires ; *l'hétérotopie* - les variables sont mesurées sur des ensembles de sites disjoints, et enfin, *l'hétérotopie partielle* - une partie des mesures est effectuée sur des sites communs aux deux variables et l'autre pas (Wackernagel Hans, 1993).

De là découle la prégnance d'un modèle de covariance sur le celui du semi-variogramme. En effet, le variogramme croisé n'est utilisable que dans le cadre d'une isotopie. Or, généralement lorsqu'il s'agit de modéliser une FAI dont les réalisations fragmentaires des $m^{l}_{(s)}$ sont défaillantes, on introduit des $m^{k}_{(s)}$ supplétives ayant des distributions spatiales plus denses et plus uniformes. Par conséquent, le recours au pseudo variogramme croisé – PVC (Wackernagel Hans, 1993) – ou au covariogramme croisé est nécessaire. Par ailleurs, il convient de remarquer que le PVC n'est pas implémenté dans ArcGis.10 (ESRI, 2013). Par conséquent tous les CKO s'effectueront dans le cadre du *modèle de corégionalisation linéaire (MCL)* par ajustement de $\hat{\gamma}^{l}(h)$, des $\hat{\gamma}^{k}(h)$ et des $\hat{\mathbb{C}}^{l,k}(h)$ à un modèle de covariance. IL convient de noter cependant que pour la mise en œuvre d'un CKO, l'utilisation des covariogrammes est vivement recommandée, en particulier dans la thèse sur les modèles linéaires géostatistiques multivariables de Wackernagel - (Wackernagel Hans, 1985), dans le cours de géostatistique de Marcotte - (Marcotte, 2008) et plus généralement par la plupart des auteurs cités dans les revues des méthodes d'interpolation spatiale multidimensionnelles (Arnaud et Emery, 2000).

L'attrait pour le modèle de covariogramme ne se fonde pas uniquement sur des considérations théoriques. En effet, si le variogramme croisé permet de s'affranchir de l'estimation de la moyenne il requiert des variables mesurées aux mêmes localisations. De plus, il est incapable de modéliser les corrélations asymétriques. Quant au PVC, bien qu'il permette d'utiliser des variables mesurées à des localisations qui ne coïncident pas forcément, il ne permet pas non plus de modéliser les corrélations négatives. Et surtout il nécessite une normalisation des données lorsque les systèmes unitaires des $m^{l}_{(s_g)}$ et les $m^{k}_{(s_g)}$ diffèrent- ce qui est presque toujours le cas. Par conséquent le choix d'un modèle de covariances croisées s'impose comme la meilleure stratégie pour le CKO.

L'estimateur du CKO est linéaire et présente une forme analogue à celui d'un KO :

$$\mathrm{M}^{\mathrm{l}}_{(\mathrm{s}_o)} = \sum_{k=1}^{N} \sum_{i=[1]}^{\left[\mathrm{n}^{\mathrm{k}}_{(\mathrm{V})}\right]} \lambda_i^k \cdot \mathrm{M}^{\mathrm{k}}_{(\mathrm{s}_i),} \quad \forall\, k = \{1, \dots, \mathrm{N}\}, l \in \{1, \dots, \mathrm{N}\}$$

L'indice : l est celui de la variable d'intérêt à interpoler et de fait le nombre de ses réalisations fragmentaires dépend de k – ce qui permet d'inclure dans l'équation les cas d'hétérotopie (Wackernagel Hans, 1993).

Il convient de notifier, que la formule tient compte, pour chaque variable, de la taille du voisinage de krigeage $\mathrm{i} \in \mathrm{V_k}(\mathrm{s_o}) = \left\{[1], \dots, \left[\mathrm{n}^{\mathrm{k}}_{(\mathrm{V})}\right]\right\}$ – afin de prendre en compte la portée des autocorrélations spatiales, les structures spatiales des données d'échantillonnage et la correction des éventuelles anisotropies géométriques et zonales (Baillargeon, 2005).

L'hypothèse de stationnarité intrinsèque conjointe inhérente au CKO implique la construction d'un estimateur spatial sans biais et de variance minimale. Le choix des pondérateurs du modèle de CKO s'effectue conditionnellement aux observations disponibles de chacune des N v.r. utilisées et assure que :

$$\begin{cases} \mathbb{E}\left[\widehat{\mathrm{M}}^{\mathrm{l}}_{(\mathrm{s_o})} - \mathrm{M}^{\mathrm{l}}_{(\mathrm{s_o})}\right] = 0 \\ \mathbb{V}\left[\widehat{\mathrm{M}}^{\mathrm{l}}_{(\mathrm{s_o})} - \mathrm{M}^{\mathrm{l}}_{(\mathrm{s_o})}\right] = \text{"minimum"} \end{cases}$$

En développant l'expression contenue dans l'opérateur espérance, la contrainte de non biais peut-être exprimée par mesure de Dirac de somme unitaire pour $\mathrm{M}^{\mathrm{l}}_{(\mathrm{s})}$ et de somme nulle pour les variables auxiliaires $\mathrm{M}^{\mathrm{k}}_{(\mathrm{s})} \forall\, k = \{1, \dots, (\mathrm{N}-1)\}/\{l\}$. De fait, la condition de non biais s'exprime de la façon suivante :

$$\sum_{i=[1]}^{\left[\mathrm{n}^{\mathrm{k}}_{(\mathrm{V})}\right]} \lambda_i^k = \delta_{lk} = \begin{cases} 1 & \text{si } k = l \\ 0 & \text{sinon} \end{cases}$$

En développant l'expression contenue dans l'opérateur variance, en formant des Lagrangiens : μ_{k} pour tenir compte des contraintes de non biais et en annulant les dérivées partielles des poids du système de CKO on obtient :

$$\sum_{k=1}^{N} \sum_{i=[1]}^{\left[\mathrm{n}^{\mathrm{k}}_{(\mathrm{V})}\right]} \lambda_i^k \cdot \mathbb{C}^{l,k}\left(\left\{\mathrm{h}_{ij} = \left\|\mathrm{s}_i - \mathrm{s}_j\right\|\right\}\right) + \mu_{\mathrm{k}} = \mathbb{C}^{l,k}(\{\mathrm{h}_{io} = \|\mathrm{s}_i - \mathrm{s}_0\|\})$$

Le système de CKO est désormais écrit. Sous forme matricielle il se présente de la façon suivante$\mathbb{C}^{lk}_{ij} \lambda^k_{\mu} = \mathbb{C}^{lk}_{io}$, où : la matrice $\mathbb{C}^{\mathrm{lk}}_{\mathrm{ij}}$est de taille $\left(\sum_{\mathrm{k}=1}^{\mathrm{N}}\left[\mathrm{n}^{\mathrm{k}}_{(\mathrm{V})}\right] + \mathrm{N}\right) \times \left(\sum_{\mathrm{k}=1}^{\mathrm{N}}\left[\mathrm{n}^{\mathrm{k}}_{(\mathrm{V})}\right] + \mathrm{N}\right)$, et les vecteurs $\lambda^{\mu}_{\mathrm{k}}$ et $\mathbb{C}^{lk}_{io}$ sont de taille : $\left(\sum_{\mathrm{k}=1}^{\mathrm{N}}\left[\mathrm{n}^{\mathrm{k}}_{(\mathrm{V})}\right] + \mathrm{N}\right)$. Sa forme est analogue à celle d'un KO mais la quantité d'informations stockées est beaucoup plus grande, soit :

$$\begin{pmatrix} \left(\hat{\mathbb{C}}^{1,1}(\mathrm{h}_{\mathrm{ij}})\right) & \cdots & \left(\hat{\mathbb{C}}^{1,\mathrm{N}}(\mathrm{h}_{\mathrm{ij}})\right) & (\mathbb{I}^1) & \cdots & (\mathbb{O}^N) \\ \vdots & \ddots \left(\hat{\mathbb{C}}^{\mathrm{k,k}}(\mathrm{h}_{\mathrm{ij}})\right) \ddots & \vdots & \vdots & \ddots (\mathbb{I}^k) \ddots & \vdots \\ \left(\hat{\mathbb{C}}^{\mathrm{N},1}(\mathrm{h}_{\mathrm{ij}})\right) & \cdots & \left(\hat{\mathbb{C}}^{\mathrm{N,N}}(\mathrm{h}_{\mathrm{ij}})\right) & (\mathbb{O}^1) & \cdots & (\mathbb{I}^N) \\ (\mathbb{I}^1)^t & \cdots & (\mathbb{O}^N)^t & (\mathbb{O}^1) & \cdots & (\mathbb{O}^N) \\ \vdots & \ddots (\mathbb{I}^k)^t \ddots & \vdots & \vdots & \ddots (\mathbb{O}^k) \ddots & \vdots \\ (\mathbb{O}^1)^t & \cdots & (\mathbb{I}^N)^t & (\mathbb{O}^1) & \cdots & (\mathbb{O}^N) \end{pmatrix} \begin{pmatrix} (\lambda^1) \\ \vdots \\ (\lambda^k) \\ \vdots \\ (\lambda^N) \\ \mu_1 \\ \vdots \\ \mu_{\mathrm{k}} \\ \vdots \\ \mu_{\mathrm{N}} \end{pmatrix} = \begin{pmatrix} \left(\mathbb{C}^{\mathrm{l},1}(\mathrm{h}_{\mathrm{i,io}})\right) \\ \vdots \\ \left(\mathbb{C}^{\mathrm{l,k}}(\mathrm{h}_{\mathrm{i,io}})\right) \\ \vdots \\ \left(\mathbb{C}^{\mathrm{l,N}}(\mathrm{h}_{\mathrm{i,io}})\right) \\ 1 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Avec $\hat{\mathbb{C}}^{\mathrm{k,k'}}(\mathrm{h}_{\mathrm{ij}})$ des matrices de variances covariances expérimentales de dimension $\left(\left[\mathrm{n}^{\mathrm{k}}_{(\mathrm{V})}\right] \times \left[\mathrm{n}^{k'}_{(\mathrm{V})}\right]\right), \forall\, \mathrm{k}, \mathrm{k'} \in \{1, \dots l, \dots, N\}$; Et $(\mathbb{I}^k)$ des vecteurs colonnes de « un » de longueur $\left(\left[\mathrm{n}^{\mathrm{k}}_{(\mathrm{V})}\right] \times 1\right)$ et $(\mathbb{O}^k)$ des vecteurs colonnes de « zéro » de longueur $\left(\left[\mathrm{n}^{\mathrm{k}}_{(\mathrm{V})}\right] \times 1\right)$; Et les vecteurs des covariances empiriques croisées entre $\mathrm{M}^{\mathrm{l}}_{(\mathrm{s})}$ et les $\mathrm{M}^{\mathrm{k}}_{(\mathrm{s})}$ au niveau de s_o $\left(\mathbb{C}^{\mathrm{l,k}}(\mathrm{h}_{\mathrm{i,io}})\right)$ sont aussi de dimension $\left(\left[\mathrm{n}^{\mathrm{k}}_{(\mathrm{V})}\right] \times 1\right)$. S'ensuit

l'ajustement des $\hat{\mathbb{C}}^{l,k}(\mathrm{h})$ à un modèle $\tilde{\mathbb{C}}^{l,k}(\mathrm{h})$ admissible. En conséquence de quoi l'estimation des vecteurs de paramètres (λ^k)de dimension $\left(\left[\mathrm{n}_{(\mathrm{V})}^{\mathrm{k}}\right] \times 1\right)$ et des scalaires $\mu_{\mathrm{k}} \in \mathbb{R}^N$ devient aisée. Les poids du CKO sont spécifiés conditionnellement aux données disponibles, tel que $\hat{\lambda}_{\mu}^{k} = \hat{\mathbb{C}}_{ij}^{\mathrm{k},\mathrm{k}'^{-1}} \tilde{\mathbb{C}}_{io}^{lk}, \forall\, \mathrm{l}, \mathrm{k}, \mathrm{k}' \in \{1, \dots, \mathrm{N}\}$. L'estimateur du CKO s'écrit :

$$\widehat{\mathrm{m}}_{(\mathrm{s}_o)}^{\mathrm{l}} = \sum_{k=1}^{N} \sum_{i=[1]}^{\left[\mathrm{n}_{(\mathrm{V})}^{\mathrm{k}}\right]} \hat{\lambda}_i^k \cdot \mathrm{m}_{(\mathrm{s}_i),}^{\mathrm{k}} \quad \forall\, k = \{1, \dots, \mathrm{N}\}, l \in \{1, \dots, \mathrm{N}\}$$

Quant à l'estimateur de la variance de CKO, il est donné par la relation :

$$\widehat{\sigma}^2_{\left(\widehat{\mathrm{m}}_{(\mathrm{s}_o)}^{\mathrm{l}}\right)} = \tilde{\mathbb{C}}^{l,l}(\{\mathrm{h}_o = \|\mathrm{s}_o - \mathrm{s}_o\|\}) - \sum_{k=1}^{N} \sum_{i=[1]}^{\left[\mathrm{n}_{(\mathrm{V})}^{\mathrm{k}}\right]} \hat{\lambda}_i^k \cdot \tilde{\mathbb{C}}^{l,k}(\{\mathrm{h}_{io} = \|\mathrm{s}_i - \mathrm{s}_o\|\}) - \hat{\mu}_{\mathrm{l}}$$

Lorsque le modèle de CKO est admissible et qu'il est correctement spécifié, il admet les propriétés suivantes : *Propriété de cohérence* : si une transformation linéaire est appliquée $\mathrm{m}_{(\mathrm{s})}^{\mathrm{l}}$ le CKO de la transformation linéaire $\mathrm{L}\left(\mathrm{m}_{(\mathrm{s})}^{\mathrm{l}}\right)$ est la transformation appliquée aux valeurs des estimés, i.e. $\widehat{\mathrm{m}}_{(\mathrm{s}_o)}^{\mathrm{l}} = \mathrm{L}^{-1}\left(\mathrm{L}\left(\widehat{\mathrm{m}}_{(\mathrm{s}_o)}^{\mathrm{l}}\right)\right)$; *Propriété d'optimisation multi-variée* : La variance d'estimation de CKO est toujours inférieure ou égale à celle d'un KO, i.e. $\widehat{\sigma}^2_{\mathrm{s}_o,\mathrm{CKO}} \leq \widehat{\sigma}^2_{\mathrm{s}_o,\mathrm{KO}}$; *Propriété de conservation* : si les densités d'échantillonnage des variables secondaires $\mathrm{m}_{(\mathrm{s})}^{\mathrm{k}}$ sont identiques ou inférieures à celles de $\mathrm{m}_{(\mathrm{s})}^{\mathrm{l}}$ et que les corrélations directes ou croisées sont proportionnelles alors le CKO est identique au KO.

En pratique, le modèle le plus couramment utilisé de CKO *est le modèle linéaire de corégionalisation à deux variables*. Un CKO prévaut sur le KO dès lors que : la densité spatiale ou l'uniformité de la structure des réalisations fragmentaires de la FAI d'intérêt : $\mathrm{m}_{(\mathrm{s})}^{\mathrm{l}}$ est défaillante, qu'une ou plusieurs autres v.r. mesurées et mesurables sont disponibles $\mathrm{m}_{(\mathrm{s})}^{\mathrm{k}}\ \forall\, \mathrm{k} = \{1, \dots, \mathrm{N}\}/\{\mathrm{l}\}$, que ces dernières sont bien échantillonnées d'un point de vue spatial et qu'elles entretiennent des corrélations simples et croisées significatives avec la variable d'intérêt (Marcotte, 2008), i.e. pour les corrélations simples que : $\left|\hat{\rho}\left(\mathrm{m}_{(\mathrm{s})}^{\mathrm{k}}\ ; \mathrm{m}_{(\mathrm{s})}^{\mathrm{l}}\right)\right| \geq 0{,}7$ et pour les corrélations croisées :

$$\left|\left\{\hat{\rho}^{\mathrm{lk}}(\mathrm{h}=0) = \frac{\hat{\mathbb{C}}^{\mathrm{l},\mathrm{k}}(\mathrm{h}=0)}{\sqrt{\hat{\mathbb{C}}^{l,l}(\mathrm{h}=0) \cdot \hat{\mathbb{C}}^{k,k}(\mathrm{h}=0)}}\right\}\right| \geq 0{,}5$$

L'utilisation du variogramme croisé est à proscrire sauf dans le cadre de l'hétérotopie et après avoir vérifié que les covariances croisées sont bien symétriques, ce qui n'est pas forcément évident. Si les conditions énoncées sont remplies l'utilisation du variogramme croisé est préférable à celle du covariogramme croisé puisqu'il permet de s'affranchir de l'estimation de la moyenne et de corriger plus facilement les dérives, en l'occurrence par le biais des modèles sans paliers (Marcotte, 2008).

Il convient aussi de remarquer que pour effectuer un CKO, i.e. pour supputer le couple $\left(\widehat{\mathrm{m}}_{(\mathrm{s}_o)}^{\mathrm{l}}, \widehat{\sigma}^2_{\mathrm{s}_o,\mathrm{CKO}}\right)$, il est nécessaire que le voisinage de CKO soit constitué d'au moins une réalisation de $\mathrm{m}_{(\mathrm{s})}^{\mathrm{l}}$ et au moins deux observations pour chacune des v.r. auxiliaires $\mathrm{m}_{(\mathrm{s})}^{\mathrm{k}}\ \forall\, \mathrm{k} = \{1, \dots, \mathrm{N}\}/\{\mathrm{l}\}$. Si le voisinage de CKO n'intègre aucune observation supplétive l'estimateur du CKO est identique à celui du KO (Marcotte, 2008).

Lors de la spécification de la taille du voisinage, il convient de prendre garde à ce que la matrice de CKO ne soit pas *quasi-singulière* par l'inclusion sporadique de trop nombreuses valeurs d'échantillonnage $\left\{\mathrm{m}_{(\mathrm{s})}^{\mathrm{k}} \cup \mathrm{m}_{(\mathrm{s})}^{\mathrm{l}}\right\}$ situées à proximité les unes des autres (Marcotte, 2008).

Enfin les contingences les plus triviales de CKO environnementaux, décrites dans la littérature et pouvant se rapporter aux cas pratiques rencontrés dans cette thèse sont :

- *sur des supports identiques* : les températures et la topographie détaillée de la zone d'investigation. Il convient de préciser que si les températures sont biens échantillonnées elles peuvent, avec la topographie, servir à interpoler le rayonnement global ;

- *sur des supports de nature différente mais présentant tout de même des caractères analogues* : les mesures chimiques de l'activité de certains radionucléides et des évaluations globales de la radioactivité environnementale dans certains milieux (Marcotte, 2008).

A l'instar d'un KO la résolution de la couche de surface résultant d'un CKO est entièrement laissée à la discrétion de l'utilisateur. Cette spécificité est capitale. Aussi afin d'évaluer la qualité des $\widehat{m}^{l}_{(s_o)}$, i.e. la qualité des estimateurs du KO ou du CKO - la stratégie statistique la plus utilisée à cet effet est connue sous la dénomination de *procédure de validation croisée*.

Qualité des estimateurs géostatistiques

Les estimateurs du krigeage et du cokrigeage sont exacts, donc leurs prédictions ne peuvent pas être comparées aux observations à l'aune d'un modèle de régression. Il existe cependant plusieurs techniques de validation permettant d'estimer la qualité de ces modèles, à commencer par la méthode « d'échantillons test » qui est la plus robuste de toutes. Mais elle n'est utilisable qu'avec des jeux de données particulièrement grands (Baillargeon, 2005). Le « boostrap paramétrique » est une autre méthode particulièrement efficace bien qu'elle ne soit presque jamais implémentée dans les logiciels SIG (Kleijnen, 2011). A ce jour, la *validation croisée* – ou cross-validation (c.v.) - est la technique la plus répandue pour apprécier la qualité d'un krigeage. Elle n'en demeure pas moins l'une des plus performantes, en particulier lorsque le nombre de données disponibles est restreint. Cette procédure est à la fois puissante et implémentée dans la plupart des SIG, en l'occurrence dans l'extension Géostat-Analyst d'ArGis.10 (ESRI, 2013). Cette phase constitue en quelque sorte un processus de vérification des paramètres de krigeage spécifiés.
Il s'agit de valider le choix du modèle de variogramme (pour le KO) ou de covariance (pour le CKO), la calibration des paramètres ainsi que les critères de voisinage spécifiés. La procédure de validation croisée est simple à comprendre. Il s'agit de ré-estimer, à partir du modèle géostatistique, toutes les observations disponibles $m^{l}_{(s_.)} = \left(m^{l}_{(s_1)}, \ldots, m^{l}_{(n^{l}_s)}\right)$. En d'autres termes, les données d'échantillonage $m^{l}_{(s_.)}$ sont supprimées successivement, une par une, tel que $\left\{m^{l}_{(s_{-g})}\right\} = \left\{m^{l}_{(s_.)}\right\} - \left\{m^{l}_{(s_g)}\right\}$. De sorte que chacune des : $\left\{m^{l}_{(s_g)}\right\}$ va être ré-éstimée à partir de toutes les autres et des paramètres choisis. De fait, on obtient une collection de données ré-estimées $m^{l,*}_{(s_.)} = \left(m^{l,*}_{(s_1)}, \ldots, m^{l,*}_{(n^{l}_s)}\right)$. Ce sont ces dernières qui vont être comparées aux données d'échantillonnage afin de mesurer la qualité du KO. Chaque itération donne lieu à une estimation de l'erreur standard de c.v. (cross-validation) : $\tilde{\sigma}^{*}_{\left(m^{l}_{(s_g)}\right)}$.
La qualité de l'estimateur krigeage est donnée par des indicateurs calculés à partir des résidus de c.v : $e_g = \left(m^{l}_{(s_g)} - m^{l,*}_{(s_g)}\right)$. Toutefois cette mesure peut être spécieuse. Les sites isolés tendent à induire des erreurs importantes lorsque les données sont réparties de façon irrégulière dans le champ. Par conséquent des résidus normalisés de c.v. interviennent : $\varepsilon_g = \left(m^{l}_{(s_g)} - m^{l,*}_{(s_g)}\right) \Big/ \tilde{\sigma}^{*}_{\left(m^{l}_{(s_g)}\right)}$ de façon à corriger cet artéfact. Par suite, des indicateurs de c.v. permettent de caractériser la qualité du modèle de krigeage ou de cokrigeage. Les plus courants sont supputés par Geostat Analys (ESRI, 2013). On dira que le modèle est *adéquat* (Marcotte, 2008), c'est-à-dire statistiquement admissible, lorsque :

$$\text{Residial Mean} \;= \frac{1}{n^{l}_s} \cdot \sum_{g=1}^{n^{l}_s} e_g = 0$$

$$\text{Root. MSE} = \sqrt{\frac{1}{n^{l}_s} \cdot \sum_{g=1}^{n^{l}_s} \left(e_g\right)^2} = \text{"smal"}$$

Average Kriging Standard Error of Cross Validation

$$\text{AKSE. cv} = \frac{1}{n_s^l} \cdot \sum_{g=1}^{n_s^l} \tilde{\sigma}^* \left(m^l_{(s_g)} \right) \approx \{\text{Root. MSE}\}$$

$$\text{Residual Mean Standardized} = \frac{1}{n_s^l} \sum_{g=1}^{n_s^l} \varepsilon_g \approx 0$$

$$\text{Root. MSE Standardized} = \sqrt{\frac{1}{n_s^l} \cdot \sum_{g=1}^{n_s^l} (\varepsilon_g)^2} \approx 1$$

$$\forall\, s_g \in \mathcal{D}(\omega_.)$$

D'une façon générale, lorsqu'il s'agit de mettre en concurrence plusieurs modèles de KO ou de CKO, les critères de prédilection sont : Root. MSE et AKSE. cv. (Marcotte, 2008).

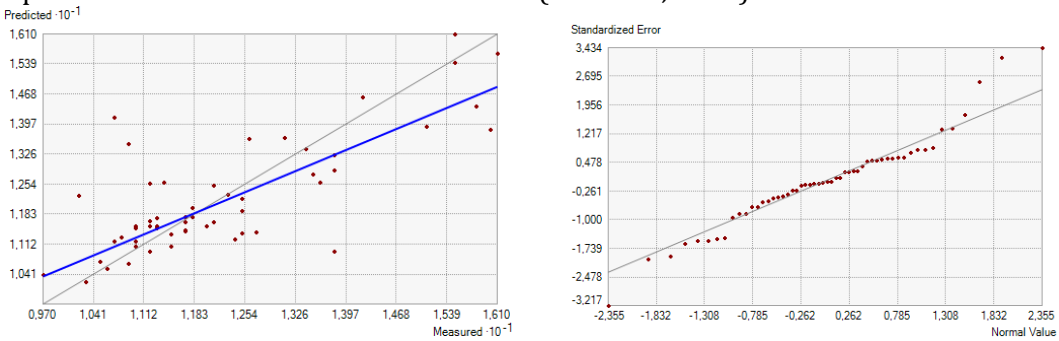


Figure 298 : Exemple d'un diagramme représentatif du nuage de points des valeurs observées et prédites (à gauche) et d'un diagramme QQ.plot Normal (à droite)

La corrélation entre les valeurs d'échantillonnage et les ré-estimées est une information intéressante qui s'apprécie par l'estimateur du coefficient de corrélation - et visuellement, sur le diagramme des résidus, par l'adéquation de la droite de régression des résidus, en bleu, et la droite standard, en pointillés noirs. Le caractère gaussien des résidus quant à lui peut être aussi estimé visuellement par le biais d'un diagramme Q-Q.plot normalisé – lorsque les quantiles des ε_g s'alignent linéairement, sur la première diagonale, (Saporta, 2006).

La validation croisée présuppose que plusieurs modélisations doivent être effectuées avant de choisir un modèle de variogramme, une technique d'ajustement des paramètres et des critères de voisinage. L'idée est bien sûr de spécifier une calibration qui optimise au mieux ces indicateurs de c.v. Il n'existe pas à ce jour de procédure automatique de sélection de modèles. D'évidence une telle stratégie serait souhaitable (Marcotte, 2008).

La mise au point de méthodes vouées à la sélection de modèles est une problématique de recherche contemporaine qui se situe au cœur de la sphère des mathématiques (Baraud, Giraud et al., 2009), dont la géostatistique fait partie.

Les scores de c.v. permettent de qualifier un modèle krigeage ou/et de cokrigeage d'adéquat. Toutefois, lorsque la procédure de c.v. permet de conjecturer que le modèle est correctement spécifié, rien ne garantit qu'il s'agisse du modèle optimal d'interpolation spatiale. Au contraire même, la probabilité de spécifier manuellement le modèle optimal est presque nulle - d'où l'intérêt d'élaborer des procédures objectives de sélection…

ANNEXE 5 : COMPLEMENTS THEORIQUES SUR LES FORETS ALEATOIRES

Les Forêts Aléatoires (FA) constituent l'aboutissement des travaux de Léo Breiman. Tout commence en 1984 : le datamining en est à ses balbutiements et Breiman, assisté par une informaticienne nommée Cluter, cristallise l'algorithme Classification And Regression Tree - CART (Breiman, Stone et al., 1984). En 1996, le développement des procédures *d'échantillonnage* type Bootstrap, couplées à la puissance des outils informatiques permet l'intégration des processus CART *en tant que méthodes d'ensemble* et un nouvel algorithme voit le jour : *Bagging predictor*, connu aussi sous le nom de *Bagging de Breiman* (Breiman. Leo, 1996). Cinq ans plus tard, des perturbations stochastiques sont implémentées dans CART. Elles randomisent l'espace des variables en intervenant au niveau de la construction des nœuds des arbres. L'intégration de cette règle CART perturbée dans la procédure de Bagging donne naissance à un nouvel algorithme : *randomForest* (Breiman, 2001).

Les performances de *randomForest* sont testées sur des jeux de données jouées et des jeux de données réelles dont les caractéristiques sont connues, et en dépit de l'absence totale de démonstration théorique, les FA font preuve d'une puissance prédictive et explicative stupéfiante. Par la suite, Léo Breiman intégra des processus de randomisation des outputs, des processus de stabilisation, des scores d'importance des variables, une stratégie de comblement des lacunes… jusqu'à ce qu'une quatrième et dernière version de *randomForest.V4* soit élaborée (Breiman, 2004). Depuis des modules complémentaires de calibration et de représentation graphique ont été introduits par différents mathématiciens. La version des FA utilisée dans le cadre cette thèse est celle du package R nommé *randomForest.V4.6-7* (Liaw, 2013).

Les FA font partie de la grande famille des *méthodes d'ensemble*, des procédures de Datamining au fond très simples mais particulièrement efficaces. L'idée des FA est de randomiser par Bootstrap un jeu de données d'apprentissage sur l'espace des individus, puis d'y appliquer une règle CART afin de générer des prédicteurs décorrélés et particulièrement performants sur une partie singulière de l'échantillon d'apprentissage ; et enfin, de former un prédicteur généralisé en agrégeant les prédictions individuelles - par une *moyenne empirique* (en régression) ou par un *vote à la majorité* (en classification). Les processus de randomisation sont la clé de voute des FA et des autres méthodes d'ensemble les plus abouties.

Avant de décliner les grands principes de l'algorithme *randomForest* il convient d'énoncer les quatre processus aléatoires les plus couramment utilisés dans les méthodes d'ensemble :
Le *Bagging* repose sur la construction d'un prédicteur unique à partir d'estimateurs individuels agrégés et construits sur des échantillons Bootstrap.
Le *Boosting* est une procédure analogue au Bagging. L'échantillonnage se fait toujours sur l'espace des individus mais *step by step*, de sorte que les tirages ne soient plus effectués dans une loi uniforme mais dans une loi de probabilité choisie *a priori.* L'idée du Boosting est de jeter le dévolu de la règle d'apprentissage sur les parties de l'espace des individus les plus difficiles à modéliser, ou les plus congruentes (Chernick, 1999).
Le Randominzing-Output consiste à altérer la *variable réponse* par des processus aléatoires contrôlés
Le Random-Subspace est en quelque sorte la pierre angulaire des FA. Ce processus stochastique perturbe aléatoirement l'espace des variables dans l'optique de créer des prédicteurs individuels biaisés mais décorrélés afin d'améliorer la variance du modèle final.

Seule la procédure *boosting* (plus performante que celle de *Bagging* mais difficilement utilisable dans la pratique) n'est pas implémentée dans les FA. Cependant *depuis qu'elles ont été introduites, les FA ont fait l'objet de plusieurs études prospectives et comparatives. Elles se sont montrées compétitives face aux méthodes intégrant le Boosting et réputées pour être une des techniques d'apprentissage des plus efficaces* (Bernard, Heutte et al., 2008).

Avant d'introduire les grands principes des FA, il convient d'énoncer sommairement ceux de l'algorithme CART. D'abord parce que CART constitue la base des FA. Et aussi, parce que l'algorithme *rpart* est utilisé dans la procédure de sélection de variables – VSURF (Genuer, Poggi et al., 2013)- lors de l'estimation du premier seuil *d'élimination des variables de bruit*.

Principes fondamentaux de la méthode CART

Classification And Regression Tree (CART) est un modèle d'inférence statistique *non paramétrique*. Cette partie s'appuie essentiellement sur la thèse de Christine Malot (Tuleau-Malot, 2005). Les principes mathématiques énoncés sont exactement ceux qui sont implémentés dans le package *rpart* du logiciel R – qui est aussi la seule version parfaitement conforme aux travaux de Léo Breiman (Therneau, Atkinson et al., 2013).

CART est un algorithme de construction d'arbres de classification et de régression. Il est donc capable de modéliser des phénomènes décrits par des variables aléatoires $Y_{\cdot}^{j}$ qualitatives ou quantitatives. Le prédicteur généré *est constant par morceau* et il prend une forme différente en fonction du contexte. En régression il s'agit d'un *modèle linéaire de type arbre de décision*; en classification il s'agit de *l'estimateur de Bayes* :

$$Y^{j} = f_{CART}^{j}\left(X_{\cdot}^{l}\right) = \begin{cases} \mathbb{E}\left[\left[\left\{Y^{j}\middle|\left(X_{\cdot}^{1}, \dots, X_{\cdot}^{p_l}\right)\right\}\right]\right], & \forall\, Y^{j} \in \mathbb{R}^{n_j} \\ \underset{\forall\, c_j \in \mathcal{C}_j}{\operatorname{argmax}}\left\{\mathbb{P}\left(Y^{j} = c_j \middle| \left(X_{\cdot}^{1}, \dots, X_{\cdot}^{p_l^j}\right)\right)\right\}, & \forall\, \mathcal{C}_j \subset \mathbb{N}^{|\mathcal{C}_j|} \end{cases}$$

Avec : $X_{\cdot}^{l} = \left(X_{i}^{1}, \dots, X_{i}^{p_l^j}\right), \forall i = \{1, \dots, n_i\}$ toutes les valeurs prises par les variables aléatoires (v.a.) potentiellement explicatives qui peuvent, elles aussi, être de nature qualitative (binaire ou multi-classe) ou quantitative (continue ou discrète) ; Et $\mathcal{C}^{j} = \{1, \dots, c_j, \dots, C_j\}$ l'ensemble des modalités de la v.a. qualitative : Y^{j}.

Le prédicteur CART : $T_{CART}^{j}\left(X_{\cdot}^{l}\right)$ est nécessairement unique et optimal. L'idée étant d'obtenir un estimateur performant à la fois *in-sample* mais aussi et surtout *out-of-sample*.
Afin de créer le prédicteur CART, l'algorithme commence par construire, en fonction du contexte induit par $\mathcal{L}_{n_i}^{j}$, un arbre maximal $T_{max}^{j}\left(x_{\cdot}^{l}\right)$. Celui-ci est *sans biais*, i.e. en régression que la somme des résidus centrés réduits et élevés au carré est nulle, et en classification que la proportion d'individus mal classés est nulle.

L'arbre maximal T_{max}^{j} se construit récursivement depuis sa racine : $\{t_k = t_1\}$ en engendrant des nœuds descendants droits $\{t_d = t_{2k+1}\}$ et gauches $\{t_g = t_{2k}\}$. Les nœuds terminaux sont appelés des feuilles et sont notés $\tilde{t}_k$; Avec $\widetilde{T}_k^{j}$ l'ensemble des feuilles de l'arbre T_k^{j} en-cours de développement – au niveau du nœud : k.

Le principe de construction des arbres CART repose sur une logique *dyadique*. A chaque nœud : t_k est associée une question binaire : $\delta_{l,k}^{*} = \{x^{l} \leq a_k^{*}\} \cup \{x^{l} > a_k^{*}\}$ permettant de partitionner, de *façon optimale*, l'espace des individus à partir de leurs coordonnées. Les e_i glissent de nœud en nœud en fonction des x_i^{l} qui leurs sont associées et au regard de seuils *décisionnels optimaux : a_k^{*}*. Le schéma suivant représente le principe de propagation des individus dans des nœuds :

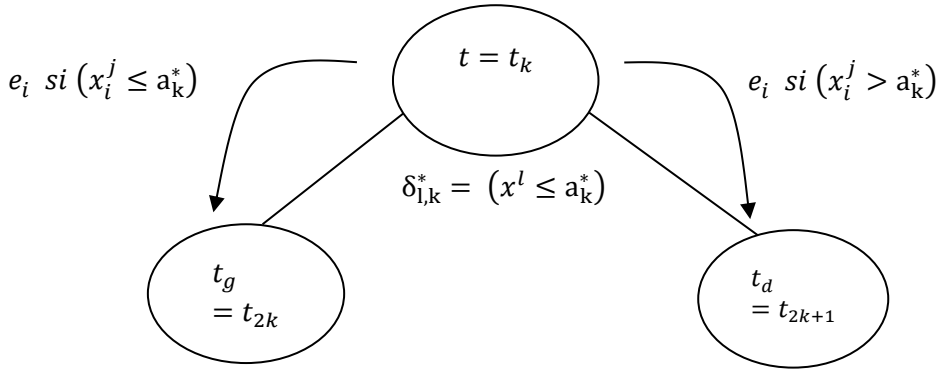


Figure 299 : Principe de propagation dyadique et descendant, des individus, dans un nœud CART

Toute la difficulté est de déterminer, à chaque nœud, une division optimale : $\delta_{l,k}^*$, i.e. la variable $x_\cdot^l$ et le seuil a_k^* qui maximisent - au regard de tous les descendants droits et gauches possibles – la *fonction d'hétérogénéité* donnée par :

$$\left\{\Delta R(t_k, \delta_{l,k}) = R(t_k, \delta_{l,k}) - \left(R(t_{2k}, \delta_{l,k}) + R(t_{2k+1}, \delta_{l,k})\right) \middle| \Delta R(t, \delta_{l,k}) > 0\right\}$$

L'hétérogénéité est définie, selon le contexte statistique, comme la différence quadratique des erreurs ou la différence des indices de Gini engendrés par les nœuds t_k et tous les descendants possibles susceptibles de former t_{2k} et t_{2k+1} :

$$\Delta R(t_k, \delta_{l,k}) = \begin{cases} \dfrac{N(t_{2k}) \cdot N(t_{2k+1})}{n_i \cdot N(t_k)} \left(\bar{y}(t_k) - \left(\bar{y}(t_{2k}) + \bar{y}(t_{2k+1})\right)\right)^2, & \text{en régression} \\ \left(r(t_k) \cdot p(t_k)\right) - r(t_{2k}) \cdot p(t_{2k}) - r(t_{2k+1}) \cdot p(t_{2k+1}), & \text{en classification} \end{cases}$$

Avec : $N(t_k)$ le nombre d'e_i dans le nœud t_k, ; $\bar{y}(t_k)$ la moyenne des réponses associées aux e_i contenus dans t_k ; Et le produit $\left(r(t_k) \cdot p(t_k)\right)$ la proportion d'e_i mal classés dans le nœud t_k

Plus le rapport : $\Delta R(t, \delta_{l,k})$ est élevé et plus les descendants générés sont purs. Par conséquent la division optimale est donnée par :

$$\delta_{l,k}^* = \underset{\forall\, \delta_{l,k} \in \mathcal{G}}{\operatorname{argmax}} \{\Delta R(t_k, \delta_{l,k})\}$$

Avec $\mathcal{G}$ l'ensemble de toutes les divisions possibles au niveau du nœud t_k :

$$\mathcal{G} = \left\{\{x_\cdot^l \leq a_k^l\} \cup \{x_\cdot^l > a_k^l\}, \ \forall l \in \{1, \dots, p_l\}\right\}, \text{avec: } a_k^l = \begin{cases} \hat{\text{moy}}(x_i^l | x_i^l \subset t_k) & \text{si: } x_\cdot^l \in \mathbb{R} \\ \underset{\forall c^l \in \mathcal{C}_l}{\operatorname{argmax}} \{\mathbb{P}(x_i^l = c^l | x_i^l \subset t_k)\} & \text{si: } x_\cdot^l \in \mathbb{N} \end{cases}$$

La division des nœuds t_k s'opère récursivement jusqu'à obtenir T_{max}^j un arbre dont toutes les feuilles sont pures, i.e. contenant des e_i auxquels est associée une valeur unique de y_i^j – quitte à n'avoir qu'un seul e_i dans chaque feuille $\tilde{t}_k$:

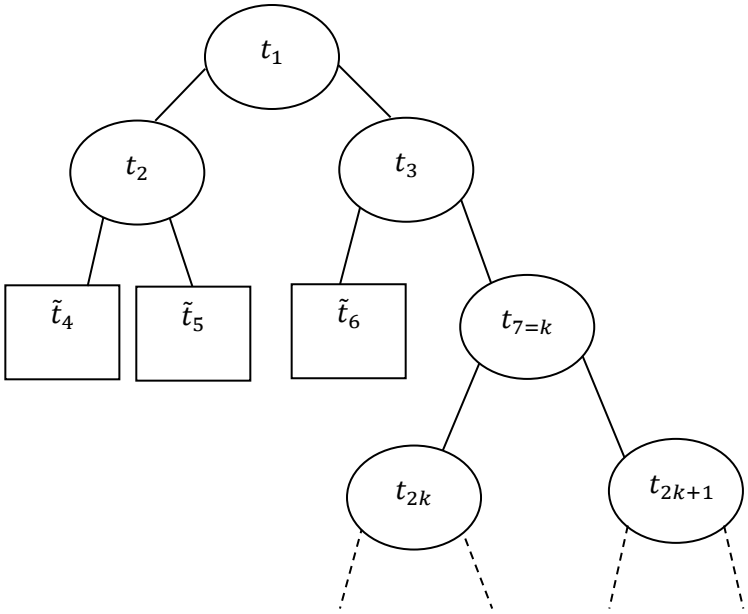


Figure 300 : Archétype de construction d'un arbre CART

A chaque nœud terminal ou feuille $\tilde{t}_k$. de l'arbre est associée une prédiction : $\tilde{t}_k(e_i) \to \hat{y}_i^j$. L'arbre maximal est sans biais, donc $R(T_{max}^j) = 0$. Par conséquent T_{max}^j prédit parfaitement ce que l'on a observé - *in sample*. En revanche ses performances *out-of-sample*, i.e. sur un jeu de données différent mais inhérent à un même phénomène sont médiocres. C'est une catastrophe en termes de variance.
Afin d'obtenir un arbre CART optimal, l'algorithme procède à *une phase d'élagage*. Autrement dit on recommence le processus en sens inverse en supprimant une à une les branches de l'arbre et en ramenant les e_i dans le nœud situé au-dessus. De cette façon on obtient *une sous-suite d'arbres élagués* :

$$\left\{T_{max}^j = \mathrm{T}_{K'}^{\mathrm{j}}\right\} \succ \mathrm{T}_{K'-1}^{\mathrm{j}} \succcurlyeq \cdots \succcurlyeq \mathrm{T}_{K'}^{\mathrm{j}} \succcurlyeq \cdots \succcurlyeq \mathrm{T}_2^{\mathrm{j}} \succ \left\{\mathrm{T}_1^{\mathrm{j}} = la\ racine\right\}$$

Cette opération récursive permet de construire la *sous-suite d'arbres élagués de Breiman* : $\mathcal{T}_{\mathrm{Breiman}}^{\mathrm{K}}$.
Plusieurs arbres peuvent comporter un même nombre de nœuds mais ils ne sont pas tous éligibles. La sous-suite de Breiman se compose uniquement des arbres qui maximisent *un compromis biais/variance* et qui s'estime par le biais d'un *critère pénalisé de complexité* :

$$\mathrm{Crit}_{\alpha}(T_k^j) = \mathrm{R}(T_k^j) + \frac{\alpha_k}{n_i} \cdot \left|\tilde{T}_k^j\right|$$

Avec $\left|\tilde{T}_k^j\right|$ = le nombre de feuilles de l'arbre à k divisions ; α_k un coefficient de pénalité inversement proportionnel à la complexité, i.e. au nombre de feuilles. Et $\mathrm{R}(T_k^j)$ l'estimateur de la variance en régression ou l'indice de Gini en classification

$$\mathrm{R}(T_k^j) = \begin{cases} \dfrac{1}{\mathrm{n_i}} \displaystyle\sum_{\mathrm{i}=1}^{\mathrm{n_i}} \left(\mathrm{y_i^j} - \hat{f}_k^j(x_i^l)\right)^2, & \text{en régression} \\ \dfrac{1}{\mathrm{n_i}} \displaystyle\sum_{\mathrm{i}=1}^{\mathrm{n_i}} \mathbb{1}_{\left\{\mathrm{y_i^j} \neq \hat{f}_k^j(x_\cdot^l)\right\}}, & \text{en classification} \end{cases}$$

Avec $\hat{f}_k^j(\cdot)$ et $\hat{f}_k^j(x_i^l)$ respectivement le prédicteur et la prédiction engendrée par l'arbre T_k^j.
La sous-suite de Breiman est constituée pendant la phase d'élagage au regard du critère pénalisé, tel que :

$$\mathcal{T}_{\text{Breiman}}^{\text{K}} = \bigcup_{\text{k}'=0}^{\text{K}'-1} \left(T_{K'-k'}^{j} \middle| \left\{\text{Crit}_{\alpha}\left(T_{K'-k'}^{j}\right) > \text{Crit}_{\alpha}\left(T_{K'-k'-1}^{j}\right)\right\}\right)$$

Chaque arbre T_k^j de la sous-suite de Breiman est unique :

$$\mathcal{T}_{\text{Breiman}}^{\text{K}} = \left\{T_{max}^{j} = T_{K}^{j}\right\} \succ T_{K-1}^{j} \succ \cdots \succ T_{k}^{j} \succ \cdots \succ T_{2}^{j} \succ \left\{T_{1}^{j} = \text{la racine}\right\}$$

L'algorithme CART est qualifié *d'instable* parce qu'il suffit de changer ne serait-ce qu'une valeur pour modifier $\mathcal{T}_{\text{Breiman}}^{\text{K}}$.

L'arbre optimal appartient nécessairement à la sous-suite de Breiman $T_{opt}^{j} \in \mathcal{T}_{\text{Breiman}}^{\text{K}}$. Pour l'identifier il existe deux méthodes : *l'échantillon test* et *la validation croisée*. La procédure de calibration par *échantillon test* n'est presque plus utilisée car elle nécessite de partitionner $\mathcal{L}_{\text{n}_\text{i}}^{\text{j}}$ sur l'espace des individus, ce qui induit une perte de puissance considérable. La technique *de validation croisée*, de l'anglais : *cross-validation (c.v.)*, est adaptée à la majorité des jeux de données et elle est beaucoup plus puissante. En l'occurrence c'est celle qui est utilisée dans la procédure de sélection de variables lors du calcul *du seuil d'élimination des variables de bruit*.

L'idée de la procédure de validation croisée est de découper $\mathcal{L}_{\text{n}_\text{i}}^{\text{j}}$ en V paquets par le biais d'un échantillonnage Boostrap sans remise. Chaque paquet contient : $\text{n}_\text{i}/\text{V}$ individus. De fait, V échantillons Boostrap sont générés : $\left(\mathcal{L}_{(1)}^{\text{j}}, \ldots, \mathcal{L}_{(\text{v})}^{\text{j}}, \ldots, \mathcal{L}_{(\text{V})}^{\text{j}}\right)$. Ils sont ensuite regroupés pour former des jeux d'apprentissage de validation croisée $\mathcal{L}_{\text{cv}}^{\text{j}} = \left(\mathcal{L}_{1}^{\text{j}}, \ldots, \mathcal{L}_{\text{v}}^{\text{j}}, \ldots, \mathcal{L}_{\text{V}}^{\text{j}}\right)$, de sorte que

$$\mathcal{L}_{v}^{\text{j}} = \bigcup_{(v)=1}^{\text{V}} \left(\left\{\mathcal{L}_{(v)}^{\text{j}} / \mathcal{L}_{(v=v)}^{\text{j}}\right\}\right)$$

Ensuite, on applique CART sur les $\mathcal{L}_{\text{cv}}^{\text{j}} = \left(\mathcal{L}_{1}^{\text{j}}, \ldots, \mathcal{L}_{\text{v}}^{\text{j}}, \ldots, \mathcal{L}_{\text{V}}^{\text{j}}\right)$ et on obtient V sous-suites de Breiman : $\mathcal{T}_{\text{Breiman}}^{\text{K}_{\text{cv}},\text{cv}} = \left\{\mathcal{T}_{\text{Breiman}}^{\text{K}_1,1}, \ldots, \mathcal{T}_{\text{Breiman}}^{\text{K}_\text{v},\text{v}} \ldots, \mathcal{T}_{\text{Breiman}}^{\text{K}_\text{V},\text{V}}\right\}$. Pour chaque arbre de chacune des $\mathcal{T}_{\text{Breiman}}^{\text{K}_\text{v},\text{v}}$ on calcule, grâce au paquet inutilisé $\mathcal{L}_{(\text{v})}^{\text{j}}$ les erreurs associées de chaque modèle $\text{R}^{\text{cv}}\left(T_{k_v}^{(v)}\right)$ et leurs $\alpha_{\text{k}_\text{v}}^{(\text{v})}$. Ce qui permet d'obtenir les collections suivantes :

$$\mathcal{T}_{\text{Breiman}}^{\text{K}_1} ;\ \mathcal{L}_{(1)}^{\text{j}} \quad \Longrightarrow \quad \begin{cases} T_{K_1,(1)}^{j} \succ T_{K_1-1,(1)}^{j} \succ \cdots \succ T_{1,(1)}^{j} \\ \alpha_{\text{K}_1}^{(\text{v}=1)} < \alpha_{\text{K}_1-1}^{(\text{v}=1)} < \ \ldots \ < \alpha_{\text{K}_1}^{(\text{v}=1)} \\ \text{R}^{\text{cv}}\left(T_{K_1,(1)}^{j}\right), \text{R}^{\text{cv}}\left(T_{K_1-1,(1)}^{j}\right), \ldots, \text{R}^{\text{cv}}\left(T_{1,(1)}^{j}\right) \end{cases}$$

$$\vdots \qquad\qquad \vdots$$

$$\mathcal{T}_{\text{Breiman}}^{\text{K}_\text{V}} ;\ \mathcal{L}_{(\text{V})}^{\text{j}} \quad \Longrightarrow \quad \begin{cases} T_{K_V,(V)}^{j} \succ T_{K_V-1,(V)}^{j} \succ \cdots \succ T_{1,(V)}^{j} \\ \alpha_{\text{K}_\text{V}}^{(\text{V})} < \alpha_{\text{K}_\text{V}-1}^{(\text{V})} < \ \ldots \ < \alpha_{1}^{(\text{V})} \\ \text{R}^{\text{cv}}\left(T_{K_V,(V)}^{j}\right), \text{R}^{\text{cv}}\left(T_{K_V-1,(V)}^{j}\right), \ldots, \text{R}^{\text{cv}}\left(T_{1,(V)}^{j}\right) \end{cases}$$

Ensuite une erreur moyenne de validation croisée $\overline{\text{R}}^{\text{cv}}\left(T_k^j\right)$, des écarts-types d'estimation $\widehat{\sigma}\left(\text{R}^{\text{cv}}\left(T_k^j\right)\right)$ et des coefficients de complexité de cv. : $\dot{\alpha}_{\text{k}}^{\text{cv}}$ (estimés par des moyennes géométriques) sont supputés pour l'ensemble des sous-arbres $T_{k_v,(v)}^{j}$ de toutes les sous-suites de Breiman. Par conséquent pour les arbres de même dimension, on peut estimer sur les V sous-sous-suites les couples :

$$\left(\overline{\text{R}}^{\text{cv}}\left(T_k^j\right);\ \widehat{\sigma}\left(\text{R}^{\text{cv}}\left(T_k^j\right)\right);\ \dot{\alpha}_{\text{k}}^{\text{cv}}\right), \qquad \forall\, T_{k_v,(v)}^{j}\ \text{k}_\text{v} = \{1, \ldots, \text{K}_\text{v}\}, \qquad \forall\, \text{v} = \{1, \ldots, \text{V}\}$$

L'arbre optimal au sens CART est celui qui *maximise le compromis biais variance*. C'est-à-dire celui dont l'erreur de validation croisée est inférieure *au seuil d'élimination de validation croisée*, pour lequel la règle est celle du *min + 1SD* et dont la complexité est la plus petite :

$$T_{opt}^{j} = \underset{\forall T_k^j \in \mathcal{T}_{\text{Breiman}}^{\text{KV}}}{\operatorname{argmin}} \left\{ \left| \tilde{T}_k^j \right| \left| \left\{ R^{cv}\left(T_k^j\right) \leq \varphi_l^{cv} \right\} \right. \right\}$$

Le *seuil d'élimination de validation croisée* s'estime de la façon suivante

$$\varphi_{ko}^{cv} = \left(R^{cv}\left(T_{ko}^j\right) + 1 \cdot \hat{\sigma}\left(R^{cv}\left(T_{ko}^j\right)\right) \middle| \left\{ R^{cv}\left(T_{ko}^j\right) = \min_{\forall k=\{1,\ldots,K\}} \left\{ R^{cv}\left(T_k^j\right) \right\} \right\} \right)$$

Le graphique suivant schématise les résultats renvoyés par l'algorithme *rpart*. Il exprime l'évolution des erreurs de cv. : $\overline{\mathrm{R}}^{\mathrm{cv}}\left(T_k^j\right)$ en fonction du nombre de feuilles $\left|\tilde{T}_k^j\right|$ des modèles et de l'évolution du coefficient de complexité $\dot{\alpha}_k^{cv}$

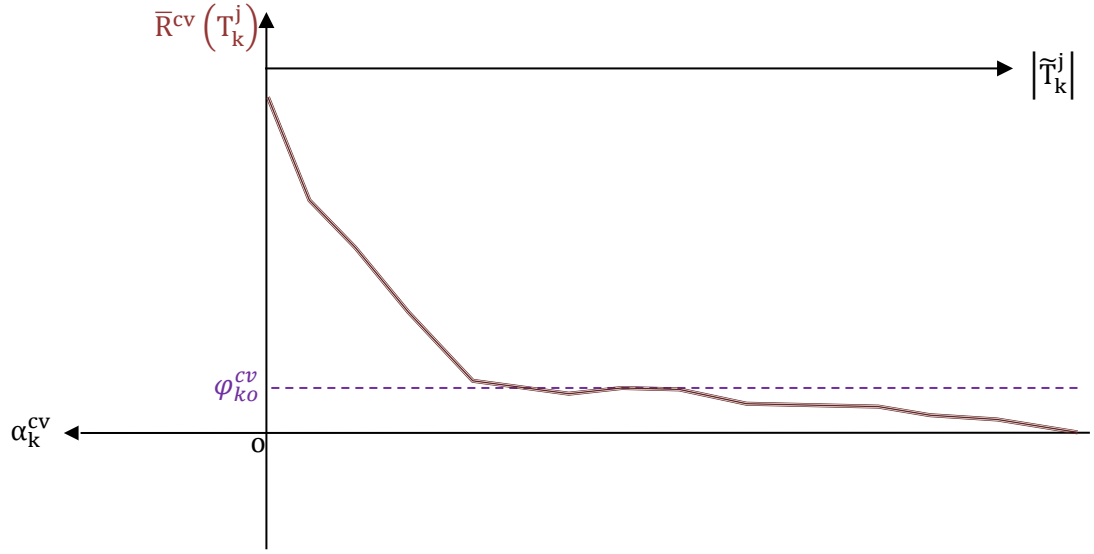


Figure 301 : Principe de sélection de l'arbre optimal par la méthode de validation croisée

Dans la pratique, afin de maximiser la calibration de l'arbre optimal il convient de récupérer la valeur $\dot{\alpha}_{\mathrm{opt}}^{\mathrm{cv}}$ associée à l'arbre optimal de validation croisée et d'élaguer à nouveau T_{max}^j en spécifiant une valeur α'_{opt} permettant, soit de *réduire le biais, soit la variance,* afin de prendre en compte les artefacts induits par *l'effet de bore* de l'algorithme. De cette façon on obtient l'estimateur CART optimal :

$$\hat{\mathrm{y}}_{\cdot}^{j} = \hat{f}_{\mathrm{CART}}^{j}\left(x_{\cdot}^{l}\right) \overset{\text{def}}{=} \widehat{\mathrm{T}}_{\mathrm{opt}}^{\mathrm{j}}\left(x_{\cdot}^{l}\right) = \begin{cases} \displaystyle\sum_{\forall \tilde{\mathrm{t}}_{\mathrm{k}} \in \widetilde{\mathrm{T}}_{\mathrm{opt}}^{\mathrm{j}}} \sum_{i=1}^{n_j} \frac{\mathrm{y}_i^j \cdot \mathbb{1}_{e_i \in \tilde{\mathrm{t}}_{\mathrm{k}}}}{\mathrm{card}(e_i | e_i \in \tilde{\mathrm{t}}_{\mathrm{k}})}, & \forall\, \mathrm{y}_i^j \in \mathbb{R}^1 \\ \underset{\forall\, \mathrm{c_j} \in \mathcal{C}_{\mathrm{j}}}{\operatorname{argmax}} \left\{ \displaystyle\sum_{\forall \tilde{\mathrm{t}}_{\mathrm{k}} \in \widetilde{\mathrm{T}}_{\mathrm{opt}}^{\mathrm{j}}} \left(\bigcup_{i=1}^{n_j} \left(\mathrm{y}_i^j = \mathrm{c_j} | e_i \in \tilde{\mathrm{t}}_{\mathrm{k}} \right) \right) \right\}, & \forall\, \mathcal{C}_{\mathrm{j}} \subset \mathbb{N}^{|\mathcal{C}_{\mathrm{j}}|} \end{cases}$$

Avec : $e_i = \left(\mathrm{y}_i^j ; x_{\cdot}^{l}\right)$ et $x_{\cdot}^{l} = \left(x_i^1, \ldots, x_i^{p_l^j}\right), \forall i \in \{1, \ldots, n_j\}$ toutes les valeurs d'échantillonnage observées pour les réalisations des v.a. potentiellement explicatives.

Les éléments essentiels de la méthode CART sont désormais présentés, il est temps maintenant passer à la présentation des Forêts Aléatoires (FA), qui sont consitutées de centaines, voire des milliers, d'arbres CART doublement perturbés.

Principes fondamentaux des Forets Aléatoires

Les Forêts Aléatoires (FA) appartiennent à famille des méthodes d'ensemble. FA est un modèle d'inférence statistique *non paramétrique* et *intégrant des processus stochastiques*. Cette partie s'appuie essentiellement sur la thèse de Robin Génuer (Genuer, 2010). Les principes mathématiques énoncés sont ceux qui sont implémentés dans le package *randomForest-V4.6-7* du logiciel R, la version officielle des Forêts Aléatoires (Liaw, 2013).
Les FA appartiennent donc aux *méthodes d'ensemble* et intègrent le processus de Bagging, donc l'espace des individus est partitionné aléatoirement par la création d'échantillons Boostrap. Sur chaque échantillon d'apprentissage Boostrap on applique une règle, en l'occurrence CART, que l'on perturbe avec du Random-Subspace. Cette seconde perturbation stochastique intervient au niveau de la construction des nœuds des arbres CART. Par défaut *randomForest* n'effectue pas du Randomizing-Output mais l'algorithme offre aussi la possibilité de perturber l'espace des réponses. Seule la technique ensembliste de Boosting n'est pas implémentée.

L'algorithme permet d'obtenir une collection de prédicteurs CART doublement randomisés qui constituent une forêt. Les arbres CART de la FA ne sont pas optimisés. Cependant le terme de biais est minimisé dans la mesure où les arbres sont développés jusqu'à leur taille maximale. Quant au terme de variance, il est optimisé d'une part par les échantillons Boostrap qui ne contiennent qu'un sous-ensemble des individus disponibles, et d'autre part grâce au Random-Subspace qui perturbe aléatoirement l'espace des variables lors de la recherche de la division optimale des nœuds. De fait, chaque arbre CART performe uniquement sur un sous-sous-espace engendré par les données d'apprentissage $\mathcal{L}_{n_i}^{j}$. De plus, le caractère stochastique des perturbations fait qu'ils sont décorrélés. Le prédicteur FA est unique, il résulte d'une agrégation ensembliste simple de tous les prédicteurs : arbres, dans la même logique que celui du Bagging de Breiman. En régression la prédiction de la FA est la moyenne empirique des prédictions de chacun des arbres perturbés, et en classification c'est la règle du vote majoritaire qui est appliquée. Les prédictions effectuées par les FA sont beaucoup plus consistantes que celles des estimateurs CART et balaient d'une eau lustrale les possibilités qu'offrait l'algorithme *rapart*.
Si les FA engendrent des prédictions robustes, elles permettent aussi de comprendre les interactions qui existent entre les réalisations du phénomène d'intérêt $y_{\cdot}^{j}$ et les mesures des variables potentiellement explicatives $(x_{\cdot}^{1}, \dots, x_{\cdot}^{p_l})$ disponibles par l'attribution *de scores d'importance des variables*.

Des scores d'importance des variables avaient aussi été proposés avec CART (Tuleau-Malot, 2005). Nonobstant, il a été montré qu'ils sont biaisés en faveur de variables de nature *qualitative multi-classes*. En revanche les scores randomForest sont *a priori* parfaitement stables et ne sont pas influencés par la nature des $x_{\cdot}^{l}$ (Genuer, 2010).
Les performances prédictives des FA permettent d'envisager l'évolution d'un phénomène au vu des variables mesurées et mesurables disponibles. Quant à leurs capacités explicatives, elles permettent de comprendre comment interagir avec le phénomène d'intérêt en identifiant des leviers, i.e. les variables influentes dans l'optique d'en infléchir l'issue.

Voici succinctement présentée l'idée des FA. Les éléments théoriques, la qualité de l'estimateur et une stratégie de calibration des FA sont décrits subséquemment.

Eléments théoriques

Par définition les FA, au sens où l'entend Léo Breiman, se composent d'une collection de d'arbres CART doublement perturbés qui engendrent des arbres prédicteurs (Breiman et Cutler, 2005).

$$\mathrm{T}_{\mathrm{cart}}^{j}\left(\mathcal{L}_{n_i}^{j}\left\{\Theta_{\mathrm{k}} \cap \left(\Xi_{\mathrm{t_v}}\right)_{1\leq \mathrm{v}\leq \mathrm{V_k}}\right\}\right) \Rightarrow f_{\mathrm{CART}}^{j}\left(\mathrm{X}\left\{\Theta_{\mathrm{k}} \cap \left(\Xi_{\mathrm{t_v}}\right)_{1\leq \mathrm{v}\leq \mathrm{V_k}}\right\}\right), \forall\, \mathrm{k} = \{1, \dots, \mathrm{K}\}$$

Les FA se constituent de K arbres et sont construites conditionnellement à un jeu de données d'apprentissage défini par :

$$\mathcal{L}_{n_i}^{j} = \left\{\left\{x_{\cdot}^{l} = \left(x_i^1, \dots, x_i^{p_l}\right), \forall i = \{1, \dots, n_i\}\right\} \cup \left\{y_{\cdot}^{j} = \left(y_i^j, \dots, y_{n_i}^j\right)\right\}\right\}$$

Avec $y_{\cdot}^{j}$, les n_i réalisations mesurables d'une forme de la variable y associée au phénomène d'intérêt-j. Le contexte d'apprentissage et de construction des arbres est fixé par la nature de $y_{\cdot}^{j}$: la classification, lorsque les réalisations sont de nature qualitative booléenne ou multi-classes ; $y_{\cdot}^{j} \in \mathcal{C}^{j} \subseteq \mathbb{N}^{|\mathcal{C}^{j}|}$, et celui de la régression lorsque la variable d'intérêt est de nature quantitative discrète ou continue ; $y_{\cdot}^{j} \in \mathbb{R}^{n_i}$. Quant aux variables potentiellement explicatives et prédictives $x_{\cdot}^{l}$, elles peuvent être de nature parfaitement hétéroclite. L'indice l fait référence au type de phénomène mesuré, mesurable et associé aux n_i réalisations de x. L'opérateur FA est très proche de celui d'un estimateur CART à la différence près que le nombre de dimensions est beaucoup plus grand :

$$\mathrm{Y}_{\cdot}^{\mathrm{j}} = f_{FA}^{j}\left(\mathrm{X}_{\cdot}^{\mathrm{l}}\right) = \begin{cases} \dfrac{1}{K}\displaystyle\sum_{k=1}^{K} f_{\mathrm{CART}}^{j}\left(\mathrm{X}_{\cdot}^{\mathrm{l}}\left\{\Theta_{\mathrm{k}} \cap \left(\Xi_{\mathrm{t_v}}\right)_{1\leq \mathrm{v}\leq \mathrm{V_k}}\right\}\right) + \varepsilon^{\mathrm{j}}, \quad \forall\, \mathrm{Y}^{\mathrm{j}} \in \mathbb{R}^{n_{\mathrm{i}}} \\ \underset{\forall\, c_j \in \mathcal{C}^{j}}{\mathrm{argmax}}\left\{\dfrac{1}{K}\displaystyle\sum_{k=1}^{K} \mathbb{1}_{f_{\mathrm{CART}}^{j}\left(\mathrm{X}_{\cdot}^{\mathrm{l}}\left\{\Theta_{\mathrm{k}} \cap \left(\Xi_{\mathrm{t_v}}\right)_{1\leq \mathrm{v}\leq \mathrm{V_k}}\right\}\right)=c_j}\right\} + \varepsilon^{\mathrm{j}}, \quad \forall\, \mathrm{Y}^{\mathrm{j}} \in \mathcal{C}^{j} \subset \mathbb{N}^{|\mathcal{C}^{j}|} \end{cases}$$

Avec ε^{j} un Bruit Blanc Faible de moyenne nulle et de variance constante. L'algorithme *randomForest* commence par confectionner des échantillons Bootstrap $\mathcal{L}_{n_i}^{\Theta_{\mathrm{k}},j}$.

$$\left\{\mathcal{L}_{n_i}^{\Theta_{\mathrm{k}},j} \subseteq \mathcal{L}_{n_i}^{j}\right\} = \left\{\left\{y_{\cdot}^{j}\{\Theta_{\mathrm{k}}\}\right\} \cup \left\{(x_{\cdot}^{1}, \dots, x_{\cdot}^{p_l})\{\Theta_{\mathrm{k}}\}\right\} \middle| \Theta_{\mathrm{k}} \sim \mathcal{M}\left((1, \dots, n_i); \frac{1}{n_i}\right)\right\}$$

Avec Θ_{k} les indices des individus constitutifs de l'échantillon Bootstrap k - tirés aléatoirement avec remise dans une multinomiale uniforme discrète.
Les *K* échantillons Boostrap sont en quelque sorte *les graines* qui vont donner naissance aux arbres CART de la forêt. Le nombre d'arbres constituant la FA est un paramètre important de *randomForest*. Il est noté **ntree** et sa valeur par défaut a été fixée par Léo Breiman à:

$$\{K \overset{\text{def}}{=} \mathbf{ntree} = 500\}$$

D'une manière générale plus *K* est grand et plus les capacités prédictives de la FA sont robustes, mais plus les temps de calculs sont longs (Liaw et Wiener, 2006).
Chaque arbre est donc construit à partir d'échantillons Bootstrap de mêmes dimensions que l'échantillon d'apprentissage. Donc certains individus sont répétés plusieurs fois. Les arbres CART sont développés de la racine jusqu'à leur taille maximale ou presque.

En effet la taille des arbres CART peut être jugulée par le paramètre **nodesize** qui permet de définir le nombre maximum d'individus présents dans chaque nœud : ns, avant que celui-ci ne soit transformé en feuille. Par défaut les valeurs implémentées dépendent du contexte.

En classification les arbres CART sont développés jusqu'à atteindre une taille maximale en revanche ce n'est pas forcément le cas en régression puisque :

$$\{ns \overset{\text{def}}{=} \textbf{nodesize}\} = \begin{cases} 5 & \text{régression} \\ 1 & \text{classification} \end{cases}$$

Le fait de juguler la taille des arbres permet de les décorréler mais en contrepartie cela augmente leur biais. Une autre façon pour contrôler la taille des arbres consiste à jouer sur le paramètre {**maxnodes**} et ainsi générer des *Extra-Trees* qui engendrent des FA particulières (Geurts, Ernst et al., 2006).
Chaque arbre CART est randomisé dans son développement, au niveau de la construction de ses nœuds, par un processus de Random-Subspace – qui prend effet à la racine. Cette seconde perturbation stochastique intervient au moment de la recherche de chaque division optimale $\delta^*_{\mathrm{t},l,k}$. Le but étant de déterminer la variable $x^l_{\cdot}$ et le seuil $\mathrm{a}^*_{\mathrm{t},l,k}$, qui permettent de maximiser la pureté du nœud. Si l'algorithme *rpart* regarde l'ensemble des coordonnées disponibles $x^l_{\cdot} = (x^1_{\cdot}, \dots, x^{p_l}_{\cdot})$, *randomForest* n'en regarde qu'une partie. Autrement dit la coupure au niveau de chaque nœud $\mathrm{t_v}$ est effectuée sur : $\boldsymbol{m}$ variables $\left\{\left\{x^l_i\{\Xi_{\mathrm{t_v}}\}, \forall\, l \in \{1, \dots, \mathrm{p}_l\} \middle| e_i \subset \mathrm{t_v}\right\}\right\}$ de sorte que $\Xi_{\mathrm{t_v}}$ sont les indices des $x^l_{\cdot}$ tirées pour chaque nœud, sans remise, de façon aléatoire dans un multinomiale uniforme discrète.

$$\left\{\Xi_{\mathrm{t_v}} \sim \mathcal{M}\left((1, \dots, p_l); \frac{1}{\mathrm{m}}\right) \middle| \dim(\Xi_{\mathrm{t_v}}) = m\right\}, \forall\, \mathrm{v} = \{1, \dots, \mathrm{V_k}\}$$

Où l'indice $\mathrm{V_k}$ représente le nombre total de nœuds de l'arbre $k \in \{1, \dots, \mathrm{K}\}$; La valeur de $\boldsymbol{m}$ est un paramètre déterminant de l'algorithme *randomForest* nommé **mtry**. Léo Breiman avait suggéré des valeurs optimales en fonction du contexte qui sont toujours celles implémentées par défaut (Liaw et Wiener, 2006)

$$\{m \overset{\text{def}}{=} \textbf{mtry}\} = \begin{cases} \left[p_l/3\right] & \text{régression} \\ \left[\sqrt{p_l}\right] & \text{classification} \end{cases}$$

Où $[.] =$ "la fonction arrondi à l'entier". Il convient de remarquer que les valeurs suggérées par Breiman sont approximatives, voire grossières, et même parfois complètement inadaptées au jeu de données d'apprentissage. La valeur de **mtry** a une influence majeure sur les performances des FA. Le paramètre m ne peut prendre qu'un nombre fini de valeurs. En fixant $\{\textbf{mtry} = \textbf{1}\}$ on obtient la forêt de Breiman la plus randomisée. A l'inverse en réglant $\{\textbf{mtry} = p_l\}$ on retombe sur le *Bagging de Breiman.* Or à l'aune de **ntree** et de **nodesize** les performances des forêts ne sont pas proportionnelles à la valeur de **mtry.** En effet, d'un côté de trop grandes perturbations $\textbf{mtry} = \textbf{1}$ dégradent les prédicteurs individuels et en conséquence l'estimateur FA est altéré. De l'autre, aucune perturbation stochastique $\{\textbf{mtry} = p_l\}$ ne diminue aussi les performances de l'estimateur FA liées à une trop forte corrélation entre des prédicteurs individuels trop performants puisque sans biais. De plus, le contexte statistique ne semble pas avoir vraiment d'impact sur ce paramètre. Par conséquent il est impératif de calibrer ce paramètre avant d'utiliser les FA à des fins explicatives ou prédictives (Genuer, 2010).

La calibration des FA s'opère au vu de *l'erreur de généralisation des FA*. Autrement dit, il s'agit de générer des centaines de FA différentes (en faisant varier les valeurs des différents paramètres) et de retenir celles dont les performances sont les meilleures. Or le nombre de paramètres implémentés dans l'algorithme *randomForest* est incommensurable. Les plus discutés dans la littérature ont été passés en revue mais il en existe beaucoup d'autres ; e.g. : **replace**, **classwt**, **cutoff**, **sampsize**, **nPerm**, **norm**. **votes**, **keep**. **inbag**, **na**. **roughfix** … qui ont plus ou moins d'influence sur la FA générée et dont les valeurs implémentées par défaut sont spécifiées dans la documentation associée à la fonction *randomForest* (Liaw, 2013).

Ainsi, le choix des paramètres de l'algorithme conditionne l'allure du prédicteur généré par l'agrégation ensembliste de tous les prédicteurs CART perturbés qui s'effectue par un *vote majoritaire* ou par l'estimation d'une *moyenne* en fonction du contexte.

$$\hat{y}_{\cdot}^{j} = \hat{f}_{FA}^{j}\left(x_{\cdot}^{l};\{m,K,ns\}\right) = \begin{cases} \dfrac{1}{\mathrm{K}}\displaystyle\sum_{\mathrm{k}=1}^{\mathrm{K}} \hat{f}_{\mathrm{CART}}^{j}\left(x_{\cdot}^{l}\left\{\Theta_{\mathrm{k}} \cap \left(\Xi_{\mathrm{t_v}}\right)_{1\le \mathrm{v} \le \mathrm{V_k}}\right\}\right) & \text{en régression} \\ \underset{\forall c_j \in \mathcal{C}^j}{\operatorname{argmax}} \left\{\dfrac{1}{\mathrm{K}} \cdot \displaystyle\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathbb{1}_{\hat{f}_{\mathrm{CART}}^{j}\left(x_{\cdot}^{l}\left\{\Theta_{\mathrm{k}} \cap \left(\Xi_{\mathrm{t_v}}\right)_{1\le \mathrm{v} \le \mathrm{V_k}}\right\}\right)=c_j}\right\} & \text{en classification} \end{cases}$$

En somme l'archétype de construction des FA qui sont utilisées dans cette thèse est décrit par le schéma suivant :

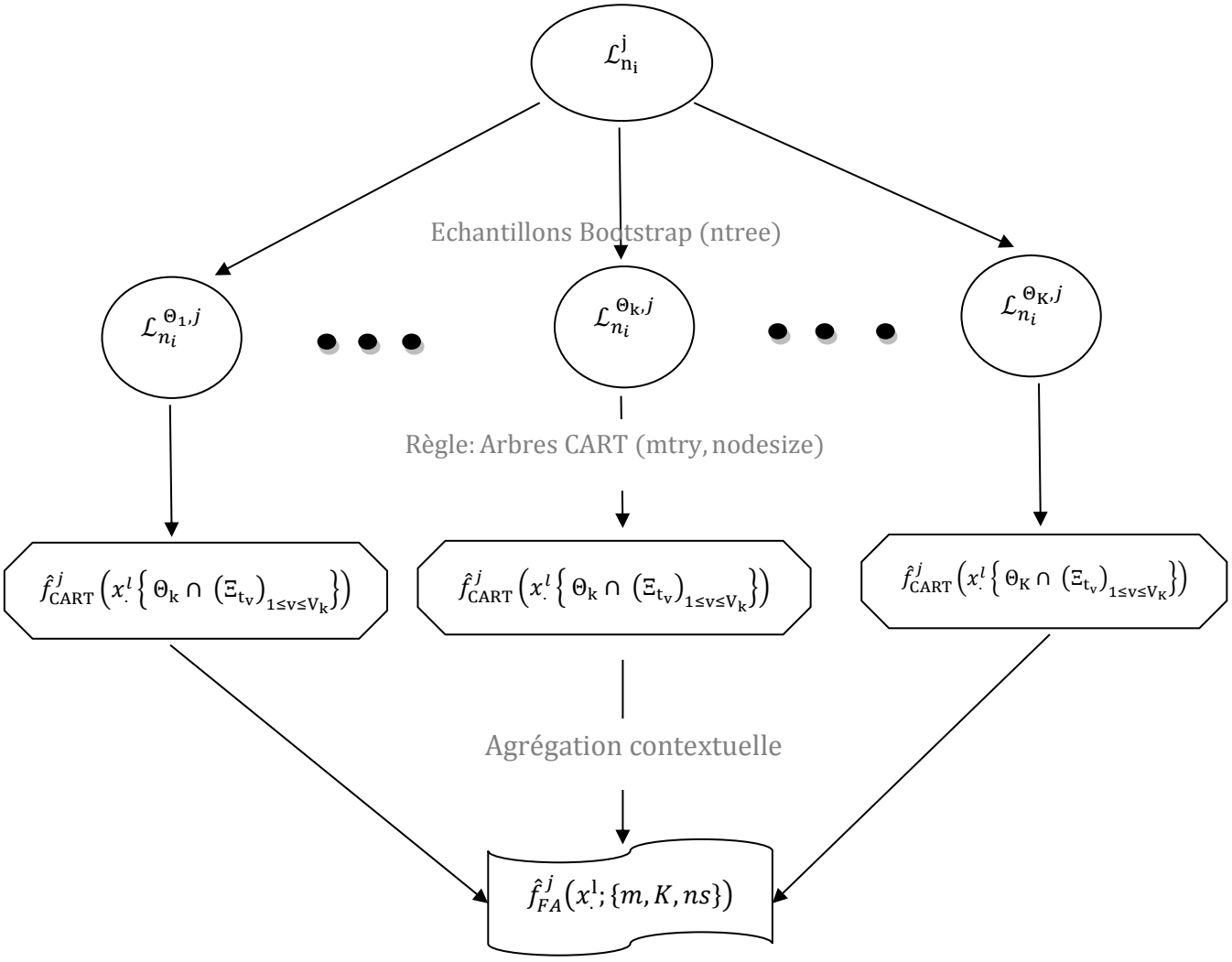


Figure 302 : Archétype de construction d'une Forêt Aléatoire

La qualité du modèle FA s'estime au vu de ses capacités prédictives *Out-Of-Bag*, par une erreur de généralisation qui s'apparente à une technique à mi-chemin entre l'échantillon test et la validation croisée.

Qualité de l'estimateur

La qualité de l'estimateur FA : $\hat{f}_{FA}^{j}(x_{\cdot}^{l})$ s'estime en fonction du contexte par l'erreur OOB – l'err.OOB – connue aussi sous le nom d'erreur de généralisation Out-Of-Bag (OOB). Cette mesure est renvoyée par l'algorithme *randomForest*. L'erreur OOB (Out-Of-Bag) signifie en dehors de l'échantillon Bootstrap. En l'occurrence, elle est calculée uniquement sur les arbres pour lesquels les e_i n'ont pas été utilisés lors de la phase de construction – de fait l'échantillon d'apprentissage n'a pas besoin d'être scindé.
L'idée de l'err.OOB consiste à effectuer, pour chaque individu e_i de $\mathcal{L}_{\mathrm{n_i}}^{\mathrm{j}}$, une seconde prédiction $\hat{y}_{\mathrm{i}}^{OOB,j}$ mais uniquement en utilisant les arbres CART perturbés construits sur des échantillons Boostrap ne contenant pas e_i, i.e. :

$$\left\{\mathcal{L}_{\mathrm{n_i}}^{\Theta_{\mathrm{k}},\mathrm{j}}/\{e_i\}\right\}, \forall\, k = \{1,\dots,K\};\ \forall\, i\ \in\ \{1,\dots,n_i\}$$

La prédiction OOB, pour chaque individu, est donnée par :

$$\hat{y}_{\mathrm{i}}^{OOB,j} = \begin{cases} \dfrac{1}{N\left(\mathrm{T}_{\mathrm{cart}}^{j}/\{e_i\}\right)}\displaystyle\sum_{k=1}^{K}\hat{f}_{\mathrm{CART}}^{j}\left(x_{\cdot}^{l}\left\{\Theta_{\mathrm{k}}\cap\ \left(\Xi_{\mathrm{t_v}}\right)_{1\le \mathrm{v}\le \mathrm{V_k}}\right\}\right)\cdot\mathbb{1}_{\left\{\{e_i\}\notin\mathcal{L}_{\mathrm{n_i}}^{\Theta_{\mathrm{k}},\mathrm{j}}\right\}} & \text{en régression} \\ \underset{\forall c_j\in\mathcal{C}^j}{\operatorname{argmax}}\left\{\dfrac{1}{N\left(\mathrm{T}_{\mathrm{cart}}^{j}/\{e_i\}\right)}\displaystyle\sum_{k=1}^{K}\mathbb{1}_{\hat{f}_{\mathrm{CART}}^{j}\left(x_{\cdot}^{l}\left\{\Theta_{\mathrm{k}}\cap\left(\Xi_{\mathrm{t_v}}\right)_{1\le \mathrm{v}\le \mathrm{V_k}}\right\}\right)=c_j}\cdot\mathbb{1}_{\left\{\{e_i\}\notin\mathcal{L}_{\mathrm{n_i}}^{\Theta_{\mathrm{k}},\mathrm{j}}\right\}}\right\} & \text{en classification} \end{cases}$$

Avec : $N\left(\mathrm{T}_{\mathrm{cart.k}}^{j}/\{e_i\}\right)$ le nombre d'arbres CART de la forêt construits sans utiliser $\mathrm{e_i}$. Conséquemment, découle l'err.OOB, ou l'erreur de généralisation OOB de la FA :

$$\mathrm{R}^{\mathrm{OOB}}(\mathrm{FA}) = \begin{cases} \dfrac{1}{\mathrm{n}_i}\displaystyle\sum_{\mathrm{i}=1}^{\mathrm{n}_i}\left(\hat{y}_{\mathrm{i}}^{OOB,j} - y_{\mathrm{i}}^{j}\right)^2 & \text{en régression} \\ \dfrac{1}{\mathrm{n}_i}\displaystyle\sum_{\mathrm{i}=1}^{\mathrm{n}_i}\mathbb{1}_{\left\{\hat{y}_{\mathrm{i}}^{OOB,j}\neq y_{\mathrm{i}}^{j}\right\}} & \text{en classification} \end{cases}$$

Dans le contexte de classification $\mathrm{R}^{\mathrm{OOB}}(\mathrm{FA})$ représente une proportion d'individus mal classés. Cet indicateur est très prégnant pour apprécier la qualité de l'estimateur FA. En revanche ce n'est pas le cas en régression. Il est difficile d'apprécier la qualité d'un modèle FA à partir de l'estimateur de la variance OOB. De fait, on préférera utiliser *le pourcentage OOB de variance expliquée*. Il s'agit tout simplement du rapport entre l'estimateur OOB de la variance des prédictions de FA et l'estimation biaisée de la variance calculée sur la cible, tel que :

$$\mathbb{var.explain}_{\mathrm{FA}}^{\mathrm{OOB}} = \frac{\mathrm{R}^{\mathrm{OOB}}(\mathrm{FA})}{\hat{\sigma}^2(y_{\cdot}^{j})}$$

Une technique alternative pour évaluer la qualité des FA est la méthode de *l'échantillon test*. Il s'agit de la même procédure que celle utilisée pour CART qui est adaptée uniquement aux échantillons de grande taille. L'idée est de fractionner l'échantillon initial en un échantillon de construction et un échantillon test. Et ensuite, d'évaluer l'erreur de généralisation de la FA mais cette fois-ci *In-Bag*, de sorte que l'estimateur soit plus robuste. Cette méthode est particulièrement consistante mais ne peut être utilisée que pour les $\mathcal{L}_{\mathrm{n_i}}^{\mathrm{j}}$ où $\{\mathrm{n_i} \ll \mathrm{p_l}\}$ ou lorsque $\{\mathrm{p_l} = \text{"petit"}\}$, i.e. une dizaine de variables (Liaw et Wiener, 2006).

Une version corrigée de l'err.OOB permet de valider la qualité prédictive des FA. Cependant elle n'est pas implémentée dans l'algorithme *randomForest* (Efron et Tibshirani, 1997).

La valeur de R^{OOB}(FA) est intimement liée à la qualité des données d'apprentissage et aux valeurs des paramètres spécifiés pour construire le prédicteur FA. Par conséquent pour optimiser la qualité du modèle il convient d'identifier la valeur des paramètres de la FA qui minimisent l'err.OOB – ce qui est justement l'objet de la phase de *calibration*.

Calibration du modèle

La calibration des paramètres de la FA est une étape nécessaire et préalable à l'analyse de l'importance des variables. Les FA permettent de caractériser les interactions entre la réponse $y_{\cdot}^{j}$ et les v.a. auxiliaires $(x_{\cdot}^{1}, \dots, x_{\cdot}^{p_I})$ en attribuant des scores d'importance aux variables. Or la robustesse des scores est intimement liée à la qualité prédictive de l'opérateur FA. Par conséquent, il est nécessaire de procéder à la calibration du modèle de sorte que *randomForest* suppute des scores qui soient statistiquement consistants pour qu'ils puissent être interprétés avec le plus de certitude possible.
Dans *randomForest* le nombre de paramètres sur lesquels il est possible de jouer est incommensurablement grand. A ce jour, il n'existe aucune procédure connue permettant d'obtenir *la FA optimale*. Néanmoins, il est possible d'optimiser les FA. Dans la pratique, pour calibrer les FA, on s'intéresse à la variation de R^{OOB}(FA) en fonction des paramètres **ntree** et **mtry** (Genuer, 2010).

S'agissant du paramètre **node**. **size,** qui est aussi discuté dans la littérature, les valeurs par défaut sont *a priori* suffisamment robustes. Cependant dans le contexte de la régression il peut être intéressant *de diminuer* la valeur par défaut tel que {nodesize < 5} (Tuleau-Malot, 2011).

La calibration du vecteur de paramètre : $\Lambda_j = \{\text{ntree} ; \text{mtry}\}$ s'effectue numériquement. Il s'agit de déterminer les valeurs de Λ_j qui minimisent l'err.OOB. Dans la mesure où le paramètre **ntree** est inversement croissant à R^{OOB}(FA) et qu'il peut prendre un nombre infini de valeurs, il convient simplement de déterminer le seuil à partir duquel le gain prédictif n'est plus significatif au regard de l'augmentation des temps de calcul. En revanche, le paramètre **mtry** est beaucoup plus difficile à calibrer car ses variations ne présupposent pas l'évolution de R^{OOB}(FA).
Pour calibrer **mtry** il est possible aussi d'utiliser le package tuneRF (Ligges, 2013). Le problème de cet *algorithme d'optimisation non linéaire* en particulier, comme de tous les algorithmes d'optimisation, est d'avoir des *performances liées à la condition initiale et au pas de calcul fixé arbitrairement par l'utilisateur*. Par conséquent, *tuneRF* n'est pas sans faille. De fait, ils peuvent être utilisés pour *corroborer ou affiner,* ceux de la technique proposée par Robin Génuer et qui est beaucoup plus robuste car elle permet de tester toute l'étendue des valeurs pouvant être prises par **mtry**.

Par ailleurs, et indépendamment de la technique de calibration, comme *randomForest* est instable l'err.OOB est généralement stabilisée sur 10 forêts, ou *runs*. Dans l'idéal il conviendrait de la stabiliser R^{OOB}(FA) sur 50 *runs*. Mais comme cette procédure est particulièrement chronophage, dans la pratique, on admet que 10 forets suffisent (Genuer, 2010).
En somme, le modèle FA qualifié d'optimal, dans le cadre cette thèse, utilise le vecteur $\widehat{\Lambda}_j$ estimé numériquement par minimisation de l'err.OOB stabilisée, tel que :

$$\widehat{\Lambda}_j = \underset{\forall\, \Lambda \in \{\Omega.\text{genuer} \cup \Omega.\text{tuneRF}\}}{\operatorname{argmin}} \left\{ \overline{R}_{\cdot}^{OOB}(\text{FA}) = \frac{1}{10} \sum_{r=1}^{10} R_r^{OOB}(\text{FA}) \right\}$$

Avec : **Ω. tuneRF** les paramètres spécifiés de la fonction *tuneRF* : valeur initiale, pas de calcul et le point d'arrêt.

Et : **Ω. genuer** l'étendue des valeurs à tester et préconisées par Robin Génuer pour **mtry** :

$$\mathbf{\Omega.\,genuer} = \left\{\left[1; \frac{\sqrt{\mathrm{p_l}}}{2}; \sqrt{\mathrm{p_l}}; 2\sqrt{\mathrm{p_l}}\,; 4\sqrt{\mathrm{p_l}}\,;\ \frac{\mathrm{p_l}}{4}; \frac{\mathrm{p_l}}{3}; \frac{\mathrm{p_l}}{2}; \frac{3}{4}\mathrm{p_l}; \mathrm{p_l}\right]\right\}$$

S'agissant de **ntree** la valeur spécifiée est celle qui n'améliore plus de façon significative les performances du prédicteur FA – en supposant auparavant que {ntree ≤ 2 500}.

Avec cette stratégie de calibration la majorité des études menées en mathématiques montre que, sur le plan prédictif, les FA diminuent de 30% l'erreur commise en termes de variance par rapport à un prédicteur CART optimal. Et sur le plan explicatif, i.e. sur l'analyse de l'importance des variables, il n'y a pas lieu de s'étendre, seul *randomForest* permet d'estimer des scores d'importance des variables statistiquement consistants (Genuer, 2010).

Importance des variables

Comprendre et expliquer les mécanismes de fonctionnement d'un phénomène revient à étudier la façon dont ses réalisations y_{i}^{j} interagissent avec des v.a. auxiliaires mesurables potentiellement explicatives $x_{\cdot}^{l} = \left(x_i^1, \ldots, x_i^{p_l}\right)$ - et aussi les liens internes qu'elles entretiennent. Pour ce faire, Léo Breiman suggère l'utilisation de scores d'importance des variables : $\mathrm{VI}(\mathrm{X^l})$. Ils permettent de mesurer l'efficience des $x_{\cdot}^{l}$, sur les observations de la réponse $y_{\cdot}^{j}$, quelles que soient leur nature et les corrélations qu'elles peuvent entretenir. Les scores d'importance des variables nourrissent le dessein d'identifier, de façon objective, les leviers – i.e. les $x_{\cdot}^{l}$ qui permettent d'interagir, voire d'infléchir, l'issue du phénomène d'intérêt.

L'idée de Léo Breiman fut de créer des scores d'importance des variables en utilisant de la façon la plus topique qui soit les propriétés des échantillons *Bootstrap* et du *Random-Subspace*. Les scores $\overline{\mathrm{VI}}_j(x^{\mathrm{l}})$ sont calculés à partir de l'err.OOB.k : $\mathrm{R}^{\mathrm{OOB}}\left(\mathrm{T_k}|\widehat{\boldsymbol{\Lambda}}_j\right)$ et d'une seconde erreur : err.OOB.k~l : $\mathrm{R}^{\mathrm{O\tilde{O}B^l}}\left(\mathrm{T_k}|\widehat{\boldsymbol{\Lambda}}_j\right)$. Comme l'err.OOB.k , l'err.OOB.k~l est estimée *Out-Of-Bag* et individuellement sur chaque prédicteur CART doublement perturbé. Pour chacun des $\mathrm{T_k}$ de la FA, la coordonnée $x_{\cdot}^{l}$ de l'individu e_i est systématiquement substituée par une autre coordonnée – dont l'indice est tiré aléatoirement dans une loi uniforme discrète. L' err.OOB.k~l est calculée, pour chaque $\mathrm{T_k}$ à partir de $\mathcal{L}_{\mathrm{n_i}}^{\mathrm{j}}$ mais uniquement avec les e_i qui n'ont pas été utilisés pour le construire – et dont le vecteur de coordonnées $\left(x_{\cdot}^{l}\right)$ est substitué par celui d'autres coordonnées choisies aléatoirement parmi $(x_{\cdot}^{1}, \ldots, x_{\cdot}^{p_l}) \ominus \left(x_{\cdot}^{l}\right)$. Ce nouvel échantillon Bootstrap perturbé OOB~l permet d'effectuer, pour chaque e_i, des prédictions OOB~l $\hat{y}_{\mathrm{i}}^{\mathrm{O\tilde{O}B^l},j}$ et par extension pour les arbres concernés d'évaluer son err.OOB.k~l, tel quel :

$$\mathrm{R}^{\mathrm{O\tilde{O}B^l}}\left(\mathrm{T_k}|\widehat{\boldsymbol{\Lambda}}_j\right) = \begin{cases} \dfrac{1}{\mathrm{n}_i}\displaystyle\sum_{\mathrm{i}=1}^{\mathrm{n}_i}\left(\hat{y}_{\mathrm{i}}^{\mathrm{O\tilde{O}B^l},j} - y_{\mathrm{i}}^{j}\right)^2 & \text{en régression} \\ \dfrac{1}{\mathrm{n}_i}\displaystyle\sum_{\mathrm{i}=1}^{\mathrm{n}_i}\mathbb{1}_{\left\{\hat{y}_{\mathrm{i}}^{\mathrm{O\tilde{O}B^l},j} \neq y_{\mathrm{i}}^{j}\right\}} & \text{en classification} \end{cases}, \qquad \forall\, k = \{1, \ldots, K\}$$

Les $\mathrm{R}_{\mathrm{r}}^{\mathrm{OOB}}\left(\mathrm{T_k}|\widehat{\boldsymbol{\Lambda}}_j\right)$ et $\mathrm{R}^{\mathrm{O\tilde{O}B^l}}\left(\mathrm{T_k}|\widehat{\boldsymbol{\Lambda}}_j\right)$ sont donc calculées arbre par arbre et non sur le prédicteur global FA, ce qui présente l'avantage d'éluder la sous-estimation induite par la phase d'agrégation ensembliste finale. De fait, l'estimation de *l'effet des permutations aléatoires* sur la variable $x_{\cdot}^{l}$ est amplifié car la valeur de $\mathrm{R}^{\mathrm{O\tilde{O}B^l}}\left(\mathrm{T_k}|\widehat{\boldsymbol{\Lambda}}_j\right)$ est un temps surestimée. Une véritable manne dans la mesure où, ce qu'il s'agit d'apprécier, est justement *l'effet permutation* de la coordonnée $x_{\cdot}^{l}$

Les scores d'importance des variables résultent de l'estimation de la différence entre toutes les $\mathrm{R}_{\mathrm{r}}^{\mathrm{OOB}}\left(\mathrm{T_k}|\widehat{\boldsymbol{\Lambda}}_j\right)$ et les $\mathrm{R}_{\mathrm{r}}^{\mathrm{O\tilde{O}B^l}}\left(\mathrm{T_k}|\widehat{\boldsymbol{\Lambda}}_j\right)$. L'algorithme *randomForest* étant instable les scores doivent faire

l'objet d'une stabilisation – qui en pratique est effectuée sur $\{n_r = 50\}$ forets. Ils s'estiment de la façon suivante :

$$\overline{\mathrm{VI}}_j(\,x^{\mathrm{l}}) = \frac{1}{\mathrm{n_r}\cdot \mathrm{K}}\cdot \sum_{\mathrm{r}=1}^{\mathrm{n_r}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\left(\mathrm{R}_{\mathrm{r}}^{\mathrm{O\tilde{O}B^l}}\left(\mathrm{T_k}|\widehat{\boldsymbol{\Lambda}}_j\right) - \mathrm{R}_{\mathrm{r}}^{\mathrm{OOB}}\left(\mathrm{T_k}|\widehat{\boldsymbol{\Lambda}}_j\right)\right),\quad \forall\, \mathrm{l} \in \{1,\dots,\mathrm{p_l}\}$$

Les $\overline{\mathrm{VI}}_{\mathrm{j}}(x^l)$ reposent sur un postulat assez intuitif. Plus les permutations OOB~l engendrent une forte augmentation de l'erreur, plus le score d'importance des variables tend à prendre des valeurs fortes, et donc plus la variable substituée est importante au sens de *randomForest*.
En théorie les $\mathrm{VI}(\mathrm{X^l}) \geq 0$. Or, dans la pratique, il arrive parfois que $\mathrm{VI}(\mathrm{X^l}) < 0$. Toutefois, les valeurs négatives restent généralement très faibles. Ce processus est connu et même couramment observé. Il est dû à l'instabilité de l'algorithme. Certains auteurs suggèrent de ramener manuellement les $\overline{\mathrm{VI}}_{\mathrm{j}}(x^l) = 0$ lorsque $\overline{\mathrm{VI}}_{\mathrm{j}}(x^l) \longrightarrow 0^-$. En revanche, si des $\overline{\mathrm{VI}}_{\mathrm{j}}(x^l) \ll 0$ sont observés, il convient d'accorder une attention particulière à l'interprétation de ces scores (Genuer, 2012).

L'estimateur biaisé de l'écart-type est renvoyé par l'algorithme *randomForest.* Il permet entre autres, de se faire une idée de la dispersion des scores et donc, par extension, de leur robustesse. A l'instar des scores il convient de les stabiliser sur $\{\mathrm{n_r} = 50\}$ runs :

$$\hat{\sigma}\left(\mathrm{VI}_j(\,x^{\mathrm{l}})\right) = \sqrt{\frac{1}{\mathrm{n_r}\cdot \mathrm{K}}\cdot \sum_{\mathrm{r}=1}^{\mathrm{n_r}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\left(\left(\mathrm{R}_{\mathrm{r}}^{\mathrm{O\tilde{O}B^l}}\left(\mathrm{T_k}|\widehat{\boldsymbol{\Lambda}}_j\right) - \mathrm{R}_{\mathrm{r}}^{\mathrm{OOB}}\left(\mathrm{T_k}|\widehat{\boldsymbol{\Lambda}}_j\right)\right) - \left(\overline{\mathrm{R}}_{\mathrm{r}}^{\mathrm{O\tilde{O}B^l}}\left(\mathrm{T}.|\widehat{\boldsymbol{\Lambda}}_j\right) - \overline{\mathrm{R}}_{\mathrm{r}}^{\mathrm{OOB}}\left(\mathrm{T}.|\widehat{\boldsymbol{\Lambda}}_j\right)\right)\right)^2}$$

L'analyse des $\hat{\sigma}\left(\mathrm{VI}_j(x^{\mathrm{l}})\right)$ n'est pas intuitive. En revanche, ils permettent de *réduire* les scores – énoncés auparavant en les ramenant à une échelle unitaire afin de mieux percevoir leurs variabilités.

$$\overline{\mathrm{VI}}_j^{\mathbb{r}}(\,x^{\mathrm{l}}) = \frac{\overline{\mathrm{VI}}_j(\,x^{\mathrm{l}})}{\hat{\sigma}\left(\mathrm{VI}_j(\,x^{\mathrm{l}})\right)},\quad \forall\, j \in \{1,\dots,p_l\}$$

Les scores réduits $\overline{\mathrm{VI}}_{\mathrm{j}}^{\mathbb{r}}(x^l)$ apportent une information complémentaire et permettent surtout de vérifier celle fournie par l'indicateur phare : $\overline{\mathrm{VI}}_{\mathrm{j}}(x^l)$. Ils permettent indirectement de s'assurer de la consistance des $\overline{\mathrm{VI}}_{\mathrm{j}}(x^l)$, des plus forts comme de ceux prenant des valeurs intermédiaires. Toutefois, la procédure de sélection des variables : VSURF se base exclusivement sur les $\overline{\mathrm{VI}}_{\mathrm{j}}(x^l)$
L'algorithme *randomForest* permet aussi de calculer des scores historiques tels qu'ils étaient supputés avec l'algorithme *rpart* : $\overline{\mathrm{VI}}_{\mathrm{j}}^{\mathrm{CART}}(x^l)$. Mais ces scores sont biaisés en faveur des variables qualitatives multi-classes *et ne présentent plus guère d'intérêt* (Genuer, 2012).

Les principes fondamentaux permettant de construire les estimateurs FA et CART tels qu'ils sont renvoyés par les algorithmes *randomForest* et *rpart* ont été décrits pour faciliter la compréhension de la stratégie de sélection de variables : VSURF. Elle est utilisée pour identifier les facteurs environnementaux liés aux phénomènes morbides d'intérêt.

ANNEXE 6 : COMPLEMENTS THEORIQUES SUR VSURF

La méthode Variable Selection Using Random Forest (VSURF) est une stratégie statistique qui permet de disjoindre par seuillage, dans des jeux de données *en grande dimension* et *a fortiori* dans des jeux de *données classiques,* les variables *de bruit* des *variables explicatives* et par suite de constituer un paquet de variables qui s'adapte à la *parcimonie prédictive*. VSURF est une proposition méthodologique récente. Les principes énoncés reprennent exactement ceux qui sont plus amplement décrits dans la thèse de Robin Genuer - (Genuer, 2010). Une version bêta de VSURF vient récemment d'être programmée dans un package.R (Genuer, Poggi et al., 2013).

A l'initialisation on, dispose d'un jeu de données d'apprentissage $\mathcal{L}_{\mathrm{n_i}}^{\mathrm{j}}$ constitué d'un vecteur *cible* $y_{\cdot}^{j}$ et d'une matrice de variables auxiliaires $x_{\cdot}^{l} = (x_{\cdot}^{1}, \dots, x_{\cdot}^{p_l})$. La méthode VSURF permet de répondre aux deux objectifs de la sélection de variables : *L'interprétation* dont le dessein est de dissocier les variables $x_{\cdot}^{l}$ qui interagissent avec la réponse étudiée, même si ces dernières sont redondantes ou corrélées - des variables $x_{\cdot}^{l}$ qui n'ont statistiquement aucun lien avec $y_{\cdot}^{j}$. L'objectif est de comprendre comment fonctionne le phénomène d'intérêt mais aussi et surtout comment interagir avec lui ;
La prédiction dont le but est trouver un ensemble *parcimonieux de variables* de $x_{\cdot}^{l}$ pour envisager l'évolution future, i.e. la plus probable, du phénomène d'intérêt. Et par suite d'essayer d'anticiper ses autres évolutions possibles lorsqu'on agit sur des leviers, i.e. sur lesdites $x_{\cdot}^{l}$ importantes.

La méthode VSURF se décline en trois phases : classement descendant des $x_{\cdot}^{l}$ en fonction de leurs scores *randomForest* et élimination des variables de bruit, sélection d'un panel de variables explicatives, et identification d'un paquet de variables qui s'adapte à la parcimonie prédictive.

Phase 1 : Hiérarchisation et élimination des variables de bruit

L'objectif est d'identifier, grâce à un seuil d'élimination $\psi_{\mathrm{bruit}}^{j}$, les variables de *bruit avérées* de celles qui sont *potentiellement explicatives.* Cette phase d'élimination des variables de bruit ne se base pas directement sur la moyenne des scores *randomForest* mais sur leurs écarts-types d'estimation. Or la robustesse de ces deux indicateurs est intimement liée aux capacités prédictives de l'estimateur FA (annexe.7).
Par conséquent il convient, avant tout, de calibrer numériquement les paramètres de la FA. Pour ce faire, on étudie les variations de l'erreur de généralisation de différents modèles, stabilisée sur 10 forêts : $\overline{\mathrm{R}}_{\cdot}^{\mathrm{OOB}}\left(\mathrm{FA}\middle|\Lambda_j\right)$ en jouant sur le vecteur de paramètres : $\Lambda_j = \{\mathbf{ntree}; \mathbf{mtry}; \mathbf{nodesize}\}$. Les valeurs de $\widehat{\Lambda}_j$ sont celles qui minimisent l'erreur de généralisation de la FA. Elles sont choisies parmi celles proposées dans l'ensemble : $\boldsymbol{\Omega}.\mathbf{genuer}$. Les valeurs spécifiées pour **mtry** et **ntree** sont calibrées numériquement et sont facilement estimables à partir de représentations graphiques. Celle de **nodesize** est celle implémentée par défaut dans *randomForest* mais peut être diminuée, *a priori* à partir de connaissances expertes, dans le contexte de la régression (annexe.7).

Une fois le modèle de FA optimisé il s'agit d'estimer la moyenne et les écarts-types des scores d'importance des variables stabilisés sur $\{\mathrm{n_r} = 50\}$ forêts afin d'obtenir :

$$\left(\overline{\mathrm{VI}}_j\left(x^{\mathrm{l}}\right);\ \widehat{\sigma}\left(\mathrm{VI_j}\left(x^{(l)}\right)\right)\right), \forall\, \mathrm{l} = \left\{1, \dots, \{\mathrm{p_l} \in \mathbb{N}^1\}\right\}$$

Proposition heuristique (i) - pour l'interprétation et l'analyse géographique :
Récupérer les scores réduits stabilisés $\overline{\mathrm{VI}}_j^{\mathbb{r}}\left(x^{\mathrm{l}}\right)$ et les scores CART stabilisés : $\overline{\mathrm{VI}}_{\mathrm{j}}^{\mathrm{CART}}\left(\mathrm{x^l}\right)$; Et présenter les valeurs $\mathrm{VI_{j,r}}\left(x^{l}\right)$, $\mathrm{VI_{j,r}^{\mathbb{r}}}\left(x^{l}\right)$, $\mathrm{VI_{j,r}^{CART}}\left(x^{l}\right)$ et leurs distributions statistiques par le biais de *diagrammes de Tukey ;* Faire apparaitre l'estimateur de la moyenne ainsi que les *outliers* (Saporta, 2006).

Une fois les scores d'importance stabilisés : $\overline{\mathrm{VI}}_{\mathrm{j}}(x^{l})$ il s'agit de hiérarchiser les variables par ordre décroissant de score et ainsi constituer $\mathcal{X}^{\mathrm{j}}_{\mathrm{ord}}$, tel que :

$$\mathcal{X}^{\mathrm{j}}_{\mathrm{ord}} = \left(\left(x_{\cdot}^{(1)}, \dots, x_{\cdot}^{(l)}, \dots, x_{\cdot}^{(p_l)} \right) \middle| \left\{ \overline{\mathrm{VI}}_{\mathrm{j}}\left(x_{\cdot}^{(1)} \right) \geq \cdots \geq \overline{\mathrm{VI}}_{\mathrm{j}}\left(x_{\cdot}^{(l)} \right) \geq \cdots \geq \overline{\mathrm{VI}}_{\mathrm{j}}\left(x_{\cdot}^{(p_l)} \right) \right\} \right)$$

Où $(l) = \mathrm{rang}\left(x_{\cdot}^{l}\right),\ \forall\, l \in \{1, \dots, p_l\}$. Ensuite, VSURF procède à *l'élimination des variables de bruit*. La stratégie proposée est dite *conservatrice* car elle permet de récupérer toutes les variables qui peuvent potentiellement interagir avec $y_{\cdot}^{j}$. De fait, certaines variables de bruit, ou très faiblement efficientes, ne sont pas éliminées. En somme, la phase d'élimination permet de scinder $\mathcal{X}^{\mathrm{j}}_{\mathrm{input}}$ en deux ensembles disjoints, les variables *potentiellement explicatives*, et les variables *de bruit avéré,* tel que

$$\mathcal{X}^{\mathrm{j}}_{\mathrm{input}} = \left\{ \left\{ \mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}} \right\} \cup \left\{ \mathcal{X}^{\mathrm{j}}_{\mathrm{bruit}} \right\} \right\}$$

Les variables *de bruit avéré* sont celles dont le score est inférieur au seuil : $\psi^{\mathrm{j}}_{\mathrm{bruit}}$. Celui-ci *ne tient pas compte directement de l'importance des variables* puisqu'il s'estime à partir de leurs écarts-types

$$\widehat{\sigma}\left(\mathrm{VI}_{\mathrm{j}}\left(x^{(l)} \right) \right) = \left(\widehat{\sigma}\left(\mathrm{VI}_{\mathrm{j}}\left(x_{\cdot}^{(1)} \right) \right), \dots, \widehat{\sigma}\left(\mathrm{VI}_{\mathrm{j}}\left(x_{\cdot}^{(p_l)} \right) \right) \right)$$

L'idée repose sur le fait que les variables de bruit ont une importance quasi nulle $\left\{ \overline{\mathrm{VI}}_{\mathrm{j}}\left(\mathrm{x}^{l} \right) \approx 0 \right\}$. Mais comme l'algorithme *randomForest* est instable, certaines peuvent, du simple fait du hasard, avoir des scores légèrement positifs. Il s'agit donc de conserver toutes les variables ayant des $\overline{\mathrm{VI}}_{\mathrm{j}}\left(x_{\cdot}^{l} \right)$ supérieurs à zéro plus un *écart-type représentatif de la variabilité des scores*. Puisque les variables explicatives ont une variabilité plus forte que les variables de bruit, alors leurs *écarts-types* doivent suggérer *un saut moyen* qu'une variable même faiblement explicative serait susceptible d'effectuer mais qu'une variable de bruit est incapable de faire.

Pour mettre en œuvre cette stratégie, le seuil : $\psi^{\mathrm{j}}_{\mathrm{bruit}}$ est proposé. Il correspond à la valeur minimale estimée par un arbre CART, i.e. un prédicteur constant par morceaux, construit à partir des rangs des variables de l'écart-type des scores. Le prédicteur CART utilisé est l'arbre de la sous-suite de Breiman dont l'erreur de validation croisée est la plus petite. En d'autres termes, le seuil d'élimination est donné par :

$$\psi^{\mathrm{j}}_{\mathrm{bruit}} = \left\{ \min_{l=\{1,\dots,p_l\}} \left\{ \widehat{\mathrm{T}}^{\widehat{\sigma}(\cdot)}_{\mathrm{CART}}((l)) \right\} \middle| \mathrm{T}^{\widehat{\sigma}\left(\mathrm{VI}_{\mathrm{j}}\left(x^{(l)}\right)\right)}_{\mathrm{CART}} \left(\mathrm{rang}\left(x_{\cdot}^{l} \right) \right) = \underset{\forall\, \mathrm{T}^{\widehat{\sigma}(\cdot)}_{\mathrm{ko}} \in \mathcal{T}^{\cdot}_{\mathrm{Breiman}}}{\mathrm{argmin}} \left\{ \mathrm{R}^{\mathrm{cv}} \left(\mathrm{T}^{\widehat{\sigma}(\cdot)}_{\mathrm{ko}}((l)) \right) \right\} \right\}$$

Une fois ψ_{bruit} estimé il ne reste plus qu'à disjoindre les variables conservées, i.e. les variables potentiellement explicatives, des variables de bruit avéré.

L'ensemble des variables potentiellement explicatives est défini, tel que :

$$\mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}} = \left(x_{\cdot}^{(1)}, \dots, x_{\cdot}^{(g_l)} \right) = \bigcup_{(l)=1}^{g_l} \left(x_{\cdot}^{(l)} \middle| \left\{ \overline{\mathrm{VI}}\left(x_{\cdot}^{(l)} \right) \geq \psi^{\mathrm{j}}_{\mathrm{bruit}} \right\} \right), \qquad \text{avec: } \{ g_l \leq p_l \}$$

L'ensemble des variables de bruit avéré éliminées est défini tel que :

$$\mathcal{X}^{\mathrm{j}}_{\mathrm{bruit}} = \left(x_{\cdot}^{(g_l+1)}, \dots, x_{\cdot}^{(p_l)} \right) = \bigcup_{(l)=g_l+1}^{p_l} \left(x_{\cdot}^{(l)} \middle| \left\{ \overline{\mathrm{VI}}\left(x_{\cdot}^{(l)} \right) < \psi^{\mathrm{j}}_{\mathrm{bruit}} \right\} \right), \qquad \text{avec: } \{ g_l \leq p_l \}$$

Remarque : l'application de VSURF à des jeux de données jouées montre que généralement $\mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}}$ contient l'intégralité des variables explicatives mais que cet ensemble contient aussi un nombre important de variables de bruit. En revanche, $\mathcal{X}^{\mathrm{j}}_{\mathrm{bruit}}$ ne contient *presque sûrement* que des variables de bruit avéré (Genuer, 2010).

Phase 2 : Sélection des variables explicatives

Il s'agit maintenant d'identifier parmi les variables conservées celles qui sont susceptibles d'interagir de façon suffisamment forte avec les réalisations du phénomène d'intérêt pour être qualifiées d'explicatives – au sens statistique du terme – et constituer ainsi le paquet :

$$\mathcal{X}^{\mathrm{j}}_{\mathrm{explic}} = \left(x_{\cdot}^{(1)}, \dots, x_{\cdot}^{(e_l)}\right), \qquad \forall\, e_l \leq g_l \leq p_l$$

L'idée *consiste à imbriquer les variables conservées,* par ordre décroissant de score moyen, dans des modèles de *forêts aléatoires* et à ne retenir que l'ensemble de variables le plus frustre qui améliore *significativement*, au sens statistique du terme, l'erreur OOB stabilisée.
Pour ce faire, on injecte *une par une* les $x_{\cdot}^{(l)}$ retenues dans la phase d'élimination, i.e. contenues dans $\mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}}$ dans des *modèles imbriqués de FA paramétrées par défaut* $FA\left(\cdot \middle| \dot{\boldsymbol{\Lambda}}\right)$ - avec $\dot{\boldsymbol{\Lambda}}$ le vecteur des paramètres par défaut - et on estime leur erreur OOB moyenne et leur écart-type OOB, stabilisés sur $\mathrm{n_r}$ forêts. Le modèle retenu est celui contenant le sous ensemble de variables le plus petit et dont l'erreur OOB stabilisée est en deçà *d'un seuil explicatif* : $\psi^{j}_{\mathrm{explic}}$.

Remarque : l'utilisation des modèles de FA par défaut est admise lorsque le nombre de variables est "petit" et que le nombre de combinaisons de variables à tester est "grand" (Genuer, 2012).

Cette technique de sélection est basée sur une estimation itérative des g_l valeurs des erreurs OOB et des écarts-types stabilisés des modèles imbriqués de $\hat{f}^{j}_{\mathrm{FA}}\left(\cdot \middle| \dot{\boldsymbol{\Lambda}}\right)$ qu'il est possible de former, en injectant les variables conservées de l'ensemble $\mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}}$. En conséquence de quoi, on obtient la collection de vecteurs suivante :

$$\left(\bar{R}^{\mathrm{OOB}}_{\left(\mathcal{X}^{\mathrm{j}}_{\mathrm{imbr}}\right)}, \hat{\sigma}_{\left(R^{OOB}_{(\mathrm{imbr})}\right)}\right) = \begin{pmatrix} \left(\bar{R}^{OOB}_{\cdot}\left(\hat{f}^{j}_{\mathrm{FA}}\left(\mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}}(1)\middle|\dot{\Lambda}\right)\right); \hat{\sigma}\left(R^{OOB}_{\cdot}\left(\hat{f}^{j}_{\mathrm{FA}}\left(\mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}}(1)\middle|\dot{\boldsymbol{\Lambda}}\right)\right)\right)\right) \\ \vdots \\ \left(\bar{R}^{OOB}_{\cdot}\left(\hat{f}^{j}_{\mathrm{FA}}\left(\mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}}(g_l)\middle|\dot{\Lambda}\right)\right); \hat{\sigma}\left(R^{OOB}_{\cdot}\left(\hat{f}^{j}_{\mathrm{FA}}\left(\mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}}(g_l)\middle|\dot{\boldsymbol{\Lambda}}\right)\right)\right)\right) \end{pmatrix}$$

Pour simplifier les notations utilisées : $\mathcal{X}^{\mathrm{j}}_{\mathrm{imbr.l'}} \overset{\mathrm{def}}{=} \mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}}(\{l' \leq g_l\}) = \cup_{l=1}^{l'}\left(\left\{x_{\cdot}^{(l)} \in \mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}}\right\}\right)$.
Idéalement, il conviendrait de conserver le paquet de variables $\mathcal{X}^{\mathrm{j}}_{\mathrm{imbr.l}}$ pour lequel l'err.OOB stabilisée des modèles par défaut imbriqués, est minimum. Or, il faut prendre aussi en compte l'instabilité de l'algorithme afin de ne pas évincer un trop grand nombre de variables réellement explicatives. Pour cela, une astuce statistique classique est utilisée. Le paquet de variables retenues est l'ensemble le plus frustre dont l'erreur OOB est inférieure au *min augmenté d'un SD*.
Le seuil de sélection explicatif : $\psi^{j}_{\mathrm{explic}}$ s'estime de la façon suivante :

$$\psi^{j}_{\mathrm{explic}} = \left(\bar{R}^{\mathrm{OOB}}_{\left(\mathcal{X}^{\mathrm{j}}_{\mathrm{imbr.lo}}\right)} + 1 \cdot \hat{\sigma}_{\left(R^{OOB}_{(\mathrm{imbr.lo})}\right)}\middle| \left\{\bar{R}^{\mathrm{OOB}}_{\left(\mathcal{X}^{\mathrm{j}}_{\mathrm{imbr.lo}}\right)} = \min_{\forall l=\{1,\dots,g_l\}}\left\{\bar{R}^{OOB}_{\cdot}\left(\hat{f}^{j}_{\mathrm{FA}}\left(\mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}}(l)\middle|\dot{\Lambda}\right)\right)\right\}\right\}\right),$$

L'ensemble des variables explicatives est défini par :

$$\mathcal{X}^{\mathrm{j}}_{\mathrm{explic}} = \left(x_{\cdot}^{(1)}, \dots, x_{\cdot}^{(e_l)}\right) = \bigcup_{l=1}^{g_l}\left(x_{\cdot}^{(l)}\middle|\left|\mathcal{X}^{\mathrm{j}}_{\mathrm{imbr.e_l}}\right| = \min_{\forall l'=\{1,\dots,g_{l'}\}}\left\{\left|\mathcal{X}^{\mathrm{j}}_{\mathrm{imbr.}l'}\right|\right\} \cap \left\{\bar{R}^{\mathrm{OOB}}_{\left(\mathcal{X}^{\mathrm{j}}_{\mathrm{imbr.}l'}\right)} < \psi_{\mathrm{explic}}\right\}\right)$$

Le paquet des variables éliminées dans cette seconde phase est noté $\mathcal{X}^{\mathrm{j}}_{\mathrm{elim}}$. Il correspond à la partie duale de $\mathcal{X}^{\mathrm{j}}_{\mathrm{explic}}$ incluse dans $\mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}}$, tel que :

$$\mathcal{X}^{\mathrm{j}}_{\mathrm{elim}} = \left(\mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}} \ominus \mathcal{X}^{\mathrm{j}}_{\mathrm{explic}}\right) = \left(x_{\cdot}^{(e_l+1)}, \dots, x_{\cdot}^{(g_l)}\right), \quad \text{avec: } e_l \leq g_l$$

Remarque : il peut y avoir lieu de discuter les résultats obtenus par la stratégie de sélection des variables explicatives. D'une part parce que son application à des jeux de données jouées a montré que si le paquet de variables conservées : $\mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}}$ ne contient que des variables réellement explicatives, certaines d'entre elles seront systématiquement éliminées lors de la constitution du paquet $\mathcal{X}^{\mathrm{j}}_{\mathrm{explic}}$.

D'autre part, il arrive parfois que le paquet de variables explicatives soit identique au paquet de variables prédictives. Or, dans ce cas, généralement plusieurs sous-ensembles de modèles imbriqués satisfont la condition au non dépassement du seuil explicatif : $\psi^{j}_{\mathrm{explic}}$ et le sous-ensemble explicatif retenu est finalement trop frustre au regard du nombre de variables réellement explicatives éliminées.

Proposition heuristique (ii) vouée à l'interprétation et à l'analyse géographique :
Introduire une règle supplétive à celles de VSURF afin de retenir un paquet de variables $\boldsymbol{\mathcal{X}^{\mathrm{j}}_{\mathbf{0.explic}}}$ plus exhaustif pour pallier le risque d'élimination de variables explicatives peu efficientes car maculées par des bruits de fond environnementaux qui complexifient la modélisation géographique des FE/FIM. Le paquet « Bourrelly » proposé se construit identiquement à $\mathcal{X}^{\mathrm{j}}_{\mathrm{explic}}$ – le paquet de « Genuer » - si ce n'est qu'est conservé le modèle imbriqué dont l'erreur OOB est en deçà de $\psi^{j}_{\mathrm{explic}}$ et dont le nombre de variables est inférieur ou égal au maximum entre le sous-ensemble le plus frustre et 60% des variables contenues dans $\mathcal{X}^{\mathrm{j}}_{\mathrm{conserv}}$, soit :

$$\mathcal{X}^{\mathrm{j}}_{0.\mathrm{explic}} = \bigcup_{l=1}^{g_l} \left(x^{(l)}_{\cdot} \middle| \left\{ \left| \mathcal{X}^{\mathrm{j}}_{\mathrm{imbr.e}_l} \right| \leq \left(\min_{\forall l = \{1,\dots,g_l\}} \left\{ \left| \mathcal{X}^{\mathrm{j}}_{\mathrm{imbr.l}} \right| \right\} \wedge \left[0{,}6 \cdot \left| \mathcal{X}^{\mathrm{j}}_{\mathrm{conser}} \right| \right] \right) \right\} \cap \left\{ \overline{R}^{\mathrm{OOB}}_{\left(\mathcal{X}^{\mathrm{j}}_{\mathrm{imbr.l}} \right)} < \psi^{j}_{\mathrm{explic}} \right\} \right)$$

Pertinence : l'application de cette règle sur des jeux de données jouées a montré que $\mathcal{X}^{\mathrm{j}}_{0.\mathrm{explic}}$ est souvent plus grand que $\mathcal{X}^{\mathrm{j}}_{\mathrm{explic}}$. Il contient généralement un nombre de variables réellement explicatives plus important mais aussi quelques variables de bruit. Sur une trentaine de répétitions le nombre de variables de bruit incluses est rarement supérieur au nombre de variables réellement explicatives ajoutées.

Intérêt : un ensemble de variables explicatives plus grand que celui de $\mathcal{X}^{\mathrm{j}}_{\mathrm{explic}}$ permet de discuter, à l'appui de connaissances expertes, de la pertinence des variables adjointes. Cette proposition trouve du sens lorsque MyVsurfGéo est appliquée à des données géographiques réelles (section.B) et, aussi et surtout, dans la proposition de BoostMyVsurfGéo appliquée aux données de la cohorte LEA, dans le cadre d'une approche individus-centrée (section.C).

Phase 3 : Identification des variables predictives

Il s'agit d'identifier parmi le paquet de *variables explicatives* $\mathcal{X}^{\mathrm{j}}_{\mathrm{explic}}$ celles qui *s'adaptent à la parcimonie prédictive*. Autrement dit, l'objectif est d'éliminer les variables corrélées ou redondantes afin d'éluder le sur-apprentissage qui dégrade les performances prédictives des modèles statistiques, en constituant le sous-ensemble :

$$\mathcal{X}^{\mathrm{j}}_{\mathrm{pred}} \subseteq \left\{\mathcal{X}^{\mathrm{j}}_{\mathrm{explic}} \cup \mathcal{X}^{\mathrm{j}}_{0.\mathrm{explic}}\right\}$$

L'idée consiste à *injecter séquentiellement* les variables explicatives par ordre décroissant de score à éliminer toutes celles qui n'apportent pas un gain de qualité significatif, i.e. supérieur au gain d'erreur OOB estimé à partir du modèle explicatif dans lequel on introduirait des variables de bruit et quelques variables explicatives *a priori* peu efficientes.

Pour mettre en œuvre cette idée la stratégie VSUF propose *un seuil d'élimination prédictif :* ψ^{j}_{pred}. Il est défini comme la moyenne des différences d'ordre un, en valeur absolue, des erreurs OOB de FA par défaut construites avec toutes les variables explicatives : $\mathcal{X}^{\mathrm{j}}_{\mathrm{explic}}$ et ce même modèle dans lequel on introduit récursivement les variables éliminées contenues dans $\mathcal{X}^{\mathrm{j}}_{\mathrm{elim}}$. Afin de prendre en compte l'instabilité de l'algorithme, la valeur de ψ^{j}_{pred} est stabilisée sur 50 forêts :

$$\psi^{j}_{\mathrm{pred}} = \frac{1}{\mathrm{n_r} \cdot (g_{\boldsymbol{l}} - e_{\boldsymbol{l}})} \cdot \sum_{r=1}^{\mathrm{n_r}} \sum_{l=e_{\boldsymbol{l}}}^{g_{\boldsymbol{l}}-1} \left| \mathrm{R}^{\mathrm{OOB}}_{\mathrm{r}}\left(\hat{f}^{j}_{\mathrm{FA}}\left(\mathcal{X}^{\mathrm{j}}_{\mathrm{use}}(l)\middle|\dot{\Lambda}\right)\right) - \mathrm{R}^{\mathrm{OOB}}_{\mathrm{r}}\left(\hat{f}^{j}_{\mathrm{FA}}\left(\mathcal{X}^{\mathrm{j}}_{\mathrm{use}}(l+1)\middle|\dot{\Lambda}\right)\right)\right|$$

Avec :

$$\mathcal{X}^{\mathrm{j}}_{\mathrm{use}}(\{l' \leq g_{\boldsymbol{l}}\}) = \left\{\bigcup_{l=1}^{e_{\boldsymbol{l}}}\left(x^{(l)}_{\cdot} \in \mathcal{X}^{\mathrm{j}}_{\mathrm{explic}}\right)\right\} \cup \left\{\bigcup_{l=e_{\boldsymbol{l}}+1}^{l'}\left(x^{(l)}_{\cdot} \in \mathcal{X}^{\mathrm{j}}_{\mathrm{elim}}\right)\right\}$$

Ensuite, il s'agit d'évaluer le gain d'erreur OOB induit par l'injection, *step by step,* des $x^{(\mathrm{l})}_{\cdot} \in \mathcal{X}^{\mathrm{j}}_{\mathrm{explic}}$ et d'éliminer toutes celles qui *n'améliorent pas significativement* la qualité du modèle. Pour ce faire une procédure récursive estime le rapport d'erreur entre le modèle de FA par défaut contenant les $x^{(\mathrm{l})}_{\cdot}$ retenues *au pas de calcul :* $(\mathrm{l}-1)$ et celui contenant les mêmes variables ainsi que celle testée $x^{(\mathrm{l}+1)}_{\cdot}$. Le rapport d'erreur OOB est stabilisé sur 50 runs :

$$\overline{\Delta}(R^{OOB}_{\cdot}) = \frac{1}{n_r}\sum_{r=1}^{n_r}\left(\mathrm{R}^{\mathrm{OOB}}_{\mathrm{r}}\left(\hat{f}^{j}_{\mathrm{FA}}\left(\left\{x^{(l)}_{\cdot} \in \mathcal{X}^{\mathrm{j}}_{\mathrm{pred}}\right\}\middle|\dot{\Lambda}\right)\right) - \mathrm{R}^{\mathrm{OOB}}_{\mathrm{r}}\left(\hat{f}^{j}_{\mathrm{FA}}\left(\left\{x^{(l)}_{\cdot} \in \mathcal{X}^{\mathrm{j}}_{\mathrm{pred}}\right\} \cup \left\{x^{(l+1)}_{\cdot}\right\}\middle|\dot{\Lambda}\right)\right)\right)$$

A l'initialisation de la procédure $x^{(1)}_{\cdot}$ est incluse dans $\mathcal{X}^{\mathrm{j}}_{\mathrm{pred}}$ – donc la variable la plus importante au sens de *randomForest* est systématiquement considérée comme prédictive. Par conséquent la première variable testée est $x^{(l+1)}_{\cdot} = x^{(2)}_{\cdot}$, elle est éliminée si le rapport d'erreur OOB stabilisé est inférieur ou égal au seuil d'élimination prédictif.

L'ensemble des variables *s'adaptant à la parcimonie prédictive* est défini par :

$$\mathcal{X}^{\mathrm{j}}_{\mathrm{pred}} = \left\{x^{(1)}_{\cdot}\right\} \cup \left\{\bigcup_{l=1}^{e_l-1}\left(x^{(l+1)}_{\cdot}\middle|\left\{\overline{\Delta}(R^{OOB}_{\cdot}) > \psi^{j}_{\mathrm{pred}}\right\}\right)\right\}$$

Cette annexe résume la façon dont VSURF est appliquée aux jeux de données géographiques. L'algorithme MyVsurGéo a été programmé en adéquation avec les principes énoncés si ce n'est que l'arbre CART de la phase d'élimination permettant d'estimer $\psi^{\mathrm{j}}_{\mathrm{bruit}}$ est plus profond que celui de VSURF dans la mesure où *le critère de complexité utilisé* $\alpha^{\mathrm{MyVsurfGeo}}_{\mathrm{min}}$, visant à réduire *l'effet de bore* de l'algorithme : *rpart*, est spécifié pour maximiser la contrainte *de non biais* (annexe.7). MyVsurGéo permet d'effectuer des prédictions *In-Sample* (IS) adaptées à la dialectique géographique (section.B).

TABLES DES INDICATEURS STAPIOTEMPORELS

La géographie des Phénomènes Morbides* (PM), des Facteurs Environnementaux* (FE) et des Facteurs Individuels et Médicaux* (FIM) - jugés pertinents* ou curieux* - est modélisée par des indicateurs spatio-temporels morbides* ou environnementaux*, notés i.st.m* ou i.st.e. Les bases de données utilisées sont déclinées dans le chapitre 1.

Modélisations géographiques morbides

Les séquelles d'intérêt sont : les cataractes (j=CATA), les tumeurs thyroïdiennes (j=THYR) et les tumeurs secondaires majeures (j=TUM2).

Variables réponses morbides			Notations		Valeurs prises
Etat de santé	Base	Variabilité spatiotemporelle	Texte	Logiciel R	
séquelle j	LEA	Prévalences spatiales pondérées	$z^{j}_{(U_k),s}$	z.prim[j]c_	⟦0,001 ; 0,192⟧
		Propensions spatiales pondérées	$z^{j}_{(U_k),q}$	z.prim[j]q	{OUI ; NON ; INCERTAIN}
		Risques d'Expositions Géographiques	$z^{REG.j}_{(U_k)}$	z_REG([j])	POSSIBLE PROBABLE INDEMONTRABLEFAIBLE

Tableau 56 : Liste des i.st.m* modélisant la géographie des phénomènes morbides

Modélisations géographiques environnementales

La géographie des expositions environnementales jugées pertinentes* est discrétisée en quatre composantes les Facteurs Individuels et Médicaux* (FIM) et les Facteurs Environnementaux* à connotation sanitaire (FE-SAN), socio-économique (FE-SOCIO.ECO) et physicochimique (FE-PHY.CHIM)

Facteurs individuels et médicaux

La géographie des caractéristiques individuelles* est modélisée à partir de données individus-centrée*. Elles caractérisent des expositions géographiques inhérentes à la localisation résidentielle des individus de la cohorte LEA.

Variables individuelles et médicales			Notations		Valeurs prises
Expositions	Base	Variabilité spatiotemporelle	Texte	Logiciel R	
Pertinentes au regard de l'historique médical disponible	LEA	Genre	$x'^{SEXE}_{(U_k)}$	x._SEXE	{Garçon ; Fille}
		Leucémies traitées	$x'^{TYPLEUC}_{(U_k)}$	x._LEUC	{LAL, LAM}
		Age au diagnostic	$x'^{AGE\_DIAG}_{(U_k)}$	x._AGE_DIAG	⟦0,5 ; 11⟧ années
		Durées de suivi	$x'^{DSUIVI}_{(U_k)}$	x._DSUIVI	⟦1 ; 19⟧ années
		Rechutes	$x'^{RECHUT}_{(U_k)}$	x._RECHUT	{OUI, NON}
		Greffes	$x'^{GREF}_{(U_k)}$	x._GREFF	{OUI, NON}
		Irradiations corporelles totales	$x'^{IRCAT}_{(U_k)}$	x._IRACT	{OUI, NON}
		Protocoles de traitement	$x'^{PROTOC}_{(U_k)}$	x._PROTOC	{11 protocoles de traitement*}
Curieuses	LEA	Activités physiques pratiquées	$x'^{ACPHY}_{(U_k)}$	x._APHY	{ACUNE ; SCOLAIRE ; EXTRA-SCOLAIRE}

Tableau 57 : Liste des i.st.e* modélisant la géographie des FIM

Facteurs environnementaux sanitaires

Les FE-SAN* évaluent l'impact géographique morbide de qualité des tissus sanitaires* territoriaux, par l'accès : aux praticiens libéraux ainsi qu'aux services spécialisés et aux Equipements Matériels Lourds* (EML*) des établissements de santé.

Variables sanitaires			Notations		Valeurs prises
Expositions	**Base**	**Variabilité spatiotemporelle**	**Texte**	**Logiciel R**	
Pertinentes ou curieuses, selon la séquelle considérée	DREES	Temps d'accès aux généralistes	$x'^{GENE}_{(U_k)}$	x._GENE	⟦0 ; 20⟧ min
		Temps d'accès aux ophtalmologues	$x'^{OPHT}_{(U_k)}$	x._OPHT	⟦0 ; 20⟧ min
		Temps d'accès aux ORL	$x'^{ORL}_{(U_k)}$	x._ORL	⟦0 ; 20⟧ min
		Temps d'accès aux pédiatres	$x'^{PEDIA}_{(U_k)}$	x._PEDIA	⟦0 ; 20⟧ min
		Temps d'accès aux radiologues	$x'^{RADIO}_{(U_k)}$	x._RADIO	⟦0 ; 80⟧ min
		Temps d'accès à un service neurologie	$x'^{NEUROs}_{(U_k)}$	x._NEUROs	⟦0 ; 145⟧ min
		Temps d'accès à un service ORL	$x'^{ORLs}_{(U_k)}$	x._ORLs	⟦0 ; 143⟧ min
		Temps d'accès à un service endocrinologie	$x'^{ENDOs}_{(U_k)}$	x._ENDOs	⟦0 ; 144⟧ min
		Temps d'accès à un service hématologie	$x'^{HEMAs}_{(U_k)}$	x._HEMAs	⟦0 ; 143⟧ min
		Temps d'accès à un service ophtalmologie	$x'^{OPTHs}_{(U_k)}$	x._OPTHs	⟦0 ; 99⟧ min
		Temps d'accès à un TEP*	$x'^{TEP}_{(U_k)}$	x._ENDOs	⟦0 ; 564⟧ min
		Temps d'accès à un scanner	$x'^{SCAN}_{(U_k)}$	x._SCAN	⟦0 ; 99⟧ min
		Temps d'accès à une caméra à scintillation	$x'^{CAM}_{(U_k)}$	x._CAM	⟦0 ; 565⟧ min
		Temps d'accès à IRM*	$x'^{IRM}_{(U_k)}$	x._IRM	⟦0 ; 143⟧ min
		Accès Potentiel Localisé aux généralistes	$x^{APL\_GEN}_{(U_k)}$	x_APL_GENE	⟦0 ; 176,35⟧ min
		Accès Potentiel Localisé aux ophtalmologues	$x^{APL\_OPHT}_{(U_k)}$	x_APL_OPHT	⟦0 ; 15.95⟧ min

Tableau 58 Liste des i.st.e* modélisant la géographie des FE-SAN

Facteurs environnementaux socio-économiques

Les FE-SOCIO.ECO* caractérisent l'impact géographique des conjonctures sociales et économiques sur les capacités collectives et les comportements individuels vis-à-vis du recours aux soins. Les contextes désavantageux induisent prédispositions géographiques morbides.

Variables socio-économiques			Notations		Valeurs prises
Expositions	**Base**	**Variabilité spatiotemporelle**	**Texte**	**Logiciel R**	
Pertinentes à des conduites à risques en fonction des niveaux de vie	INSEE	Taux de foyers fiscaux imposables	$x_{(U_k)}^{tx.FoyFisc}$	x_txFoyFisc	⟦0,227 ; 0,765⟧%
		Revenus fiscaux médians par personne ou ménage	$x_{(U_k)}^{RevMed.p/m}$	x_ RevMedp/m	p.⟦6268; 18774⟧€ m.⟦13886; 51339⟧€
		Répartitions des richesses par personne ou ménage	$x_{(U_k)}^{GINI.p/m}$	x_GINIp/m	p.⟦0,28; 0,50⟧SU m.⟦0,30; 0,51⟧ SU
		Taux de chômage	$x_{(U_k)}^{tx.CHOM}$	x_txCHOM	⟦0,034; 0,244⟧ %
		Taux d'emplois ouvriers	$x_{(U_k)}^{tx.OUV}$	x_txOUV	⟦0,06; 0,52⟧%
Pertinentes aux comportements collectifs liés à la qualité des politiques publiques	INSEE	Taux de mortalité	$x_{(U_k)}^{tx.MORT}$	x_txMORT	⟦1,8; 40,6⟧ ‰
		Taux d'accroissement naturel	$x_{(U_k)}^{txAccNAT}$	x_txAccPOP	⟦−0,178; 0,161⟧ %
		Taux accroissement démographique (géométrique)	$x_{(U_k)}^{txAccPOP}$	x_txAccPOP	⟦−0,027; 0,074⟧ %
		Taux d'individus diplômés au moins du baccalauréat	$x_{(U_k)}^{tx.BAC}$	x_txBAC	⟦0,015; 0,125⟧ %
Pertinentes aux substances toxiques liées à la spécialisation socio-économique	INSEE	Taux d'Emplois A Risque	$x_{(U_k)}^{tx.EAR}$	x_txEAR	⟦0,236; 0,818⟧ %
		Taux de Surfaces Agricoles Utilisées	$x_{(U_k)}^{tx.SAU}$	x_tx.SAU	⟦0 ; 0,97⟧ %
		Intensités des activités agricoles	$x_{(U_k)}^{int.AGRI}$	x_int.AGRI	⟦0 ; 79,84⟧ SU
Pertinentes aux contextes conjoncturels morbides	INSEE	Indices de défaveur sociale 1999	$x_{(U_k)}^{FDep99}$	x_FDep99	⟦−3,3 ; 6,42⟧ SU
		Indices de défaveur sociale 2009	$x_{(U_k)}^{FDep09}$	x_FDep09	⟦−6,08 ; 4,72⟧ SU
		Indices de défaveur sociale probable entre 1997-2011	$x_{(U_k)}^{FDepXX}$	x_FDepXX	⟦−0,97 ; 0,84⟧ SU
Curieuses au stresse d'origine sociale généré par l'insécurité	ONDRP	Intensités des atteintes aux biens	$x_{(U_k)}^{att.BIENS}$	x_attBIEN	⟦0,03 ; 587⟧ infractions/ha
		Intensités des atteintes à l'intégrité physique	$x_{(U_k)}^{att.PHY}$	x_attPHY	⟦0,004 ; 97,7⟧ infractions/ha
		Indices composites d'insécurité	$x_{(U_k)}^{INSECU}$	x_INSECU	⟦−0,22 ; 12,31⟧ SU

Tableau 59 : Liste des i.st.e* modélisant la géographie des FE-SOCIO.ECO

Facteurs environnementaux physicochimiques

Les FE-PHY.CHIM* modélisent la géographie des expositions chroniques à de faibles doses de substances nocives, particulières ou combinées, et omniprésentes dans les milieux de vie. Ces expositions ont longtemps été controversées. Elles sont désormais reconnues comme scientifiquement probantes - *a fortiori* sur les populations prédisposées.

Variables physicochimiques			Notations		Valeurs prises
Expositions	**Base**	**Variabilité spatiotemporelle**	**Texte**	**Logiciel R**	
Pertinentes aux paramètres géophysiques	Météo-France	Doses annuelles de rayonnement global cumulé	$x_{(U_k)}^{RAY}$	x_RAY	⟦362760; 574814⟧joule/cm²
		Températures annuelles moyennes	$x_{(U_k)}^{TEMP}$	x_TEMP	⟦10,0; 16,1⟧ Celsius
		Nombres de Jours Annuels Pluvieux :	$x_{(U_k)}^{NJAP}$	x_NJAP	⟦50,8; 135,6⟧ jours
	Géofla	Altitudes communales moyennes :	$x_{(U_k)}^{TOPO}$	x_TOPO	⟦1; 2009⟧mètres

Pertinentes à la radioactivité environnementale d'origine naturelle et artificielle	RNM	Doses efficaces totales de rayonnement γ dans l'air	$x_{(U_k)}^{GAMMA}$	x_GAMMA	⟦64,9; 153,8⟧NanoSv/h
		Activités volumiques totales des particules α dans les eaux douces	$x_{(U_k)}^{ALPHA}$	x_ALPHA	⟦0,01; 0,145⟧Bq/litre
		Activités volumiques totales des particules β dans les eaux douces	$x_{(U_k)}^{BÊTA}$	x_BETA	⟦0,01; 0,106⟧ Bq/litre
		Activités volumiques des isotopes radioactifs du Tritium dans les eaux douces	$x_{(U_k)}^{3H}$	x_3H	⟦0,01; 198,2⟧ Bq/litre
		Activités volumiques du Plutonium 238 dans les sols fins	$x_{(U_k)}^{238Pu}$	x_238Pu	⟦0,02; 0,82⟧ Bq/kg.mat.sec
		Activités volumiques du Césium 137 dans les sols fins	$x_{(U_k)}^{137Cs-sol}$	x_137Cs_SOL	⟦0,37; 30,66⟧ Bq/kg.mat.sec
		Activités volumiques de l'antimoine 125 dans les sols fins	$x_{(U_k)}^{125Sb}$	x_125Sb	⟦0,23; 1,66⟧ Bq/kg.mat.sec
		Activités volumiques du Strontium 90 dans le lait	$x_{(U_k)}^{90Sr}$	x_90Sr	⟦0,00; 0,33⟧ Bq/litre
		Activités volumiques du cesium137 dans le lait	$x_{(U_k)}^{137Cs-biol}$	x_137Cs_BIOL	⟦0,00; 0,36⟧ Bq/litre
		Activités volumiques de l'iode 131 dans le lait	$x_{(U_k)}^{131i}$	x_131i	⟦0,05; 5,62⟧ Bq/litre
Expositions (pertinentes) Géographiques à des Radionucléides Artificiels	ASN	Nombres d'INB* en activité et pondérés par leur proximité spatiale	$x_{(U_k)}^{EGRA}$	x_EGRA	⟦0; 15⟧ U/m
Pertinentes à la radioactivité tellurique dans les habitations	IRSN	Classes IRSN d'exposition au radon	$x_{(U_k)}^{RADON}$	x_RADON	{B, V, J, O, R}
		Activités Volumiques spatiotemporelles du radon	$x''^{RADON}_{(U_k)}$	x.._RADON	⟦7,9; 1276,7⟧ Bq/m3
Pertinentes multi-milieux par inhalation ou ingestion d'Eléments Métalliques Traces*	INERIS	Doses journalières de chrome	$x_{(U_k)}^{DEJ(Cr)}$	x_DJECr	⟦0,05; 0,4⟧ mg/kg
		Indices du risque spatial des expositions au plomb	$x_{(U_k)}^{irs(Pb)}$	x_irsPb	⟦0,05; 0,69⟧ SU
		Indices du risque spatial des expositions au nickel	$x_{(U_k)}^{irs(Ni)}$	x_irsNi	⟦0,06; 0,26⟧ SU
		Indices du risque spatial des expositions au cadmium	$x_{(U_k)}^{irs(Cd)}$	x_irsCd	⟦0,12; 0,87⟧ SU
		Proxy du Risque d'Exposition à des substances Chimiques toxiques	$x_{(U_k)}^{PREC}$	x_PERC	⟦0; 15850⟧ SU
Pertinentes à des substances nocives d'après l'occupation biophysique des sols	CLC	Proportions de zones sur lesquelles des pesticides sont répandus	$x_{(U_k)}^{PEST}$	x_PEST	⟦0; 0,911⟧ %
		Proportions de zones urbanisées ou industrialisées	$x_{(U_k)}^{URIN}$	x_URIN	⟦0; 0,995⟧ %
Curieuses d'après l'occupation biophysique des sols	CLC	Proportions de zones incendiées par des feux de forêt	$x_{(U_k)}^{FEFO}$	x_FEFO	⟦0; 0,866⟧ %
		Proportions de zones supposées préventives	$x_{(U_k)}^{PREV}$	x_PREV	⟦0; 15850⟧ %

Tableau 60 : Liste des i.st.e* modélisant la géographie des FE-PHY.CHIM

TABLES DES FIGURES

TABLES DES SCHEMAS SYNOPTIQUES

TABLES DES TABLEAUX

TABLES DES MATIERES

PARTIE.II : MODELISATIONS GEOGRAPHIQUES ENVIRONNEMENTALES ET IDENTIFICATION DES DETERMINANTS DE SANTE