Evaluation des méthodes statistiques en épidémiologie spatiale : cas des méthodes locales de détection d'agrégats

Aline Guttmann

Université d'Auvergne

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

Année 2014 N° d'ordre :

Thèse

pour l'obtention du grade de

DOCTEUR D'UNIVERSITE

Spécialité

Biostatistique

Présentée et soutenue publiquement par

Aline GUTTMANN

Le 27 novembre 2014

Évaluation des méthodes statistiques en épidémiologie spatiale
Cas des méthodes locales de détection d'agrégats

Sous la direction de

M. Jean-Yves Boire

M. Lemlih Ouchchane

Jury :

M. Jean-Yves Boire	Directeur	PU-PH, Université d'Auvergne
M. Lemlih Ouchchane	Directeur	MCU-PH, Université d'Auvergne
M. Gerbaud Laurent	Examinateur	PU-PH, Université d'Auvergne
M. Demongeot Jacques	Examinateur	PU-PH, Université de Grenoble
M. Salamon Roger	Rapporteur	PU-PH, Université Bordeaux 2
M. Giorgi Roch	Rapporteur	PU-PH, Université d'Aix-Marseille

UMR 6284 UdA – CNRS

Université d'Auvergne

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

Année 2014 N° d'ordre :

Thèse

pour l'obtention du grade de

DOCTEUR D'UNIVERSITE

Spécialité

Biostatistique

Présentée et soutenue publiquement par

Aline GUTTMANN

Le 27 novembre 2014

Évaluation des méthodes statistiques en épidémiologie spatiale
Cas des méthodes locales de détection d'agrégats

Sous la direction de

M. Jean-Yves Boire

M. Lemlih Ouchchane

Jury :

M. Jean-Yves Boire	Directeur	PU-PH, Université d'Auvergne
M. Lemlih Ouchchane	Directeur	MCU-PH, Université d'Auvergne
M. Gerbaud Laurent	Examinateur	PU-PH, Université d'Auvergne
M. Demongeot Jacques	Examinateur	PU-PH, Université de Grenoble
M. Salamon Roger	Rapporteur	PU-PH, Université Bordeaux 2
M. Giorgi Roch	Rapporteur	PU-PH, Université d'Aix-Marseille

UMR 6284 UdA – CNRS

A mon fils, Samuel

La vie n’aurait pas pu m’offrir de plus merveilleux privilège que celui de te voir grandir.
Il n’y a aucune montagne que je ne soulèverais pour toi.

« *Le petit Palémon, grand de huit ans à peine,*
Maintient en vain le bouc qui résiste et l'entraîne,
Et le force à courir à travers le jardin,
Et brusquement recule et s'élance soudain.
Ils luttent corps à corps ; le bouc fougueux s'efforce ;
Mais l'enfant, qui s'arc-boute et renverse le torse,
Étreint le cou rebelle entre ses petits bras,
Se gare de la corne oblique et, pas à pas,
Rouge, serrant les dents, volontaire, indomptable,
Ramène triomphant le bouc noir à l'étable.
Et Lysidé, sa mère aux belles tresses d'or,
Assise au seuil avec un bel enfant qui dort,
Se réjouit à voir sa force et son adresse,
L'appelle et, souriante, essuie avec tendresse
Son front tout en sueur où collent ses cheveux ;
Et l'orgueil maternel illumine ses yeux. »

Le petit Palémon.
Albert Samain (1858-1900)

Remerciements

Je voudrais tout d'abord exprimer mes sincères remerciements au Professeur JEAN-YVES BOIRE pour avoir accepté de diriger cette thèse. A ces remerciements s'ajoute ma plus sincère gratitude pour m'avoir accueillie dans son service depuis mes premiers pas en tant qu'interne et m'avoir, par la suite, infailliblement soutenue, encouragée et parfois poussée lorsque c'était nécessaire. J'espère toujours me montrer digne de la confiance qu'il m'a accordée.

Je tiens à remercier chaleureusement le Docteur LEMLIH OUCHCHANE pour avoir codirigé cette thèse. Il a su me faire partager ses connaissances scientifiques et son implacable sens du détail, ce qui a grandement amélioré mes connaissances et fortement participé à la qualité de cette thèse. Je ne peux évidemment pas le remercier sans mentionner sa patience et sa persévérance face à mes multiples « étourderies orthographiques ».

Je voudrais remercier les Professeurs ROGER SALAMON et ROCH GIORGI d'avoir accepté d'être les rapporteurs de cette thèse. Je suis flattée de l'intérêt qu'ils ont porté à ce travail.

Je voudrais également exprimer mes remerciements au Professeur JACQUES DEMONGEOT d'avoir accepté de faire partie du jury. Ses remarques, aussi pertinentes que précises, ont permis d'améliorer nettement la qualité des publications issues de ce travail. J'associe à ces remerciements le Docteur JEAN GAUDART, ce travail concrétise la collaboration entre Marseille et Clermont-Ferrand que nous espérons tous durable. Je tiens également à remercier le Professeur LAURENT GERBAUD pour avoir accepté de faire partie du jury ainsi que pour son soutien et ses conseils depuis le début de mon internat en 2006.

Je remercie YAN GÉRARD et FABIEN FESCHET pour leurs précieuses contributions à nos travaux.

Je tiens à réserver une place toute particulière dans ces remerciements pour SYLVIE ROUX sans qui ces presque cinq années passées dans le service auraient été bien différentes. Les méandres administratifs de notre bonne institution n'ont pas de secrets pour elle et c'est une aide qui n'a pas de prix. Toujours disponible, elle a su rendre encore meilleurs les bons moments et adoucir les mauvais.

Je tiens à remercier tous mes maîtres de stage et mes enseignants. Ils ont fait de moi la professionnelle que je suis et continueront d'influencer grandement celle que je deviendrai. J'espère leur faire honneur.

Je souhaite remercier tous mes amis et collègues pour leur présence et leur soutien : merci aux Claire, Gabriela, Mathias, Amélie, Anne, Laurent, Émilie, Xinran, Sophie et bien d'autres encore…

Je remercie toute ma famille et notamment mes parents qui m'ont toujours soutenue dans mes études et dans la vie en générale. Leur soutien pendant les derniers moments de cette thèse s'est avéré irremplaçable. Je vous dois énormément. Du fond du cœur : merci.

Enfin, je remercie JULIEN et FLORENT. Même si Samuel est un peu petit encore pour l'exprimer lui-même, soyez sûrs qu'il se joindrait à moi pour vous dire que vous êtes les meilleurs papas du monde.

Résumé

L'évaluation des performances des méthodes de détection d'agrégats de maladie est fondamentale dans le domaine de l'épidémiologie spatiale et, paradoxalement, on déplore une absence de consensus quant à sa conduite. Cette problématique est d'autant plus importante que les nouvelles technologies de partage d'informations promettent une évolution importante des signaux disponibles pour l'épidémiologie et la veille sanitaire.

Les spécialistes du domaine ont adopté un mode d'évaluation fondé sur l'utilisation concomitante de plusieurs indicateurs de performances complémentaires tels que des indicateurs dérivés de l'évaluation des méthodes diagnostiques ou encore diverses définitions de puissance conditionnelle. Cependant, ces évaluations issues de schémas de simulation classiques reposent sur le choix de quelques hypothèses alternatives particulières et ne permettent qu'une interprétation limitée à ces hypothèses. De plus, la démultiplication des indicateurs évaluant la performance, différents selon les protocoles, gêne la comparaison des études entres elles et complique l'interprétation des résultats.

Notre travail propose et évalue plusieurs indicateurs de performance prenant en compte à la fois puissance et précision de localisation. Leur intérêt dans l'évaluation spatiale systématique des méthodes est illustré par la création de cartes de performance. En complément de l'évaluation des performances lorsqu'une détection est attendue, nous proposons également une méthode d'évaluation de la répartition spatiale de l'erreur de type I complétée par la construction d'une nouvelle inférence statistique testant l'éventualité d'un effet de bord.

Mots clés : épidémiologie spatiale, méthodes locales de détection d'agrégats, performance

Summary

Evaluation of statistical methods in spatial epidemiology: the case of cluster detection tests

Although performance assessment of cluster detection tests is a critical issue in spatial epidemiology, there is a lack of consensus regarding how it should be carried out. Nowadays, with the spread of new technologies in network systems, data sources for epidemiology are undergoing radical changes that will increase the need for performance evaluation.

Field specialists are currently evaluating cluster detection tests with multiple complementary performance indicators such as conditional powers or indicators derived from the field of diagnostic tools evaluation. These evaluations are performed following classical protocols for power assessment and are often limited to a few number of simulated alternative hypotheses, thus restricting results interpretation and scope. Furthermore, with the use of multiple varying indicators, comparisons between studies is difficult at best.

This work proposes and compares different global performance indicators that take into account both usual power and location accuracy. Their benefit for cluster detection tests evaluation is illustrated with a systematic spatial assessment enabling performance mapping. In addition to the evaluation of performance when clusters exist, we also propose a method for the spatial evaluation of type I error, together with a new statistical test for edge effect.

Key words: spatial epidemiology, cluster detection tests, performance evaluation

Table des matières

Table des Figures

Table des Tableaux

Glossaire

AE	agrégat éligible
AUC_{EP}	Aire sous la courbe de puissance étendue
FN	Faux Négatif
FP	Faux Positif
PPH	Processus de poisson homogène
PPI	Processus de poisson inhomogène
PRNG	« Pseudo Random Number Generator » générateur de nombres pseudo-aléatoires
RV^+	Ratio de Vraisemblance positif
Se	Sensibilité
SIG	Système d'Information Géographique
Sp	Spécificité
US	Unité spatiale
v.a.	Variable aléatoire
VN	Vrai Négatif
VP	Vrai Positif
VPP	Valeur Prédictive Positive

Introduction générale

Ce travail s'appuie sur un contexte local de création ou de développement d'outils de surveillance et de veille sanitaire. Initialement, la demande provenait du Centre d'Etude des Malformations Congénitales en Auvergne (CEMC Auvergne) qui désirait se doter d'un outil de veille par la mise en place d'une procédure d'analyse systématique de ces données. Par la suite, le projet GINSENG (projet ANR - 2010) a ouvert la perspective d'obtenir des données de surveillance d'une qualité supérieure à celle que l'on peut voir habituellement (absence de délai entre la création de l'information et sa mise à disposition, rapprochement automatisé de données multi-sources, dédoublonnement et identito-vigilance automatisés). Un système de veille fondé sur ce type de signal offrirait des perspectives de performances intéressantes.

Cependant, vis-à-vis de la veille sanitaire, l'authentification de l'apport d'un système du type de celui proposé dans le projet GINSENG doit s'appuyer sur l'évaluation d'un gain de performance. Une des principales composantes de cette évaluation concerne les performances des méthodes statistiques de détection d'agrégats intégrées à ce système. D'un point de vue méthodologique, les outils disponibles pour l'évaluation des performances des méthodes de détection d'agrégats spatiaux ou spatiaux-temporels sont encore mal adaptés. En effet, s'il est admis que l'évaluation des performances de ces méthodes dépasse la seule estimation du risque β (sous une hypothèse alternative proche de la réalité de terrain) ou du risque α (vérifiant ainsi sa conformité à la valeur pré-définie de 5%), les solutions adoptées en pratique courante restent insuffisantes.

Les méthodologies d'évaluations sont très hétérogènes rendant les études de simulation difficilement comparables entre elles. Les indicateurs de performance utilisés sont fondés soit sur l'utilisation d'indicateurs dérivés des méthodes d'évaluation des outils diagnostiques (sensibilité/spécificité et valeurs prédictives pour les méthodes locales - courbes ROC pour les méthodes de cartographie du risque), soit sur des définitions plus ou moins restrictives de performance de localisation de l'agrégat simulé permettant de calculer des puissances conditionnelles. L'utilisation d'indicateurs n'offrant qu'une évaluation partielle de la performance oblige à démultiplier leur nombre lorsqu'une évaluation et une interprétation globales sont souhaitées. Cette masse d'informations devient vite difficile à résumer et à interpréter lorsque le nombre de simulations est important. Ainsi, la plupart de ces évaluations sont issues de schémas de simulation classiques et reposent sur le choix d'une ou de quelques hypothèses alternatives particulières. Elles n'offrent donc qu'une évaluation de la performance limitée à ces hypothèses.

Jusqu'à présent, une seule référence a proposé un indicateur global, la puissance étendue [1], qui se construit sous la forme d'une courbe et prend en compte à la fois les dimensions de rejet de l'hypothèse nulle et de « précision » de localisation de l'agrégat réel.

Objectif de la thèse

L'objectif de ce travail de thèse est de développer des outils d'évaluation globale des performances des méthodes locales de détection d'agrégats. Ces outils se doivent d'évaluer à la fois la puissance usuelle et la performance de la localisation, sans imposer de définition restrictive à cette dernière. Ils doivent de plus être aisément interprétables et simples à calculer.

Concernant la puissance, il s'agit dans un premier temps, comme suggéré par Tango et Takahashi [1], de développer un indicateur unique à partir de l'indicateur qu'ils proposent : la puissance étendue. Associé à un design de simulation spatiale systématique d'hypothèses alternatives, cet indicateur global permet d'aboutir à une cartographie de la performance.

Concernant l'erreur de type I, il s'agit du développement d'un indicateur spatial de contribution à l'erreur de type I permettant une cartographie de cette contribution à l'erreur de type I.

Déroulement du mémoire

Ce mémoire est constitué de quatre parties :

La première partie effectue une synthèse de la méthodologie statistique en épidémiologie spatiale en ne détaillant que les méthodes locales de détection d'agrégat qui sont l'objet de ce travail.

La deuxième partie présente les fondements théoriques de l'évaluation de ces méthodes et effectue une synthèse bibliographique des indicateurs de performance utilisés dans ce domaine.

La troisième partie est consacrée à la présentation des résultats de ce travail sous la forme d'articles, précédés d'une introduction/synthèse générale.

Enfin, ce mémoire se termine par une conclusion générale incluant les perspectives du travail de thèse.

Partie 1

Méthodes statistiques en épidémiologie spatiale

1 Répartition spatiale des cas

1.1Caractérisation des profils de répartition

Pour comprendre les hypothèses sous-jacentes aux méthodes statistiques d'analyse de données spatialisées, il est nécessaire de connaître et comprendre quelques définitions permettant de classifier les différents profils de répartition spatiale des cas [2].

Pour illustrer ces définitions, considérons une zone hypothétique de 100 km^2 dans laquelle sont répartis 100 cas d'une maladie ou d'un phénomène quelconque. La Figure 1 montre une répartition spatiale de cas dite « complètement aléatoire ». Ce profil de répartition implique que la position de chacun des cas est complètement aléatoire, et que par conséquent, la position de n'importe quel cas est indépendante de la position de tous les autres cas. Ce profil est très souvent utilisé comme hypothèse nulle représentant l'absence d'agrégation spatiale dans les tests d'inférence.

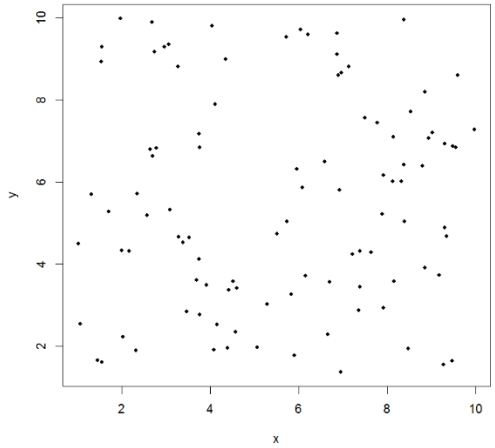

Figure 1: répartition spatiale des cas dite « complètement aléatoire »

Un profil de répartition particulier est celui où la distance entre tous les cas est identique. Ce profil, représenté dans la Figure 2, est appelé complètement régulier.

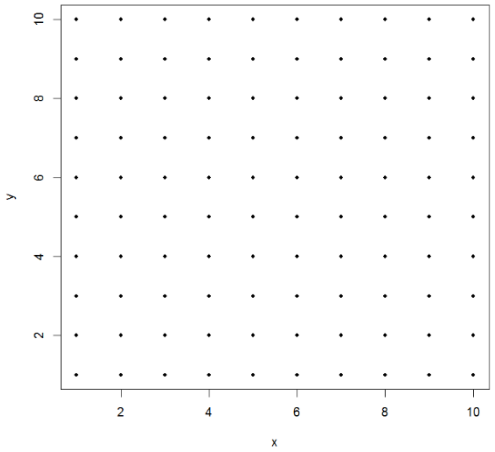

Figure 2: répartition spatiale des cas dite « complètement régulière »

La Figure 3 représente un cas proche du précédent profil : il s'agit de la répartition spatiale régulière. Cette répartition implique que les cas sont situés « à peu-près » à la même distance les uns des autres. Ce type de répartition est observé notamment lorsqu'il existe un comportement territorial des individus pour le phénomène considéré (*e.g.*, sièges occupés dans une salle d'attente). Dans ce cas, les individus cherchent typiquement à rester les plus éloignés les uns des autres. (Pour reprendre l'exemple de la salle d'attente, lorsque le nombre de places est égal à deux fois le nombre d'individus, le profil de répartition peut même être complétement régulier, tous les individus étant séparés par un siège.).

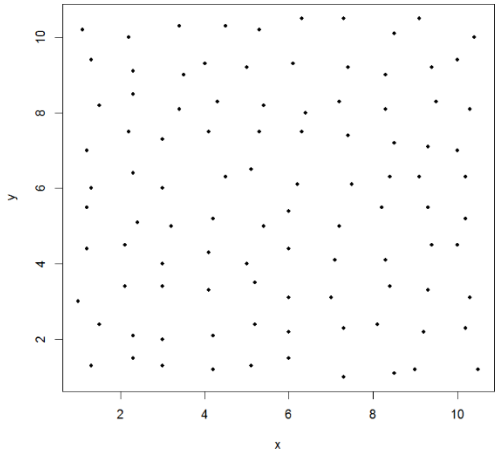

Figure 3: répartition spatiale des cas dite « régulière »

Enfin, le dernier profil représenté dans la Figure 4 est la répartition agrégée des cas. Dans cette figure, les cas semblent effectivement concentrés préférentiellement dans deux zones appelées agrégats.

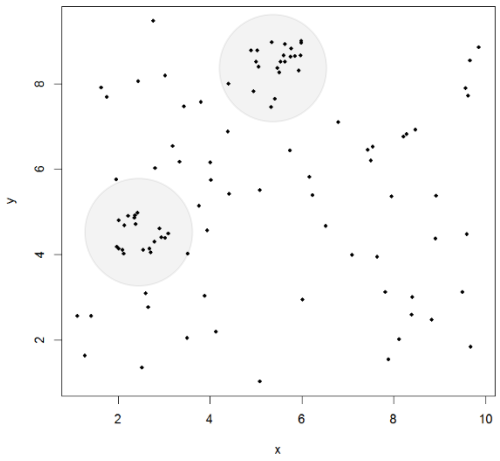

Figure 4: répartition spatiale agrégée

1.2Processus ponctuels spatiaux

1.2.1 Répartition régulière des cas

La répartition régulière implique que les cas se situent approximativement à la même distance les uns des autres, ou autrement dit, le nombre moyen de cas par unité de surface est à peu près constant. Si l'on considère le nombre de cas par unité de surface comme une variable aléatoire (v.a.), alors son comportement peut être modélisé par une loi binomiale. En effet, la variance d'une telle distribution est inférieure à son espérance.

1.2.2 Répartition agrégée des cas

Au contraire de la répartition régulière, la répartition agrégée est marquée par une tendance des cas à être regroupés. Il y a donc des zones où les cas sont nombreux et d'autres où ils sont relativement rares. Si on considère la même v.a. « nombre de cas par unité de surface », alors sa loi de distribution sera caractérisée par une variance plus importante que son espérance. On pourra modéliser ce profil de répartition par une loi binomiale négative.

1.2.3 Répartition aléatoire des cas

Comme indiqué plus haut, ce profil de répartition est très souvent utilisé comme hypothèse nulle représentant l'absence d'agrégation dans les tests d'inférence. Il est donc important de détailler le modèle probabiliste qui permet de le formaliser.

Dans le cas d'une répartition aléatoire, la variance (toujours de la v.a. « nombre de cas par unité de surface ») est plus importante que pour une répartition régulière mais moins importante que pour répartition agrégée.

En effet, dans une répartition aléatoire, on peut observer des zones où les individus sont regroupés et d'autres zones plus vides. Cependant, les zones d'agrégation seront moins importantes (en nombre de cas) que pour une répartition agrégée et les zones « vides » seront moins vides que pour une distribution régulière. La loi de Poisson permet de modéliser cette répartition où la variance est égale à l'espérance.

1.2.3.1 Processus de Poisson homogène

Dans un processus de Poisson homogène (PPH), la présence d'un cas à un endroit donné n'affecte par la probabilité d'apparition d'autres cas à proximité et il n'y a pas de zone où les cas sont plus susceptibles d'apparaître.

De façon plus formelle, Diggle [3] écrit que l'hypothèse de répartition spatiale complètement aléatoire implique que :

(1) le nombre n d'évènements dans une région A (espace polonais) d'aire $|A|$ suit une distribution de Poisson de moyenne $\lambda|A|$ où la constante λ, strictement positive, est l'intensité ou nombre moyen d'évènements par unité de surface ;

(2) et qu'étant donné n évènements $\{x_i, i = 1, .., n\}$ observés dans la région A, les x_i sont un échantillon aléatoire indépendant de distribution uniforme dans A.

Selon (1), une répartition complètement aléatoire implique donc bien que l'intensité des évènements est constante dans A. Selon (2), cette répartition implique également qu'il n'y a pas de lien entre les évènements.

De plus, le PPH est

(3) stationnaire, c'est-à-dire invariant par translation dans l'espace ;

(4) et isotropique, c'est-à-dire que le processus est invariant par rotation autour de son origine.

Stationnarité et isotropisme impliquent que toute relation entre deux évènements dépend seulement de la distance qui les sépare. De plus selon (1), l'intensité constante du processus implique que λ, *i.e.* le nombre moyen d'évènements par unité de surface, peut être estimé par

$$\hat{\lambda} = \frac{n}{|A|}$$

Où n est le nombre d'évènements dans *A*. L'intensité du processus est appelé mesure de premier ordre car elle décrit la moyenne de ce processus. La mesure de second ordre est la variance qui reflète la tendance des évènements à être agrégés, ou au contraire, régulièrement répartis. Dans un PPH, cette mesure est constante.

1.2.3.2 Processus de Poisson inhomogène

Cependant, modéliser l'hypothèse nulle par un PPH n'est pas réaliste, sauf à supposer que la répartition de la population à risque soit homogène, ce qui est rarement le cas. Pour pallier ce problème, le processus de poisson inhomogène (PPI), généralisation du PPH, permet de modéliser une intensité variable dans l'espace.

Le PPI d'intensité variable $\lambda(z)$ (z est une localisation), est défini tel que :

(1) le nombre n d'évènements survenant dans une zone $a \subset A$ est une v.a. suivant une distribution de Poisson de moyenne $\int_a \lambda(z) d_z$;

(2) étant donné un nombre n d'évènements $\{x_i, i = 1, .., n\}$ observés dans la région A, les x_i sont indépendants, distribués dans A selon une densité proportionnelle à $\lambda(z)$.

1.3 Système d'information géographique

1.3.1 Définitions

Un Système d'Information Géographique (SIG) est un système d'information capable d'organiser, de présenter et d'analyser des données alphanumériques spatialement référencées [4]. Un SIG comprend les outils logiciels, les données, le matériel et les compétences nécessaires à l'utilisation de tous ces éléments. Les données d'un SIG se répartissent en 3 catégories :

- Les objets géographiques qui sont définis par des données géométriques et des données graphiques.
- Les données attributaires qui permettent d'associer des informations à un objet géographique particulier (e.g., le nombre de cas est une donnée attributaire associée à un objet géographique comme une commune, un département, *etc.*)
- Les métadonnées qui fournissent les informations relatives à l'ensemble des données (système de projection et étendue géographique, date et méthodes d'acquisition, propriétaire des données, *etc.*).
- Il existe deux modes de représentation de la réalité en objets géographiques :
    - Le mode matriciel (ou raster) où des matrices sont utilisées pour réaliser une partition régulière de l'espace. Chaque maille de la grille, cellule ou pixel, est associée à une intensité de gris ou de couleur. La juxtaposition des points créés permet d'obtenir une représentation de l'espace (comme dans une image scannée). Par exemple, un lac sera représenté par un ensemble de points d'intensité identique.
    - Le mode vectoriel où les objets sont représentés à partir de points, de lignes ou de polygones. Par exemple un lac sera représenté par un polygone, une rivière par une ligne, *etc.*

Les objets géographiques sont localisés dans l'espace terrestre à l'aide d'un système de coordonnées sphérique ou projectif.

Un système de coordonnées géographique sphérique (Figure 5) représente un point par une distance (altitude) et deux angles (latitude et longitude) exprimés en degrés et minutes.

Un système de coordonnées projectif (Figure 6) permet de déterminer la position d'un point dans l'espace à partir d'un repère. Le système de coordonnées cartésiennes est un système de coordonnées projectif utilisant un repère cartésien (donnée conjointe d'un point d'origine et de trois vecteurs non coplanaires). Dans un système de coordonnées cartésiennes, un point est donc repéré par l'association d'une abscisse, d'une ordonnée et d'une cote.

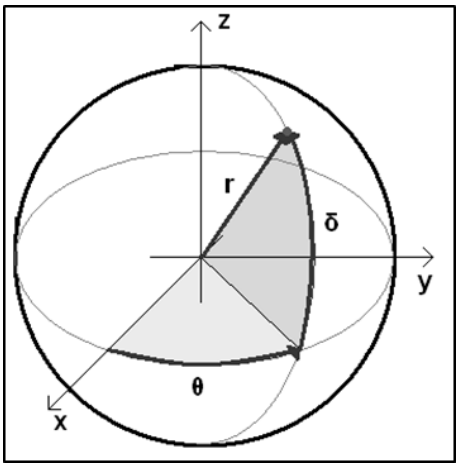

Figure 5: coordonnées sphériques ; rayon (r), longitude (θ), latitude (δ)

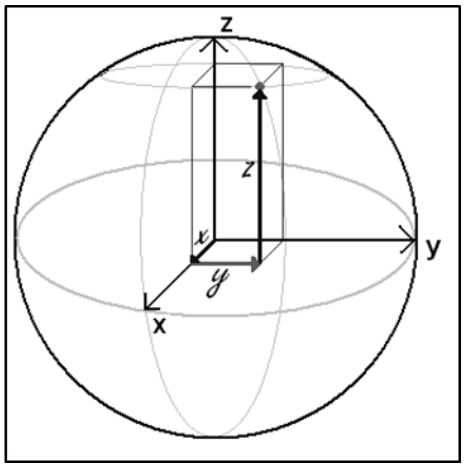

Figure 6: coordonnées cartésiennes ; abscisse (x), ordonnée (y), cote (z)

Dans le domaine de l'épidémiologie, il est rare que l'altitude soit un paramètre pertinent. Dans la grande majorité des cas, les objets géographiques sont représentés seulement par un couple de coordonnées (latitude et longitude).

1.3.2 SIG et données

1.3.2.1 Données ponctuelles

Un système d'information géographique fondé sur la géolocalisation des individus, qu'ils soient des cas ou appartiennent à la population à risque, permettrait de disposer de toute l'information spatiale d'intérêt pour le phénomène étudié. Les données disponibles seraient de type ponctuel où chaque observation (individu) correspondrait à un point au sein de la région d'étude.

Ce type de données spatiales est en réalité rarement disponible en dehors de grandes cohortes populationnelles [5–8].

1.3.2.2 Données groupées

Pour des raisons de confidentialité, la plupart des SIG utilisés en épidémiologie et veille sanitaire se fondent sur des SIG administratifs préexistants. Ils sont issus de représentations vectorielles des régions d'études, où le découpage territorial (communes, zones postales,…) forme un ensemble de polygones. Ces polygones déterminent des surfaces qui constituent des unités spatiales (US). Les cas ou évènements sont géolocalisés par le centroïde (*i.e.* centre de masse ou barycentre) de leur

US d'affectation. Les données qui en sont issues ne sont plus à proprement parler des données ponctuelles mais des données agrégées[1] par US. De très nombreuses études utilisent ce type de données dont on ne donnera ici que quelques références [9–12] à titre d'exemple.

1.3.2.3 Cas-contrôles

Comme dit plus haut, la localisation des individus est en général difficile à obtenir. Cependant, notamment lors d'étude ad'hoc disposant de financements et de moyens logistiques conséquents, ce type d'information peut être recueilli. Il paraît difficilement envisageable de recueillir la même information pour l'ensemble de la population à risque, tout comme il paraît inopportun de se passer entièrement d'information sur cette population. Une solution intermédiaire est de ne recueillir la localisation que d'un échantillon d'individus de cette population. Ces individus jouent le rôle de contrôles et permettent l'exploitation de toute la finesse de l'information spatiale. Les études utilisant ce type de données [13–15] s'appuient le plus souvent sur des registres de pathologies existants.

2 Méthodes d'analyse

Pour l'ensemble de ce chapitre, les notations adoptées pour la présentation des différentes méthodes sont :

- G : la région d'étude
- N : le nombre d'US au sein de la région d'étude
- P : le nombre total d'individus (population à risque)
- P_i : le nombre d'individus dans l'US i pour $\forall i$, $i = 1, ..., N$
- C : la v.a. représentant le nombre total de cas dans la région d'étude
- C_i : la v.a représentant le nombre de cas dans l'US i pour $\forall i$, $i = 1, ..., N$
- c : le nombre total de cas observé dans la région d'étude
- c_i : le nombre de cas observé dans l'US i pour $\forall i$, $i = 1, ..., N$
- d_{ij} : la distance euclidienne entre l'US i et l'US j, pour $\forall i$, $i = 1, ..., N$ et pour $\forall j$, $j = 1, ..., N$

[1] Nous les appellerons par la suite « données groupées » pour éviter tout risque de confusion avec les notions d'agrégation spatiale ou d'agrégat.

2.1 Classification des méthodes

Un très grand nombre de tests d'agrégation spatiale existent, répondant à différents objectifs. Les travaux successifs de Besag et Newell [16] puis Kulldorff [17] ont permis d'établir une classification des méthodes statistiques d'analyse de l'agrégation spatiale qui est ici résumée dans la Figure 7. Une autre grande famille de méthodes, fréquemment utilisée en épidémiologie spatiale, rassemble les méthodes de cartographie du risque [18, 19]. Ces méthodes ne seront pas abordées dans ce travail car elles sont purement descriptives et ne proposent pas d'inférence à proprement parler.

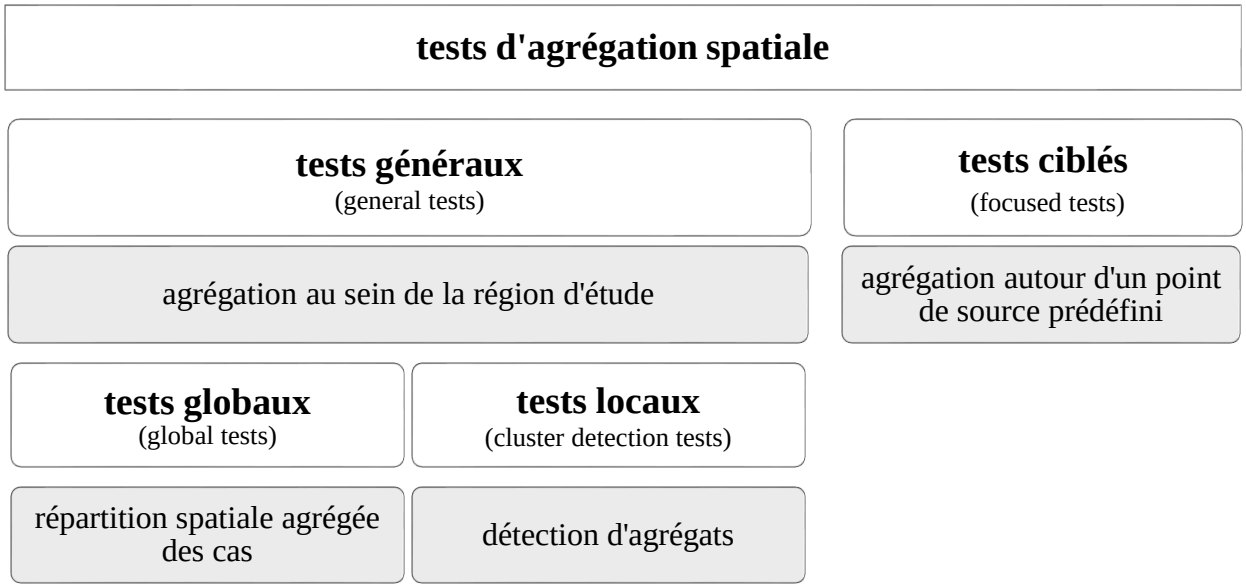

Figure 7: classification des tests d'agrégation spatiale

Les tests locaux qui font l'objet de ce travail se caractérisent par le fait qu'ils apportent une double conclusion : (1) ils individualisent (localisent) la ou les zones de la région d'étude qui constituent le (les) agrégats le (les) plus vraisemblable(s), puis (2) ils donnent une portée statistique en évaluant leur significativité par une inférence.

Nous allons ici résumer les grands principes et définitions des deux grandes familles de tests locaux : les méthodes de balayages et les statistiques d'autocorrélation spatiale.

2.2 Statistiques de balayage

Les méthodes de balayage, comme leur nom l'indique, ont pour principe de balayer la région d'étude à l'aide d'une fenêtre dont les positions successives servent à déterminer autant d'agrégats potentiels qui seront ensuite testés localement pour déterminer leur significativité. Les méthodes diffèrent en général soit par l'algorithme permettant de construire cette fenêtre et déterminant ces positions successives, soit par la façon dont est mené le test d'inférence.

Historiquement, les méthodes de balayage dérivent de la « Geographical Analysis Machine » proposée par Openshaw (1987) [20]. Par la suite, en 1995, Kulldorff [21] proposera la statistique de scan spatial qui est aujourd'hui la méthode de balayage sans doute la mieux connue et la plus utilisée au monde.

2.2.1 Scan spatial de Kulldorff

Le principe développé par Kulldorff [21, 22] consiste à balayer la région d'étude avec une fenêtre circulaire de taille variable. De cette façon, aucune hypothèse préalable sur la localisation ou la taille de l'agrégat n'est posée.

Pour des données ponctuelles, chaque fenêtre circulaire possède un centre dont la position passe successivement sur tous les points de la carte et un rayon dont la valeur varie de zéro jusqu'à une longueur maximale prédéterminée.

Lorsque les données sont groupées, chaque fenêtre possède un centre situé sur le centroïde d'une US et un rayon dont la valeur varie de zéro à une longueur maximale prédéterminée. Les US incluses dans la fenêtre sont celles dont le centroïde se trouve à l'intérieur du cercle.

La longueur maximale du rayon du cercle est le plus souvent déterminée par l'inclusion d'une fraction maximale de la population à risque mais peut également concerner une fraction du nombre de cas attendus, ou encore une taille maximale d'agrégat (en unité de surface). Cette limite est fondée sur le raisonnement qu'au-delà d'une certaine « taille » (que l'on interprétera selon la nature de la limite choisie), un agrégat représenterait plutôt la répartition spatiale « de base » de la région et l'extérieur de cet agrégat représenterait une zone de risque anormalement faible.

La statistique est fondée sur le rapport de vraisemblances observé pour chaque cercle, *i.e.* le rapport de la vraisemblance sous l'hypothèse alternative (le risque à l'intérieur du cercle est supérieur à celui à l'extérieur) sur la vraisemblance sous l'hypothèse nulle d'homogénéité des risques. Le cercle ayant le rapport de vraisemblances le plus élevé définit l'agrégat le plus probable. C'est sur ce ratio de vraisemblance que portera ensuite l'inférence.

La distribution de Bernouilli s'applique pour des données ponctuelles (l'individu est malade ou non malade) alors que celle de Poisson est utilisée lorsque les données sont groupées (un nombre de malades par US).

On détaillera ici uniquement la construction de la statistique pour le modèle de Poisson.

On définit une zone Z représentant la fenêtre mobile d'étude et une collection Ω des m zones Z possibles (selon la définition données plus haut de la fenêtre mobile) telle que $\Omega = \{Z_i \subset G \mid i = 1, \ldots, m\}$.

On note P_{z_i} le nombre d'individus dans la zone Z_i et c_{z_i} le nombre de cas parmi les P_{z_i}.

Soit A une zone incluse dans G, P_A le nombre d'individus de A et C_A la v.a. du « nombre de cas dans A ». Pour $\forall\, A \subset G$, C_A est généré par un PPI tel que

$$C_A \sim Poisson\left(\pi P_{(A \cap Z_i)} + \delta P_{(A \cap \overline{Z_i})}\right)$$

où π est la probabilité d'être un cas dans Z_i, $P_{(A \cap Z_i)}$ le nombre d'individus à l'intersection des zones A et Z_i, δ la probabilité d'être un cas à l'extérieur de Z_i et $P_{(A \cap \overline{Z_i})}$ le nombre d'individus dans la zone A et en dehors de la zone Z_i.

Sous l'hypothèse nulle (H_0) d'homogénéité des risques : $\pi = \delta$ et $C_A \sim Poisson\ (\pi P_A)\ \forall A$.

Sous l'hypothèse alternative (H_1) : $\pi > \delta$.

Etape 1 : Calculer la fonction de vraisemblance pour toutes les zones Zi.

Dans la région G connaissant la fenêtre Z_i, la probabilité d'observer c cas est telle que :

$$Prob(c) = \frac{e^{-\left(\pi P_{Z_i} + \delta\left(P - P_{Z_i}\right)\right)} \times \left(\pi P_{Z_i} + \delta\left(P - P_{Z_i}\right)\right)^c}{c!}$$

Chaque cas x parmi les c cas, possède une probabilité d'être dans une zone donnée de G qui dépend de la population à risque de cette zone.

On définit une fonction *f(x)* de densité de probabilité des cas dans *G* sachant Z_i telle que :

$$f(x) = \frac{\pi P_x}{\pi P_{Z_i} + \delta\left(\mathrm{P} - P_{z_i}\right)} I(x \in Z_i) + \frac{\delta P_x}{\pi P_{Z_i} + \delta\left(P - P_{z_i}\right)} I(x \in \bar{Z}_i)$$

$I(x \in Z_i)$ et $I(x \in \bar{Z}_i)$ étant des indicatrices dont la valeur est égale à 1 si $x \in Z_i$ (respectivement $x \in \bar{Z}_i$) et 0 si non.

La fonction de vraisemblance pour la fenêtre Z_i est égale à :

$$L(Z_i, \pi, \delta) = Prob(c) \times \prod_{j=1}^{\mathrm{c}} f\left(x_j\right)$$

En développant

$$L(Z_i, \pi, \delta) = \frac{e^{-\left(\pi P_{Z_i} + \delta\left(\mathrm{P} - P_{Z_i}\right)\right)} \times \left(\pi P_{Z_i} + \delta\left(\mathrm{P} - P_{Z_i}\right)\right)^{\mathrm{c}}}{\mathrm{c}!} \times \prod_{k \in c_{z_i}} \frac{\pi P_{x_k}}{\pi P_{Z_i} + \delta\left(\mathrm{P} - P_{z_i}\right)}$$

$$\times \prod_{l \in \left(\mathrm{c} - c_{z_i}\right)} \frac{\delta P_{x_l}}{\pi P_{Z_i} + \delta\left(\mathrm{P} - P_{z_i}\right)}$$

On pose $\lambda = \pi P_{Z_i} + \delta\left(\mathrm{P} - P_{Z_i}\right)$

$$L(Z_i, \pi, \delta) = \frac{e^{-\lambda} \times \lambda^c}{\mathrm{c}!} \times \prod_{k \in c_{z_i}} \frac{\pi P_{x_k}}{\lambda} \times \prod_{l \in \left(\mathrm{c} - c_{z_i}\right)} \frac{\delta P_{x_l}}{\lambda}$$

$$L(Z_i, \pi, \delta) = \frac{e^{-\lambda}}{c!} \times \pi^{c_{Z_i}} \delta^{\left(\mathrm{c} - c_{Z_i}\right)} \times \prod_{j=1}^{\mathrm{c}} P_{x_j}$$

Etape 2 : Trouver la zone Z_i qui maximise la fonction de vraisemblance.

Pour cela on pose :

Lorsque $\frac{c_{z_i}}{P_{z_i}} > \frac{\mathrm{c} - c_{z_i}}{\mathrm{P} - P_{z_i}}$

$$L(Z) \stackrel{\text{def}}{=} \sup_{\pi > \delta} L\left(Z_i, \pi, \delta\right) = \frac{e^{-c}}{\mathrm{c}!} \times \left(\frac{c_{Z_i}}{P_{Z_i}}\right)^{c_{Z_i}} \left(\frac{\mathrm{c} - c_{Z_i}}{\mathrm{P} - P_{Z_i}}\right)^{\left(c - c_{Z_i}\right)} \times \prod_{j=1}^{\mathrm{c}} P_{x_j}$$

Si non

$$L(Z) = \frac{e^{-c}}{\mathrm{c}!} \times \left(\frac{\mathrm{c}}{\mathrm{P}}\right)^{\mathrm{c}} \times \prod_{j=1}^{\mathrm{c}} P_{x_j}$$

L'estimateur $\widehat{Z}$ du maximum de vraisemblance est tel que :

$$\hat{Z} = \{Z_s \in \Omega: \mathrm{L}(\mathrm{Z_s}) \geq \mathrm{L}(\mathrm{Z_i}) \; \forall \; \mathrm{Z_i} \in \; \Omega\} \qquad (2\text{-}1)$$

Le ratio de vraisemblance s'écrit :

$$LR = \frac{L(\hat{Z})}{L_0}$$

Avec

$$L_0 \stackrel{\text{def}}{=} \sup_{p=q} L\,(Z_i, \pi, \delta) = \frac{e^{-c}}{\mathrm{c}!} \times \left(\frac{\mathrm{c}}{\mathrm{P}}\right)^{\mathrm{c}} \times \prod_{j=1}^{\mathrm{c}} P_{x_j}$$

Etape 3 : Inférence

La significativité est testée par inférence de Monte Carlo, la distribution de la statistique n'étant pas connue.

En rejetant l'hypothèse nulle, on accepte l'hypothèse alternative qu'il existe une zone à l'intérieur de laquelle il est plus probable d'être un cas qu'à l'extérieur. Cette zone est la fenêtre Z_s (définie dans (2-1)) qui maximise la fonction de vraisemblance et constitue donc le cluster le plus vraisemblable.

2.3 Autocorrélation spatiale

Les tests globaux d'autocorrélation spatiale sont construits à partir de la relation entre les valeurs locales observées à un endroit particulier et les valeurs de voisinage. Ces tests supposent qu'une relation de voisinage ait donc été définie entre les US. Le point faible de ces tests est qu'ils ne testent que l'indépendance entre les distributions marginales des valeurs observées au sein des US. Si une telle indépendance implique l'absence d'autocorrélation spatiale, l'inverse n'est pas forcément vrai.

A partir de ces statistiques similaires à des coefficients de corrélation, des versions locales ont été dérivées (voir [23, 24] pour plus de détails sur le comportement de ces statistiques), grâce aux méthodes standard d'identification des observations ayant une forte influence sur la pente.

2.3.1 Matrice de proximité

La matrice de proximité est l'outil de la prise en compte de relations de voisinage. Autrement dit, elle permet de formaliser la relation spatiale entre les US. Il existe de nombreuses façons de considérer cette relation en prenant en compte les relations de voisinage (dichotomique : l'US *i* est ou n'est pas voisine/adjacente à l'US *j*) ou encore la distance entre deux US.

La matrice de proximité permet d'attribuer un poids (w_{ij}) à chaque couple d'US (*i*, *j*). Les éléments de cette matrice sont obtenus grâce à une fonction de pondération dépendant de la structure spatiale de la région d'étude (voisinage des US, distances) et qui peut être ajustée sur tout paramètre pertinent comme, par exemple, la superficie des US [25]. La fonction de pondération la plus fréquemment rencontrée est une fonction exponentielle décroissante utilisant la distance entre les US combinée à un paramètre d'échelle :

$$W_{ij} = e^{-\left(\frac{d_{ij}}{r}\right)}$$

où *r* est le paramètre d'échelle. Plus *r* est grand plus les couples d'US éloignées conservent des poids importants. Le choix du paramètre d'échelle détermine donc la sensibilité du test à une taille d'agrégats : plus *r* est élevé plus le test sera sensible aux agrégats étendus et inversement.

Quelle que soit la relation de voisinage prise en compte (distance, frontière commune, …), la matrice de voisinage fait généralement l'objet d'une standardisation selon diverses méthodes. Les plus utilisées sont :

- La méthode de standardisation dite « en ligne », où la somme de chaque ligne de la matrice est égale à 1, i.e. $\sum_{j=1}^{n} w_{ij} = 1, \forall i$ et, par conséquent, $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} = N$.
- Les méthodes de standardisation globales, où la somme totale des w_{ij} est égale à *N* ou 1 selon la méthode.
- Les méthodes de standardisation destinées à stabiliser la variance, telle que celle proposée par Tiefelsdorf en 1997 [26].

2.3.2 Coefficient de Moran

La statistique globale de Moran [27, 28] est un coefficient d'autocorrélation spatiale introduit dans les années 1950, encore très largement utilisé. Le coefficient *I* de Moran s'écrit :

$$I = \frac{N * \sum_{ij}^{N} w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{ij}^{N} w_{ij} * \sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

où $Y_i = \frac{c_i}{P_i}$ est la proportion de cas observés dans l'US *i*, $\bar{Y} = \frac{\sum_{i=1}^{N} \frac{c_i}{P_i}}{N}$ est la moyenne des proportions de cas sur l'ensemble de la région et w_{ij} est l'élément de la matrice de proximité correspondant au couple d'US *i* et *j*.

Les statistiques locales de Moran (I_i), sont les *N* composantes de *I*

$$I_i = \frac{(Y_i - \bar{Y}) \times \sum_{j}^{N} w_{ij}(Y_j - \bar{Y})}{\frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N}}$$

Le principal point faible de cette statistique est qu'il est supposé que la moyenne des proportions de cas est une représentation adéquate du phénomène d'intérêt. Cette statistique suit approximativement une loi normale lorsque les cas sont indépendamment distribués selon une loi normale. L'inférence pose une double difficulté : d'une part les conditions de normalités sont en pratique rarement respectées et d'autre part il y a autant de tests qu'il y a de statistiques et l'inflation du risque alpha doit être prise en compte. Ainsi, l'inférence s'effectue par méthode de Monte Carlo (voir chapitre 2.4), puis une méthode d'ajustement des p-values est utilisée (voir chapitre 2.5).

2.3.3 Getis and Orb

Getis et Orb [29] ont introduit en 1992 deux statistiques *G* et G^* et leurs versions locales G_i et G_i^*.

$$G_i(d) = \frac{\sum_{j}^{N} w_{ij}(d) c_j}{\sum_{j}^{N} c_j} \ , j \neq i \qquad (2\text{-}2)$$

où c_j est le nombre de cas observés dans l'US *j* et l'ensemble des $w_{ij}(d)$ constitue une matrice de proximité valant 1 pour tous les liens définis comme étant à une distance inférieure à *d* pour une

US i donnée. Tous les autres liens, y compris le lien à l'US i elle-même, valent 0. La statistique $G_i(d)$ est comprise entre 0 et 1.

Si l'on pose

$$\bar{c}_i = \frac{\sum_j^N c_j}{(N-1)} \; et \; s_i^2 = \frac{\sum_j^N c_j^2}{(N-1)} - \bar{c}_i{}^2$$

et que l'on note W_i la somme des $w_{ij}(d)$ sur j pour $j \neq i$, G_i peut être redéfinie en considérant (1) sa différence avec la valeur attendue $E[G_i] = W_i/(N-1)$ divisée par son écart-type, (2) des poids non binaires pour la matrice de proximité et (3) en incluant $w_{ii} \neq 0$. La statistique standardisée G_i^* se définit par :

$$G_i^*(d) = \frac{\sum_j^C w_{ij}(d)\, c_j - W_i^* \bar{c}}{s\sqrt{\left(NS_i^* - W_{ij}^{*\,2}\right)/(N-1)}} \;, \forall j$$

où $W_i^* = W_i + w_{ii}, S_i^* = \sum_j w_{ij}^2 \; \forall j$, $\bar{c}$ et s^2 sont respectivement l'espérance et la variance de l'échantillon.

La distribution de cette statistique est normale lorsque la distribution des cas est normale. Lorsque cela n'est pas le cas et que la distribution est fortement asymétrique, la distribution de la statistique s'approche de la normalité lorsque d est grand. Cette approximation est cependant moins bonne et converge plus lentement (vers une distribution normale) pour les US au bord de la région d'étude [23]. De la même façon que pour le test local de Moran, l'inférence peut être obtenue par la méthode de Monte Carlo et l'inflation du risque alpha maitrisée par ajustement des p-values.

2.4 Inférence de Monte Carlo

N'importe quel test d'inférence nécessite de connaître la distribution de la statistique sous l'hypothèse nulle formulée.

Dans certaines situations, il est difficile de formuler une distribution sous l'hypothèse nulle (que l'on appellera distribution nulle), même asymptotique car les conditions d'applications ne sont pas remplies voire inappropriées. Cette situation est fréquente dans le domaine des statistiques spatiales où, pour ne donner qu'un exemple, les hypothèses de normalité faites pour les statistiques locales

d'autocorrélation spatiale peuvent être facilement réfutées. C'est la raison pour laquelle l'inférence de Monte Carlo [30] joue un rôle important dans les statistiques spatiales.

Cette méthode est une approche non paramétrique fondée sur le principe des tests de permutation.

Les tests de permutation ou tests exacts sont fondés sur l'idée que, sous H_0, une partie des données est interchangeable. Considérons comme exemple, un jeu de données observées comprenant deux variables : un groupe de traitement et un critère de jugement sur lequel porte la statistique de test. Le principe des tests de permutation est que, sous l'hypothèse nulle, l'étiquette de traitement peut-être interchangée. La statistique de test est calculée pour toutes les permutations possibles (que l'on appellera réplicats). La p-value est obtenue à partir du rang de la statistique observée parmi les statistiques calculées pour les réplicats.

Lorsque la taille de l'échantillon est importante, il est difficile de considérer toutes les permutations possibles. L'échantillonnage aléatoire (échantillonnage de Monte Carlo) permet de générer une distribution de référence à partir d'un nombre relativement faible de réplicats.

La méthode de Monte Carlo est donc une méthode permettant d'estimer la p-value. La précision de cette estimation dépend du nombre de réplicats. Un grand nombre de réplicats permet une meilleure précision mais aux dépens d'une augmentation du temps de calcul qui peut être rédhibitoire. De plus, contrairement aux tests de permutations, pour le même jeu de données observée, la p-value estimée peut varier d'une inférence à une autre puisque seul un échantillon aléatoire des réplicats possibles est utilisé.

Il s'agit là d'une considération importante dans une étude de simulation pour le choix des résultats qui seront recueillis et sauvegardés car la p-value ne pourra pas être retrouvée *a posteriori*.

2.5 Inflation du risque alpha et statistiques locales d'autocorrélation spatiale

La méthode de maitrise de l'inflation du risque alpha la plus connue est sans doute celle de Bonferroni, citée par Dunn en 1961 [31]. Elle est cependant connue pour être trop conservatrice et engendrer une baisse de puissance importante. La méthode de Bonferroni et les méthodes dérivées (dont celle de Holm [32]) appartiennent à la même famille de méthodes destinées à contrôler la

probabilité de commettre une erreur de type I sur un ensemble de comparaisons. D'autres méthodes, moins conservatrices ont depuis été proposées, dont la méthode de Benjamini et Hochberg [33] fondée sur le contrôle de la proportion d'erreurs commises en rejetant à tort l'hypothèse nulle. Cette méthode, que l'on appellera procédure de contrôle FDR (« False Detection Rate »), est équivalente aux méthodes précédentes dans le cas où l'hypothèse nulle est toujours vraie, et moins conservatrice sinon.

La procédure de contrôle FDR est la suivante :

Considérons m tests d'inférence H_1, H_2, ..., H_m pour lesquels on obtient les p-values correspondantes p_1, p_2, ..., p_m. On pose $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ les p-values ordonnées et $H_{(i)}$ l'hypothèse nulle correspondante à $p_{(i)}$. La procédure de contrôle est telle que pour k étant la plus grande valeur de i pour laquelle $p_{(i)} \leq \frac{i}{m}\alpha$, alors les $H_{(i)}$ $i = 1, 2, \ldots, k$ sont rejetées.

A partir de cette méthode, il est possible d'obtenir des p-values ajustées. Sous R [34], la fonction « p.adjust » (le paramètre « *method* » étant spécifié comme « fdr ») corrige les N p-values obtenues pour les statistiques locales d'autocorrélation spatiale de la façon suivante :

On considère les N p-values ordonnées de façon croissante constituant ainsi l'ensemble $p_i \in \{p_1 < p_2 < \cdots < p_{N-1} < p_N\}$. Les p'_i p-values ajustées sont telles que :

$$\begin{cases} p'_j = min\left(\frac{p_j \times N}{j}, p'_{j+1}, \ldots, p'_{N-1}, p'_N\right) \\ p'_N = \frac{p_N \times N}{N} = p_N \end{cases}$$

Partie 2

Évaluation des méthodes statistiques en épidémiologie spatiale

1 Simulation de données spatialisées

L'évaluation de méthodes statistiques, qu'elles traitent de données spatialisées ou non, implique d'approcher la « réalité » du phénomène objet de ces méthodes, afin de pouvoir la comparer aux résultats effectivement obtenus.

Dans le domaine de l'épidémiologie spatiale, comme dans bien d'autres domaines d'application des statistiques spatiales, les phénomènes que l'on souhaite pouvoir mettre en évidence peuvent être extrêmement complexes. La première étape de l'évaluation des méthodes d'analyse est de définir précisément pour quel phénomène leur performance est évaluée. On appellera par la suite « scénario » l'ensemble des caractéristiques permettant de définir le phénomène d'intérêt.

La validité de l'évaluation, en particulier de l'interprétation que l'on pourra en faire, dépend de la qualité des données analysées. L'évaluation de la performance sera d'autant plus valide (et donc utile) que les données représenteront correctement le scénario étudié. Le risque à éviter est d'aboutir à des conclusions concernant un scénario qui n'est pas le scénario initialement ciblé.

L'idéal, bien sûr, serait de disposer de données réelles pour lesquelles on connaîtrait la réalité du scénario qui les a générées, ce qui est impossible en pratique. De plus, l'utilisation de données réelles implique de n'évaluer le test que sur une seule réalisation du scénario d'intérêt. Il est donc souvent nécessaire d'avoir recours à la simulation de données. La simulation garantit la portée des résultats dans le sens où le scenario est parfaitement connu et le test évalué sur un grand nombre de réalisations. Néanmoins, un scénario simulé ne reflètera jamais complètement la complexité de la réalité et cette simplification limite forcément la portée des résultats.

La première étape de la simulation de données spatialisée consiste à définir le(s) scénario(s) étudié(s) et en particulier le modèle d'agrégation retenu ainsi que le modèle de distribution des cas.

1.1Modèles d'agrégation

1.1.1 Types d'agrégat

1.1.1.1 Agrégat de type « hot-spot »

Les agrégats de type « hot-spot » (Figure 8) se caractérisent par une élévation constante du risque de maladie dans la zone qu'ils occupent. En dehors de cette zone, le risque de maladie correspond au risque nominal[2] souvent qualifié de risque de base.

1.1.1.2 Agrégat clinal

Les agrégats de type clinal (Figure 9) sont caractérisés par un décroissement continu du risque de maladie autour du point de source jusqu'à retrouver le niveau de base.

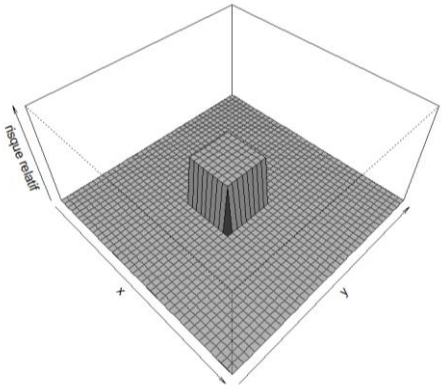

Figure 8: agrégat de type "hot-spot"

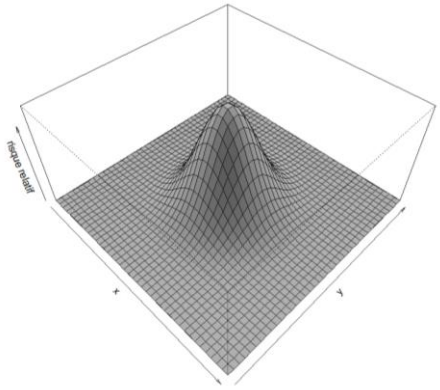

Figure 9: agrégat clinal

1.1.1.3 Autres paramètres

Définir un modèle d'agrégation implique également de définir la localisation, la taille, le nombre et la forme précise du ou des agrégat(s) simulé(s). À ces paramètres que l'on qualifiera de géographiques, s'ajoutent des paramètres épidémiologiques (tels que l'incidence de base ou le risque relatif au sein de l'agrégat) qui concerneront le phénomène ou la maladie étudiée.

[2] Au sens de conforme à une référence prédéterminée, aux erreurs de mesure près, en parlant d'une grandeur, un état ou d'un processus.

L'ensemble de ces paramètres sont pris en compte dans la formulation des hypothèses de distribution des cas.

1.2Hypothèses de distribution

1.2.1 Simulation d'agrégats de type hot-spot

1.2.1.1 Scénario simple

Lorsque l'on travaille à partir de données groupées, il est aisé de simuler le nombre de cas par US grâce à un PPI dont l'intensité devra au moins être fonction de la taille de la population lorsque cette dernière est connue.

Il est également tout à fait possible de prendre en compte de multiples facteurs épidémiologiques dans le PPI à partir du moment où ces facteurs sont à la fois pertinents et que leur répartition spatiale est connue.

Le nombre des cas dans chaque US pour un agrégat de type hot-spot sont des v.a. indépendantes telles que :

$$\begin{cases} H_0 : E(C_i) = \varepsilon_i, C_i \sim Poisson(\varepsilon_i), i = 1, \dots, N \\ H_1 : E(C_i) = \pi_i, C_i \sim Poisson(\pi_i), \pi_i = \mathbb{I}_i \theta \varepsilon_i + \varepsilon_i(1 - \mathbb{I}_i), i = 1, \dots, N \end{cases}$$

où C_i est le nombre de cas observés, ε_i et π_i représentent, respectivement, le nombre attendu de cas dans l'US i sous l'hypothèse nulle d'homogénéité des risque (H_0) et l'hypothèse alternative simulée (H_1). θ est le risque relatif, et $\mathbb{I}_i$ est une indicatrice binaire valant 1 si l'US i est à l'intérieur du (des) agrégat(s) simulé(s) et 0 autrement. Le nombre de cas attendus ε_i est égal à λP_i où λ est l'incidence de base au sein de la région d'étude et P_i est la taille de la population à risque dans l'US i.

Lorsque l'incidence de base n'est pas connue, C (le nombre total de cas) est une statistique suffisante. Par conséquent, une statistique de test dont la distribution est indépendante du paramètre λ inconnu, peut être conditionnée par le nombre total de cas observés c. Étant donné c, l'hypothèse nulle peut-être formulée par :

$$H_0 : C_i \sim Multinomial\left(c, \rho_i = \frac{P_i}{\sum_{j=1}^{N} P_j}\right), i = 1, \ldots, N$$

Et l'hypothèse alternative par :

$$H_1 : C_i \sim Multinomial\left(c, \nu_i = \frac{P_i[1 + \mathbb{I}_i(\theta - 1)]}{\sum_{j=1}^{N} P_j\left[1 + \mathbb{I}_j(\theta - 1)\right]}\right), i = 1, \ldots, N$$

1.2.1.2 Prise en compte de facteurs de confusion

D'autres facteurs épidémiologiques (*e.g:* sexe, âge) peuvent être pris en compte dans les hypothèses de distribution en utilisant des méthodes de standardisation indirecte, également appelées méthodes de standardisation interne car elles utilisent les données observées. Pour cela on définit autant de strates *k* que de combinaisons de niveaux des facteurs à prendre en compte. Pour chacune de ces strates on dénombre les cas observés et la taille de la population à risque par US *i*. Pour une distribution de Poisson, on définit le ratio

$$r_k = \frac{\sum_i c_{ik}}{\sum_i P_{ik}},$$

correspondant au ratio de la somme des cas observés sur la taille de la population à risque dans la strate *k* pour l'ensemble de la région d'étude. Le nombre de cas attendus dans l'US *i* est alors défini par

$$\varepsilon_i = \sum_k P_{ik} r_k.$$

Le raisonnement est superposable pour une distribution multinomiale :

$$\rho_{ik} = \frac{P_{ik}}{\sum_{j=1}^{N} P_{jk}},$$

La probabilité ρ_i est alors définie par :

$$\rho_i = \frac{\sum_k c_k \rho_{ik}}{c}$$

Où c_k est le nombre de cas observé dans la région pour la strate *k*.

On peut également, par ces méthodes de standardisation indirecte, modéliser une interaction entre agrégation spatiale et facteur(s) épidémiologique(s) (*e.g.*, un facteur de risque

environnemental dont l'effet serait d'autant plus important qu'il concernerait un sexe et/ou une tranche d'âge particuliers).

Pour une distribution de Poisson, la formulation de l'hypothèse alternative peut être de la forme :

$$H_1 : \begin{cases} E(C_i) = \pi_i, C_i \sim Poisson(\pi_i) \\ \pi_i = \mathbb{I}_i \left(\sum_k \theta_k P_{ik} r_k \right) + (1 - \mathbb{I}_i) \sum_k P_{ik} r_k \\ i = 1, \dots, N \end{cases}$$

Où θ_k est le risque relatif au sein de la strate *k*.

Pour une distribution multinomiale, l'hypothèse alternative peut être de la forme :

$$H_1 : \begin{cases} C_i \sim Multinomial(c, \nu_i), i = 1, \dots, N \\ \nu_{ik} = \dfrac{P_{ik}[1 + \mathbb{I}_i(\theta_k - 1)]}{\sum_{j=1}^{N} P_{jk} \left[1 + \mathbb{I}_j(\theta_k - 1)\right]} \\ \nu_i = \dfrac{\sum_k c_k \nu_{ik}}{c} \end{cases}$$

1.2.2 Simulation d'agrégats clinaux

La simulation d'agrégats de type clinal peut se faire simplement en reprenant les hypothèses alternatives proposées pour les agrégats de type hot-spot mais en formulant θ comme un vecteur de risque relatif θ_i.

Ainsi on obtient :

$$\begin{cases} H_0 : E(C_i) = \varepsilon_i, C_i \sim Poisson(\varepsilon_i), i = 1, \dots, N \\ H_1 : E(C_i) = \pi_i, C_i \sim Poisson(\pi_i), \pi_i = \theta_i \varepsilon_i, i = 1, \dots, N \end{cases}$$

Où θ_i vaut 1 lorsque l'US i n'appartient pas à l'agrégat simulé et est supérieur à 1 autrement, avec une valeur maximale pour la (les) US au centre de l'agrégat.

De la même façon, pour une distribution multinomiale :

$$H_1 : C_i \sim Multinomial\left(c, \nu_i = \frac{P_i \theta_i}{\sum_{j=1}^{N} P_j \, \theta_j}\right), i = 1, \dots, N$$

Les valeurs de θ_i peuvent être déterminées de multiples façons. L'idéal étant évidemment de disposer de données de terrain (e.g. relevés de pollution atmosphérique,…) pour quantifier

l'exposition. L'association entre l'exposition et la maladie peut cependant rarement être quantifiée de façon suffisamment complète pour pouvoir traduire directement la mesure d'exposition en risque relatif. Il est possible d'utiliser une fonction gaussienne à deux dimensions pour estimer un risque relatif au point de coordonnées *(x,y)*. La fonction gaussienne à deux dimensions est de la forme :

$$f(x,y) = Ae^{-(a(x-x_0)^2+2b(x-x_0)(y-y_0)+c(y-y_0)^2)}$$

Où (x_0, y_0) sont les coordonnées du centre de l'agrégat, A est la valeur du risque relatif au centre de l'agrégat (son sommet). Les paramètres a, b et c sont définis tels que :

$$\begin{cases} a = \dfrac{\cos^2\vartheta}{2\sigma_x^2} + \dfrac{\sin^2\vartheta}{2\sigma_y^2} \\ b = -\dfrac{\sin(2\vartheta)}{4\sigma_x^2} + \dfrac{\sin(2\vartheta)}{4\sigma_y^2} \\ c = \dfrac{\sin^2\vartheta}{2\sigma_x^2} + \dfrac{\cos^2\vartheta}{2\sigma_y^2} \end{cases}$$

L'écartement par rapport à l'axe central (la variance) est défini par σ_x^2 et σ_y^2. La base de l'agrégat forme une ellipse orientée dans le sens horaire d'un angle ϑ.

Cependant, cette fonction n'est adaptée qu'à des données spatiales ponctuelles. Néanmoins, un processus de définition du vecteur de risques relatifs pour données groupées pourrait être résumé par :

1) Déterminer la fonction gaussienne à 2 dimensions la plus appropriée aux données de terrain.
2) Déterminer la matrice de risques relatifs correspondante.
3) Attribuer un risque relatif à chaque US, par exemple en attribuant la moyenne de l'ensemble des risques relatifs calculés pour la surface de l'US.

La prise en compte de facteurs de confusion peut se faire de façon similaire à celle évoquée pour les agrégats de type hot-spot en remplaçant l'indicatrice $\mathbb{I}_i$ et le scalaire θ par un vecteur θ_i. Un niveau de complexité supplémentaire peut être introduit en construisant une matrice θ_{ik} afin de modéliser une interaction éventuelle entre les strates et l'exposition.

1.3 Objectifs et Stratégies de simulation

1.3.1 Scénario simple

Lorsque tous les paramètres aussi bien géographiques qu'épidémiologiques sont fixés, le scénario est simulé un certain nombre de fois. On appellera ces simulations successives des réalisations, le terme de simulation étant en général utilisé pour désigner le scénario simulé. Pour chaque réalisation (voir Figure 10), les données sont analysées par le (les) test(s) évalué(s) et les résultats sont collectés soit de façon « brute » (l'agrégat potentiel et la p-value associée), soit directement par le recueil de la mesure d'intérêt (*e.g.* une variable binaire indiquant si oui ou non l'agrégat détecté contient au moins une US de l'agrégat simulé). Cette mesure d'intérêt sera ensuite résumée par un indicateur de performance (le taux de réalisations ayant résulté en la détection d'un agrégat contenant au moins une US de l'agrégat simulé). Dans tous les cas, les résultats de ce type de simulations se rapportent strictement au scénario simulé et toute généralisation des conclusions doit être considérée avec prudence.

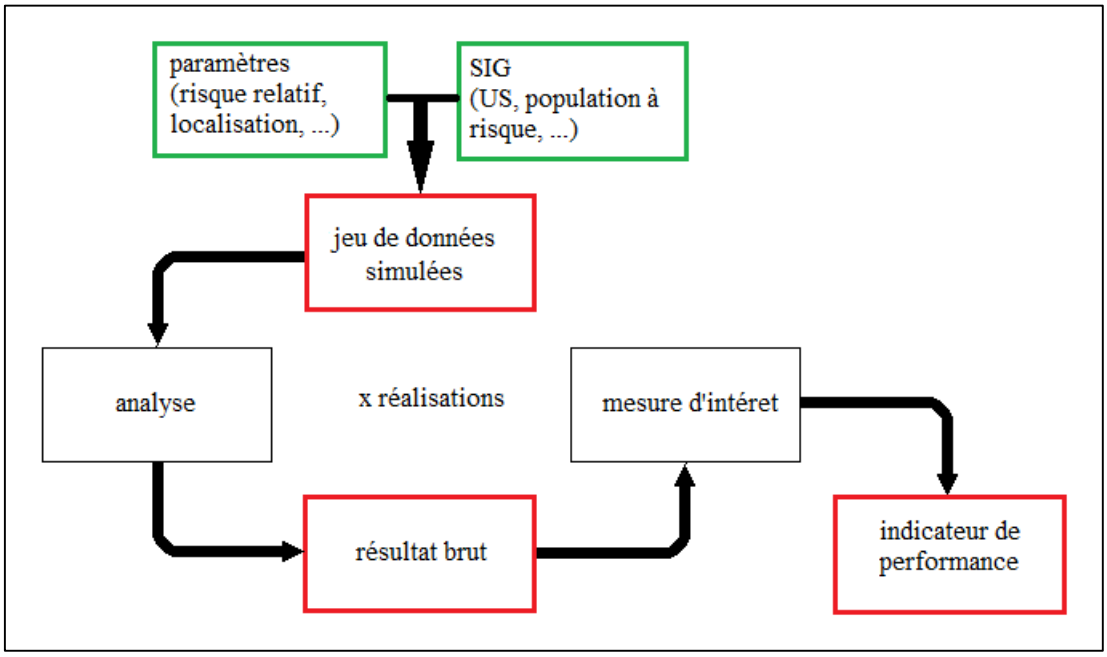

Figure 10: scénario de simulation simple

1.3.2 Évaluation d'un facteur

Lorsque l'on cherche à évaluer l'effet d'un paramètre (géographique ou épidémiologique) sur les performances d'un test, la stratégie de simulation consiste à fixer tous les autres paramètres et à faire varier la valeur du paramètre évalué au sein d'une gamme de valeurs cibles (voir Figure 11, e.g. [35–38]). Ce paramètre est appelé facteur pour le différencier des autres paramètres dont les valeurs sont fixes. Un certain nombre de réalisations du scénario est consacré à l'évaluation de chacune des différentes valeurs étudiées.

L'avantage de cette stratégie, est qu'il est possible d'imputer directement un éventuel effet sur les performances au facteur étudié puisqu'il constitue la seule source de différence systématique entre les différentes réalisations de la simulation. L'interprétation reste cependant limitée ici aussi au scénario étudié, et en particulier, il n'est pas possible de dégager l'effet propre du facteur étudié sur la performance. En effet, si l'on prend l'exemple du risque relatif au sein de l'agrégat (que l'on appelle parfois force de l'agrégat), puisque tous les autres paramètres sont fixés et ne varient pas, il ne sera pas possible de dégager l'effet de l'incidence ou de la taille de la population de celui du risque relatif sur la performance du test évalué. Le deuxième inconvénient de ce type de stratégie est la démultiplication du nombre de réalisations nécessaires à l'évaluation.

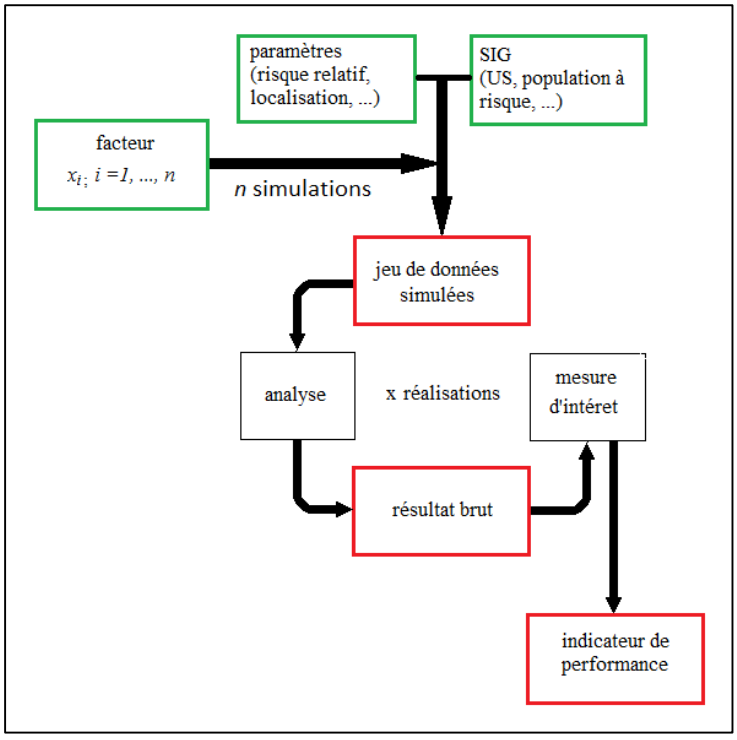

Figure 11: stratégie de simulation pour un facteur

1.3.3 Évaluation de plusieurs facteurs

L'étude concomitante de plusieurs facteurs peut être réalisée selon un plan factoriel complet (voir Figure 12) et permet d'étudier leurs effets propres et leurs éventuelles interactions, toujours bien sûr en gardant à l'esprit que cette évaluation est faite tout autre paramètre fixé et qu'il n'est donc pas garanti qu'aucun phénomène de confusion ne subsiste. De plus, ce type de stratégie [39–43] nécessite un grand nombre de réalisations, ce qui est difficilement envisageable si l'on ne dispose de matériel informatique dédié.

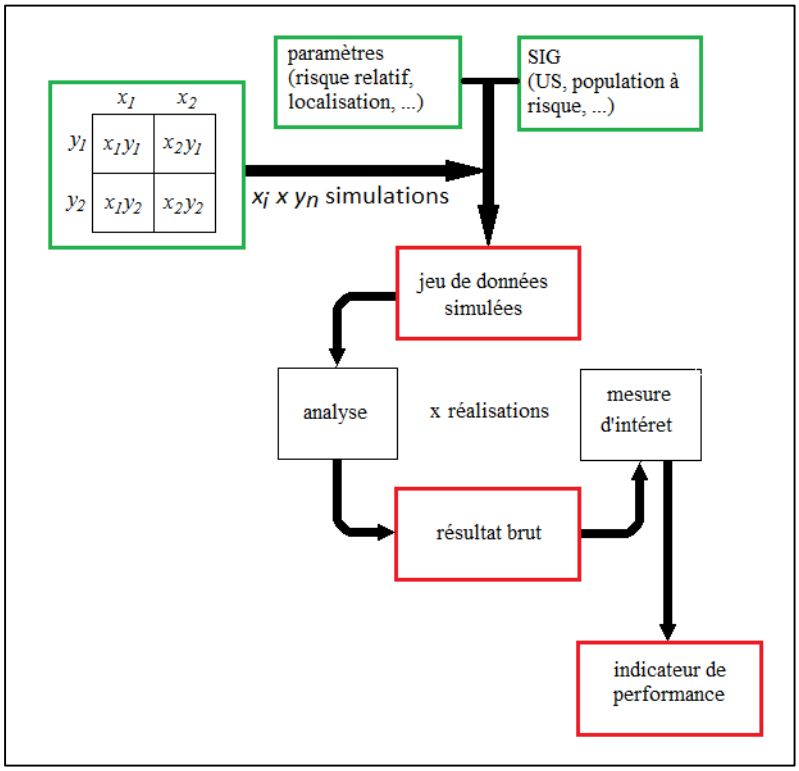

Figure 12: stratégie de simulation selon un plan factoriel

1.4 Sondage aléatoire

Le principe est, au contraire des stratégies décrites ci-dessus, de ne pas fixer de valeur pour les paramètres définissant le contexte de la simulation mais d'effectuer un sondage aléatoire au sein des valeurs possibles de ces paramètres pour chaque réalisation de la simulation (voir Figure 13).

Le(s) facteur(s) évalué(s) est (sont) fixé(s) et une simulation est réalisée pour chaque valeur (ou combinaison de valeurs) de ce(s) facteur(s). Ce type de stratégie (*e.g.* Kulldorff *et al.* 2012 [44])

est employé lorsque l'on cherche à évaluer la performance d'un test tout en prenant en compte d'éventuels phénomènes de confusion induits par certains paramètres (comme la localisation ou la force de l'agrégat par exemple). Par exemple si l'on souhaite évaluer l'effet du risque relatif au sein de l'agrégat en tenant compte de l'incidence de base, de la localisation de l'agrégat, de sa taille, etc., la première étape est de définir les valeurs du risque relatif que l'on souhaite explorer. Pour chacune de ces valeurs une simulation est réalisée dans laquelle tous les paramètres que l'on cherche à prendre en compte se voient affecter une valeur selon un sondage aléatoire de leur distribution théorique. Ainsi, on pourra mesurer l'effet propre du risque relatif indépendamment de l'effet potentiel de l'incidence de base, de la localisation de l'agrégat, de sa taille et de tout autre paramètre que l'on aura pris en compte.

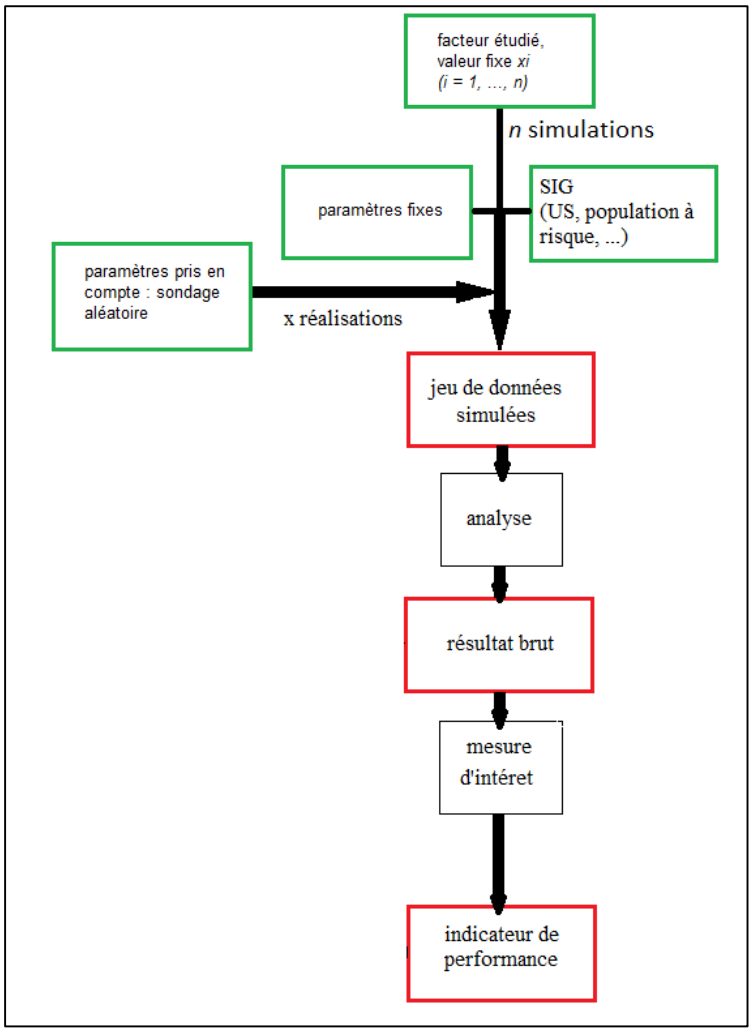

Figure 13: stratégie de simulation par sondage aléatoire

1.5 Programmation statistique

L'ensemble du travail présenté est réalisé sous R [34] qui est à la fois un langage de programmation (fondé sur la quatrième version de S : S4) et un environnement développé dans les laboratoires Bell (anciennement AT&T, maintenant Lucent Technologies) par John Chambers et ses collègues.

C'est un projet GNU[3] donc totalement libre et gratuit. (La version payante du langage étant S-PLUS.) Le langage R est donc un véritable langage de programmation. De type interprété et fonctionnel, ce langage est particulièrement adapté à la programmation statistique, à laquelle il est dédié.

1.5.1 Générateur de nombres aléatoires

Chaque réalisation d'une simulation implique de générer des nombres qui correspondent à un sondage aléatoire d'autant de valeurs que souhaitées (en général une valeur par US) d'une distribution conforme à celle de l'hypothèse simulée (en général l'hypothèse alternative).

Les logiciels de programmation statistique ne sont pas capables de générer des nombres aléatoires. Cependant, il existe de nombreuses façons de générer des nombres pseudo-aléatoires. Les générateurs les plus « sûrs » reposent sur des phénomènes physiques (*e.g.* passage d'un photon à travers une surface semi-réfléchissante).

Les programmes informatiques permettant de générer des nombres pseudo-aléatoires utilisent des algorithmes plus ou moins complexes [45, 46]. Les générateurs de nombre pseudo-aléatoires (PRNG pour « Pseudo Random Number Generator » ou générateur de nombres pseudo-aléatoires), puisqu'exécutés sur ordinateurs, sont *de facto* déterministes.

Les PRNG sont amorcés par une racine qui détermine l'état initial du système. Ils sont caractérisés par une période définie par le temps nécessaire au système pour retrouver son état initial. Autrement dit, la période est la longueur de la chaîne de nombres pseudo-aléatoires générés avant que celle-ci ne se répète. Parmi les qualités attendues d'un PRNG, on peut citer (en dehors d'une période suffisamment longue) : une période identique quelle que soit la racine, une distribution uniforme quelle que soit la racine, une indépendance entre les valeurs successives.

[3] GNU est un acronyme récursif pour « GNU's not Unix ! »

L'installation de base de R offre tous les outils nécessaires à la génération de nombres pseudo-aléatoires pour une utilisation dans le cadre de ce travail et plus généralement dans le cadre de simulations telles que celles réalisées dans le domaine de la Biostatistique.

Sous R, le PRNG utilisé par défaut[4] est celui inventé par Makoto Matsumoto et Takuji Nishimura en 1997 [47]. Appelé algorithme de Mersenne Twister, il est encore aujourd'hui particulièrement réputé pour sa qualité (bien que les améliorations les plus récentes ne soient accessibles que par la librairie *randtoolbox* [46]). Ce PRNG est caractérisé par une période de 2^{19937}-1. Il génère des entiers de 32 bits, c'est-à-dire compris dans l'intervalle [0, 2^{32}-1] selon une loi uniforme.

A partir de ces entiers, les lois de distributions plus complexes sont ensuite générées par des méthodes adaptées : la méthode d'inversion [48] pour la distribution normale, la méthode de Ahrens et Dieter [49] pour la distribution de Poisson, la méthode d'inversion ou la méthode de Kachitvichyanukul et Schmeiser [50] pour la distribution binomiale (et par extension pour une distribution multinomiale).

Les travaux de Burton et al. [51] sur les protocoles des études de simulation en médecine mettent en avant deux qualités importantes que sont la reproductibilité et l'absence de corrélation entre les réalisations. Outre les qualités intrinsèques du PRNG choisi, ces deux exigences peuvent être satisfaites de plusieurs façons.

Pour garantir la reproductibilité, il suffit de sauvegarder la racine ayant servi à générer les données pour pouvoir retrouver les même données lors d'une nouvelle réalisation. Une autre solution, très simple, est de sauvegarder les données générées pour chaque réalisation. Reproduire les résultats d'une étude impliquera de plus longs temps de calcul avec la première solution et nécessitera plus d'espace de stockage avec la seconde.

La corrélation entre les différentes réalisations d'une simulation n'est pas un problème[5] avec le PRNG Mersenne Twister en raison de ses qualités intrinsèques et en particulier de sa très grande période.

[4] C'est également le PRNG par défaut en Python, Ruby, PHP et sur MATLAB.

[5] Seuls quelques champs d'application, comme la physique nucléaire, peuvent avoir besoin d'algorithmes plus performants.

2 Les outils d'évaluation : synthèse bibliographique

L'évaluation de la performance des méthodes locales de détection d'agrégats comprend différents aspects dont certains correspondent à des aspects généraux d'évaluation de la performance et d'autres sont spécifiques de ces méthodes.

Un des aspects généraux de la performance concerne les problématiques de mise en œuvre des méthodes. Parmi les points les plus importants, on peut citer :

- La simplicité d'utilisation et l'accessibilité, qui se traduisent par l'existence ou non d'une solution logicielle libre ou payante.
- Les temps de calculs, qui constituent un critère difficile à évaluer mais peuvent limiter l'utilisation de méthodes statistiques à des projets de recherche conduits dans des structures bénéficiant de ressources adéquates (au contraire d'une utilisation « de terrain » sur du matériel de type ordinateur portable).
- La complexité de la méthode en elle-même qui peut limiter son utilisation à des spécialistes qui pourront garantir une mise en œuvre correcte et la bonne interprétation des résultats.

Les méthodes locales de détection d'agrégats, du simple fait qu'elles délivrent une double conclusion (la localisation du (des) agrégat(s) potentiels et l'inférence correspondante), font appel à une évaluation qui doit prendre en compte chacun de ces aspects (performance de localisation et puissance statistique).

Quel que soit l'aspect concerné, la performance peut être entendue comme la quantification de l'atteinte d'un objectif de référence. Si, pour la puissance et la localisation de l'agrégat, cet objectif peut être aisément déterminé (puissance égale à 100% et détection de toutes les US de l'agrégat et uniquement ces US), ce n'est pas forcément le cas pour ce qui concerne la mise en œuvre des méthodes pour lesquelles la performance cible ou de référence sera extrêmement dépendante du contexte d'application.

Par exemple, les temps de calculs sont non seulement dépendants du matériel utilisé, mais également du langage de programmation utilisé et de la nature même de la méthode. De plus, pour des exécutions sur plusieurs machines de configurations matérielles identiques, en utilisant le

même logiciel et le même programme, les temps de calculs peuvent varier du simple fait de l'environnement logiciel et de la charge de travail des machines.

Ainsi, l'interprétation d'un temps de calcul comme mesure de performance est toujours dépendante à la fois du contexte dans lequel il a été mesuré et du contexte dans lequel l'utilisateur potentiel pourra utiliser la méthode. Il en va de même pour tous les autres aspects de performance ayant trait à la mise en œuvre des méthodes.

Dans la littérature scientifique, cet aspect est traité indirectement grâce à la description précise de toutes les étapes de la mise en œuvre pratique, les utilisateurs potentiels ayant à charge l'interprétation de ces éléments pour déterminer la performance. Même si la performance liée à la mise en œuvre des méthodes prend souvent une position secondaire au regard de la performance de leurs résultats, elle peut aisément conditionner la diffusion et l'utilisation des méthodes.

Les outils d'évaluation utilisés dans la littérature sont très hétérogènes et il n'existe pas, à l'heure actuelle, de consensus méthodologique dans ce domaine. La performance des résultats des méthodes locales de détection d'agrégats peut faire l'objet de mesures partielles qui n'en considèrent qu'un seul aspect (puissance ou localisation), ou au contraire de mesures globales.

La construction de ces indicateurs est explicitée ci-dessous. Le Tableau 4 synthétise les méthodologies retrouvées dans la littérature.

2.1 Indicateurs partiels

2.1.1 Mesures d'intérêt

A chaque réalisation, les résultats fournis par la (les) méthode(s) font l'objet d'une première manipulation destinée à recueillir l'information nécessaire au calcul de l'indicateur de performance qui portera sur l'ensemble des réalisations. Cette mesure peut caractériser chaque US de la région d'étude ou uniquement l'agrégat détecté.

2.1.1.1 Mesures portant sur les US

A. Qualification des US

Lorsqu'un agrégat est détecté, les US de la région d'étude sont réparties en quatre catégories définies dans le Tableau 1 : les vrais positifs (VP), faux positifs (FP), vrai négatifs (VN) et enfin les faux négatifs (FN).

Tableau 1: classification des US selon l'agrégat détecté et l'agrégat simulé

	US de l'agrégat simulé	US hors de l'agrégat simulé
US de l'agrégat détecté	VP	FP
US hors de l'agrégat détecté	FN	VN

Cette classification constitue parfois la mesure d'intérêt utilisée mais le plus souvent, cette classification est une étape intermédiaire permettant d'obtenir la mesure d'intérêt.

B. Mesures quantitatives

Le Tableau 2 résume les différentes mesures d'intérêt construites à partir de cette classification et utilisées dans la littérature.

Tableau 2: mesure de performance à partir des US

Mesure	Expression	Définition
sensibilité	$Se = \frac{VP}{VP + FN}$	Probabilité qu'une US appartienne à l'agrégat détecté si elle fait partie de l'agrégat simulé.
spécificité	$Sp = \frac{VN}{VN + FP}$	Probabilité qu'une US n'appartienne pas à l'agrégat détecté si elle ne fait pas partie de l'agrégat simulé.
valeur prédictive positive[6]	$VPP = \frac{VP}{VP + FP}$	Probabilité qu'une US appartienne à l'agrégat simulé si elle fait partie de l'agrégat détecté.
efficacité (« accuracy »)	$Acc = \frac{VP + VN}{VP + VP + FP + FN}$	Probabilité qu'une US soit correctement classée comme appartenant ou non à l'agrégat simulé
Ratio de vraisemblance positif	$RV^{+} = \frac{Se}{1 - Sp}$	Une US de l'agrégat simulé à RV+ fois plus de chance d'être positive qu'une US en dehors de l'agrégat simulé.
Taux d'erreur	$TE = \frac{FP + FN}{VP + FP + FN}$	Sur l'ensemble des US de l'agrégat simulé et détecté, part des US classées incorrectement.

2.1.1.2 Mesures portant sur les agrégats

Ces mesures sont des mesures qualitatives binaires qui qualifient l'agrégat dans son ensemble. La mesure la plus simple est celle de l'absence/présence d'un agrégat détecté quelles que soient ces caractéristiques. Cette mesure permet de calculer la puissance usuelle. Les autres mesures se définissent comme la présence/absence d'agrégat détecté répondant à un certain nombre de critères destinés à évaluer la performance de la localisation de l'agrégat simulé (*cf.* Tableau 3). Ces mesures, qui portent sur les agrégats et non plus les US, permettent de calculer des puissances

[6] Les valeurs prédictives, telles qu'entendues en évaluation de méthodes diagnostique, dépendent de la prévalence de la maladie que l'on pourrait assimiler dans le contexte de ce travail à la proportion d'US appartenant à l'agrégat sur l'ensemble des US de la région d'étude. La définition donnée ici, bien que critiquable, est celle retrouvée dans la littérature.

conditionnelles. Là encore, il n'existe pas de consensus et plusieurs définitions sont retrouvées dans la littérature. La définition la plus restrictive considère uniquement les agrégats détectant parfaitement les agrégats simulés, c'est-à-dire ne générant ni FP ni FN.

Tableau 3: mesures d'intérêt portant sur les agrégats

Type	Définition
1	l'agrégat détecté correspond parfaitement à l'agrégat simulé
2	l'US centrale de l'agrégat détecté est également celle de l'agrégat simulé
3	l'agrégat détecté contient au moins une US de l'agrégat simulé
4	le centre de l'agrégat détecté est situé à moins d'un rayon du centre de l'agrégat simulé
5	l'agrégat détecté contient entièrement l'agrégat simulé
6	l'agrégat détecté contient le centre de l'agrégat simulé

2.1.2 Construction de l'indicateur de performance

2.1.2.1 Statistique de résumé

L'ensemble des mesures obtenues concerne les résultats d'une seule réalisation. Pour résumer l'information sur l'ensemble de la simulation, plusieurs méthodes sont utilisées.

La moyenne est la seule statistique de résumé utilisée. Dans la littérature, elle est rarement assortie d'un intervalle de confiance et jamais d'une mesure de dispersion. La moyenne porte le plus souvent sur des mesures d'intérêt quantitatives (sensibilité, spécificité, …), mais on la trouve également calculée d'abord sur les quatre catégories d'US (VP, VN, FP, FN), dont les effectifs moyens servent ensuite à calculer l'indicateur de performance.

Une autre façon de résumer l'information fournie par la mesure d'intérêt est d'utiliser les sommes cumulées des US appartenant à chacune des quatre catégories sur l'ensemble des réalisations prises en compte pour calculer un indicateur de performance tel que la sensibilité par exemple.

Enfin, les mesures qualitatives portant sur les agrégats sont résumées sous forme de proportions dont le dénominateur est le plus souvent le nombre total de réalisations pour la simulation. Ces proportions constituent des puissances conditionnelles.

2.1.2.2 Prise en compte des réalisations sans détection

A. Mesures portant sur les US

Les mesures portant sur les US (sensibilité, spécificité,…) sont souvent destinées à évaluer la performance de localisation indépendamment de la puissance usuelle. Pour cela, ne seront considérées que les réalisations ayant aboutis à la détection d'un agrégat. Ainsi, deux tests ayant des puissances très différentes peuvent cependant aboutir à des mesures de sensibilité (*e.g.*) identiques.

Pour prendre en compte la puissance dans ces mesures, il est possible de considérer les réalisations n'ayant pas abouti au rejet de l'hypothèse nulle comme des « non-détections », c'est-à-dire en considérant qu'il n'y a pas de positif (VP + FP = 0). L'ensemble des réalisations sont ainsi prises en compte. Cependant, cela peut poser des difficultés d'interprétation pour certaines mesures. Par exemple, si l'on considère la spécificité, un test ayant une puissance nulle aura une spécificité absolue soit 100%, de même qu'un test ayant une puissance parfaite (égale à 1) sans jamais générer le moindre FP. Ces deux situations ne représentent pourtant évidemment pas les mêmes performances.

Enfin, certains auteurs considèrent tous les agrégats potentiels, que l'hypothèse nulle ait été rejetée ou non. Cela permet de prendre en compte toutes les réalisations pour l'évaluation de la performance de localisation. Cependant, cela a également l'inconvénient d'être plus difficilement interprétable dans le contexte de l'utilisation pratique des méthodes (qui ne s'intéresse finalement qu'aux agrégats détectés).

B. Mesures portant sur les agrégats

Les mesures portant sur les agrégats sont le plus souvent utilisées pour calculer des puissances conditionnelles. Par définition toutes les réalisations sont prises en compte dans le dénominateur. Cependant, il est tout à fait possible de ne considérer que les réalisations ayant abouti à la détection d'un agrégat afin de construire un indicateur de performance ne prenant en compte que la localisation (e.g., proportion parmi les agrégats détectés, des agrégats recouvrant parfaitement l'agrégat simulé). Il est également possible de considérer tous les agrégats, sans tenir compte de l'inférence, même si cela est rarement rencontré dans la littérature.

Tableau 4: indicateurs de performance dans la littérature

Auteur	Année	Mesure sur les US	Statistiques de résumé*	Réalisations considérées**	Puissance usuelle	Type de puissance conditionnelle	Erreur de type I
Kulldorff [52]	2003	non			oui	non	non
Aamodt [40]	2006	Se, Sp, Acc	m	S	non	non	non
Waller [36]	2006	non			oui	2†, 6‡	non
Duczmal [53]	2006	non			oui	non	non
Assunçao [54]	2006	non			oui	non	non
Ozonoff [43]	2007	Se, Sp	?	AD	oui	4	non
Cook [55]	2007	non			non	5	oui
Huang [56]	2007	Se, VPP	m	AD	oui	non	non
Huang [39]	2008	Se, VPP	m + IC	AD	oui	non	non
Jacquez [35]	2009	Se, Sp, VPP	m	AD	oui	non	non
Jackson [37]	2009	non			oui	non	non
Zhang [57]	2009	non			oui	2	oui
Jung [58]	2010	Se, VPP	m	AD	oui	non	non
Li [41]	2011	Se, VPP	m	AD	oui	non	non
Goujon [59]	2011	Se, VPP	m	A	oui	1	oui
Zhang [60]	2011	non			oui	1	oui
Jung [42]	2012	Se, VPP	m	AD	oui	non	non
Lemke [38]	2013	Se, Sp, VPP, RV^+	m, IC, s	S	oui	non	non
Wang [61]	2013	Se, TE	?	?	oui	non	oui

* m = moyenne ; IC = Intervalle de confiance, s = indicateur obtenu à partir de la somme cumulées des catégories d'US

** S = toutes les réalisations, A = tous les agrégats potentiels sans tenir compte de l'inférence statistique, AD = uniquement les agrégats détectés

† Waller *et al.* considèrent également « la proportion parmi tous les agrégats détectés ou non, de ceux qui localisent l'US centrale du cluster simulé »

‡ Waller *et al.* considèrent également « la proportion parmi tous les agrégats détectés ou non, de ceux qui incluent l'US centrale du cluster simulé »

2.2 Indicateur global – Puissance étendue

La puissance étendue a été proposée par Tango et Takahashi [1, 62] comme un indicateur de performance amélioré, capable de prendre en compte la puissance usuelle tout en évaluant la performance de localisation de façon globale, c'est-à-dire sans avoir à qualifier la performance par l'intermédiaire d'une définition plus ou moins restrictive (tel que pour le calcul d'une puissance conditionnelle).

Pour un agrégat simulé, la puissance étendue se définit comme une somme cumulée pondérée de la contribution à la performance de chaque agrégat détecté sur l'ensemble des réalisations.

Pour un agrégat simulé de s US, si l'hypothèse nulle est rejetée, la taille l (nombre d'US) de l'agrégat détecté et le nombre s^* d'US correspondant à des vrais positifs sont recueillis.

Les auteurs imposent un nombre arbitraire L correspondant à la taille maximum acceptable de l'agrégat détecté avant que celui ne soit considéré comme trop grand pour pouvoir contribuer à la performance. Ainsi, si $l > L$, l'agrégat détecté est ignoré et la réalisation considérée comme n'ayant pas abouti à une détection.

Tango [62] propose de fixer cette limite L à un quart ou un tiers de la région d'étude (en nombre d'US), argumentant qu'il est raisonnable de penser qu'un agrégat réel serait plus petit.

Tous les agrégats potentiels ayant à la fois abouti au rejet de l'hypothèse nulle et étant de taille inférieure à L sont pris en compte dans la performance. Ces agrégats sont appelés agrégats éligibles (AE). Leurs caractéristiques l et s^* sont recueillies et pour chaque combinaison de valeurs de ces caractéristiques, la proportion (sur toutes les réalisations) des AE correspondants ($P_{(l,s^*)}$) se voit attribuer un poids $W_{(l,s^*,q)}$.

Ce poids est fonction d'un paramètre q, continu sur [0,1], permettant une évaluation quantitative de la performance de localisation et donc évitant d'avoir à définir ce que l'on considère comme une performance suffisante. Ce poids est défini comme :

$$W_{(l,s^*,q)} = \sqrt{\left[1 - min\left\{\frac{1}{s}(s - s^*), 1\right\}\right]\left[1 - min\left\{\frac{q}{s}(l - s^*), 1\right\}\right]}$$

Pour chaque valeur de q, la puissance étendue est telle que :

$$EP_q = \sum_{l=1}^{L} \sum_{s^*=0}^{s} W_{(l,s^*,q)} P_{(l,s^*)}$$

La pondération permet de favoriser la sensibilité par rapport à la spécificité, *i.e.* d'être plus tolérant envers les faux positifs que les faux négatifs.

Considérant l_0 le nombre de faux positifs au sein d'un AE, la taille de cet agrégat est alors $l=s^*+l_0$. Lorsque $q=1$, les AE ayant $l_0 \geq s$ faux positifs se voient attribuer un poids $W_{(l\geq s^*+s,s^*,q=1)} = 0$.

Lorsque $q = 0.5$, la contribution des AE est étendue car seuls les AE ayant $l_0 \geq 2s$ faux positifs se voient attribuer un poids $W_{(l\geq s^*+2s,s^*,q=1)} = 0$.

Lorsque $q = 0$, il n'y a plus de pénalité pour les faux positifs et parmi les AE, tout agrégat détecté contribue à la puissance étendue, quelle que soit la valeur de l_0.

La performance globale peut être représentée par la courbe de puissance étendue pour q allant de 0 à 1. En tout point de cette courbe, la puissance étendue est, par construction, comprise entre 0 et 1. De plus, la puissance étendue est une fonction monotone décroissante de q.

Bien que cette solution n'ait encore jamais été utilisée en pratique, Tango et Takahashi [1] suggèrent d'utiliser l'aire sous la courbe de puissance étendue comme indicateur global de performance. L'aire sous la courbe de puissance étendue se définit comme :

$$AUC_{EP} = \int_{q=0}^{1} EP_q dq$$

L'AUC_{EP} est comprise entre 0 et 1, 0 correspondant à un test inopérant (s^* toujours nul ou puissance usuelle nulle) et 1 correspondant à un test parfait (puissance usuelle de 100% et agrégat détecté correspondant toujours parfaitement à l'agrégat simulé).

Partie 3

Résultats

1 Présentation / synthèse :

1.1 Vers un indicateur de performance globale pour les statistiques spatiales

La communauté scientifique s'accorde à reconnaître que les études de puissance classiques sont mal adaptées pour évaluer la performance des méthodes de détection d'agrégats. La plupart des travaux sur la méthodologie d'évaluation de ces méthodes restent cantonnés à la question de la simulation de données spatialisées, soit dans l'idée de proposer à la communauté des données de références [52], soit afin de proposer une base théorique commune (en particulier pour la simulation de distribution spatiales globalement agrégées qui pose un certain nombre de défis, et on peut citer les travaux de Kulldorff *et al.* sur les modèles de chaîne d'agrégation globale [52, 63], ou encore le travail de Jackson *et al.* [37] sur la simulation d'agrégation globale non stationnaire).

Seuls Tango *et al.* se sont intéressés aux mesures de performance en elles-mêmes. D'après leurs réflexions, si l'intérêt d'une mesure de performance prenant en compte à la fois puissance et localisation est indiscutable, le principal écueil dans la construction d'un tel indicateur est la difficulté à définir ce qu'est la performance de localisation.

S'il existe en effet de nombreuses définitions de puissances conditionnelles prenant en compte à la fois puissance et localisation, celles-ci se fondent sur des définitions binaires de la performance de localisation. Ces indicateurs n'évaluent donc la performance que de façon très restrictive. Les travaux de Tango *et al.* ont abouti à la création de la puissance étendue qui permet de décrire de façon complète et globale la performance des méthodes locales de détection d'agrégat à travers une courbe de puissance étendue. Les auteurs suggèrent d'utiliser l'aire sous la courbe de puissance étendue comme indicateur global de performance mais n'ont jamais par la suite publié de travaux mettant en pratique cette idée.

Notre premier travail [64] (voir section 0 de cette partie) a poursuivi cette démarche et en a illustré les avantages et les inconvénients *via* une étude de simulation réalisant une évaluation spatiale systématique de la performance.

De ce travail, nous retenons que l'aire sous la courbe de puissance étendue est un indicateur de performance pertinent qui reflète bien les propriétés connues du scan spatial de Kulldorff,

notamment la forte influence de la taille de la population à risque sur les performances du test. De plus, il a été constaté dans ce travail, que dans certaines situations très similaires (même risque de base, même risque relatif au sein de l'agrégat, taille de population à risque similaire), les performances observées montrent toujours une certaine variabilité. Cette capacité de discrimination des performances peut être utile à l'étude du comportement des méthodes locales de détection d'agrégats et des facteurs qui l'influence, avec pour conséquence directe une meilleure appréhension des voies d'amélioration possible pour ces méthodes.

Cependant, l'aire sous la courbe de puissance étendue présente deux principaux inconvénients. Le premier est inhérent au fait de résumer l'information de performance au sein d'un indicateur unique. Dans le processus, une part de l'information est forcément perdue et par conséquent, il est possible, pour deux situations correspondant à des comportements très différents, d'aboutir malgré tout à la même mesure de performance. Le deuxième inconvénient de cet indicateur est sa complexité qui peut freiner sa diffusion et son utilisation par la communauté. En effet, le lecteur d'une étude de performance s'intéresse avant tout à la confrontation entre la nature du scénario simulé et la performance obtenue pour ce scénario. L'ajout d'une difficulté supplémentaire sous la forme d'une mesure de performance difficile à interpréter peut aisément aboutir à une limitation de la diffusion des résultats à un groupe restreint de spécialistes.

Notre second travail (voir la section 3 de cette partie), portant aussi sur le développement d'un indicateur global de performance, s'est intéressé au coefficient de Tanimoto en raison de son utilisation extensive dans d'autres domaines ayant des préoccupations similaires (*e.g.* biochimie, imagerie médicale), c'est-à-dire ne pouvant se contenter de considérer la performance de « localisation » de façon restrictive (*i.e.* sur un seul aspect tel que la sensibilité ou la spécificité ou de façon binaire comme dans les puissances conditionnelles).

Cet indicateur reconnu a l'avantage, au contraire de l'aire sous la courbe de puissance étendue, de pouvoir s'interpréter de façon extrêmement simple. La difficulté dans l'utilisation de cet indicateur pour les études de simulation est qu'il est par construction une mesure d'intérêt se calculant pour une seule réalisation.

Nous avons donc proposé une statistique permettant d'en résumer l'information de façon pertinente sans perdre l'avantage de sa simplicité d'interprétation. Le coefficient de Tanimoto cumulé a effectivement l'avantage de s'interpréter très simplement comme l'intersection du

« volume » des agrégats détectés (constitué par l'empilement des agrégats détectés ou non à chaque réalisation) sur la réunion du volume des agrégats simulés et du volume des agrégats détectés. Il est donc plus simple pour un lecteur de se représenter ce qu'est une performance mesurée par le coefficient de Tanimoto cumulé qu'une performance mesurée par l'aire sous la courbe de puissance étendue.

Le coefficient de Tanimoto cumulé, du fait qu'il résume l'information de performance, présente le même défaut que l'aire sous la courbe de puissance étendue, des comportements différents d'un test pouvant aboutir à la même mesure de performance. Par exemple, un test ayant une puissance usuelle de 100% mais ne donnant que des agrégats détectés dont l'ensemble des US recouvrent strictement la moitié des US de l'agrégat simulé se verra affecter la même mesure de performance qu'un test dont la puissance est de 50% mais dont les agrégats détectés recouvrent toujours parfaitement l'agrégat simulé, soit un coefficient de Tanimoto cumulé égal à 0.5.

1.2 Effets de bord et optimisation des protocoles de simulation

Une piste d'optimisation des protocoles de simulation est, en dehors de la simulation d'hypothèses nulles ou alternatives, la prise en compte de l'effet de bord inhérent à la restriction des analyses à une région d'étude.

En effet, s'il existe une autocorrélation spatiale des données, alors, les données observées dans la région d'étude, en particulier à son bord, seront en partie liées aux données non connues de l'extérieur de la région d'étude [65, 66]. L'effet de bord peut être potentiellement important dans l'analyse de données groupées où le nombre de cas dans les US (particulièrement lorsque les US sont petites) peut dépendre du nombre de cas des US voisines.

Plusieurs méthodes ont été développées pour prendre en compte cet effet de bord lors de l'analyse de données réelles, mais peu d'études de performance le prennent en compte dans leurs protocoles de simulation, à moins que son évaluation soit un objectif spécifique de l'étude [67]. En effet, dans les études de performance, la simulation des données est effectuée comme si la région d'étude était totalement isolée et indépendante spatialement des régions voisines non simulées. En particulier, lorsque ces études s'intéressent à la simulation d'agrégats, ceux-ci sont tous intégralement situés à l'intérieur de la région d'étude.

De ce fait, dans ces études, le seul effet de bord qui peut être pris en compte est celui lié aux méthodes d'analyse elles-mêmes. Si cet effet de bord peut être considérablement atténué dans la simulation d'hypothèses alternatives, en particulier lorsque l'agrégat est simulé avec une forte intensité, ce n'est pas le cas lors de la simulation d'hypothèses nulles. Ainsi, si l'on peut supposer dans ce cas que le taux global de l'erreur de type I est effectivement égal à la valeur α prédéfinie, il est fort probable que la distribution spatiale des agrégats détectés à tort soit soumise à un effet de bord.

Notre troisième travail [68] (*cf.* section 4 de cette partie) s'est donc consacré à l'évaluation de la répartition spatiale de l'erreur de type I et à la mise en évidence de l'effet de bord.

Pour réaliser cette évaluation, deux difficultés ont dû être surmontées : premièrement, le développement d'un indicateur de participation à l'erreur de type I pour les US, et deuxièmement, le développement d'une méthode d'inférence afin de donner une portée statistique à l'existence d'une répartition spatiale de cette participation de type « effet de bord ».

Les résultats ont confirmé et illustré l'existence d'un effet de bord important avec une répartition spatiale de l'erreur de type I plus importante au centre qu'au bord de la région d'étude.

2 Carte de performance utilisant l’aire sous la courbe de Puissance étendue

INTERNATIONAL JOURNAL OF HEALTH GEOGRAPHICS

METHODOLOGY Open Access

Performance map of a cluster detection test using extended power

Aline Guttmann[1,2*], Lemlih Ouchchane[1,2], Xinran Li[2], Isabelle Perthus[3], Jean Gaudart[4,5], Jacques Demongeot[6] and Jean-Yves Boire[1,2]

Abstract

Background: Conventional power studies possess limited ability to assess the performance of cluster detection tests. In particular, they cannot evaluate the accuracy of the cluster location, which is essential in such assessments. Furthermore, they usually estimate power for one or a few particular alternative hypotheses and thus cannot assess performance over an entire region. Takahashi and Tango developed the concept of extended power that indicates both the rate of null hypothesis rejection and the accuracy of the cluster location. We propose a systematic assessment method, using here extended power, to produce a map showing the performance of cluster detection tests over an entire region.

Methods: To explore the behavior of a cluster detection test on identical cluster types at any possible location, we successively applied four different spatial and epidemiological parameters. These parameters determined four cluster collections, each covering the entire study region. We simulated 1,000 datasets for each cluster and analyzed them with Kulldorff's spatial scan statistic. From the area under the extended power curve, we constructed a map for each parameter set showing the performance of the test across the entire region.

Results: Consistent with previous studies, the performance of the spatial scan statistic increased with the baseline incidence of disease, the size of the at-risk population and the strength of the cluster (i.e., the relative risk). Performance was heterogeneous, however, even for very similar clusters (i.e., similar with respect to the aforementioned factors), suggesting the influence of other factors.

Conclusions: The area under the extended power curve is a single measure of performance and, although needing further exploration, it is suitable to conduct a systematic spatial evaluation of performance. The performance map we propose enables epidemiologists to assess cluster detection tests across an entire study region.

Keywords: Cluster detection test, Performance map, Extended power, Simulation study

* Correspondence: aline.guttmann@udamail.fr
[1]Department of Biostatistics, Medical Informatics and Communication Technologies, Clermont University Hospital, Clermont-Ferrand F-63000, France
[2]ISIT, UMR CNRS UDA 6284, Auvergne University, Clermont-Ferrand F-63001, France
Full list of author information is available at the end of the article

Résumé

Contexte: Les études de puissance ont montré leurs limites dans l'évaluation des performances des tests de détection d'agrégats. En raison de la nécessité de prendre en compte à la fois la capacité du test à rejeter l'hypothèse nulle et à localiser correctement l'agrégat, la puissance usuelle ne peut refléter la véritable performance de ces tests. De plus, ces évaluations ne traitent en général qu'un nombre limité d'hypothèses alternatives ignorant donc le comportement de ces tests sur l'ensemble d'une région d'étude. Takahashi et Tango ont proposé le concept de puissance étendue qui, au-delà de la puissance usuelle, reflète également la précision de localisation de l'agrégat. Nous proposons une méthode d'évaluation systématique, fondée ici sur la puissance étendue, pour produire une carte offrant une visualisation synoptique des performances des tests de détection d'agrégats sur l'ensemble d'une région.

Méthodes: De façon à explorer le comportement d'un test de détection d'agrégats sur un même type d'agrégat pour toutes les localisations possibles, nous avons fixé quatre jeux de paramètres spatiaux et épidémiologiques, de façon à simuler quatre collections d'agrégats, chacune couvrant l'ensemble de la région d'étude. Mille jeux de données ont été simulés pour chaque agrégat et soumis au scan spatial de Kulldorff. A partir de l'aire sous la courbe de puissance étendue, nous avons produit une carte de performance pour chaque jeu de paramètres.

Résultats: Conformément aux précédentes études, la performance du scan spatial croît avec l'incidence de base de la maladie, la taille de la population à risque et la force de l'agrégat (i.e., le risque relatif). Cependant, même pour des agrégats très similaires, la performance du test est hétérogène, suggérant l'influence potentielle d'autres facteurs.

Conclusions: L'aire sous la courbe de puissance étendue est une mesure unique de performance et, bien qu'elle nécessite des évaluations plus poussées, elle convient à l'évaluation spatiale systématique de la performance. La carte de performance que nous proposons autorise les épidémiologistes à évaluer les tests de détection d'agrégats sur l'ensemble d'une région d'étude.

Background

Spatial clusters can be detected using a wide range of statistical tests [1,2], many of which are available in free software packages such as R [3,4]. Epidemiologists use local methods to detect clusters without a priori knowledge of their location, and to determine their significance. Because these cluster detection tests (CDTs) must reveal both the presence and location of clusters, performance studies have been constrained by the limitations of conventional estimation techniques. For example, a CDT may have maximum power for rejecting the null hypothesis (cluster absence), yet be incapable of accurately locating the simulated cluster. CDT performance is also a function of epidemiological and geographical context [1,5-11]. Furthermore, because epidemiological (e.g., incidence and relative risk) and geographical (e.g., spatial unit size and shape) factors tend to be intrinsically linked, their proper or common effects are difficult to evaluate. When evaluating the behavior of these CDTs in a particular region, limited knowledge can consequently be gleaned by simulating one or a few clusters in that region, and even less knowledge can be accrued from studies on other region.

Takahashi and Tango have proposed the concept of extended power (EP) [12,13] as a more accurate measure of CDT performance. This measure assesses both the probability that the null hypothesis is rejected and the accuracy of the cluster location. As such, it overcomes the inadequacy of conventional power measures. However, EP cannot eliminate the need to define what is meant by "an accurate" or "sufficiently accurate" location. The level of spatial accuracy depends upon context; for instance, an epidemiologist will require higher spatial accuracy for an ad hoc study than for a survey system. Takahashi and Tango therefore introduced a quantitative indicator of spatial accuracy, and summarized CDT performance using an EP curve in conjunction with this spatial accuracy indicator.

In this work, we propose a method that integrates the area under the EP curve (AUC_{EP}) in order to produce maps that provide a global overview of CDT performance over an entire study region.

Methods

Clustering model

To explore CDT behavior on same-class clusters in all possible locations, we set common spatial and epidemiological characteristics for four cluster collections covering the entire study region. The study region was the Auvergne region (France), divided into $n = 221$ spatial

units (SUs) equivalent to U.S. ZIP codes. The exhaustive collection of approximately circular clusters with four SUs was identified within the study region. To achieve this outcome, the 221 SUs were successively associated with their three nearest neighbors as defined by Euclidian distances between the SU centroids. To obtain four cluster collections, we applied four combinations of two baseline risks (incidences) and two relative risks to the same at-risk population, whose size was estimated by mean annual number of live births.

For a realistic analysis, we used data archived in CEMC (birth defects registry for the Auvergne region) and INSEE (National Institute of Statistics and Economic Studies) databases. We collected two categories of data from 1999 to 2006: all birth defects and cardiovascular birth defects. Both datasets were sorted by SU. The number of live births was approximated by the number of birth declarations in the at-risk population. Global annual incidences of all birth defects (I_{all}) and cardiovascular birth defects (I_{CV}) were estimated as 2.26% and 0.48% of births, respectively. In the analysis, we constructed risk combinations of these two incidences at relative risks of 3 and 6.

Datasets

For each cluster within the four categories (221 × 4), we generated 1,000 datasets, i.e., a total of 884,000 datasets. Each dataset consisted of 221 rows and 5 columns. The rows contained SU coordinates (longitude and latitude), observed number of cases, size of the at-risk population (i.e., the number of live births) and expected number of cases in the specified SU. This last quantity was the product of the global incidence (I_{all} or I_{CV}) and the at-risk population size in the SU. The observed case numbers were assumed as independent Poisson variables such that

$$\begin{cases} H_0 : E(N_i) = \varepsilon_i, N_i \sim Pois(\varepsilon_i), i = 1, ..., n \\ H_1 : E(N_i) = \pi_i, N_i \sim Pois(\pi_i), \pi_i = \mathbb{I}\theta\varepsilon_i + \varepsilon_i(1-\mathbb{I}), i = 1, ..., n \end{cases}$$

where N_i is the observed number of cases, ε_i denote the expected number of cases in the ith SU under the null hypothesis of risk homogeneity (H_0) and π_i the expected number of cases in the ith SU under the alternative hypothesis of one simulated cluster (H_1). θ is the relative risk, and $\mathbb{I}$ is a binary indicator set to 1 if the ith SU is within the simulated cluster, and 0 otherwise.

Measure of performance

The extended power was proposed by Takahashi and Tango as an improved measure of CDT performance. For a particular cluster, global performance is the weighted cumulative sum of the contribution of each detected cluster in all submitted datasets. Here, we summarize the construction of the performance indicator. For a more detailed description, the reader is referred to Takahashi and Tango [12,13].

Within a simulated cluster of s SUs, if the null hypothesis is rejected, the size l of a detected cluster and its s^* SUs (where s^* denotes a subset of s) are recorded. A maximum cluster size L is imposed, such that if $l > L$, the detected cluster is discarded. This limit prevents very large, meaningless clusters from contributing to CDT global performance. In this work, L was set to 30 SUs.

All eligible detected clusters (EDCs), i.e. with $l \leq L$, are counted and sorted by l and s^*. For each combined value of l and s^*, the proportion of corresponding detected clusters ($P_{(l,s^*)}$) in all submitted datasets is assigned a weight $W_{(l,s^*)}$. This weight is also a function of the detection accuracy (i.e., the correct location of the simulated cluster). Thus, Takahashi and Tango define $W_{(l,s^*,w+,w-)}$ as

$$W_{(l,s^*,w^+,w^-)} = \sqrt{[1-min\{w^-(s-s^*),1\}][1-min\{w^+(l-s^*),1\}]}$$

where w^- and w^+ are penalties for false negative and false positive SUs, respectively. The penalties w^- and w^+ are determined according to the following constraints. For w^-, detected clusters that generate no false negative must fully contribute to global performance, and those that induce s false negatives must be discarded. These constraints are satisfied when

$$w^- = 1/s$$

For w^+, detected clusters that generate no false positive must fully contribute to global performance, and those that induce at least l_0 false positives must be discarded. These constraints are satisfied when

$$w^+ = 1/l_0$$

So that l_0 is not assigned arbitrarily, Takahashi and Tango specify the ratio

$$q = w^+/w^-$$

To favor sensitivity over specificity (as is usually preferred), w^- is greater than or equal to w^+; thus $l_0 \geq s$ because $1/s \geq 1/l_0$. For example, when:

- $l_0 = s, w^- = w^+$ and $q = 1$;
- $l_0 = 2s, w^- = 2w^+$ and $q = 0.5$;
- $l_0 \to \infty, w^+ = 0$ and $q = 0$.

For each value of q, the extended power is the cumulative sum of $W_{(l,s^*,q)} \times P_{(l,s^*)}$, where l runs from 1 to L and s^* runs from 0 to s. CDT global performance in detecting a particular cluster is then represented by the extended power curve with q running from 0 to 1. At any point

on this curve, the extended power is, by construction, between 0 and 1. Furthermore, we note that the extended power is a monotonically decreasing function of q. Consequently, the area under the extended power curve (AUC_{EP}), defined by

$$AUC_{EP} = \int_{q=0}^{1} \left(W_{(l,s^*,q)} \times P_{(l,s^*)}\right) dq$$

is between 0 and 1, with 0 signifying an inoperative CDT (s^* always null) and 1 a perfect CDT (H_0 always rejected, with all detected clusters exactly overlaying the simulated cluster). As suggested by Takahashi and Tango [13], we used the area under the extended power curve as the measure of CDT performance.

Performance mapping

Global performance was visualized over the entire region using maps representing the measured AUC_{EP} for each collection of clusters.

The AUC_{EP} is a measure of a cluster and thus associated with four SUs. In order to obtain a global overview on a single map, we assigned the AUC_{EP} value of each cluster, to its central SU. Thus, we affected a single measure of AUC_{EP} to each SU of the map. As we defined four cluster collections for four risks combination (incidence and relative risks), we produced four performance maps.

Kulldorff's Spatial scan statistic

In this study, we selected Kulldorff's spatial scan statistic [14,15], a well-known and widely used CDT whose performance has been studied by many authors [1,6,10,16]. The spatial scan statistic detects the most likely cluster based on locally observed statistics of likelihood ratio tests. The scan statistic considers all possible zones z defined by two parameters: a center that is successively placed on the centroid of each SU, and a radius varying between 0 and a predefined maximum. The true geography being delineated by administrative tracts, i.e., each zone z defined by all SUs whose centroids lie within the circle, is irregularly shaped. Let N_z and n_z be the size of the at-risk population and the number of cases counted in zone z (over the entire region, these quantities are the total population size N and the total number of cases n, respectively). The probabilities that an at-risk case lies inside or outside zone z are respectively defined by $p_z = n_z/N_z$ and $q_z = (n - n_z)/(N - N_z)$. Given the null hypothesis H_0: $p_z = q_z$ versus the alternative H_1: $p_z > q_z$ and assuming a Poisson distribution of cases, Kulldorff defined the likelihood ratio statistics as proportional to

$$\left(\frac{n_z}{\lambda N_z}\right)^{n_z} \left(\frac{n-n_z}{\lambda(N-N_z)}\right)^{n-n_z} I[n_z > \lambda N_z]$$

where λ is global incidence, and the indicator function I equals 1 when the number of observed cases in zone z exceeds the expected number under H_0, and 0 otherwise. The circle yielding the highest likelihood ratio is identified as the most likely cluster. The p-value is obtained by Monte Carlo inference.

Software

Data simulation and analysis (see Data and Script in the Additional files 1 and 2) were performed in R 2.14.0 [3,17-19] using AUVERGRID [20].

Results

The Auvergne region is characterized by low and medium mountains situated around a central plain. The at-risk population (see Methods) was heterogeneously distributed throughout sparsely populated areas (mainly borderland and mountainous) and highly populated urban areas. Figure 1 shows the size of the at-risk population in each cluster, which was assigned to its central SU.

Figure 2 demonstrates how CDT performance improved with increasing risk level. Clearly, the CDT could not detect clusters within regions with low number of births. For these clusters, performance only marginally improved, even at the highest risk combination (Figure 3).

CDT performance increased monotonically with the at-risk population size (Figure 3). We noted a stronger heterogeneity of CDT performance for the clusters with the largest populations, especially at intermediate risk levels (Figure 3); by this, we mean that clusters with nearly the same population size led to slightly different test performance behaviors. For example, Figure 4 shows test performance in detecting three clusters centered on SUs "43770" (red cluster in the figure), "03700" (blue cluster) and "03420" (green cluster), which had population sizes of 544, 558 and 545 births (mean number over 8 years), respectively. At the lowest risk level, the red cluster was the only one even marginally detected, whereas under other configurations, the blue cluster was best detected. The worst detection performance was exhibited with respect to the green cluster, particularly at intermediate risk levels. We note that the green cluster was the only borderland cluster.

Some summary statistics of the AUC_{EP} distributions are displayed in Table 1. Figure 5 shows two different extended power curves (and thus two different CDT behaviors) that have nearly equal AUC_{EP}. One of these clusters was centered on SU "03160", the other on SU "63112".

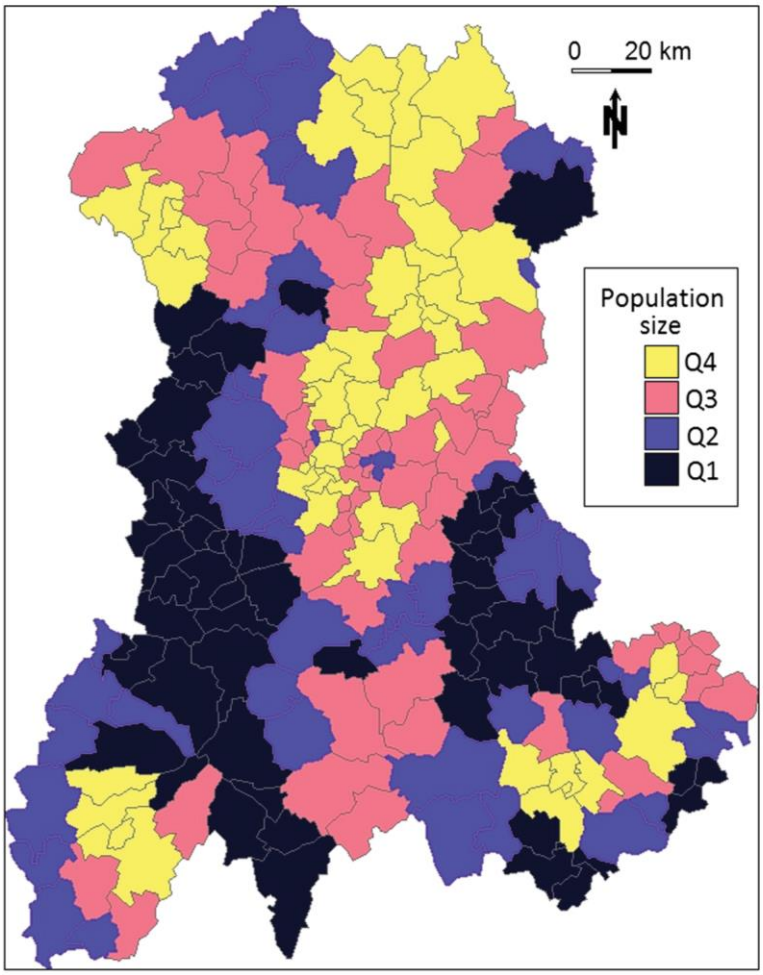

Figure 1 Size of the at-risk population for each cluster in the Auvergne region, as defined by mean number of live births per year between 1999 and 2006 (source: INSEE). Q1: ≤ 102; Q2: > 102 and ≤ 175; Q3: > 175 and ≤ 293; Q4: >293.

Generation of one performance map from 221,000 datasets required about 5 days of computational time using the AUVERGRID grid.

Discussion

Takahashi and Tango [13] have suggested using the AUC_{EP} to compare performance between CDTs. We used this synthetic indicator, suitable for compiling maps, to describe CDT performance. It thus fulfills our primary goal of realizing a systematic performance assessment of a CDT over an entire study area, rather than over only a few clusters. This mapping method, although using Takahashi and Tango's extended power, is not dependent on this concept. Our method can use any other indicator that meets the requirements of being a scalar (i.e., a single measure of performance) indicating both the spatial accuracy of the detection and the capacity of cluster detection tests to reject the null hypothesis.

Interpretation of the AUC_{EP} requires further exploration, however. Although a higher AUC_{EP} clearly signifies stronger CDT performance, quite different behaviors can yield the same AUC_{EP}. As shown in Figure 5, different curves can possess very similar AUC_{EP} values. This figure shows the extended power curves "03160" and "63112", whose AUC_{EP} values are nearly equal (0.931 and 0.932, respectively), but which reflect different CDT behaviors. The procedures used to construct these curves are described in detail within separate spreadsheets (see EP curve in the Additional file 3).

The curve "63112" is nearly horizontal, indicating that the EDCs (H_0 rejected, and cluster size $l <$ maximum cluster size L) located the simulated cluster with high accuracy. As q increases, less tolerance is given to false positives until, eventually, only EDCs with at least one true positive and less than s false positives can contribute to the extended power. A near zero slope thus indicates that the same detected clusters, all of which

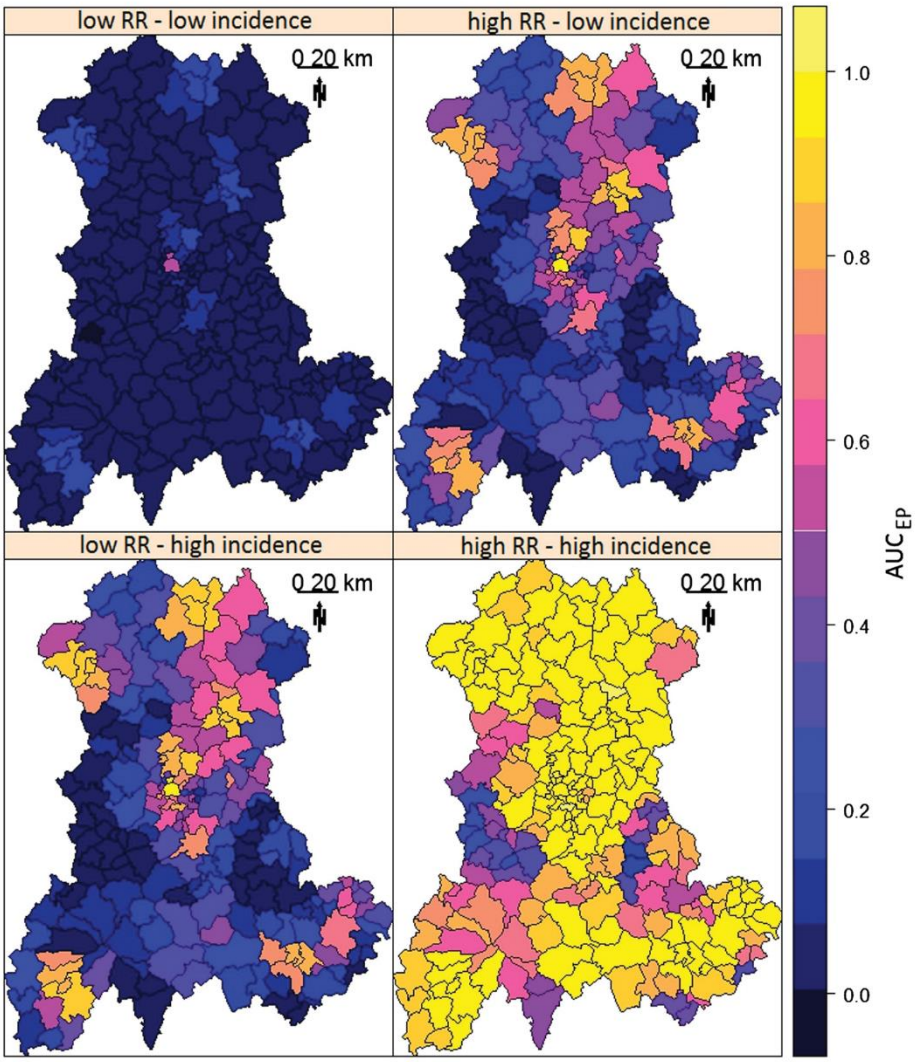

Figure 2 AUC_{EP} of Kulldorff's spatial scan. AUC_{EP} was measured for four combinations of two relative risk (RR) and two annual incidence of birth defects: low RR = 3 and high RR = 6; low incidence = 0.48% births and high incidence = 2.26% births.

contain less than s false positives, contribute to the extended power, regardless of q.

The intercept of curve "63112" is 0.939, meaning that eligible clusters ($l < L$), all of which contribute to the extended power (i.e., all clusters contain at least one true positive), were detected in 93.9% of the tests (H_0 rejected).

To summarize curve "63112", the simulated cluster was not always detected (no H_0 rejection or EDC without true positive); however, provided that an EDC identified at least one true positive, the location was accurate (i.e., less than s false positives existed in the cluster).

In contrast, the curve "03160" yields the same AUC_{EP}, but is negatively sloped with an intercept of 0.951. Thus, the associated CDT produced more EDCs containing at least one true positive. The negative slope indicates that a higher proportion of these EDCs generated at least s false positives.

To summarize curve "03160", the test rejected H_0 more often and/or produced more EDCs, but located the simulated cluster with less accuracy (i.e., this analysis produced more than s false positives).

One particular curve has intercept equal to 1 ($q = 0$) and a zero slope. An intercept equal to 1 implies that the CDT always rejects H_0 and that no false negatives exist in the EDCs. All detected clusters entirely overlap the simulated cluster, as in all other cases the weighting function $W_{(l,\ s^*,\ q=0)}$ is less than one. In addition, the zero slope indicates the perfect test that always exactly locates the simulated cluster. A perfect test always rejects H_0, and detected clusters always satisfy $l = s^* = s$ (i.e., generate no false positive or negative). The AUC_{EP} of a perfect

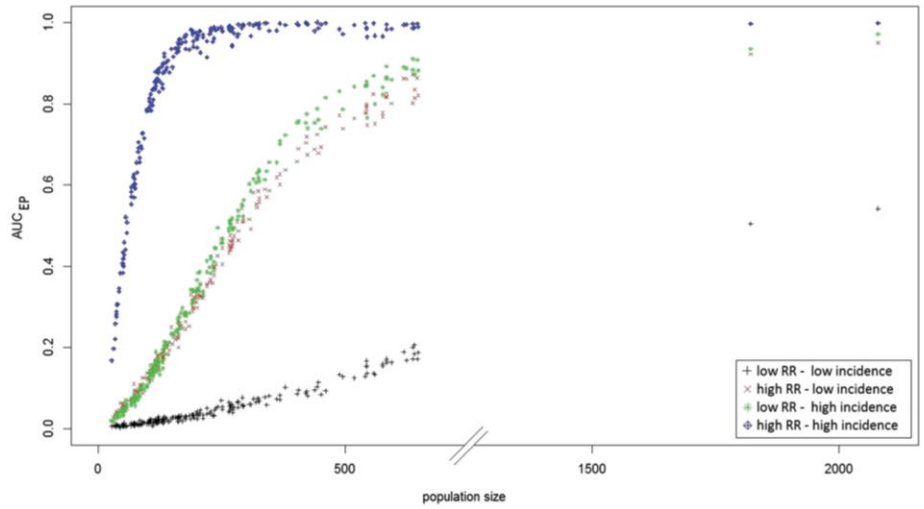

Figure 3 AUC_{EP} **of Kulldorff's spatial scan based on the size of the at-risk population for four combinations of two relative risk (RR) and two annual incidence of birth defects: low RR = 3 and high RR = 6; low incidence = 0.48% births and high incidence = 2.26% births.**

test equals one, because in all other cases $W_{(l, s^*, q)}$ is less than one.

The intercept of an extended power curve can be regarded as a "quantitative" feature of CDT performance (all EDCs generating true positives contribute to the extended power), whereas the slope may be thought of as a "qualitative" feature of CDT performance, assessing location accuracy. The parameter q can, in fact, be regarded as a continuous indicator reflecting to what extent a detected cluster must accurately locate the simulated cluster to contribute to the performance measure.

As shown in Figure 5, however, if an entire curve is condensed into a single measure (such as the AUC), some information is lost, because CDTs with different behaviors (i.e., curves with different shapes) can yield the same performance value.

Consequently, the impact of CDT behavior on the extended power curve must be thoroughly explored, and behaviors relevant to a particular research or application need to be defined. Through such exploration, the extent to which the AUC_{EP} is a relevant performance measure, and the purposes for which it is most suited, can be determined.

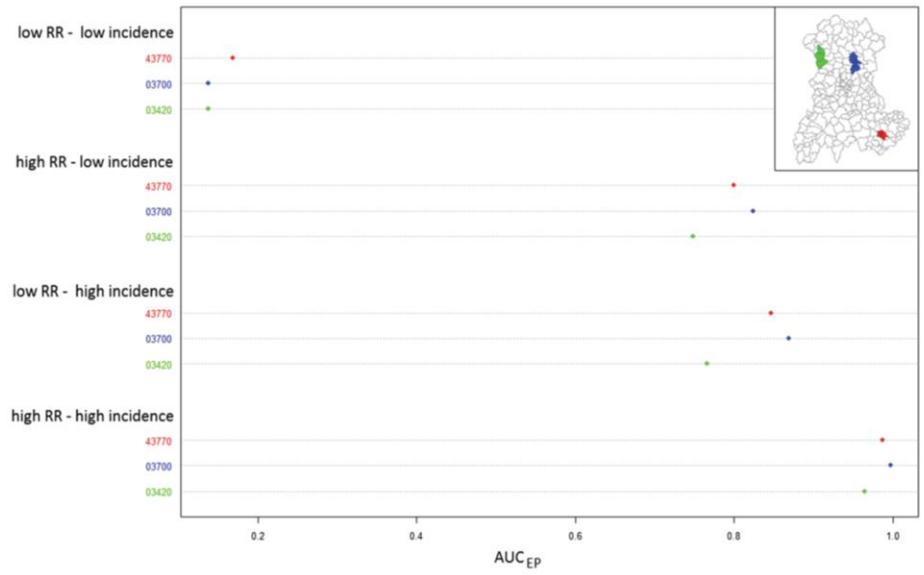

Figure 4 AUC_{EP} **of Kulldorff's spatial scan and locations of three simulated clusters for four combinations of two relative risk (RR) and two annual incidence of birth defects: low RR = 3 and high RR = 6; low incidence = 0.48% births and high incidence = 2.26% births.**

Table 1 AUC_{EP} distribution for each risk combination and category of at-risk population size

Risk combination	Number of births[a]	AUC_{EP}	
		Mean (SD)	*Min - Max*
l_{CV} and RR = 3	≤ 102	0.010 (0.003)	0.003 - 0.020
	[102, 175]	0.021 (0.006)	0.007 - 0.033
	[175, 293]	0.043 (0.013)	0.023 - 0.077
	> 293	0.133 (0.089)	0.055 - 0.542
l_{all} and RR = 3	≤ 102	0.070 (0.028)	0.019 - 0.138
	[102, 175]	0.183 (0.038)	0.119 - 0.268
	[175, 293]	0.382 (0.075)	0.246 - 0.543
	> 293	0.713 (0.117)	0.492 - 0.950
l_{CV} and RR = 6	≤ 102	0.061 (0.025)	0.016 - 0.110
	[102, 175]	0.185 (0.047)	0.114 - 0.297
	[175, 293]	0.412 (0.083)	0.277 - 0.553
	> 293	0.768 (0.113)	0.524 - 0.971
l_{all} and RR = 6	≤ 102	0.511 (0.162)	0.168 - 0.787
	[102, 175]	0.874 (0.050)	0.783 - 0.959
	[175, 293]	0.970 (0.019)	0.915 - 0.995
	> 293	0.990 (0.010)	0.964 - 1

[a]mean number between 1999 and 2006.

The EP has the advantage of requiring only one arbitrarily set parameter. In this work, the parameter L, that determines the maximum allowed size for EDCs, has been set to 30 SUs. Takahashi and Tango [12] initially proposed to set the limit L to one fourth or one third of region size (in numbers of SUs). The authors stated that it was not unreasonable to assume that an actual cluster size will be less than such a limit. Such arguments are often open to dispute but in any case, it is an arbitrary decision. In our view, it would be more correct to set L according to the size s of the simulated cluster because, in the simulation, it is the "real" cluster. By construction, the consequences of this arbitrary setting are limited to the lowest values of q. Indeed, low values of q mean that EDCs with false positives are less penalized, and thus large clusters are allowed to contribute to EP. In our case ($L = 30$), only values of extended power for $q \leq 0.15$ could be underestimated, and only if we consider that detected clusters more than 7.5 times larger than the simulated cluster (4 SUs) are still meaningful. At last, compared with L set to 30, computing AUC_{EP} with L equal to 221 (i.e. without an arbitrary limit) yields a difference in AUC_{EP} always less than 10^{-5} in this work.

In producing our performance map, we chose to assign the AUC_{EP} value of a single cluster of four SUs to a single SU. Because two clusters centered on neighboring

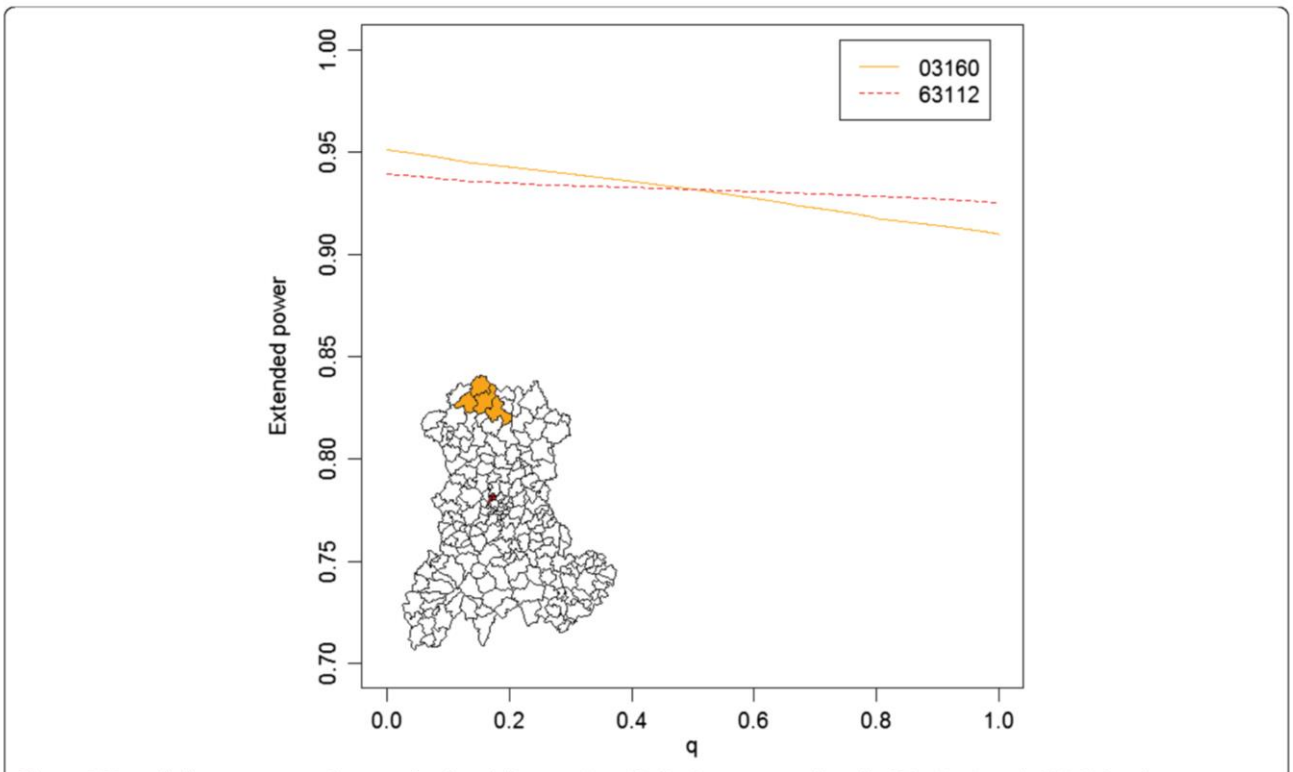

Figure 5 Extended power curves for two simulated clusters. Line 03160: cluster centered on the SU with zip code 03160 (northwest Auvergne); line 63112: cluster centered on the SU with zip code 63112 (central Auvergne). Both clusters were simulated with a relative risk of 6 and a baseline incidence of birth defects set to 2.26%.

SUs likely contain common SUs, and the AUC_{EP} evaluates the detection of the entire cluster, visualizing performance on a single map can only be done in two ways. On the one hand, the AUC_{EP} of a cluster can be assigned to each of its SUs, or on the other hand, it can be assigned to a single, albeit arbitrarily chosen, SU. In the first solution, as each SU has a strong probability to be associated with more than one cluster, it is then necessary to compute a summary statistic, such as the mean, to produce a single map. In our view, it seems more comprehensible to arbitrary assign the performance measure for the whole cluster on a single SU. As we simulated more or less circular clusters, the central SU of the cluster was naturally chosen for this assignment. When simulating different cluster shapes, this choice will clearly be less obvious. We nevertheless recommend assigning the performance measure to the SU where the centroid of the cluster is located.

Authors who have studied CDT behavior mentioned its dependence on epidemiological and geographical factors [1,5-11]. Consistent with previously published results, the performance of Kulldorff's spatial scan, and more generally, all local CDTs, improves in study regions of small SUs, large populations, high incidence of the studied phenomenon and for clusters with strong relative risk. Furthermore, as shown in Figure 4 and Table 1, the variation in AUC_{EP} among very similar simulated clusters (identical length, shape, population size and risk association) suggests that other factors influence CDT performance. To our knowledge, no other simulation study has been performed to both assess and visualize CDT performance over an entire region. Until now, authors have always considered a limited set of simulated clusters with particular epidemiological or geographical characteristics of interest. Consider the typical example of population size effect. To assess this effect, clusters are generally simulated in only a few arbitrarily chosen locations where a CDT behavior is assumed to be representative of its behavior in any other "similar" location. Usually, clusters in rural areas are compared with clusters in urban areas. Such studies are not sufficient to assess this factor that, as we have shown (Figure 3), has a strong relationship with CDT performance. Furthermore, population size cannot explain in itself all the variability in CDT performance.

However, some authors [21] have assessed performance on many randomly located clusters, which is a way to take into account the effect of spatial location without assessing it. It enabled them to assess the effect of factors such as relative risk or spatial resolution without the potential confounding effect of the spatial location. Still, this approach, while accounting for this effect, cannot quantify it.

Our systematic evaluation allows us to assess exactly when heterogeneity is most important, and thus within what population size range we can expect any other potential factor to have a maximum effect. In this work, we used predefined values for incidence and clustering characteristics (relative risk, shape, size and number) to generate performance maps. Epidemiologists should use reasonable values if a priori knowledge is available for some factors. However, the proper effect of any factor on CDT performance can be studied with this systematic evaluation, provided it uses suitable measure such as the AUC_{EP}.

Conclusion

Given that CDT performance depends on geographical and epidemiological context, the performance of these methods should be explored prior to monitoring a particular phenomenon in a given region. This work enables epidemiologists to study global CDT performance over an entire region. Furthermore, from a research viewpoint, our method seems beneficial for unraveling the proper effect of many factors, particularly geographical ones, on CDT performance.

Additional files

Additional file 1: Script: This file is an r script (script.r) containing a complete procedure to define the collection of clusters, simulate the datasets, perform the test and plot the corresponding performance map.

Additional file 2: Data: This is a zip file (Data.zip) containing the population data in an r format (Pop.rda) and a folder with the shapefiles for the Auvergne region.

Additional file 3: EP curve: This file is an Excel spreadsheet (EP curve.xls) containing two worksheets. Sheets "03160" and "63112" describe step-by-step construction of EP curves for clusters centered on SU "03160" and SU "63112", respectively. In both constructions, the relative risk is set to 6 and the baseline incidence of birth defects is assumed to be 2.26%. To toggle between the corresponding procedures for calculating EP, the user need only alter the value of q in cell D41.

Abbreviations

AUC_{EP}: Area under the curve of extended power; CDT: Cluster detection test; EDC: Eligible detected cluster; EP: Extended power; H_0: Null hypothesis; H_1: Alternative hypothesis; I_{all}: Incidence of all birth defects; I_{cv}: Incidence of cardiovascular birth defects; RR: Relative risk.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AG and LO conceived the design, performed the study and drafted the manuscript. AG was responsible for statistical programming and data analysis. JD, JG, IP, XL and JYB contributed to manuscript revision. All authors read and approved the final manuscript.

Acknowledgments

The authors are very grateful to Dr. Francannet who granted access to the CEMC database. We thank Paul De Vlieger who provided access and technical support for AuverGrid on behalf of the particle physics laboratory, Blaise Pascal University.

Author details
[1]Department of Biostatistics, Medical Informatics and Communication Technologies, Clermont University Hospital, Clermont-Ferrand F-63000, France. [2]ISIT, UMR CNRS UDA 6284, Auvergne University, Clermont-Ferrand F-63001, France. [3]PEPRADE, EA 4681, Clermont-Ferrand F-63000, France. [4]SESSTIM, UMR 912 INSERM IRD AMU, Aix-Marseille University, Marseille F-13005, France. [5]Biostatistics Unit, Assistance Publique Hôpitaux de Marseille, Marseille F-13005, France. [6]AGIM, FRE CNRS 3405, J. Fourier University, La Tronche University School of Medicine, Grenoble F-38700, France.

Received: 31 July 2013 Accepted: 15 October 2013
Published: 25 October 2013

References

1. Kulldorff M, Tango T, Park PJ: **Power comparisons for disease clustering tests.** *Comput Stat Data Anal* 2003, **42**:665–684.
2. Sankoh OA, Becher H: **Disease cluster methods in epidemiology and application to data on childhood mortality in rural Burkina Faso.** *Inform Biom Epidemiol Med Biol* 2002, **33**:460–472.
3. Gomez-Rubio V, Ferrándiz J, López A: **Detecting clusters of diseases with R.** *Proc DSC* 2003:2.
4. Robertson C, Nelson TA: **Review of software for space-time disease surveillance.** *Int J Heal Geogr* 2010, **9**:16.
5. Aamodt G, Samuelsen SO, Skrondal A: **A simulation study of three methods for detecting disease clusters.** *Int J Heal Geogr* 2006, **5**:15.
6. Ozonoff A, Jeffery C, Manjourides J, White LF, Pagano M: **Effect of spatial resolution on cluster detection: a simulation study.** *Int J Heal Geogr* 2007, **6**:52.
7. Jeffery C, Ozonoff A, White LF, Nuño M, Pagano M: **Power to detect spatial disturbances under different levels of geographic aggregation.** *J Am Med Informatics Assoc JAMIA* 2009, **16**:847–854.
8. Olson KL, Grannis SJ, Mandl KD: **Privacy protection versus cluster detection in spatial epidemiology.** *Am J Public Heal* 2006, **96**:2002–2008.
9. Puett R, Lawson A, Clark A, Aldrich T, Porter D, Feigley C, Hebert J: **Scale and shape issues in focused cluster power for count data.** *Int J Heal Geogr* 2005, **4**:8.
10. Goujon-Bellec S, Demoury C, Guyot-Goubin A, Hémon D, Clavel J: **Detection of clusters of a rare disease over a large territory: performance of cluster detection methods.** *Int J Heal Geogr* 2011, **10**:53.
11. Jacquez GM: **Cluster morphology analysis.** *Spat Spatio-Temporal Epidemiol* 2009, **1**:19–29.
12. Tango T, Takahashi K: **A flexibly shaped spatial scan statistic for detecting clusters.** *Int J Heal Geogr* 2005, **4**:11.
13. Takahashi K, Tango T: **An extended power of cluster detection tests.** *Stat Med* 2006, **25**:841–852.
14. Kulldorff M: **A spatial scan statistic.** *Commun Stat Theor M* 1997, **26**:1481–1496.
15. Kulldorff M, Nagarwalla N: **Spatial disease clusters: detection and inference.** *Stat Med* 1995, **14**:799–810.
16. Ribeiro SHR, Costa MA: **Optimal selection of the spatial scan parameters for cluster detection: a simulation study.** *Spat Spatio-Temporal Epidemiol* 2012, **3**:107–120.
17. Cici C, Kim AY, Ross M, Wakefield J, Venkatraman ES: *SpatialEpi: Performs various spatial epidemiological analyses. R package version 1.1*; 2013. http://CRAN.R-project.org/package=SpatialEpi.
18. Team RC: *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2012. http://www.R-project.org/.
19. Keitt TH, Bivand Pebesma E, Rowlingson B: *Rgdal: Bindings for the Geospatial Data Abstraction Library*. 2012. http://CRAN.R-project.org/package=rgdal.
20. *AuverGrid*. http://www.auvergrid.fr/.
21. Jones SG, Kulldorff M: **Influence of spatial resolution on space-time disease cluster detection.** *PLoS One* 2012:7.

doi:10.1186/1476-072X-12-47
Cite this article as: Guttmann *et al.*: Performance map of a cluster detection test using extended power. *International Journal of Health Geographics* 2013 **12**:47.

3 Étude de la valeur informationnelle du coefficient de Tanimoto dans les études de simulation

(Article soumis à *PLOS ONE*)

Résumé

L'utilisation de tests locaux de détection d'agrégats est fréquente dans l'analyse de l'agrégation spatiale de maladies. De nouvelles méthodes sont régulièrement proposées et leurs performances se doivent d'être évaluées. Comme la capacité du test à localiser correctement l'agrégat réel doit être prise en compte, l'évaluation des performances dépasse la question de l'estimation des erreurs de type I ou II. Cependant, aucun consensus n'existe quant à la méthodologie d'évaluation de ces performances et les méthodes employées, très hétérogènes, rendent les études de performance difficilement comparables. Un indicateur global de performance évaluant à la fois performance de localisation et puissance usuelle faciliterait l'étude du comportement des méthodes locales de détection d'agrégat et améliorerait la comparabilité des études.

Le coefficient de Tanimoto est un indicateur bien connu, capable d'évaluer la performance de localisation de l'agrégat simulé pour la réalisation d'un seul test. Dans une étude de simulation, de nombreuses réalisations sont conduites. A partir du coefficient de Tanimoto, nous proposons deux statistiques, le coefficient de Tanimoto moyen et le coefficient de Tanimoto cumulé. Nous étudions les propriétés de ces deux indicateurs et démontrons la supériorité du coefficient de Tanimoto cumulé dans l'évaluation de la performance. L'utilisation de ces indicateurs est illustrée dans une évaluation spatiale systématique de la performance visualisable par une carte de performance.

Cluster detection tests in spatial epidemiology

a global indicator for performance assessment

Aline Guttmann · Xinran Li · Fabien Feschet · Jean Gaudart · Jacques Demongeot · Jean-Yves Boire · Lemlih Ouchchane

the date of receipt and acceptance should be inserted later

Abstract In disease cluster detection, the use of local cluster detection tests (CDTs) is common. These methods aim to both locate likely clusters and test for their statistical significance. New or improved CDTs are regularly proposed to epidemiologists and must be subjected to performance assessment. Because location accuracy has to be taken into account, performance evaluation goes beyond the estimation of type I or II errors. As no consensus exists for performance evaluations, very heterogeneous methods are used, and studies are thus rarely comparable. A global indicator of performance, assessing both spatial accuracy and usual power, would facilitate exploration of CDTs behavior and improve comparability between studies. The well known Tanimoto coefficient (TC) is an indicator of location accuracy but can only assess performance for one detected cluster. In a simulation study, performance is measured for many tests. From the TC, we propose two statistics, the averaged TC and the cumulated TC, as indicators able to provide a global overview of CDTs performance for both usual Power and location accuracy. We demonstrate the properties of these two indicators and the superiority of the cumulated TC in assessing performance. We use these indicators to make a systematic spatial assessment displayed through performance maps.

Keywords spatial statistics, disease clustering, performance, simulation study, Tanimoto coefficient

A. Guttmann
Clermont University Hospital, Department of Biostatistics,F-63000, France
Auvergne University, UMR CNRS UDA 6284 ISIT, Clermont-Ferrand, F-63001, France
Tel.: +334-73-718195
E-mail: aline.guttmann@udamil.fr

X. Li
Auvergne University, UMR CNRS UDA 6284 ISIT, Clermont-Ferrand, F-63001, France

F. Feschet
Auvergne University, EA 7282 IGCNC, Clermont-Ferrand, F-63001, France

J. Gaudart
Aix-Marseille University, UMR INSERM 912 SESSTIM, Marseille, F-13005, France
Assistance Publique Hôpitaux de Marseille, Biostatistic and Modelisation, Marseille, F-13005, France

J. Demongeot
J. Fourier University, Faculty of Medicine of Grenoble, FRE CNRS 3405 AGIM, F-38700, France

J-Y. Boire · L. Ouchchane
Clermont University Hospital, Department of Biostatistics,F-63000, France
Auvergne University, UMR CNRS UDA 6284 ISIT, Clermont-Ferrand, F-63001, France

1 Introduction

Assessing performance of local cluster detection tests (CDTs) is a complex and necessary task. For development of new statistical methods, simulation studies are obviously essential. In field investigation, they provide useful knowledge for interpretation of real data and decision making [3]. From a methodological point of view, simulation studies in spatial epidemiology still have no commonly accepted protocol. Evaluations are often incomplete in that they are realized only on a few clustering models defined by arbitrary settings that cannot reflect all the possible clustering configurations. Furthermore, performance, a critical aspect of which is the location accuracy, cannot be assessed just by usual

power because it only measures null hypothesis rejection. To address this issue, many different indicators of performance have been proposed and used.

Power and location accuracy are sometimes evaluated separately with indicators purely dedicated to assess the location accuracy. These indicators are based on the SUs classification in four types resulting from the confrontation between the detected cluster (positives or negatives SUs) and the simulated cluster (the *gold standard* leading to classification in true/false positives or negatives SUs). From this classification indicators such as sensitivity and positive predictive value (for example see [10,16,1,11,8]) are computed. Their mathematical definitions are very heterogeneous, however. Some authors evaluate all clusters whether the null hypothesis is rejected or not [8], others only the detected clusters (*i.e.* with null rejection) [10,11] and, finally, some authors also evaluate power by considering all analysis without null rejection as "no detected cluster" (*i.e.*, all SUs are false or true negatives)[1]. Other studies equally proposed concomitant evaluation using conditional power such as power to detect at least one spatial unit of the true cluster or power to detect exactly the true cluster (for example see [26,8,28]). As they are based on very restrictive definitions, these indicators only partially measure performance.

As only partial performance indicators are available, performance is usually evaluated using a more or less large set of indicators complementing each other. Depending on the set of performance indicators used, interpretations and comparisons between studies might be difficult.

If the use of multiple indicators can provide very detailed information on CDTs behavior, it also limits the number of clustering models that can be simulated. Indeed, a large number of clustering models results in a huge amount of information to treat and interpret which, in turn, makes it difficult to provide a comprehensible overview of performance. Even when clustering models are restricted (by setting some parameters, such as relative risk and baseline incidence in realistic ranges regarding the disease under study), global overview of performance is better feasible with the measure of a single indicator. Such indicator should obviously assess both power and location accuracy. What can be considered a sufficiently accurate test is quite ambiguous and depends on context, however. For example, one will need far better accuracy for a secondary investigation than for a surveillance system. Thus, location accuracy should be measured with a quantitative indicator. In [9], we proposed the area under the curve of extended Power [24]. This indicator, while accounting for both usual Power and location accuracy, can be complex to comprehend and is not fast to compute.

This work is based on the coefficient developed by Tanimoto [25] (see also [23]). The Tanimoto coefficient (TC) is an easily comprehensible, fast computed indicator extensively used in image science [12,7,5] and biochemistry [27,17]. The TC is a measure of similarity comparing two sample sets by using the ratio of the intersecting set to the union set. It is thus well suited to assess location accuracy for one detected cluster (the result of one test). In order to assess CDTs performance, we propose two statistics of the TC, each taking into account both location accuracy and usual power in simulation studies. We conduct a systematic spatial assessment that, combined with these global measures, enables the construction of performance maps.

The structure of this paper is as follows : in Section 2, we describe each procedure of this simulation study following guidelines proposed by [4] when relevant. In Section 3, we present the performance of Kulldorff's spatial scan statistic such as measured by the proposed statistics. Finally, in Section 4, we briefly compare these indicators with the area under the extended Power curve, discuss the behavior of these two statistics derived from the TC and argue the recommendation of the cumulated TC.

2 Methods

2.1 Clustering model

The study region is the Auvergne region (France), divided into n = 221 spatial units (SUs) equivalent to U.S. ZIP codes. For a realistic analysis, we used data archived in CEMC (birth defects registry for the Auvergne region) and INSEE (National Institute of Statistics and Economic Studies) databases. We collected two categories of data from 1999 to 2006: all birth defects and cardiovascular birth defects. For each SU, the number of live births (*i.e.*, the size of the at-risk population) was approximated by the number of birth declarations in the at-risk population. Global annual incidences of all birth defects (I_{all}) and cardiovascular birth defects (I_{cv}) were estimated as 2.26% and 0.48% of births, respectively.

We applied these two baseline risks (incidences) of birth defects to the same at-risk population, which size was approximated by mean annual number of live births. (The distribution of the at-risk population is shown in Figure 1.) For each baseline incidence ($I = 2.26\%$ of births or $I = 0.48\%$), we defined two cluster collections by applying two relative risks (3 and 6) to the same pattern of location and cluster size. The relative risks were

chosen in order to observe all the range of performance. Each cluster collection contains 221 clusters of four SUs (one central SU and its three nearest neighbors in euclidean distances) successively centered on each SU of the region.

2.2 Datasets

We generated 1000 datasets for each combination of baseline risk, relative risk and cluster location, i.e. a total of 884 000 datasets.

Each dataset is a table of 221 rows and 5 columns. The rows contain the coordinates (longitude and latitude) of a SU, the observed number of cases, the size of the at-risk population (i.e., the number of live births) and the expected number of cases in the specified SU assuming an inhomogeneous Poisson process for the cases distribution. The expected number of cases is the product of the global incidence ($I = 2.26\%$ or $I = 0.48\%$) and the size of the at-risk population in the SU. The observed case numbers are assumed as independent Poisson variables such that

$$\begin{cases} H_0 : N_i \sim \text{Pois}\left(\varepsilon_i\right), i = 1, ..., n \\ H_1 : N_i \sim \text{Pois}\left(\pi_i\right), \pi_i = \varepsilon_i \left[1 + \mathbb{I}\left(\theta - 1\right)\right], i = 1, ..., n \end{cases}$$

where N_i is the observed number of cases, ε_i denotes the expected number of cases in the ith SU under the null hypothesis of risk homogeneity (H_0) and π_i the expected number of cases in the ith SU under the alternative hypothesis of one simulated cluster (H_1), θ is the relative risk, and $\mathbb{I}$ is a binary indicator set to 1 if the ith SU is within the simulated cluster, and 0 otherwise.

We used the R function "rpois"[2] with the default Mersenne-Twister pseudo-random number generator developed by Matsumoto [18]. For reproducibility purpose, all datasets were archived.

2.3 Statistical programming

Statistical programming was done with R 3.0.2 64 bits using the "SpatialEpi" library [6] and the "kulldorff" function to perform the analysis.

In order to optimize computational time, we used parallel programming through the function "foreach" of package "Foreach" [21] with the parallel backend provided by the package "DoSNOW" [20]. Computation were done on a Dell T7600 (processor Intel(R) Xeon CPU ES-2620 2 GHz and 32 Go RAM).

2.4 Kulldorff's spatial scan statistic

In this study, we selected Kulldorff 's spatial scan statistic [14,13] as a well-known and widely used CDT which performance has been studied by many authors [8,22, 19,15]. The spatial scan statistic detects the most likely cluster on locally observed statistics of likelihood ratio tests. The scan statistic considers all possible zones z defined by two parameters: a center that is successively placed on the centroid of each SU, and a radius varying between 0 and a predefined maximum. The true geography being delineated by administrative tracts, each zone z, defined by all SUs which centroids lie within the circle, is irregularly shaped. Let N_z and n_z be the size of the at-risk population and the number of cases counted in zone z, respectively (over the whole region, these quantities are the total population size N and the total number of cases n). The probabilities that a case lies inside and outside zone z are defined by $p_z = \dfrac{n_z}{N_z}$ and $q_z = \dfrac{(n - n_z)}{(N - N_z)}$, respectively. Given the null hypothesis of risk homogeneity $H_0 : p_z = q_z$, versus the alternative $H_1 : p_z > q_z$ and assuming a Poisson distribution of cases, the likelihood ratio statistics are defined as proportional to $\left(\frac{n_z}{\lambda N_z}\right)^{n_z} \left(\frac{n-n_z}{\lambda(N-N_z)}\right)^{n-n_z} \mathbb{I}\left[n_z > \lambda N_z\right]$, where λ is the global incidence I (here equal to 2.26% or 0.48%) and the indicator function $\mathbb{I}$ equals 1 when the number of observed cases in zone z exceeds the expected number under H_0 of risk homogeneity, and 0 otherwise. The circle yielding the highest likelihood ratio is identified as the most likely cluster. The p-value is obtained by Monte Carlo inference.

Over the 884 000 simulated datasets, each test was performed with a maximum size of zone z set to 50% of the total at-risk population, a number of 999 Monte Carlo samples for significance measures, and alpha set to 5%.

2.5 Measure of performance

For each simulation, in order to compute the performance measures, we stored the identifiers of the SUs in the most likely cluster and the corresponding estimated *p-value*. As Monte Carlo hypothesis testing is based on simulations, there is no guaranty that *p-values* would be the same for successive analyses of the same datasets. For reproducibility purpose, the aforementioned results were thus archived along with the original datasets.

2.5.1 Tanimoto coefficient

The TC was computed for each analyzed dataset. This coefficient measures the similarity between the simulated cluster and the detected cluster. The superimposing of these two clusters leads to the definition of four types of SUs. The SUs both within the simulated and the detected cluster are true positives (TP), the SUs only within the detected cluster are false positives (FP), the SUs only within the simulated cluster are false negatives (FN) and, finally, the SUs within neither cluster are true negatives (TN). When no cluster was detected, i.e. *p-value* higher than 0.05, all 221 SUs were considered negatives and the analysis resulted in $TP = 0$, $FP = 0$, $TN = 217$, $FN = 4$.

The TC, computed for each analyzed dataset, is such that $TC = \frac{TP}{TP+FP+FN}$. For each simulated cluster, 1000 datasets were analyzed, and thus 1000 TC were computed.

We defined two statistics of TC in order to obtain two performance measures for each simulated cluster (with a total of 884 clusters).

2.5.2 Averaged Tanimoto coefficient

This first summary statistic of TC, referred to as TC_a is the arithmetic mean of all TC over the m simulated datasets. It is defined as

$$TC_a = \frac{1}{m} \times \sum_{i=1}^{m} \frac{TP_i}{TP_i + FP_i + FN_i}.$$

2.5.3 Cumulated Tanimoto coefficient

The second summary statistic, the TC_c, is the cumulated TC over the m simulated datasets, and is defined as

$$TC_c = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i + FP_i + FN_i}.$$

Both the TC_c and TC_a range between 0 and 1.

2.6 Performance mapping

Following a previous study [9], global performance is visualized over the entire region using maps representing the TC_a and TC_c for each collection of clusters.

Each of these measures correspond to a measure of a cluster and thus is associated with four SUs. In order to obtain a global overview on a single map, we assigned the performance measure for one cluster to its central SU. We thus affected a single measure of performance to each SU of the map. As we defined four cluster collections for four risks combination (incidence and relative risks), we produced four performance maps for each indicator.

3 Results

3.1 Performance maps

The results of this simulation study can be seen in Figures 2 and 3. No matter the indicator, the performance is heterogeneously distributed, in close relationship with the size of the at-risk population (Figure 4). The distributions of the TC_a and TC_c for each risks level are described in Figure 5.

3.2 Averaged Tanimoto coefficient versus cumulated Tanimoto coefficient

The TC_c is generally more severe in assessing performance than the TC_a (see Figure 6-d). For $RR = 6$ with $I = 2.26\%$, $RR = 3$ with $I = 2.26\%$ and $RR = 6$ with $I = 0.48\%$, the TC_c is lower than the TC_a in 100 %, 74.7% and 75.6 % of simulations, respectively. On the contrary, for $RR = 3$ with $I = 0.48\%$, i.e. the lowest risks level, the TC_c is higher than TC_a in 97.3% of simulations.

Figures 6-a and 6-b show TC_c and TC_a compared with the usual Power. Usual Power was always higher than both the TC_c and TC_a, as was expected. Indeed, each detected cluster (most likely cluster with significant *p-value*) always contributes for 1 in the usual Power, but it contributes for 1 in the TC_c or TC_a only if the detected cluster is exactly the same as the simulated cluster and less than 1 otherwise.

With both TC_c and TC_a, the spatial scan shows comparable performance on the two intermediate levels of risks ($RR = 3$ with $I = 2.26\%$ and $RR = 6$ with $I = 0.48\%$) and poor performance on the lowest level of risks ($RR = 3$ with $I = 0.48\%$). The TC_c shows more variability than TC_a when the spatial scan is the most efficient in terms of usual Power (see Figure 6-a and 6-b).

4 Discussion

Both performance indicators enable the construction of performance maps, providing global overview of Kulldorff's spatial scan performance.

In a previous study [9], we used the area under the curve of extended Power (AUC_{EP}), of which concept and construction are described in Appendix A of the

Supplementary Materials. Compared to this study (Figure 7), the present results are very similar, even if we note that both TC_a and TC_c are more severe with the test (see Figure 6-e and 6-f).

The TC_c is more severe than either the TC_a or the AUC_{EP} (see Figure 6-d and 6-e), except for the lowest risks level where this order relation is reversed.

In order to understand this behavior, we consider the functions $f(s)$ and $g(s)$ representing the computation at simulation s of respectively TC_a and TC_c, where $s \geq 1$ and the resulting *p-value* of simulation s is less than that of simulation $s+1$, for all s. The simulations are thus sorted according to their results, by increasing *p-value*, from $s = 1$ corresponding to the simulation resulting in the lowest *p-value* to $s = m'$ corresponding to the simulation resulting in the highest *p-value*.

Figure 8 shows two examples of curves defined by $f(s)$ and $g(s)$. Figure 8-a corresponds to the simulated cluster with the maximum value of TC_a - TC_c and figure 8-b corresponds to the one with the minimum value of TC_a - TC_c.

Considering the probability P of H_0 rejection (*i.e.*, the usual Power in our simulations) over the m' simulated datasets, the first $q = m' \times P$ values of $f(s)$ and $g(s)$ are thus corresponding to the behavior of the indicators when a cluster is detected (*p-value* < 0.05) and the last $m' - q$ values to their behavior when no cluster is detected.

At the simulation q, $f(q)$ is equal to

$$\begin{aligned} f(q) &= \frac{\sum_{s=1}^{q} \frac{TP_s}{TP_s+FN_s+FP_s}}{q} \\ &= \frac{\sum_{s=1}^{q} \frac{TP_s}{D+FP_s}}{q} \\ &= \sum_{s=1}^{q} \frac{TP_s}{qD+qFP_s}, \end{aligned}$$

where D is the number of SUs in the simulated cluster (by definition D is constant in our simulations). The value of $g(s)$ at the simulation q is equal to

$$g(q) = \frac{\sum_{s=1}^{q} TP_s}{\sum_{s=1}^{q} TP_s + FN_s + FP_s} = \frac{\sum_{s=1}^{q} TP_s}{qD + \sum_{s=1}^{q} FP_s}.$$

We first note that for every simulation in this study, $f(q)$ is strictly greater than the corresponding $g(q)$. This relationship can be explained by partitioning the q simulations in three disjoint sets : $S_0 = \{s|TP_s = 0\}$, $S_1 = \{s|FP_s = 0\}$ and $S_2 = \{s|TP_s \neq 0 \; and \; FP_s \neq 0\}$. (In the first q simulations, a cluster is always detected and thus true and false positives can never be both null.) We can then write

$$f(s) = \sum_{S_1} \frac{TP_s}{qD} + \sum_{S_2} \frac{TP_s}{qD + qFP_s} \quad (1)$$

and

$$g(s) = \frac{\sum_{S_1} TP_s + \sum_{S_2} TP_s}{qD + \sum_{S_0} FP_s + \sum_{S_2} FP_s},$$

or equivalently

$$g(s = q) = \frac{\sum_{S_1} TP_s}{qD + \sum_{S_0} FP_s + \sum_{S_2} FP_s} + \frac{\sum_{S_2} TP_s}{qD + \sum_{S_0} FP_s + \sum_{S_2} FP_s} \quad (2)$$

It is then easy to show graphical proof that the first terms of the sums in equations (1) and (2), referred to as *A1* and *C1* respectively in Figure 9, determine the order relation between $f(q)$ and $g(q)$. (The second terms of the sums in equations (1) and (2) are referred to as *A2* and *C2* respectively.) In fact, simulations where there is no *TP* do not impact $f(q)$ but decrease $g(q)$ all the more so due to the *FP*. As the mean number of *FP* is 14.32 (median 6) when there is no *TP* and 2.66 (median 0 and third quartile 1) when there is at least one *TP*, our observation ($f(q) > g(q)$) is explained.

Our second observation is that TC_c (*i.e.* $g(s = m')$) is less than TC_a (*i.e.*, $f(s = m')$), except for the lowest risks level. To explain this, let now consider any simulation s, where $s > q$. As no cluster is detected, there are neither false nor true positives and the quantities $M = \sum_{s=1}^{q} \frac{TP_s}{D+FP_s}$, $A = \sum_{s=1}^{q} TP_s$ and $B = \sum_{s=1}^{q} FP_s$ are equal to $\sum_{s=1}^{m'} \frac{TP_s}{D+FP_s}$, $\sum_{s=1}^{m'} TP_s$ and $\sum_{s=1}^{m'} FP_s$, respectively. Thus, we can write

$$f(s > q) = \frac{M}{s}$$

and

$$g(s > q) = \frac{A}{B + qD + (s-q)D} = \frac{A}{B + sD}.$$

The asymptotic behavior of the ratio of $f(s)$ to $g(s)$, is then

$$\begin{aligned} \lim_{s\to\infty} \frac{f(s)}{g(s)} &= \lim_{s\to\infty} \frac{M}{s} \times \frac{B+sD}{A} \\ &= \lim_{s\to\infty} \left\{ \frac{M}{A} \times \left(\frac{B}{s} + D \right) \right\} \\ &= \frac{MD}{A}, \end{aligned}$$

as $\frac{B}{s}$ tends to 0. As M can only be less than or equal to $\frac{A}{D}$, then $\lim_{s\to\infty} \frac{f(s)}{g(s)}$ is less than or equal to 1. When there is at least one *FP* in the first q simulations, then $\frac{A}{D}$ is strictly greater than M and $\lim_{s\to\infty} \frac{f(s)}{g(s)}$ is strictly less than 1. That is, TC_a is less impacted by simulations where no cluster are detected (*p-value* ≥ 0.05),

explaining the higher final values of TC_c compared to TC_a for the lowest risk levels where usual Power is of 11.7 % on average.

The absence of TP when a cluster is detected reflects poor performance and should negatively impact the indicators. As the contributions of these simulations are much stronger in TC_c than in TC_a, TC_c better distinguishes low accuracy in cluster location. Furthermore, even if TC_a is generally lower than TC_c when the usual Power is very low, the range of values reflects unambiguously low performance. Finally, TC_c can be directly interpreted like the original Tanimoto coefficient: it is a measure of similarity comparing two sample sets by using the ratio of the intersecting set to the union set where the two sets are the stacked results of the simulations. For these reasons, we recommend the use of TC_c for evaluating CDTs performance.

This type of study is generally undertaken for research purpose or in preparation for the deployment of a health monitoring system. In this context, long computational time can be tolerated as there is no need for repeating the study. Nevertheless, a systematic spatial assessment of a CDT performance in detecting a type of cluster (fixed shape, size and epidemiological factors) is bound to take time. The simulation and analysis of the 221 000 datasets necessary for the construction of one map required about 43 hours of computation. Most of this time was taken by the analysis of the datasets by the CDT. Once obtained the characteristics of the detected clusters, computation of the performance indicators and construction of the maps were relatively short (less than half an hour).

Many statistical methods are available to analyze spatial and temporal data. Quality of monitoring system or epidemiological research does not depend on the performance of these methods per se, but on how well their performance is known. Indeed, such knowledge is key in choosing appropriate methods and interpreting results. Every new or improved CDT is proposed along with a study of its performance. There is no consensus or even commonly used methodology for performance evaluation, however. Consequently, studies are rarely comparable and each new performance assessment must repeat evaluation of the same reference CDTs in order to dispose of interpretable results. Considerable gain could be obtained by homogenization of evaluation methods. One step is to dispose of benchmark datasets such as proposed by Kulldorff *et al.* [15]. But another essential progress would be to dispose of a global performance indicator enabling the evaluation of numerous scenario and thus the exploration of CDT behavior. We propose a global performance indicator taking into account both usual Power and location accuracy and easy to compute and interpret. Furthermore, the cumulated Tanimoto coefficient can be used as is for assessment of performance on temporal data and can be easily adapted to spatio-temporal data.

5 Supplementary Material

The reader is referred to the on-line Supplementary Materials for technical appendices. Software in the form of R code, together with simulated data sets and complete documentation is available on request from the corresponding author.

Acknowledgements Data have been provided by the CEMC (birth defect registry of Auvergne), with the participation of the Regional Health Agency of Auvergne, InVS (National institute for Health Surveillance) and INSERM (National institute of health and medical research).

Conflict of Interest: None declared.

References

1. Aamodt, G., Samuelsen, S.O., Skrondal, A.: A simulation study of three methods for detecting disease clusters. International journal of health geographics **5**, 15 (2006)
2. Ahrens, J.H., Dieter, U.: Computer generation of poisson deviates from modified normal distributions. ACM Transactions on Mathematical Software (TOMS) **8**(2), 163–179 (1982)
3. Bellec, S., Hémon, D., Clavel, J.: Answering cluster investigation requests: the value of simple simulations and statistical tools. European journal of epidemiology **20**(8), 663–671 (2005)
4. Burton, A., Altman, D.G., Royston, P., Holder, R.L.: The design of simulation studies in medical statistics. Statistics in medicine **25**(24), 4279–4292 (2006)
5. Chatzichristofis, S.A., Boutalis, Y.S.: Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In: Computer Vision Systems, pp. 312–322. Springer (2008)
6. Chen, C., Kim, A.Y., Ross, M., Wakefield, J., Venkatraman, E.S.: SpatialEpi: Performs various spatial epidemiological analyses (2013). URL http://CRAN.R-project.org/package=SpatialEpi. R package version 1.1
7. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. John Wiley & Sons (2012)
8. Goujon-Bellec, S., Demoury, C., Guyot-Goubin, A., Hémon, D., Clavel, J., et al.: Detection of clusters of a rare disease over a large territory: performance of cluster detection methods. International journal of health geographics **10**(1), 53 (2011)
9. Guttmann, A., Ouchchane, L., Li, X., Perthus, I., Gaudart, J., Demongeot, J., Boire, J.Y.: Performance map of a cluster detection test using extended power. International journal of health geographics **12**(1), 47 (2013)
10. Huang, L., Pickle, L.W., Das, B.: Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases. Statistics in medicine **27**(25), 5111–5142 (2008)

11. Jacquez, G.M.: Cluster morphology analysis. Spatial and spatio-temporal epidemiology **1**(1), 19–29 (2009)
12. Kara, L.B., Stahovich, T.F.: An image-based, trainable symbol recognizer for hand-drawn sketches. Computers & Graphics **29**(4), 501–517 (2005)
13. Kulldorff, M.: A spatial scan statistic. Communications in Statistics-Theory and methods **26**(6), 1481–1496 (1997)
14. Kulldorff, M., Nagarwalla, N.: Spatial disease clusters: detection and inference. Statistics in medicine **14**(8), 799–810 (1995)
15. Kulldorff, M., Tango, T., Park, P.J.: Power comparisons for disease clustering tests. Computational Statistics & Data Analysis **42**(4), 665–684 (2003)
16. Li, X.Z., Wang, J.F., Yang, W.Z., Li, Z.J., Lai, S.J.: A spatial scan statistic for multiple clusters. Mathematical biosciences **233**(2), 135–142 (2011)
17. Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K., Moos, W.H.: Measuring diversity: experimental design of combinatorial libraries for drug discovery. Journal of medicinal chemistry **38**(9), 1431–1436 (1995)
18. Matsumoto, M., Nishimura, T.: Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Transactions on Modeling and Computer Simulation (TOMACS) **8**(1), 3–30 (1998)
19. Ozonoff, A., Jeffery, C., Manjourides, J., White, L.F., Pagano, M.: Effect of spatial resolution on cluster detection: a simulation study. International journal of health geographics **6**(1), 52 (2007)
20. Revolution Analytics: doSNOW: Foreach parallel adaptor for the snow package (2013). URL http://CRAN.R-project.org/package=doSNOW. R package version 1.0.9
21. Revolution Analytics and Steve Weston: foreach: Foreach looping construct for R (2013). URL http://CRAN.R-project.org/package=foreach. R package version 1.4.1
22. Ribeiro, S.H.R., Costa, M.A.: Optimal selection of the spatial scan parameters for cluster detection: a simulation study. Spatial and spatio-temporal epidemiology **3**(2), 107–120 (2012)
23. Rogers, D.J., Tanimoto, T.T.: A computer program for classifying plants. Science **132**(3434), 1115–1118 (1960). DOI 10.1126/science.132.3434.1115. URL http://www.sciencemag.org/content/132/3434/1115.short
24. Takahashi, K., Tango, T.: An extended power of cluster detection tests. Statistics in Medicine **25**(5), 841–852 (2006). DOI 10.1002/sim.2419
25. Tanimoto, T.: IBM internal report, nov. 17, 1957 (1957)
26. Waller, L.A., Hill, E.G., Rudd, R.A.: The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. Statistics in Medicine **25**(5), 853–865 (2006)
27. Willett, P.: Similarity-based approaches to virtual screening. Biochemical Society Transactions **31**(Pt 3), 603–606 (2003)
28. Zhang, T., Zhang, Z., Lin, G.: Spatial scan statistics with overdispersion. Statistics in medicine **31**(8), 762–774 (2012)

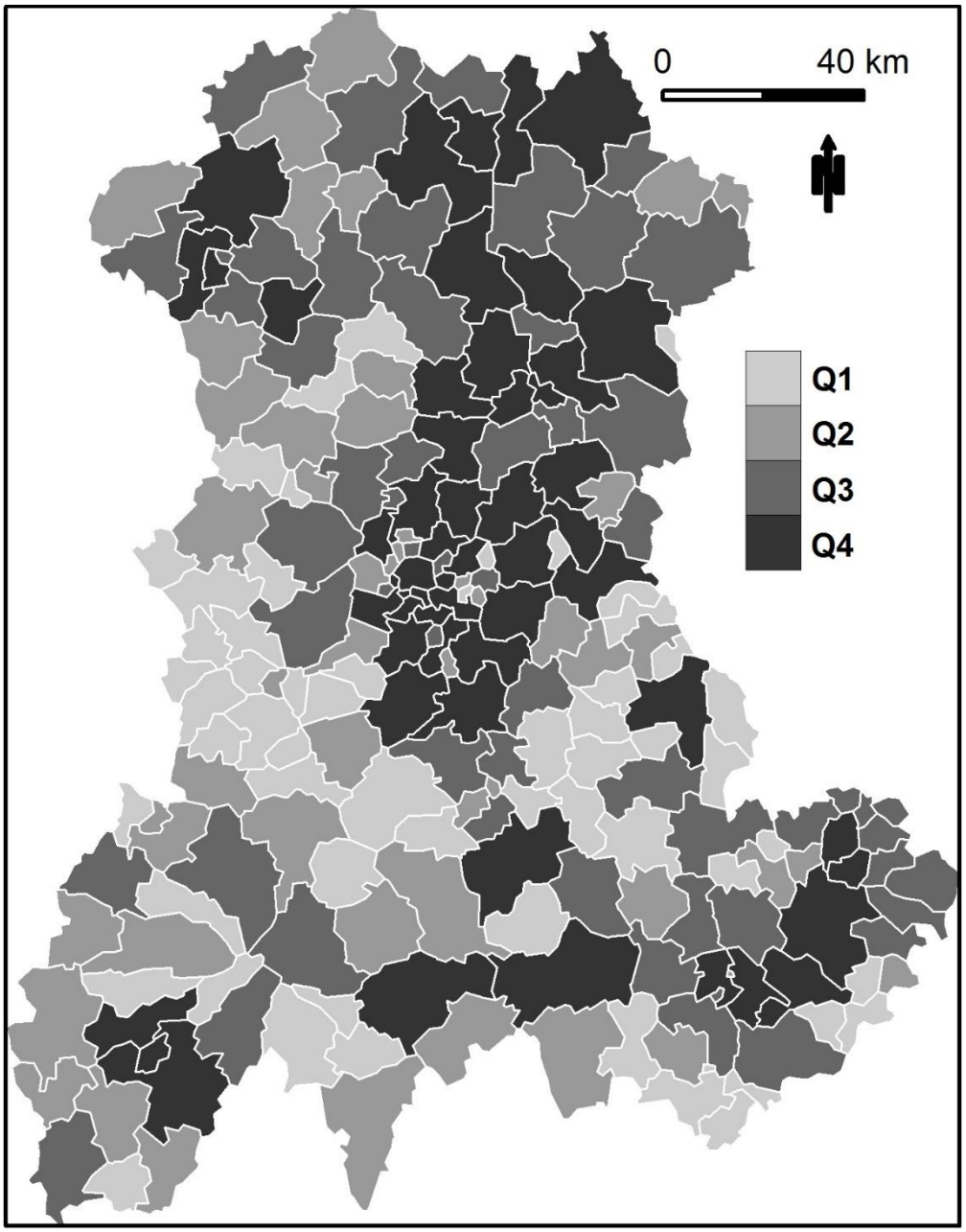

Fig. 1 Size of the at-risk population for each SU in the Auvergne region, as defined by mean number of live births per year between 1999 and 2006 (source: INSEE). $Q1 :\leq 17$; $Q2 :> 17$ and ≤ 35; $Q3 :> 35$ and ≤ 70; $Q4 :> 70$.

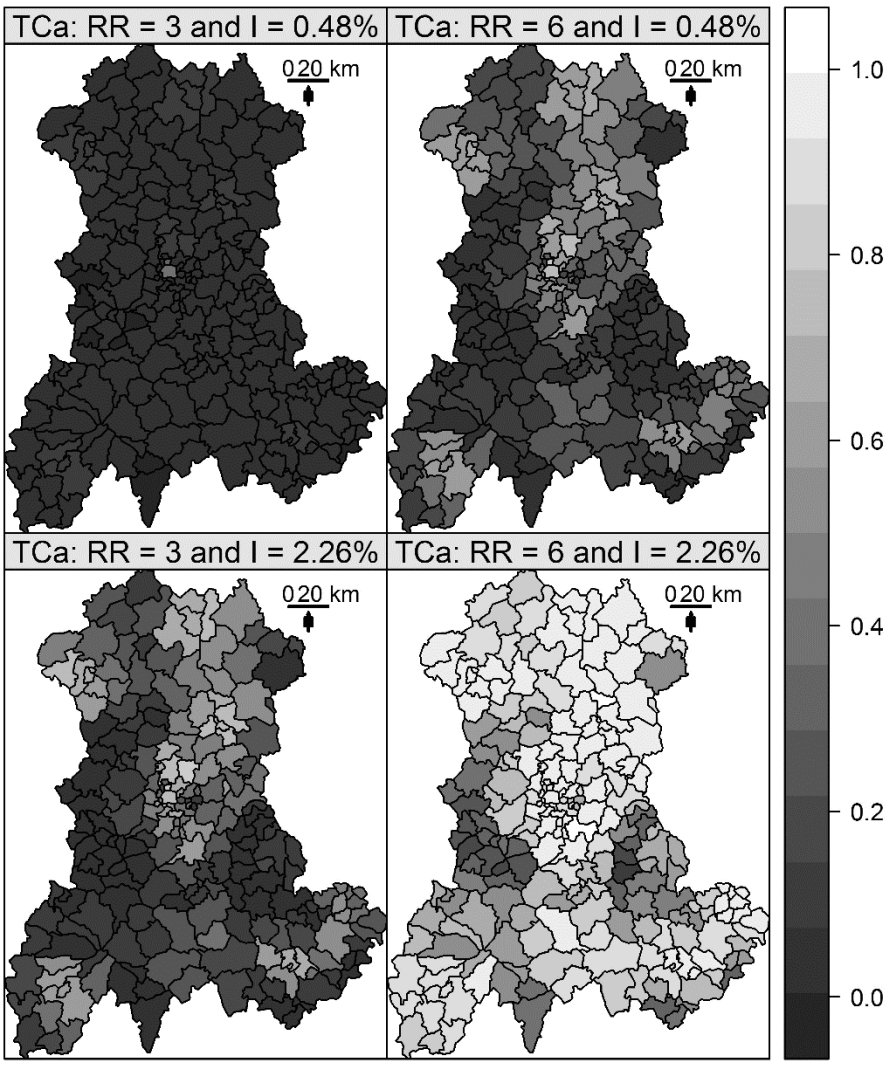

Fig. 2 TC_a of Kulldorffs spatial scan. TC_a measured for four combinations of two relative risks (RR) and two annual incidences of birth defects: low RR = 3 and high RR = 6; low incidence = 0.48% births and high incidence = 2.26% births.

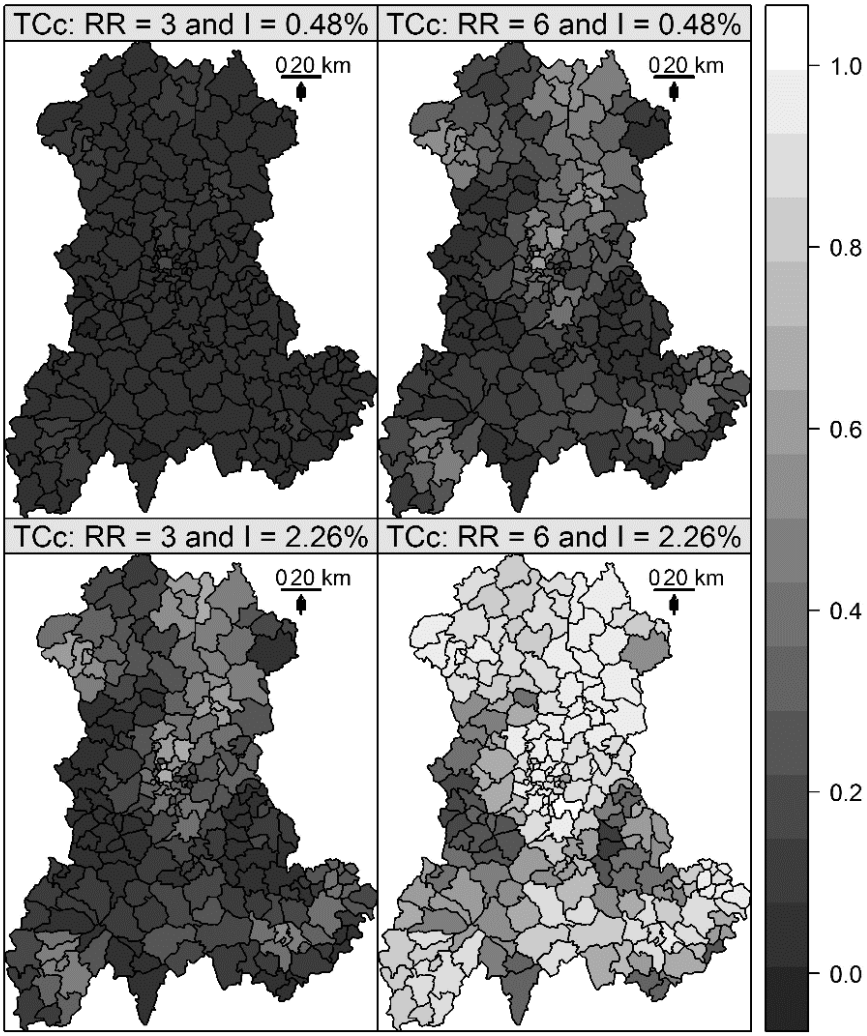

Fig. 3 TC_c of Kulldorffs spatial scan. TC_c measured for four combinations of two relative risks (RR) and two annual incidences of birth defects: low RR = 3 and high RR = 6; low incidence = 0.48% births and high incidence = 2.26% births.

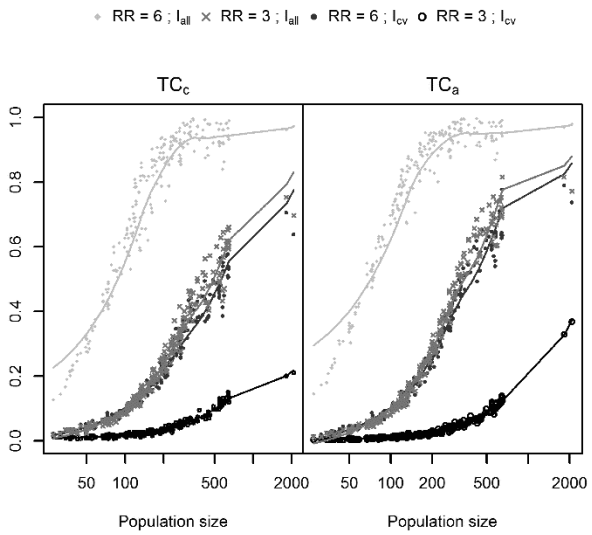

Fig. 4 Performance indicators and size of at-risk population. Indicators are measured for four combinations of two relative risks (RR) and two annual incidences of birth defects: low RR = 3 and high RR = 6; low incidence = 0.48% births and high incidence = 2.26% births.

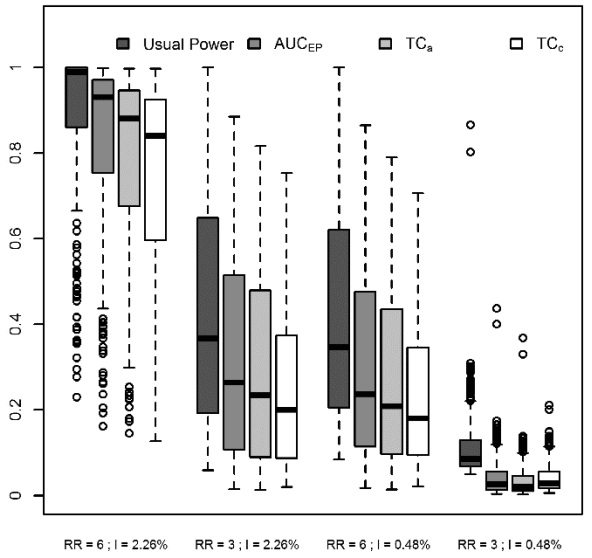

Fig. 5 Summary statistics of usual Power, AUC_{EP}, TC_a and TC_c for four combinations of two relative risks (RR) and two annual incidences of birth defects: low RR = 3 and high RR = 6; low incidence = 0.48% births and high incidence = 2.26% births.

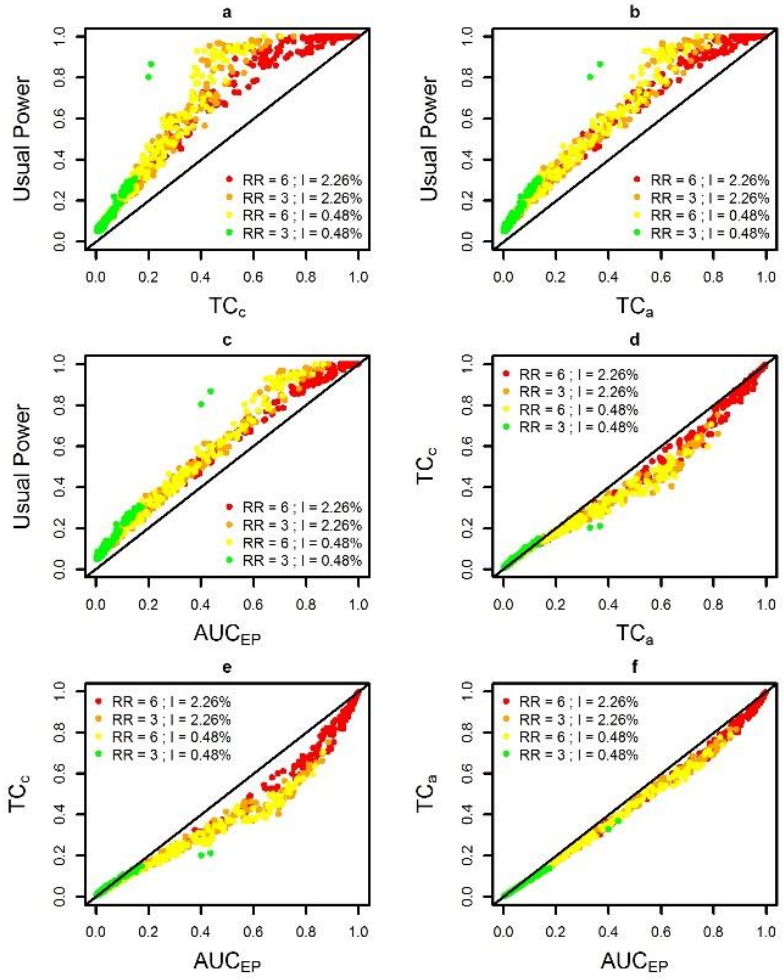

Fig. 6 Performance measures for four combinations of two relative risks (RR) and two annual incidences of birth defects: low RR = 3 and high RR = 6; low incidence = 0.48% births and high incidence = 2.26% births. (a) Usual Power and TC_c, (b) Usual Power and TC_a, (c) Usual Power and AUC_{EP}, (d) TC_c and TC_a, (e) TC_c and AUC_{EP}, (f) TC_a and AUC_{EP}

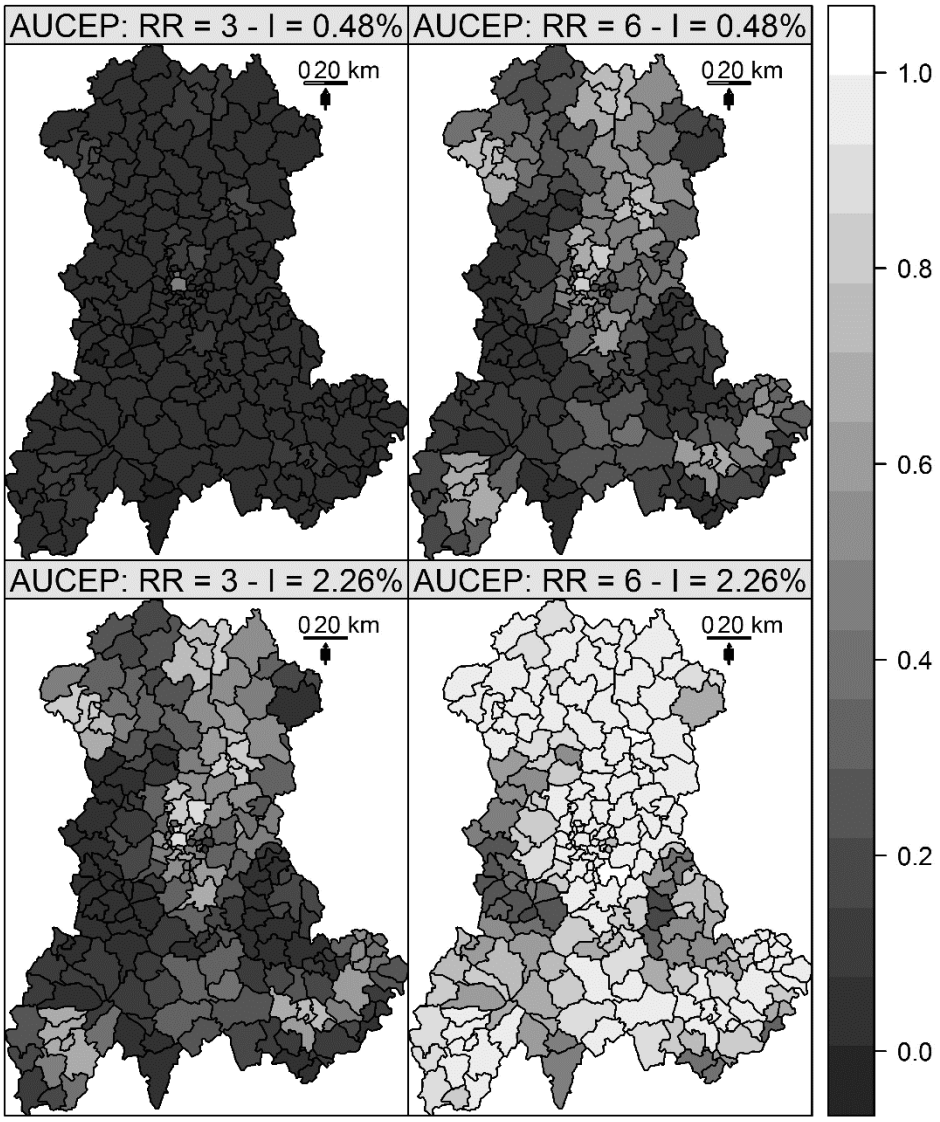

Fig. 7 AUC_{EP} of Kulldorffs spatial scan. AUC_{EP} was measured for four combinations of two relative risks (RR) and two annual incidences of birth defects: low RR = 3 and high RR = 6; low incidence = 0.48% births and high incidence = 2.26% births.

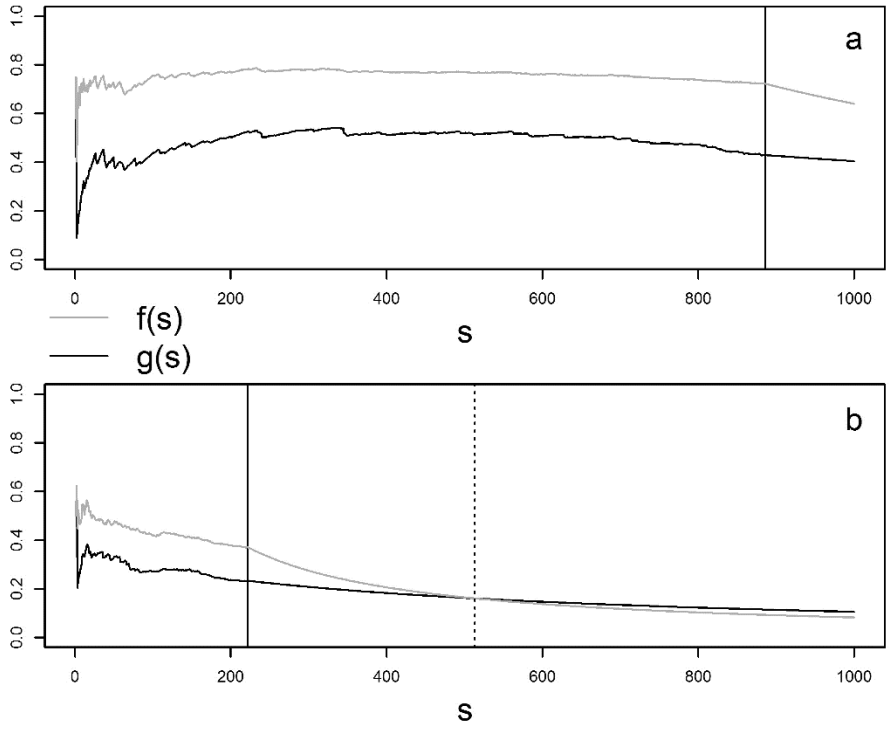

Fig. 8 Values of $f(s)$ and $g(s)$ for simulation $s = 1 : 1000$. The s simulations are sorted by increasing value of *p-value*. The functions $f(s)$ and $g(s)$ represent respectively the computation of TC_a and TC_c over the s simulations. The vertical plain line correspond to the last simulation leading to a detected cluster ($p - value < 0.05$). (a) simulated cluster with the maximum value of TC_a - TC_c and (b) simulated cluster with the minimum value of TC_a - TC_c.

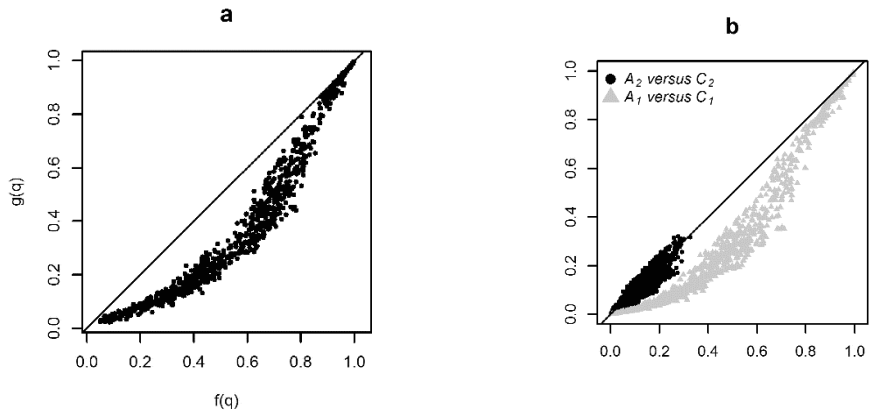

Fig. 9 Relationship between $f(s)$ and $g(s)$ at simulation $s = q$. (a) $f(q)$ versus $g(q)$ and (b) Contribution of each term of the sums $f(q) = A_1 + A_2$(in ordinate) and $g(q) = C_1 + C_2$ (in abscissa). With $A_1 = \sum_{S_1} \frac{TP_s}{qD}$, $C_1 = \frac{\sum_{S_1} TP_s}{qD + \sum_{S_0} FP_s + \sum_{S_2} FP_s}$, $A_2 = \sum_{S_2} \frac{TP_s}{qD + qFP_s}$ and $C_2 = \frac{\sum_{S_2} TP_s}{qD + \sum_{S_0} FP_s + \sum_{S_2} FP_s}$.

4 Étude la répartition spatiale de l’erreur de type I et effet de bord

INTERNATIONAL JOURNAL OF HEALTH GEOGRAPHICS

METHODOLOGY Open Access

Spatial heterogeneity of type I error for local cluster detection tests

Aline Guttmann[1,2*], Xinran Li[2], Jean Gaudart[3,4], Yan Gérard[2], Jacques Demongeot[5], Jean-Yves Boire[1,2] and Lemlih Ouchchane[1,2]

Abstract

Background: Just as power, type I error of cluster detection tests (CDTs) should be spatially assessed. Indeed, CDTs' type I error and power have both a spatial component as CDTs both detect and locate clusters. In the case of type I error, the spatial distribution of wrongly detected clusters (WDCs) can be particularly affected by edge effect. This simulation study aims to describe the spatial distribution of WDCs and to confirm and quantify the presence of edge effect.

Methods: A simulation of 40 000 datasets has been performed under the null hypothesis of risk homogeneity. The simulation design used realistic parameters from survey data on birth defects, and in particular, two baseline risks. The simulated datasets were analyzed using the Kulldorff's spatial scan as a commonly used test whose behavior is otherwise well known. To describe the spatial distribution of type I error, we defined the participation rate for each spatial unit of the region. We used this indicator in a new statistical test proposed to confirm, as well as quantify, the edge effect.

Results: The predefined type I error of 5% was respected for both baseline risks. Results showed strong edge effect in participation rates, with a descending gradient from center to edge, and WDCs more often centrally situated.

Conclusions: In routine analysis of real data, clusters on the edge of the region should be carefully considered as they rarely occur when there is no cluster. Further work is needed to combine results from power studies with this work in order to optimize CDTs performance.

Keywords: Cluster detection test, Type I error, Simulation study, Edge effect, Spatial scan

Résumé

Contexte: Les tests de détection de clusters (CDT) permettent à la fois de détecter et de localiser les clusters. Au même titre que pour la puissance, il est donc nécessaire d'étudier la répartition spatiale de l'erreur de type I de ces CDT. Dans le cas de l'erreur de type I, la répartition spatiale des clusters détectés à tort (WDC) peut être particulièrement concernée par un effet de bord. Cette étude de simulation a pour objectif de décrire la distribution spatiale des WDCs et de confirmer et quantifier la présence de cet effet de bord.

(Continued on next page)

* Correspondence: aline.guttmann@udamail.fr
[1]Department of Biostatistics, Medical Informatics and Communication Technologies, Clermont University Hospital, Clermont-Ferrand F-63000, France
[2]UMR CNRS UDA 6284 ISIT, Auvergne University, Clermont-Ferrand F-63001, France
Full list of author information is available at the end of the article

(Continued from previous page)

Méthodes: Ce travail s'appuie sur la synthèse de 40 000 jeux de données simulant l'hypothèse nulle d'homogénéité spatiale des risques. Les simulations étaient fondées sur les paramètres réels de données d'un registre de malformations congénitales, et notamment sur deux risques de base réels. La description de la distribution spatiale de l'erreur de type I nous a conduits à définir le concept de taux de participation de chaque unité spatiale de la région. Cet indicateur a ensuite été intégré pour la construction d'un nouveau test statistique destiné à confirmer et quantifier l'effet de bord.

Résultats: La valeur globale de l'erreur de type I à 5% a bien été retrouvée. Les résultats montraient un très net effet de bord avec un gradient décroissant du taux de participation depuis le centre vers le bord, les WDC étant plus souvent situés en zone centrale.

Conclusions: Lors de la mise en œuvre des CDT sur données réelles, les détections de clusters près du bord d'une région d'étude doivent être examinées avec la plus grande attention, ces dernières étant très rares en l'absence de cluster réel. Il est maintenant nécessaire d'orienter de futurs développements vers la combinaison de ces résultats à ceux des études de puissance, et ce dans le but d'optimiser les performances des CDT.

Background

Spatial clusters can be detected using a wide range of statistical tests [1,2] many of which are available in free software such as R [3,4]. Epidemiologists use cluster detection tests (CDTs) to detect clusters without *a priori* knowledge either of their number or their location, and to determine their significance. CDTs performance being a function of epidemiological and geographical context [1,5-11], it is recommended to perform power studies before using these tests in a particular region for a given phenomenon. However, statistical power is not the only test characteristic determining performance. Performance at large depends on two type of risks: type I and type II errors.

In presence of clusters, usual statistical power (1-β) is not sufficient to assess CDT performance to reject the null hypothesis of risk homogeneity. At worst, a CDT could have a maximum power to reject this null hypothesis of risk homogeneity but never correctly locate the true cluster. Similar concern can be raised for type I error. A CDT could, under the null hypothesis of no cluster, generate wrongly detected clusters (WDC) preferentially localized in particular zones of the studied region. The overall type I error could effectively be equal to its predefined value usually set to 5%, but the interpretation of the analyses would certainly not be the same for detected clusters inside or outside such zones.

In the case of statistical power, authors have since used either evaluation of power and location by different indicators [6,12-14] or concomitant evaluation of both with a single measure such as the extended power [15,16]. The development of single measure of performance taking into account both power and location accuracy has enabled systematic spatial evaluation of performance on entire regions [15]. The question of the spatial evaluation of CDT is, so far, not totally answered with regards to power because evaluation of factors such as relative risks or cluster shape and size are still assessed by a non-systematic approach based on more or less arbitrary settings in simulation designs.

The question of relative risks and clustering characteristics is not relevant in the spatial evaluation of type I error, other factors have to be taken into account, however. First, there is still one epidemiological factor that requires setting: the baseline risk. For an applicative purpose, the use of the baseline incidence of the studied disease is the evident choice, but for research, a systematic evaluation over a wide range of this factor should be carried out. Second, simulation studies evaluating type I error are much more likely to be influenced by edge effect [17-19] than power studies. Indeed, in the majority of simulation studies assessing power, edge effect is largely lessened by designs simulating clusters wholly within the studied region.

We aimed to evaluate CDTs regarding the spatial distribution of type I error. Such description was carried-out at the level of the spatial unit (SU) introducing the concept of SU's participation rate. We proposed a statistic to quantify and test for edge effect which was of particular interest. We used Kulldorff spatial scan statistic as an example of CDT, whose behavior is otherwise well known, and performed a simulation study using realistic parameters from survey data on birth defects.

Methods

Disease modeling

The study region was the Auvergne region (France), divided into $n = 221$ spatial units (SUs) equivalent to U.S. ZIP codes. We applied two baseline risks (incidences) of birth defects to the same at-risk population, whose size was approximated by mean annual number of live births.

For a realistic analysis, we used data archived in CEMC (birth defects registry for the Auvergne region) and INSEE (National Institute of Statistics and Economic Studies) databases. We collected two categories of data from 1999

to 2006: all birth defects and cardiovascular birth defects. Both datasets were sorted by SU. The number of live births was approximated by the number of birth declarations in the at-risk population. Global annual incidences of all birth defects (I_{all}) and cardiovascular birth defects (I_{cv}) were estimated at 2.26% and 0.48% of births, respectively.

Datasets

We generated 20 000 datasets for each baseline risk, *i.e.* a total of 40 000 datasets.

Each dataset is entered as a table of 221 rows and 5 columns. The rows contain the coordinates (longitude and latitude) of a SU, the observed number of cases, the size of the at-risk population (*i.e.*, the number of live births) and the expected number of cases in the specified SU. This last quantity is the product of the global incidence (I_{all} or I_{cv}) and the at-risk population size in the SU. The observed case numbers are assumed as independent Poisson variables such that

$$E(N_i) = \mu_i, N_i \sim Pois(\mu_i), i = 1, \ldots, n$$

where N_i is the observed number of cases, and μ_i denotes the expected number of cases in the ith SU under the null hypothesis of risk homogeneity.

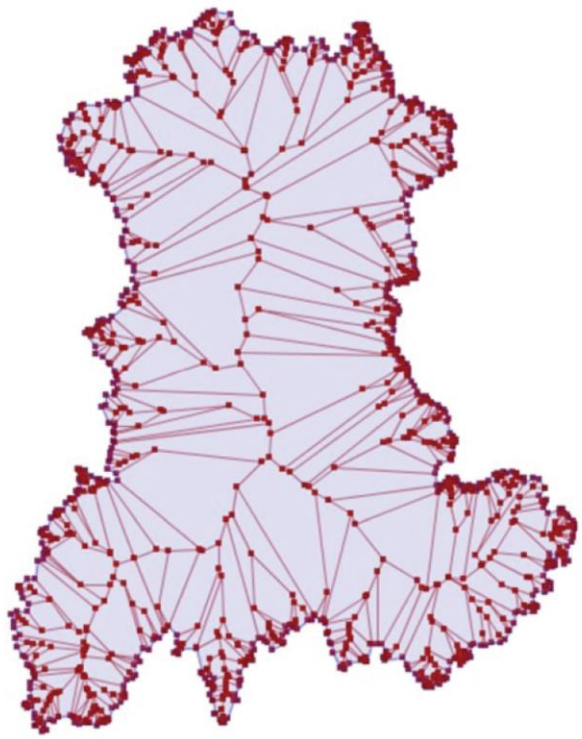

Figure 1 Representation of the medial axis of the Auvergne region.

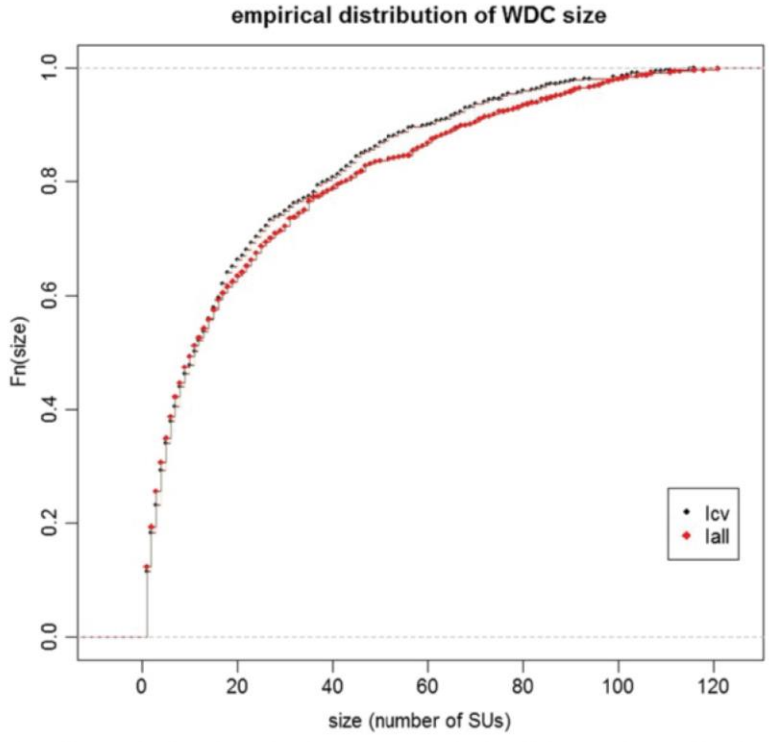

Figure 2 Empirical distribution of WDC size (number of SUs) over 20 000 simulated datasets for two baseline incidences of birth defects: 2.26% (Iall) and 0.48% (Icv).

Assessment of type I error

Overall rate

The global type I error rate was estimated by the proportion of WDC over the 20 000 datasets for each baseline risk.

Spatial distribution

SU participation rate: Participation rate of each WDC in the overall type I error is equal to *1/m*, with *m* the number of WDCs. Participation rate of each SU in the overall type I error was estimated by a weighted sum of the number of times each SU was included in a WDC. This weight is a function of *m* and the length of each WDC (number of SUs within). For each SU *i* among the *n* SUs of the region, the participation rate P_i in the overall type I error is such that

$$\mathrm{P_i} = \sum_{\mathrm{j=1}}^{\mathrm{m}} \mathbb{I}_{\mathrm{ij}} \left(\mathrm{ml_j}\right)^{-1}$$

where *m* is the number of WDCs, l_j is the length of the *j*th WDC and $\mathbb{I}_{ij}$ a binary indicator equal to 1 when the *i*th SU is within the *j*th WDC and 0 otherwise. By construction, $P_i \geq 0$ and $\sum_{i=1}^{n} P_i = 1$, where *n* is the number of SUs in the region.

Edge effect: The edge effect is defined here as an inhomogeneous distribution of P_i characterized by a gradient from the medial axis (or cut locus or skeleton) of the region to its edge. This gradient can either be ascending or descending. The medial axis is the set of all points having more than one closest point on the region's edge [20-23]. The Figure 1 shows the medial axis of the region under study[a]. For such a simple polygon, the medial axis is a tree whose leaves are the vertices and whose edges are straight segments reflecting local symmetries of the shape.

To confirm the presence of an edge effect, we propose a test whose statistic, referred to as *E*, is such that

$$\begin{cases} E = \sum_{\mathrm{i=1}}^{\mathrm{n}} \varepsilon_{\mathrm{i}} \left(\mathrm{P_i - n^{-1}}\right) \\ \varepsilon_{\mathrm{i}} = \left(1 - \frac{2\mathrm{d_i}}{\mathrm{D}}\right) \end{cases}$$

Where d_i is the minimal Euclidian distance between the centroid of the *i*th SU and the edge of the region, *D* the maximum Euclidian distance between any point of the medial axis and the region closest edge, and *n* the number of SUs in the region. By construction, as $0 \geq d_i \geq D$, $-1 \geq \varepsilon_i \geq +1$. The coefficient ε_i is a continuous indicator quantifying how much a point can be considered "on the edge" of the region. It is referred to as "the edge

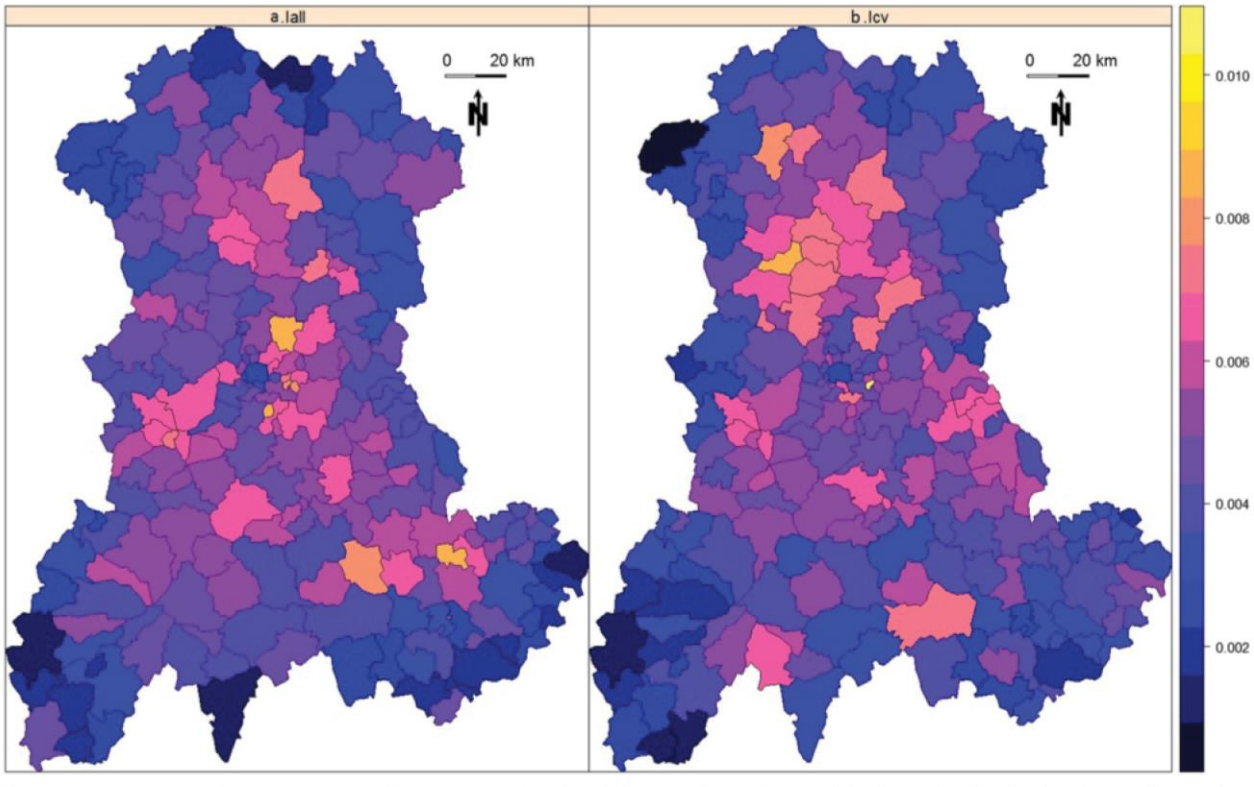

Figure 3 SUs participation rates computed over 20 000 simulated datasets for each map. (a) Observed values for baseline incidence of birth defects set to 2.26% (lall). **(b)** Observed values for baseline incidence of birth defects set to 0.48% (lcv).

coefficient" in the remainder of this paper. For any point in the region, the closer to the edge, the higher the edge coefficient, and the closer to the medial axis, the smaller the edge coefficient. The edge coefficient ranges from -1 for the most "central/medial" points of the region to +1 for points on the edge. For a study region divided into census tract, each SU is attributed the edge coefficient of its centroid. All SUs with the same edge coefficient are at the same distance to the edge and the closer to the medial axis, the smaller the edge coefficient, tending to -1 for the most "central" SUs of the region.

The test hypotheses are expressed by

$$\begin{cases} H_0 : E = 0 \\ H_1 : E \neq 0 \end{cases}$$

The quantity n^{-1} is the expected participation rate for all SUs under the null hypothesis of spatial homogeneity in type I error. When P_i is higher than expected towards the edge of the region, by construction, it is lower towards the center (as $\sum_{i=1}^{n} P_i = 1$) and there is an ascending gradient. On the contrary, when P_i is higher towards the center of the region, there is a descending gradient. The statistic E is positive when there is an ascending gradient of P_i and negative when the gradient is descending. Indeed, in case of an ascending gradient

- central SUs will tend to have

$$\varepsilon_i < 0, \left(P_i - n^{-1}\right) < 0 \quad (1)$$

- border SUs will tend to have

$$\varepsilon_i > 0, \left(P_i - n^{-1}\right) > 0 \quad (2)$$

and E will tend to be highly positive.

In case of a descending gradient

- central SUs will tend to have

$$\varepsilon_i < 0, \left(P_i - n^{-1}\right) > 0 \quad (3)$$

- border SUs will tend to have

$$\varepsilon_i > 0, \left(P_i - n^{-1}\right) < 0 \quad (4)$$

and E will tend to be highly negative.

Finally, under H_0 of spatial homogeneity of type I error, the sum of all P_i, equal to 1, is homogeneously distributed among the n SUs with an expected participation rate

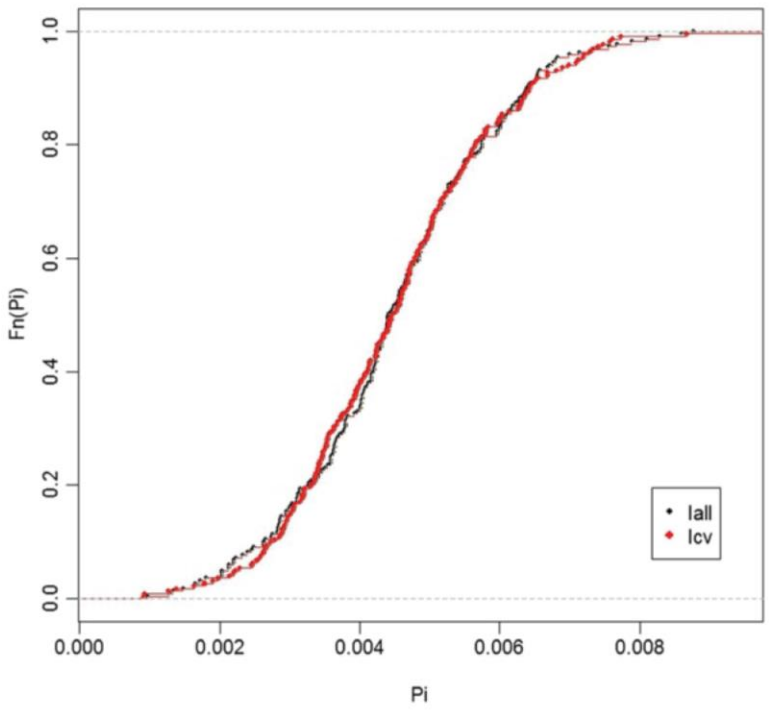

Figure 4 Empirical distribution of SUs participation rates computed over 20 000 simulated datasets for two baseline incidences of birth defects: 2.26% (Iall) and 0.48% (Icv).

equal to n^{-1}. Under this null hypothesis, the expected value of $(P_i - n^{-1})$ is null and independent to ε_i. Consequently, under null hypothesis, the expected value of E is null.

Since the variance of the E statistic under H_0 (spatial homogeneity of type I error) is unknown, we used Monte Carlo simulation where the n observed P_i were randomly distributed 99 999 times among the n SUs in the region. The p-value was the proportion of elements among the collection of simulated and observed statistics which were greater than or equal to the observed value. The precision of this p-value was thus of 10^{-5} digits.

Kulldorff's spatial scan statistic

In this study, we selected Kulldorff's spatial scan statistic [24,25] as a well-known and widely used CDT whose performance has been studied by many authors [1,6,10,26]. The spatial scan statistic detects the most likely cluster on locally observed statistics of likelihood ratio tests. The scan statistic considers all possible zones z defined by two parameters: a center that is successively placed on the centroid of each SU, and a radius varying between 0 and a predefined maximum. The true geography being delineated by administrative tracts, each zone z defined by all SUs whose centroids lie within the circle, is irregularly shaped. Let N_z and n_z be respectively the size of the at-risk population and the number of cases counted in zone z (over the whole region, these quantities are the total population size N and the total number of cases n). The probabilities that an at-risk case lies inside and outside zone z are respectively defined by $p_z = n_z/N_z$ and $q_z = (n-n_z)/(N-N_z)$. Given the null hypothesis of risk homogeneity H_0: $p_z = q_z$, versus the alternative H_1: $p_z > q_z$ and assuming a Poisson distribution of cases, Kulldorff defined the likelihood ratio statistics as proportional to

$$\left(\frac{n_z}{\lambda N_z}\right)^{n_z}\left(\frac{n-n_z}{\lambda(N-N_z)}\right)^{n-n_z} I[n_z > \lambda N_z],$$

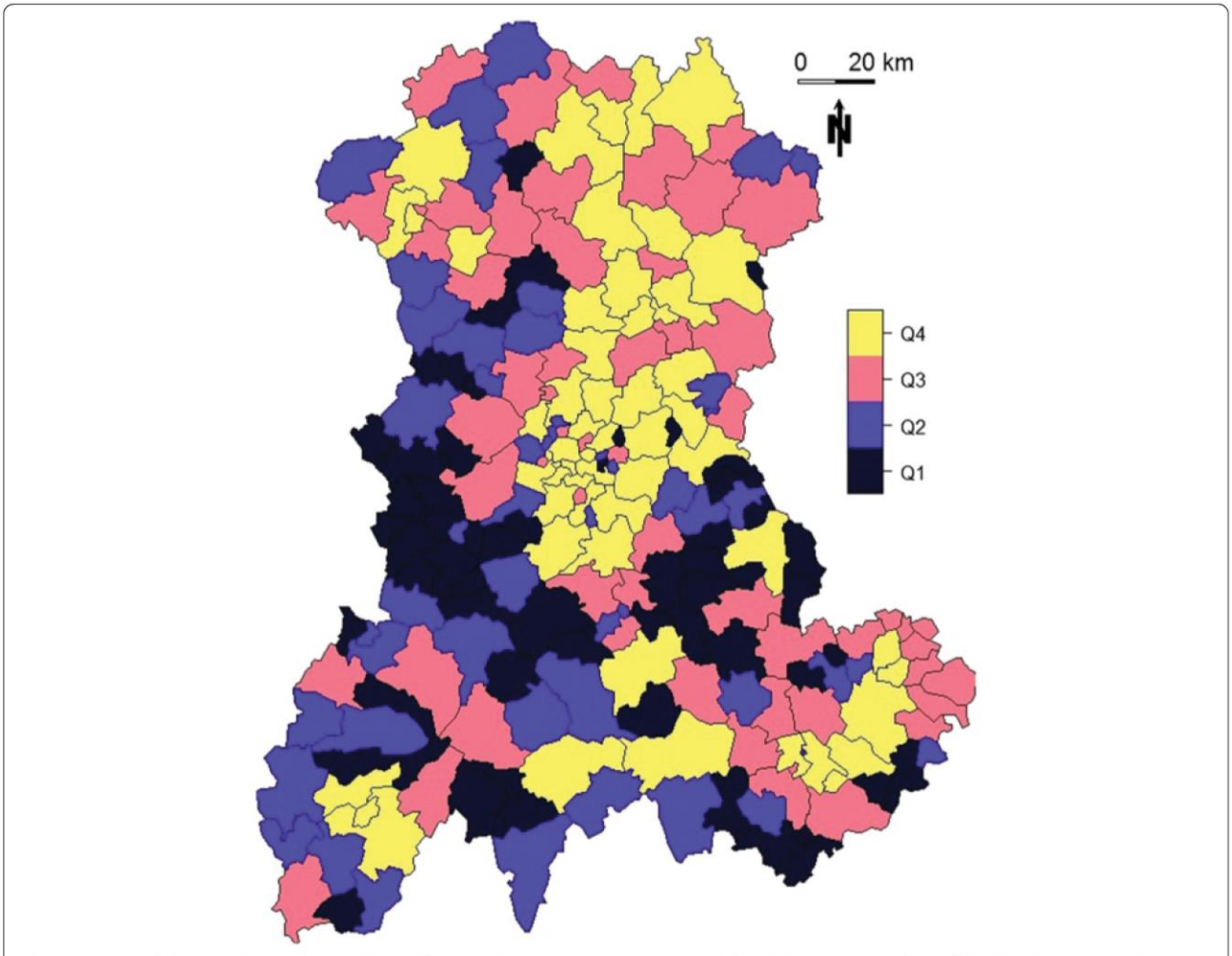

Figure 5 Size of the at-risk population for each SU in the Auvergne region, as defined by mean number of live births per year between 1999 and 2006 (source: INSEE). Q1: ≤ 17; Q2: > 17 and ≤ 35; Q3: > 35 and ≤ 70; Q4: > 70.

where λ (here equal to I_{all} or I_{cv} depending on the case considered) is the global incidence and the indicator function I equals 1 when the number of observed cases in zone z exceeds the expected number under H_0 of risk homogeneity, and 0 otherwise. The circle yielding the highest likelihood ratio is identified as the most likely cluster. The p-value is obtained by Monte Carlo inference.

Over the 40 000 simulated datasets, each test was performed with a maximum size of zone z set to 50% of the total at-risk population, a number of 999 Monte Carlo samples for significance measures, and an alpha level set to 5%.

Software

Data simulation and analysis were performed on R 2.14.0 [3,27-29], using the function "kulldorff" of the SpatialEpi package [27] to perform the Kulldorff's spatial scan.

Results

Overall rate and WDC characteristics

The overall type I error rate was 5.11% (1021 WDC over 20 000 datasets; CI 95% [4.80%, 5.42%]) for I_{all} and 5.06% (1012 WDC over 20 000 datasets; CI 95% [4.76%, 5.38%]) for I_{cv}. The average size of WDCs was 21.4 SUs (minimum 1SU, median 11 SUs, maximum 116 SUs) and 23.4 SUs (minimum 1SU, median 11 SUs, maximum 132 SUs), respectively. The Figure 2 shows the empirical distribution of the WDC size for each baseline risk.

SUs participation rates

Figure 3 shows the SUs participation rates for baseline risks I_{all} (Figure 3a) and I_{cv} (Figure 3b). The expected participation rate (n^{-1}) for each SU is equal to 0.452%. With 0.452% ± 0.147% (mean ± standard deviation) for I_{all} and 0.452% ± 0.148% for I_{cv}, the two observed distributions of participation rates were very close to each other (Figure 4). The observed values varied from 0.097% to 0.877% for I_{all} and from 0.091% to 1.03% for I_{cv}.

We sought for a correlation between P_i and size of the at-risk population (Figure 5) by Spearman's rank test. Both coefficients were negative but none resulted in significant relationship (r = -0.13 with p-value = 0.056 for I_{all} and r = -0.11 with p-value = 0.1 for I_{cv}).

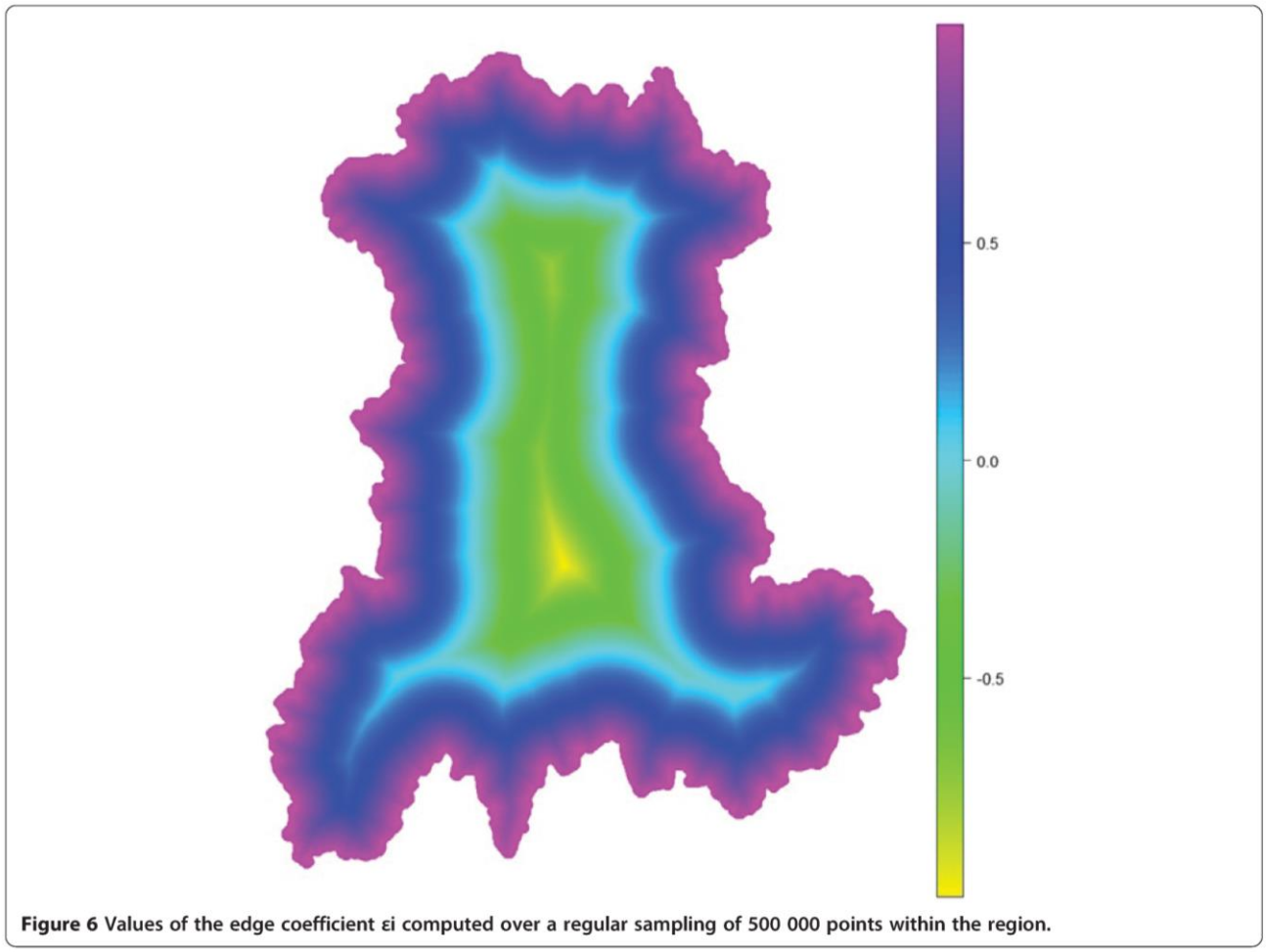

Figure 6 Values of the edge coefficient εi computed over a regular sampling of 500 000 points within the region.

Edge effect

Figure 6 shows the value of the edge coefficient ε_i computed for a regular sampling of 500 000 points within the region. Figure 7 shows the value of the edge coefficient computed for the $n = 221$ SUs within the region.

With E equal to -0.086 for I_{all} and -0.074 for I_{cv}, both simulations resulted in descending gradient of P_i, *i.e.* higher P_i for central SUs. As shown by E values, this gradient was stronger for I_{all} than for I_{cv}.

As shown in Figure 8, the SUs contributing to the overall type I error for more than n^{-1} ($P_i > n^{-1}$) were mostly located away from the border of the region. The black line delineates a central zone where the edge coefficient is negative and a complementary zone where the edge coefficient is positive. Within the central zone, red SUs contribute negatively to E (see Equation 3), on the contrary, outside the central zone, red SUs contribute positively to E (see Equation 4).

Both tests were highly significant, with Monte Carlo p-values both equal to 10^{-5} (99 999 replicates). Figure 9 shows the simulated null distributions of E and the observed values for the two simulated baseline risks.

Discussion

We have shown that type I error is heterogeneously distributed with a descending gradient from center to edge. Even if global type I error is very near the predefined 5%, WDCs are rarely located on the edge of the map. In a survey system, where sensitivity matters over specificity, it could be argued that since global type I error is preserved, the global cost in unfruitful secondary investigation is not affected by the spatialization of type I error.

Our work did not aim to test for clustering in type I error rate and thus we did not used CDTs to analyze the spatial distribution of P_i. We note, however, that methods such as Bayesian smoothing could be of interest in the description of the spatial distribution of type I error. As the presence of an edge effect with descending gradient was obviously expected, our contribution aimed to describe, quantify and test for this edge effect. Furthermore, within a given region, the spatial description of type I error makes possible to see with precision which detected clusters should be carefully considered because they are less likely to coincide with false alarm.

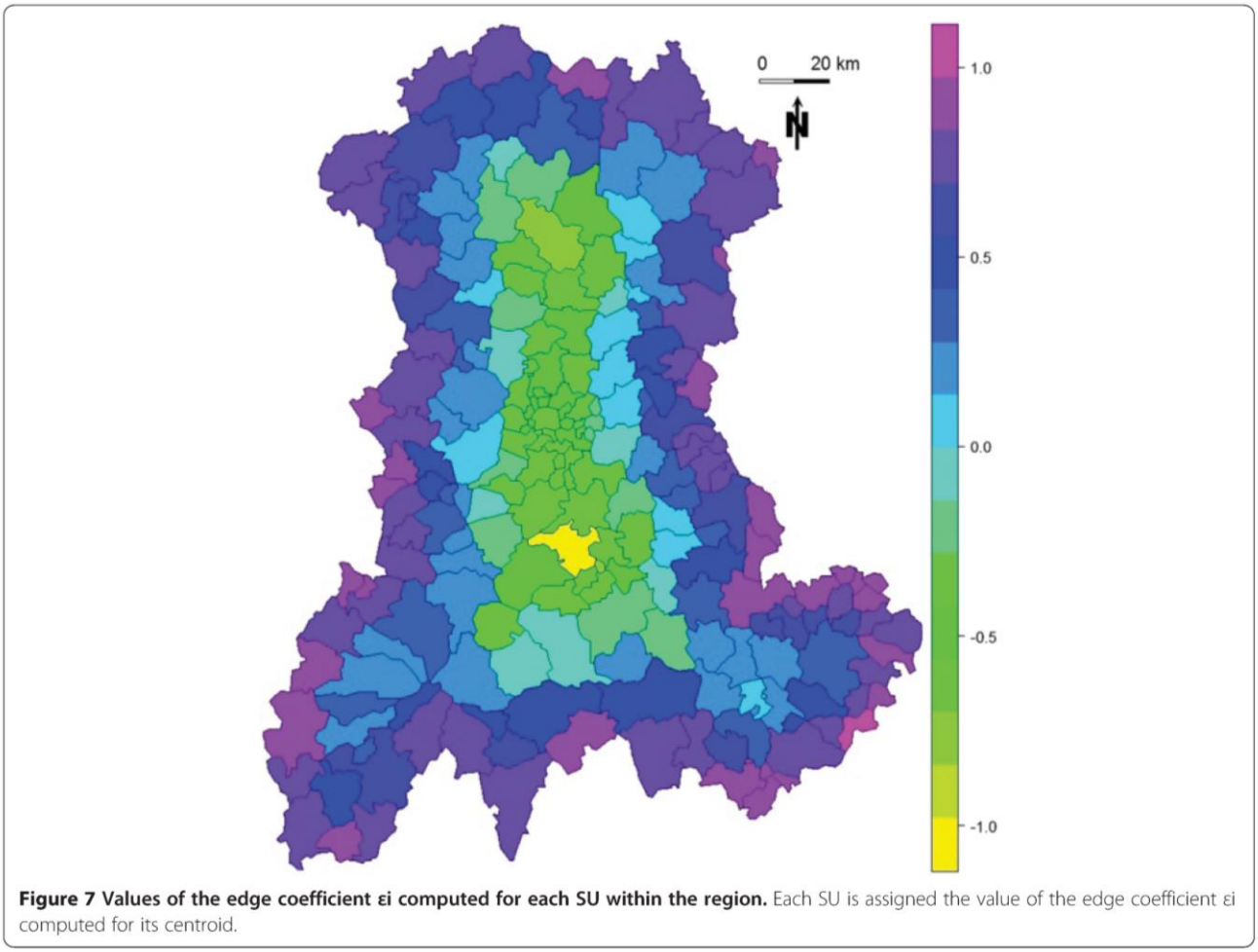

Figure 7 Values of the edge coefficient εi computed for each SU within the region. Each SU is assigned the value of the edge coefficient εi computed for its centroid.

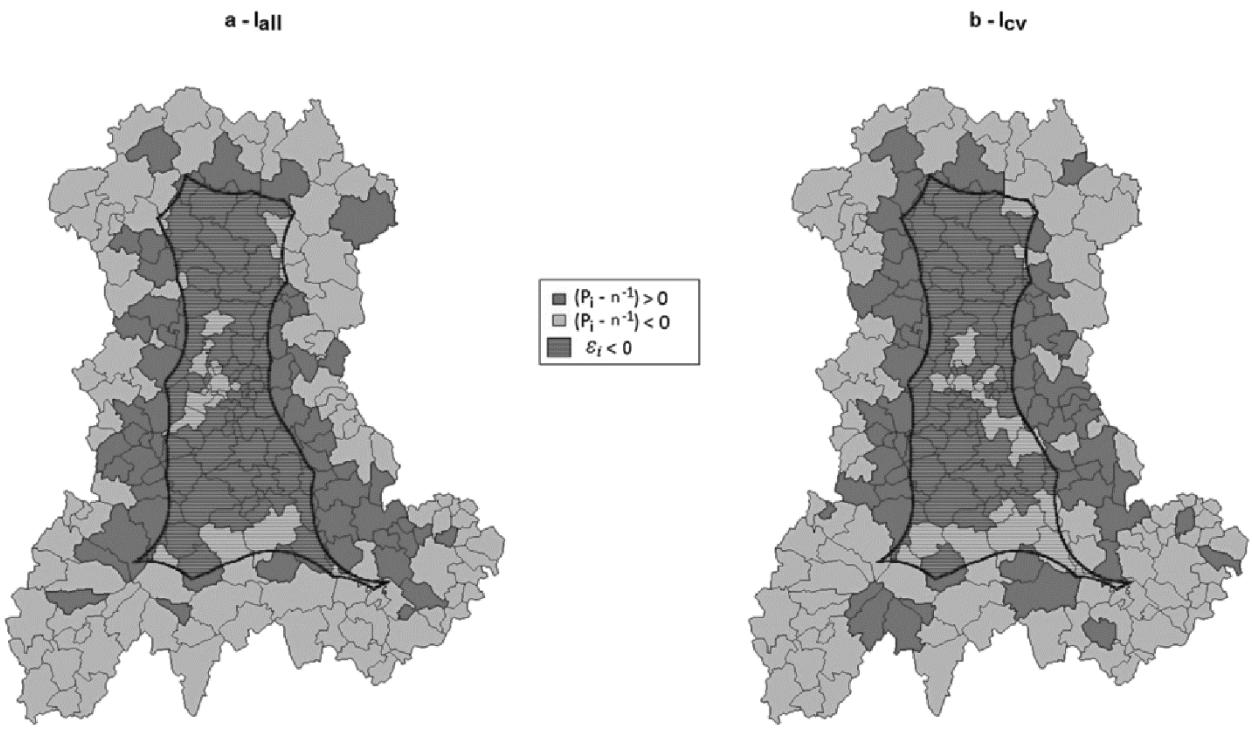

Figure 8 Overlaying of the sign of the difference between observed and expected values of SUs participation rates (computed over 20 000 simulated datasets for each map) and the sign of the edge coefficient εi (εi negative in hatched area). (a) Baseline incidence of birth defects set to 2.26%. **(b)** Baseline incidence of birth defects set to 0.48%.

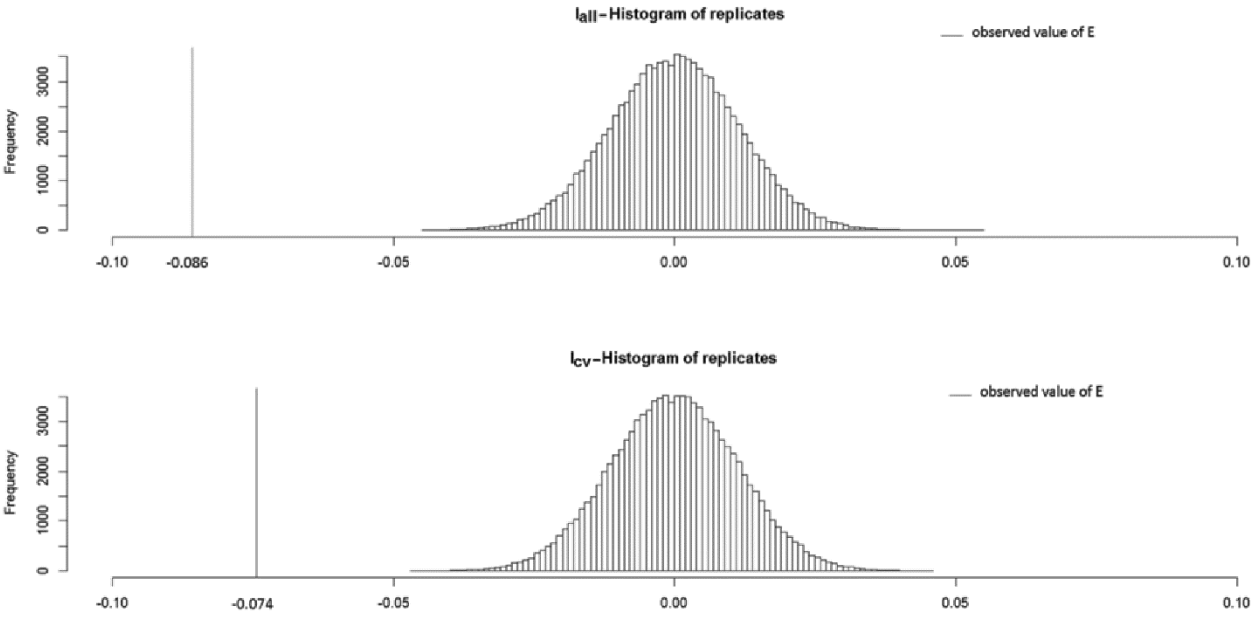

Figure 9 Histograms of the null distribution (99 999 replicates) and observed values of E for two baseline incidences of birth defects: 2.26% (Iall) and 0.48% (Icv).

The edge effect was present and strong, no matter the baseline risk. Only two levels have been tested for this risk. One could wonder about a possible correlation between edge effect and the level of baseline risk. Levels at regular interval between these two baseline risks are currently being explored and there is no evidence of such a correlation so far (data not shown).

The edge effect is indisputable in this study (Figure 8) and the statistic E has consequently resulted in a highly significant test. This statistic is based on the edge coefficient ε_i that defines what is "on the edge" of the map and what is not. By using medial axis, we proposed a distance-based definition, but other parameters could be considered. For instance, it could be useful to distinguish between two SUs at the same distance to the edge but in different configurations with one in a "peninsula" (between two edges) and thus more isolated than the others. To be accounted for, this factor needs geometrical tools to characterize the spatial isolation.

Aside from a purely geometrical definition of what is an edge, confounding factors should also be taken into account. Suppose that the at-risk population is heterogeneously distributed, with more populated areas centrally localized. Then, suppose again that the at-risk population size is negatively correlated to participation rate (this was not the case in our study). Our test for edge effect might turn out to be significant, concluding in an ascending gradient of P_i from center to edge, only due to this confounding factor. In our simulations, the at-risk population is effectively more centrally localized. If the negative correlation between population size and P_i had been significant, we would have an even stronger evidence for a descending edge effect regarding P_i from center to edge, because our results, that turned out to be significant, would have actually been underestimated.

Even if we did not find any relationship between population size and participation rate, other factors (such as the number of neighbors, the accessibility by road or rail system, *etc.*) should be evaluated. The best way to deal with these confounding factors might be to integrate them in the construction of ε_i for geographical factors or to replace the constant n^{-1} by a vector of expected participation rates for epidemiological factors. For the E statistic to be equal to 0 under H_0 (spatial homogeneity of type I error), this last adaptation should be done in such a way that the sum of all expected participation rates stays equal to 1.

Our results highlight the edge effect in type I error, and thus can help the interpretation of real data analysis. It could be even more useful to provide a way to integrate spatial heterogeneity of type I error in the analysis itself. Furthermore, adjustment in CDT behavior should be done to address this issue only if it does not impede the tests' power. In a previous simulation study on CDT performance, we proposed a method to build performance map based on a systematic spatial evaluation [15]. The now available data for both H_1 (single clusters of 4SUs in this previous study) and H_0 (risk homogeneity) in similar settings (same baseline risk and population size) will enable us to study whether and how it could be gainful to add a spatial adjustment of type I error.

Conclusion

Spatial heterogeneity of type I error should be considered when interpreting analysis of real data, because of the strong edge effect. This work clearly shows that a detected cluster on the edge of the region of interest is less common when no alarm should be raised. To explore all avenues, assessment of edge effect and its factors, as well as development of tools to integrate it in routine health survey, should be considered.

Endnotes

[a]Computation of the straight skeleton was performed using [30] and the results were imported and displayed with JTS Topology Suite [31], a software under GNU license.

Abbreviations

WDC: Wrongly detected cluster; CDT: Cluster detection test; H_0: Null hypothesis; H_1: Alternative hypothesis; I_{all}: Incidence of all birth defects; I_{cv}: Incidence of cardiovascular birth defects.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AG and LO conceived the design, performed the study and drafted the manuscript. AG was responsible for statistical programming and data analysis. YG contributed to the construction of the E test. JD, JG, YG, XL and JYB contributed to manuscript revision. All authors read and approved the final manuscript.

Acknowledgments

Data have been provided by the CEMC (birth defect registry of Auvergne), with the participation of the Regional Health Agency of Auvergne, InVS (National Institute for Health Surveillance) and INSERM (National Institute of Health and Medical Research).

Author details

[1]Department of Biostatistics, Medical Informatics and Communication Technologies, Clermont University Hospital, Clermont-Ferrand F-63000, France. [2]UMR CNRS UDA 6284 ISIT, Auvergne University, Clermont-Ferrand F-63001, France. [3]UMR 912 SESSTIM (INSERM IRD AMU), Aix-Marseille University, Marseille F-13005, France. [4]Assistance Publique Hôpitaux de Marseille, Biostatistic and Modelisation, Marseille F-13005, France. [5]La Tronche University School of Medicine, FRE CNRS 3405 AGIM, J. Fourier University, Saint-Martin-d'Hères F-38700, France.

Received: 10 March 2014 Accepted: 17 May 2014
Published: 27 May 2014

References

1. Kulldorff M, Tango T, Park PJ: **Power comparisons for disease clustering tests.** *Comput Stat Data Anal* 2003, **42**:665–684.
2. Sankoh OA, Becher H: **Disease cluster methods in epidemiology and application to data on childhood mortality in rural Burkina Faso.** *Inform Biom Epidemiol Med Biol* 2002, **33**:460–472.
3. Gomez-Rubio V, Ferrandiz J, Lopez A: **Detecting clusters of diseases with R.** *J Geogr Syst* 2003, **7**:189–206.

4. Robertson C, Nelson TA: **Review of software for space-time disease surveillance.** *Int J Health Geogr* 2010, **9**:16.
5. Aamodt G, Samuelsen SO, Skrondal A: **A simulation study of three methods for detecting disease clusters.** *Int J Health Geogr* 2006, **5**:15.
6. Ozonoff A, Jeffery C, Manjourides J, White LF, Pagano M: **Effect of spatial resolution on cluster detection: a simulation study.** *Int J Health Geogr* 2007, **6**:52.
7. Jeffery C, Ozonoff A, White LF, Nuño M, Pagano M: **Power to detect spatial disturbances under different levels of geographic aggregation.** *J Am Med Informat Assoc* 2009, **16**:847–854.
8. Olson KL, Grannis SJ, Mandl KD: **Privacy protection versus cluster detection in spatial epidemiology.** *Am J Public Health* 2006, **96**:2002–2008.
9. Puett R, Lawson A, Clark A, Aldrich T, Porter D, Feigley C, Hebert J: **Scale and shape issues in focused cluster power for count data.** *Int J Health Geogr* 2005, **4**:8.
10. Goujon-Bellec S, Demoury C, Guyot-Goubin A, Hémon D, Clavel J: **Detection of clusters of a rare disease over a large territory: performance of cluster detection methods.** *Int J Health Geogr* 2011, **10**:53.
11. Jacquez GM: **Cluster morphology analysis.** *Spat Spatiotemporal Epidemiol* 2009, **1**:19–29.
12. Li X-Z, Wang J-F, Yang W-Z, Li Z-J, Lai S-J: **A spatial scan statistic for multiple clusters.** *Math Biosci* 2011, **233**:135–142.
13. Wang T-C, Yue C-SJ: **A binary-based approach for detecting irregularly shaped clusters.** *Int J Health Geogr* 2013, **12**:25.
14. Jones SG, Kulldorff M: **Influence of spatial resolution on space-time disease cluster detection.** *PLoS One* 2012, **7**:e48036.
15. Guttmann A, Ouchchane L, Li X, Perthus I, Gaudart J, Demongeot J, Boire J-Y: **Performance map of a cluster detection test using extended power.** *Int J Health Geogr* 2013, **12**:47.
16. Takahashi K, Tango T: **An extended power of cluster detection tests.** *Stat Med* 2006, **25**:841–852.
17. Griffith DA: **The boundary value problem in spatial statistical analysis.** *J Reg Sci* 1983, **23**:377–387.
18. Dreassi E, Biggeri A: **Edge effect in disease mapping.** *J Ital Stat Soc* 1998, **7**:267–283.
19. Meter EMV, Lawson AB, Colabianchi N, Nichols M, Hibbert J, Porter DE, Liese AD: **An evaluation of edge effects in nutritional accessibility and availability measures: a simulation study.** *Int J Health Geogr* 2010, **9**:40.
20. Blum H: **A transformation for extracting descriptors of shape.** In *Models Percept Speech Vis Forms.* Boston: MIT Press; 1967:362–380.
21. Thom R: **Sur le cut-locus d'une variété plongee.** *J Differ Geom* 1972, **6**:577–586.
22. Blum H: **Biological shape and visual science I.** *J Theor Biol* 1973, **38**:205–287.
23. Wolter F-E: *Cut Locus and Medial Axis in Global Shape Interrogation and Representation*, Sea Grant College Program, Massachusetts Institute of Technology. 1993.
24. Kulldorff M: **A spatial scan statistic.** *Commun Stat Theor M* 1997, **26**:1481–1496.
25. Kulldorff M, Nagarwalla N: **Spatial disease clusters: detection and inference.** *Stat Med* 1995, **14**:799–810.
26. Ribeiro SHR, Costa MA: **Optimal selection of the spatial scan parameters for cluster detection: a simulation study.** *Spat Spatiotemporal Epidemiol* 2012, **3**:107–120.
27. Chen C, Kim AY, Ross M, Wakefield J, Venkatraman ES: *SpatialEpi: Performs Various Spatial Epidemiological Analyses.* 2013.
28. Team RC: *R: A Language and Environment for Statistical Computing.* Vienna, Austria: 2012.
29. Keitt TH, Bivand R, Pebesma E, Rowlingson B: *Rgdal: Bindings for the Geospatial Data Abstraction Library.* 2012.
30. **Straight Skeleton Builder.** http://polyskeleton.appspot.com/.
31. **JTS Topology Suite.** http://tsusiatsoftware.net/jts/main.html.

doi:10.1186/1476-072X-13-15
Cite this article as: Guttmann *et al.*: Spatial heterogeneity of type I error for local cluster detection tests. *International Journal of Health Geographics* 2014 **13**:15.

Discussion générale et perspectives

Vis-à-vis des méthodes locales de détection d'agrégats, l'utilisation d'un indicateur global de performance oblige à résumer l'information et, par cette « réduction », implique de renoncer à une part de discrimination. L'utilisation d'indicateurs partiels, adaptés à l'évaluation de comportements spécifiques, demeure donc indispensable et complémentaire à l'évaluation des méthodes dans des contextes où ces comportements particuliers sont recherchés. Par exemple, un système de veille sanitaire sera plus intéressé par la sensibilité ou la valeur prédictive positive que par la spécificité.

Cependant, parce qu'ils constituent une mesure unique et synthétique de performance, les indicateurs globaux offrent l'avantage de permettre l'évaluation de très nombreuses simulations dont la masse résultante d'information devient suffisamment restreinte pour être interprétable. Nos évaluations spatiales systématiques de performance ont impliqué la conduite de 221 simulations (une par US), chaque simulation comptant 1 000 réalisations, et ce pour chacune des quatre combinaisons de niveaux de risques étudiés, soit un total 884 000 jeux de données analysés. L'utilisation d'un indicateur de performance unique rend possible la visualisation de l'ensemble de ces résultats sur quatre cartes.

Ce type d'évaluation spatiale systématique n'a, à notre connaissance, été réalisé que par Waller *et al.* [36] mais dans cette étude le choix d'indicateurs partiels de performance résulte en l'édition d'autant de cartes de performance que d'indicateurs utilisés (soit cinq dans cette étude). Ainsi, une évaluation semblable à celle que nous avons réalisée (quatre combinaisons de niveaux de risques pour chaque agrégat) aurait impliqué de démultiplier d'autant le nombre de cartes de performance (soit 20 cartes au total) rendant l'interprétation des résultats beaucoup plus difficile.

Enfin, les indicateurs globaux de performance permettent, en s'affranchissant de la barrière du nombre de simulations réalisables, d'envisager de modéliser les performances des méthodes locales de détection d'agrégats.

Cependant, même en tenant compte du développement considérable des capacités de calcul des ordinateurs actuels, envisager la conduite d'un très grand nombre de simulations nécessite de maîtriser les temps de calculs. En particulier pour les méthodes locales de détection d'agrégats où l'inférence statistique est souvent elle-même obtenue par méthode de Monte Carlo, ces temps de calculs peuvent être excessivement longs.

Du point de vue de la programmation statistique pure, l'utilisation du logiciel R [34] impose une programmation parallèle afin d'obtenir des temps de calculs acceptables. Cependant, le levier

le plus important dans l'optimisation des temps de calculs n'est pas dépendant du langage de programmation mais des protocoles de simulations eux-mêmes.

En effet, les temps de calculs sont directement dépendants du nombre de réalisations exécutées pour chaque simulation. Le choix du nombre de réalisations est dans la presque totalité des études complètement arbitraire et rarement justifié. La plupart des auteurs conduisent 1 000 réalisations pour chaque simulation mais parfois ce nombre diminue, en particulier lorsque le nombre de simulations est important. Ainsi certains auteurs ne conduisent que 250 réalisations [59] alors que d'autres en conduisent jusqu'à 100 000 [52].

L'estimation du nombre nécessaire de réalisations pour obtenir une bonne précision de la mesure de performance est rendue difficile par le fait que ce nombre est forcément dépendant de la performance elle-même. En effet, les mesures de performance sont soit des proportions (puissance usuelle et puissances conditionnelles, coefficient de Tanimoto cumulé), soit des moyennes de proportions (moyenne de sensibilité, spécificité, *etc.*), soit des moyennes pondérées de proportion (puissance étendue). Le nombre de réalisations pour estimer précisément la performance sera d'autant plus faible que celle-ci s'approche de 1 ou de 0. Au contraire, ce nombre sera maximum pour une performance égale à 0.5. L'attitude la plus conservatrice serait de calculer le nombre nécessaire de réalisations pour la situation la plus défavorable (performance cible de 0.5), mais cela au prix de temps de calculs beaucoup plus important que strictement nécessaires pour tous les autres cas.

Une recherche sur la convergence des indicateurs de performance apparaît utile voire indispensable pour, à défaut d'obtenir un nombre de réalisations nécessaires, proposer un protocole d'arrêt de la simulation lorsque la convergence de l'estimation de cette performance est satisfaisante.

Le travail mené sur la répartition spatiale de l'erreur de type I et l'effet de bord a permis de soulever plusieurs réflexions.

Premièrement, lors de l'analyse de données réelles, nos résultats indiquent directement que la détection à tort d'un agrégat au bord de la région d'étude est peu fréquente/probable. Cependant, à cet effet de bord dû à la méthode d'analyse elle-même, s'ajoute celui dû aux données non observées hors des limites de la région d'étude. Il est donc recommandé de mettre en œuvre des méthodes de prise en compte de l'effet de bord telles que les méthodes utilisant des zones de gardes internes

[69] ou externes (ou externes (où les données sont traitées comme des données manquantes et soumises à méthodes d'imputation [70]), ou encore des méthodes de pondération/correction des données en fonction de leur proximité au bord [71].

Ensuite, l'évaluation de l'effet de bord dans les études de performance nécessite des protocoles de simulation adaptés qui sont encore à développer. Plusieurs pistes sont à explorer dont la plus triviale consisterait à simuler des données pour une région plus grande que la région d'étude et incluant entièrement cette dernière, l'analyse ne portant ensuite que sur les données de la région d'étude.

Quels que soient leurs objectifs, les analyses portant sur des données spatiales, ne sont en fait jamais que purement spatiales. En effet, les données réelles analysées concernent une période de temps particulière. Lorsque l'épidémiologiste ne souhaite pas étudier les variations temporelles d'un phénomène, les périodes analysées sont en général longues afin de lisser d'éventuels phénomènes saisonniers. Pour certaines pathologies dont l'incidence est très faible, les périodes analysées sont également longues du simple fait de la rareté des cas que l'on souhaitera accumuler afin d'en minimiser la variance [39]. L'utilisation de méthodes purement temporelles est privilégiée lorsque la répartition spatiale des cas revêt une importance secondaire face à leur quantification. C'est par exemple le cas lorsqu'une pathologie est surveillée dans l'objectif du déclenchement d'une alerte épidémiologique en cas de franchissement d'un seuil épidémique. Dans ces systèmes, les méthodes spatiales sont le plus souvent utilisées en seconde ligne lors d'investigations secondaires menées lors d'alertes épidémiques. Dans ces investigations, les périodes analysées sont courtes, correspondant à la période épidémique en cours d'investigation. Les méthodes spatio-temporelles sont réservées le plus souvent aux situations intermédiaires, lorsque le phénomène étudié, soumis à des variations temporelles que l'on souhaite analyser, est également susceptible d'être fortement spatialisé pouvant masquer une épidémie lors d'une analyse purement temporelle.

La méthodologie d'évaluation des méthodes d'analyse de série temporelles est plus développée que pour les méthodes purement spatiales. Des indicateurs de performances spécifiques existent et leurs qualités et limites sont bien connues. Il s'agit des ARL (pour « average run length ») [72] ou nombre moyen d'observations avant déclenchement d'une alerte lorsque le processus est sous-

contrôle (absence d'épidémie – correspond à l'erreur de type I) ou hors-contrôle (en phase épidémique – correspond à l'erreur de type II).

L'évaluation des méthodes spatio-temporelles rencontre les mêmes difficultés que l'évaluation des méthodes purement spatiales avec une méthodologie encore peu développée à l'heure actuelle et des indicateurs de performances très hétérogènes. Le développement d'indicateurs globaux de performance pour ces méthodes nécessite de prendre en compte, en plus de la puissance usuelle, non seulement la précision de localisation de l'agrégat dans l'espace mais également dans le temps. Dans ce cadre, le coefficient de Tanimoto cumulé offre une perspective intéressante car il est aisément adaptable à l'ajout de la dimension temporelle.

Conclusion

L'épidémiologie spatiale, bien qu'en développement depuis les années 50, est un domaine qui a encore besoin d'étoffer ses propres méthodologies d'évaluations. La complexité des méthodes disponibles pour l'analyse de données spatiales ainsi que leur nombre sans cesse grandissant, imposent un choix de plus en plus difficile pour l'épidémiologiste quant à l'utilisation de méthodes adaptées aux données dont il dispose et aux objectifs qui lui sont fixés. Un tel choix doit reposer sur une base de connaissances solides qui dépasse la seule connaissance du fondement théorique des méthodes disponibles.

Puisqu'aujourd'hui plusieurs méthodes sont envisageables quel que soit le contexte d'utilisation, connaitre leurs performances respectives et être capable de les discriminer est devenu essentiel. Sans cette base de connaissance, le choix des méthodes ne repose plus, en pratique, que sur leur facilité d'utilisation (logiciel libre d'utilisation simple et résultats facilement interprétables). Les méthodes locales de détection d'agrégats sont très utilisées car elles permettent de localiser un agrégat potentiel sans point de source prédéfini et en donne une portée statistique *via* une méthode d'inférence. De plus, ces méthodes sont pour la plupart disponibles dans des logiciels libres, dont certains (SaTScan™ [73] ou FleXScan [74]) ne nécessitent pas de programmation statistique car ils possèdent des interfaces graphiques dédiées aux méthodes qu'ils implémentent.

La revue de la littérature conduit actuellement au double constat (1) d'un défaut de consensus des méthodes d'évaluation et de comparaison des performances de ce type de tests et (2) du défaut informationnel des indicateurs employés pour la quantification de ces performances tous incapables d'appréhender à la fois la puissance et la juste localisation de l'agrégat simulé.

En l'absence d'une méthodologie reconnue, les évaluations proposées sont entachées d'une hétérogénéité importante. Elles sont le plus souvent partielles, tant sur le plan de ce qui est évalué (hypothèses d'agrégation restrictives, effet de bord, *etc.*) que sur le protocole d'évaluation (indicateurs de performances). De plus, certains éléments de ces protocoles sont rarement justifiés (nombre de réalisations) et d'autres parfois omis (statistique de résumé de la mesure d'intérêt).

Nous avons donc proposé de nouveaux indicateurs de quantification de la performance palliant les défauts habituels, dont un, l'aire sous la courbe de puissance étendue, est issu de travaux en statistiques spatiales [1] et l'autre, le coefficient cumulé de Tanimoto, dérive notamment des techniques d'analyse en imagerie [75].

Nous proposons également une approche d'évaluation intensive et systématique qui pourrait contribuer à la création d'un standard en termes d'évaluation des tests de détection d'agrégats puisqu'elle permet une évaluation sur l'intégralité des zones géographiques concernées et sur une gamme potentiellement infinie de configurations géographiques et épidémiologiques, y compris les plus réalistes et, donc, les plus utiles.

Enfin, l'évaluation des méthodes locales de détection d'agrégats ne se limitent pas à leurs performances lorsqu'un ou plusieurs agrégats sont présents. L'étude de l'erreur de type I pour ces tests a nécessité la création d'outils permettant d'en évaluer la répartition spatiale. Lors de l'étude de l'hétérogénéité spatiale de l'erreur de type I, nous avons développé une statistique, ainsi que la méthode d'inférence associée, permettant de tester la significativité d'un « pattern » d'hétérogénéité assimilable à un effet de bord, ce dernier étant une préoccupation très fréquente en statistiques spatiales.

Evidemment, même si les développements que nous proposons constituent une contribution utile à notre communauté scientifique, la problématique de l'évaluation des méthodes d'analyse en épidémiologie spatiale reste encore ouverte. D'une part du fait de l'importance des méthodes de simulation de données en elles-mêmes (dans leur capacité à représenter une réalité complexe) et d'autre part du fait que ces évaluations dépassent le cadre des méthodes locales de détection d'agrégats étudiées dans ce travail.

En conclusion, et de façon synthétique, notre travail de thèse s'est attaché à fournir quelques éléments de réponse à cette problématique d'évaluation. Le développement d'indicateurs globaux de performance est une étape nécessaire dont on peut espérer qu'elle favorisera la comparabilité des études de performance tout en levant certains écueils inhérents à l'évaluation de contextes épidémiologiques complexes. Une partie des travaux réalisés s'appuie sur des travaux reconnus qu'il a fallu adapter et développer afin qu'ils puissent répondre à nos objectifs. Enfin, une autre partie de nos travaux est purement originale puisqu'aucune étude de la répartition spatiale de l'erreur de type I n'avait jusqu'ici été publiée.

Bibliographie

1. Takahashi K, Tango T: **An extended power of cluster detection tests**. *Stat Med* 2006, **25**:841–852.

2. Tango T: *Statistical Methods for Disease Clustering*. New York, NY: Springer New York; 2010. [*Statistics for Biology and Health*]

3. Diggle PJ: *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition*. Édition : 3. Chapman and Hall/CRC; 2013.

4. Pornon H: *Systèmes D'information Géographique, Pouvoir et Organisations: Géomatique et Stratégies D'acteurs*. Editions L'Harmattan; 1998.

5. Poljak Z, Dewey CE, Rosendal T, Friendship RM, Young B, Berke O: **Spread of porcine circovirus associated disease (PCVAD) in Ontario (Canada) swine herds: Part I. Exploratory spatial analysis**. *BMC Vet Res* 2010, **6**:59.

6. Vieira VM, Hart JE, Webster TF, Weinberg J, Puett R, Laden F, Costenbader KH, Karlson EW: **Association between residences in U.S. northern latitudes and rheumatoid arthritis: A spatial analysis of the Nurses' Health Study**. *Environ Health Perspect* 2010, **118**:957–961.

7. Gruebner O, Khan MMH, Lautenbach S, Müller D, Kraemer A, Lakes T, Hostert P: **A spatial epidemiological analysis of self-rated mental health in the slums of Dhaka**. *Int J Health Geogr* 2011, **10**:36.

8. Fritz CE, Schuurman N, Robertson C, Lear S: **A scoping review of spatial cluster analysis techniques for point-event data**. *Geospatial Health* 2013, **7**:183–198.

9. D' Orsi E, Carvalho MS, Cruz OG: **Similarity between neonatal profile and socioeconomic index: a spatial approach**. *Cad Saúde Pública* 2005, **21**:786–794.

10. Nelson EJ, Hughes J, Kulasingam SL: **Spatial patterns of human papillomavirus-associated cancers within the state of Minnesota, 1998-2007**. *Spat Spatio-Temporal Epidemiol* 2014, **9**:13–21.

11. Torabi M, Rosychuk RJ: **An examination of five spatial disease clustering methodologies for the identification of childhood cancer clusters in Alberta, Canada**. *Spat Spatio-Temporal Epidemiol* 2011, **2**:321–330.

12. Lin H, Ning B, Li J, Ho SC, Huss A, Vermeulen R, Tian L: **Lung Cancer Mortality Among Women in Xuan Wei, China: A Comparison of Spatial Clustering Detection Methods**. *Asia-Pac J Public Health Asia-Pac Acad Consort Public Health* 2012.

13. Avilés LA, Alvelo-Maldonado L, Padró-Mojica I, Seguinot J, Jorge JC: **Risk factors, prevalence trend, and clustering of hypospadias cases in Puerto Rico**. *J Pediatr Urol* .

14. Baastrup Nordsborg R, Meliker JR, Kjær Ersbøll A, Jacquez GM, Raaschou-Nielsen O: **Space-time clustering of non-hodgkin lymphoma using residential histories in a Danish case-control study**. *PloS One* 2013, **8**:e60800.

15. Hennebelle JH, Sykes JE, Carpenter TE, Foley J: **Spatial and temporal patterns of Leptospira infection in dogs from northern California: 67 cases (2001-2010)**. *J Am Vet Med Assoc* 2013, **242**:941–947.

16. Besag J, Newell J: **The Detection of Clusters in Rare Diseases**. *J R Stat Soc Ser A Stat Soc* 1991, **154**:143–155.

17. Kulldorff M: **Statistical methods for spatial epidemiology: tests for randomness**. In *GIS Health*. *Volume 6*; 1998:49.

18. Tango T: **Disease Mapping: Visualization of Spatial Clustering**. In *Stat Methods Dis Clust*. New York, NY: Springer New York; 2010. [*Statistics for Biology and Health*]

19. Bivand RS, Pebesma E, Gómez-Rubio V: **Disease Mapping**. In *Appl Spat Data Anal R*. Springer New York; 2013:319–361. [*Use R!*, vol. 10]

20. Openshaw S, Charlton M, Wymer C, Craft A: **A mark 1 geographical analysis machine for the automated analysis of point data sets**. *Int J Geogr Inf Syst* 1987, **1**:335–358.

21. Kulldorff M, Nagarwalla N: **Spatial disease clusters: detection and inference**. *Stat Med* 1995, **14**:799–810.

22. Kulldorff M: **a spatial scan statistic**. *Commun Stat Theor M* 1997, **26**:1481–1496.

23. Ord JK, Getis A: **Local Spatial Autocorrelation Statistics: Distributional Issues and an Application**. *Geogr Anal* 1995, **27**:286–306.

24. Scrucca L: **Clustering multivariate spatial data based on local measures of spatial autocorrelation**. *Quad Dipartimento Econ Finanza E Stat* 2005, **20**:1–11.

25. Song C, Kulldorff M: **Tango's maximized excess events test with different weights**. *Int J Health Geogr* 2005, **4**:32.

26. Tiefelsdorf M, Griffith DA, Boots B: **A variance-stabilizing coding scheme for spatial link matrices**. *Environ Plan A* 1999, **31**:165 – 180.

27. MORAN PAP: **Notes on continuous stochastic phenomena**. *Biometrika* 1950, **37**:17–23.

28. Anselin L: **Local Indicators of Spatial Association—LISA**. *Geogr Anal* 1995, **27**:93–115.

29. Getis A, Ord JK: **The analysis of spatial association by use of distance statistics**. *Geogr Anal* 1992, **24**:189–206.

30. Dwass M: **Modified Randomization Tests for Nonparametric Hypotheses**. *Ann Math Stat* 1957, **28**:181–187.

31. Dunn OJ: **Multiple Comparisons among Means**. *J Am Stat Assoc* 1961, **56**:52–64.

32. Holm S: **A simple sequentially rejective multiple test procedure**. *Scand J Stat* 1979:65–70.

33. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *J R Stat Soc Ser B Methodol* 1995, **57**:289–300.

34. R Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2014.

35. Jacquez GM: **Cluster Morphology Analysis**. *Spat Spatio-Temporal Epidemiol* 2009, **1**:19–29.

36. Waller LA, Hill EG, Rudd RA: **The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations**. *Stat Med* 2006, **25**:853–865.

37. Jackson MC, Huang L, Luo J, Hachey M, Feuer E: **Comparison of tests for spatial heterogeneity on data with global clustering patterns and outliers**. *Int J Health Geogr* 2009, **8**:55.

38. Lemke D, Mattauch V, Heidinger O, Pebesma E, Hense H-W: **Detecting cancer clusters in a regional population with local cluster tests and Bayesian smoothing methods: a simulation study**. *Int J Health Geogr* 2013, **12**:54.

39. Huang L, Pickle LW, Das B: **Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases**. *Stat Med* 2008, **27**:5111–5142.

40. Aamodt G, Samuelsen SO, Skrondal A: **A simulation study of three methods for detecting disease clusters**. *Int J Health Geogr* 2006, **5**:15.

41. Li X-Z, Wang J-F, Yang W-Z, Li Z-J, Lai S-J: **A spatial scan statistic for multiple clusters**. *Math Biosci* 2011, **233**:135–142.

42. Jung I, Lee H: **Spatial cluster detection for ordinal outcome data**. *Stat Med* 2012:n/a–n/a.

43. Ozonoff A, Jeffery C, Manjourides J, White LF, Pagano M: **Effect of spatial resolution on cluster detection: a simulation study**. *Int J Health Geogr* 2007, **6**:52.

44. Jones, SG, Kulldorff, M: **Influence of Spatial Resolution on Space-Time Disease Cluster Detection**. *PLoS ONE* 2012, **7**:e48036.

45. Viega J: **Practical random number generation in software**. In *Comput Secur Appl Conf 2003 Proc 19th Annu*. IEEE; 2003:129–140.

46. Christophe D, Petr S: *Randtoolbox: Generating and Testing Random Numbers*. 2014.

47. Matsumoto M, Nishimura T: **Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudo-random Number Generator**. *ACM Trans Model Comput Simul* 1998, **8**:3–30.

48. Wichura MJ: **Algorithm AS 241: The percentage points of the normal distribution**. *Appl Stat* 1988:477–484.

49. Ahrens JH, Dieter U: **Computer Generation of Poisson Deviates from Modified Normal Distributions**. *ACM Trans Math Softw* 1982, **8**:163–179.

50. Kachitvichyanukul V, Schmeiser BW: **Binomial random variate generation**. *Commun ACM* 1988, **31**:216–222.

51. Burton A, Altman DG, Royston P, Holder RL: **The design of simulation studies in medical statistics**. *Stat Med* 2006, **25**:4279–4292.

52. Kulldorff M, Tango T, Park PJ: **Power comparisons for disease clustering tests**. *Comput Stat Data Anal* 2003, **42**:665–684.

53. Duczmal L, Kulldorff M, Huang L: **Evaluation of spatial scan statistics for irregularly shaped clusters**. *J Comput Graph Stat* 2006, **15**.

54. Assunção R, Costa M, Tavares A, Ferreira S: **Fast detection of arbitrarily shaped disease clusters**. *Stat Med* 2006, **25**:723–742.

55. Cook AJ, Gold DR, Li Y: **Spatial Cluster Detection for Censored Outcome Data**. *Biometrics* 2007, **63**:540–549.

56. Huang L, Kulldorff M, Gregorio D: **A spatial scan statistic for survival data**. *Biometrics* 2007, **63**:109–118.

57. Zhang T, Lin G: **Spatial scan statistics in loglinear models**. *Comput Stat Data Anal* 2009, **53**:2851–2858.

58. Jung I, Kulldorff M, Richard OJ: **A spatial scan statistic for multinomial data**. *Stat Med* 2010, **29**:1910–1918.

59. Goujon-Bellec S, Demoury C, Guyot-Goubin A, Hémon D, Clavel J: **Detection of clusters of a rare disease over a large territory: performance of cluster detection methods**. *Int J Health Geogr* 2011, **10**:53.

60. Zhang T, Zhang Z, Lin G: **Spatial scan statistics with overdispersion**. *Stat Med* 2012, **31**:762–774.

61. Wang T-C, Yue C-SJ: **A binary-based approach for detecting irregularly shaped clusters**. *Int J Health Geogr* 2013, **12**:25.

62. Tango T, Takahashi K: **A flexibly shaped spatial scan statistic for detecting clusters**. *Int J Health Geogr* 2005, **4**:11.

63. Song C, Kulldorff M: **Power evaluation of disease clustering tests**. *Int J Health Geogr* 2003, **2**:9.

64. Guttmann A, Ouchchane L, Li X, Perthus I, Gaudart J, Demongeot J, Boire J-Y: **Performance map of a cluster detection test using extended power**. *Int J Health Geogr* 2013, **12**:47.

65. Lawson AB: **Scales of Measurement and Data Availabitlity**. In *Stat Methods Spat Epidemiol*. 2nd edition. Wiley; 2006.

66. Cressie N: *Statistics for Spatial Data*. Édition : Revised Edition. New York: Wiley-Blackwell; 1993.

67. Meter EMV, Lawson AB, Colabianchi N, Nichols M, Hibbert J, Porter DE, Liese AD: **An evaluation of edge effects in nutritional accessibility and availability measures: a simulation study**. *Int J Health Geogr* 2010, **9**:40.

68. Guttmann A, Li X, Gaudart J, Gérard Y, Demongeot J, Boire J-Y, Ouchchane L: **Spatial heterogeneity of type I error for local cluster detection tests**. *Int J Health Geogr* 2014, **13**:15.

69. Ripley BD: *Statistical Inference for Spatial Processes*. Édition : Reprint. Cambridge England; New York: Cambridge University Press; 1991.

70. Tanner MA: *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Édition : 3rd ed. 1996. Corr. 2nd printing 1998. New York: Springer-Verlag New York Inc.; 1997.

71. Lawson AB: **Exploratory Approaches, Parametric Estimation and Inference**. In *Stat Methods Spat Epidemiol*. 2nd edition. Wiley; 2006.

72. Hawkins DM, Olwell DH: *Cumulative Sum Charts and Charting for Quality Improvement*. Édition : Softcover reprint of the original 1st ed. 1998. New York, NY: Springer-Verlag New York Inc.; 1998.

73. Kulldorff M, Information Management Services, Inc.: *Software for the Spatial and Space-Time Scan Statistics.* 2009.

74. Takahashi K, Yokoyama T, Tango T: *Software for the Flexible Scan Statistic*. Japon: National Institute of Public Health; 2010.

75. Rogers DJ, Tanimoto TT: **A Computer Program for Classifying Plants**. *Science* 1960, **132**:1115–1118.

Liste des publications & communications

Publications

A. Guttmann, L. Ouchchane, X. Li, I. Perthus, J. Gaudart, J. Demongeot, et J.-Y. Boire, « *Performance map of a cluster detection test using extended power* », Int. J. Health Geogr., vol. 12, no 1, p. 47, oct. 2013.

A. Guttmann, X. Li, J. Gaudart, Y. Gérard, J. Demongeot, J.-Y. Boire, et L. Ouchchane, « *Spatial heterogeneity of type I error for local cluster detection tests* », Int. J. Health Geogr., vol. 13, no 1, p. 15, mai 2014.

Communications orales

A. Guttmann, X. Li, J. Gaudart, J. Demongeot, J.-Y. Boire, et L. Ouchchane, « *Détection d'agrégats : carte de performance utilisant le coefficient de Tanimoto* », Rev. D'Épidémiologie Santé Publique, vol. 62, p. S196, sept. 2014.

Communications affichées

Guttmann A, Ouchchane L, Gaudart J, Demongeot J, Perthus I, Boire JY. *Détection d'agrégats : carte de puissance d'une méthode de balayage* (Clermont-Ferrand, France). Poster congrès ADELF EMOI. 2012, Dijon.

Annexes

1 Programmes R : puissance étendue et Tanimoto cumulé

Le programme présenté n'est pas celui utilisé dans les travaux : les chemins d'accès aux fichiers ont été remplacés par des chemins factices ne conservant que l'organisation des dossiers. Le programme original est composé de plusieurs scripts replacés ici dans un seul script.

Ce programme permet la simulation spatiale systématique d'un agrégat, l'analyse et la visualisation des résultats sur une carte de performance.

```
library(rgdal)
library(foreach)
library(SpatialEpi)
library(doSNOW)

## désignation des agrégats : 1 agrégat par US, 1 agrégat = 1 US et ses 3 plus proches voisines
Auvergne <- readOGR("C/carte/code_postaux_juil09_region.shp","code_postaux_juil09_region")
coords <- as.data.frame(coordinates(Auvergne))
coords <- cbind(coords,Auvergne$CPCOND)
colnames(coords) <- c("x","y","CPCOND")
listVoisins <- knearneigh(coordinates(Auvergne), k=3, longlat = NULL)
indexVoisins <- listVoisins$nn
LocClusters <- data.frame(centre = coords$CPCOND, V1 = rep(NA,221), V2 = rep(NA,221), V3 = rep(NA,221))
CPCOND <- as.character(coords$CPCOND)
for (j in 1:3){
  for (i in 1:221) {LocClusters[i,j+1] <-  CPCOND[indexVoisins[i,j]]}
}
LocClusters$centre <- as.character(LocClusters$centre)cl <- makeCluster(4)

##Génération des jeux de données
#chargement des données sources (taille de la population à risque par US et par année)
load('data sim/Cas.rda')
tableCas <- merge(tableCas,coords,by = c("CPCOND"), all.x = TRUE)
```

```
Population <- aggregate(tableCas[,"Population"],by = list(CPCOND = tableCas[,"CPCOND"]),FUN = mean)
colnames(Population) <- c("CPCOND","Population")
Sim <- merge(coords,Population, by = "CPCOND")
for (a in 1:221) {
  SauvSim <- data.frame()
  for (b in 1:1000) {
    ## un jeu de données
    for (i in 1:221){
      ifelse(Sim[i,"CPCOND"]%in% LocClusters[a,], Sim[i,"Observed"] <- rpois(1,Rclust*Sim[i,"Population"]),Sim[i,"Observed"] <- rpois(1,Rbase*Sim[i,"Population"]))
    }
    Expected <- expected(Sim$Population,Sim$Observed, 1)
    Simx <- cbind(Sim, Expected)
    indice <-  cbind(Simx,indice = rep(b,221) )
    SauvSim <-  rbind(SauvSim, indice)
    save(SauvSim,file = paste("C :/Datasets/0.0226_3R /Dataset", a , ".rda", sep = ""))
}

##Analyses
registerDoSNOW(cl)
ptm1 <- Sys.time()
foreach(a= c(1:221),.combine = list) %do% {
   load(paste("C:/Datasets/ 0.0226_3R/Dataset",a,".rda", sep = ""))
   ListIDmlClust <- foreach(b=1:1000, .inorder=FALSE, .packages="SpatialEpi") %dopar% {
      ## récupération du jeu de données
      ptm <- Sys.time()
      Simx <- SauvSim[SauvSim[,"indice"]== b,]
      ## Kulldorff using Poisson likelihoods
      scanKN <- kulldorff(Simx[,c(2,3)],Simx$Observed, Simx$Population, Simx$Expected,0.5,999,0.05,FALSE)
      ## traitement des résultats du scan de KN
      cluster1 <- scanKN$most.likely.cluster$location.IDs.included
      PowerClust <- matrix(rep(0,5*221),nrow = 221, ncol = 5, byrow = TRUE,dimnames = list(l = c(1:221),s = c(0:4)))
      s <- 0
      if (scanKN$most.likely.cluster$p.value < 0.05){
```

```
      for (i in 1:length(cluster1)){
        ifelse(Simx[cluster1[i],"CPCOND"]%in% LocClusters[a,], s <- s+1, s <- s)}
      PowerClust[length(cluster1),s+1] <- PowerClust[length(cluster1),s+1] + 1}
          result <- list(cluster1, scanKN$most.likely.cluster$p.value,format(Sys.time() - ptm), PowerClust)
          }
  matrice <- foreach(b=1:1000, .inorder=FALSE, .combine = "+") %dopar% {ListIDmlClust[[b]][[4]]}
  #addmargins(matrice)
  save(ListIDmlClust, file = paste("C:/Users/alguttma/6 - PhD/3_firstresults/resultsim/ListIDmlClust",a,".rda", sep = ""))
  save(matrice,  file = paste("C:/resultsim/RR3_RbAll/matrice",a,".rda", sep = ""))
}
tEnd <- format(Sys.time() - ptm1)
stopCluster(cl)

## traitement des résultats : aire sous la courbe de puissance étendue
Risks <-  "RR3_RbAll/"
WD <- paste("C:/resultsim/",Risks, sep = "")
Wx <- function(r,lx,sx){sqrt( (1 - min(c(0.25*(4-sx),1))) * (1 - min(c((r/4)*(lx-sx),1))))}
W <- matrix(rep(0,5*221),nrow = 221, ncol = 5, byrow = TRUE,dimnames = list(l = c(1:221),s = c(0:4)))
W.r <- list()
for (i in seq(0,1,0.001)){
   for (lx in c(1:221)){for (sx in c(0:4)) {ifelse(sx > lx,  W[lx,sx+1] <- 0, W[lx,sx+1] <- Wx(i,lx,sx))}}
   W.r[[1+i*1000]] <-  W
}

listfichier <- list.files(WD, pattern = ("^matrice"))
a <- foreach(i = 1:length(listfichier)) %do% {
  n <- strsplit(strsplit(listfichier[i], split = ("^(matrice)"))[[1]][2],split = (".rda$"))[[1]]
  load(paste(WD,listfichier[i],sep = ""))
  I.ls <- rep(NA,1001)
  for (i in seq(0,1,0.001)){I.ls[1+ i*1000] <- sum((matrice/1000) * W.r[[1+ i*1000]])}
  list(n,I.ls)
}
IDclust <- foreach(b=seq_along(a), .combine = c) %do% {a[[b]][[1]]}

##matrice contenant les valeurs de puissance étendue pour construire la courbe
```

```
EPcurve <- foreach(b=seq_along(a),.combine = cbind) %do% {a[[b]][[2]]}
#nommer les colonnes selon l'id du cluster et les trier par ordre d'id cluster
colnames(EPcurve) <- as.numeric(IDclust)
dd <- cbind(as.numeric(dimnames(EPcurve)[[2]]),1:ncol(EPcurve))
dd <- dd[order(dd[,1]),]
EPcurve <- EPcurve[ ,order(as.numeric(dimnames(EPcurve)[[2]]))]

##calcul de l'auc pour chaque cluster
auc <- foreach(i = 1:ncol(EPcurve),.combine = rbind) %do% {
  I.ls <- EPcurve[,i]
  Q <- rep(NA,1000)
  for (x in 1:1000) {Q[x] <- 0.001*I.ls[x+1] + (0.001*(I.ls[x]-I.ls[x+1]))/2}
  cbind(as.numeric(colnames(EPcurve)[i]),sum(Q))
}
auc <- as.data.frame(auc)
colnames(auc) <- c("NumClust","auc")

## traitement des résultats : coefficient de Tanimoto
dico <- data.frame(NumUS = c(1:221), CPCOND = coords$CPCOND)
x <- apply(LocClusters[,],1, function(x) x[1] == dico$CPCOND | x[2] == dico$CPCOND | x[3] == dico$CPCOND | x[4] == dico$CPCOND)
MALADES <- x*1

listfichier <- list.files(paste("C:/resultsim/",Risks, sep = ""), pattern = ("^ListIDmlClust"))

a <- foreach(i = 1:221,.combine = rbind) %do% {
  n <- strsplit(strsplit(listfichier[i], split = ("^(ListIDmlClust)"))[[1]][2],split = (".rda$"))[[1]]
  n <- as.integer(n)
  load(paste("C:/resultsim/",Risks,listfichier[i],sep = ""))
  b <- foreach(j = 1:1000,.combine = rbind) %do% {
    POSITIF <- rep(0,221)
    POSITIF[ListIDmlClust[[j]][[1]]] <- 1
    MALADE <- MALADES[,n]
    VP <- sum(MALADE == 1 & POSITIF  == 1)
    FN <- sum(MALADE == 1 & POSITIF  == 0)
    FP <- sum(MALADE == 0 & POSITIF  == 1)
```

```
    VN <- sum(MALADE == 0 & POSITIF  == 0)
    data.frame(NumClust = n, Numsim = j, Tanimoto = VP/(VP+FN+FP) ,Pval = ListIDmlClust[[j]][[2]], VP = VP, FP = FP, VN = VN, FN = FN)
  }
  save(b, file = paste(WD,"b",n,".rda",sep = ""))
}

Risks <-  "RR3_RbAll/"
 WD <- paste("C:/ resultsim /",Risks, sep = "")
  BrutRes <- foreach(i = 1:221, .combine = rbind) %do% {
   load(file = paste(WD,"b",i,".rda",sep = ""))
   b_ <- b
   b_[b_$Pval >= 0.05,c(5:6)] <-  0
   b_[b_$Pval >= 0.05,c(7)] <-  217
   b_[b_$Pval >= 0.05,c(8)] <-  4
   csum <- cumsum(b_$VP) / cumsum(b_$VP+b_$FN+b_$FP)
   data.frame(Risks = Risks, NumClust = i,NumSim = 1:1000, Tsum = csum)
 }
 save(BrutRes, file = "C:/BrutRes.rda")

## Traitement des résultats : carte de performance
echelle = list("SpatialPolygonsRescale", layout.scale.bar(),offset = c(710000,2182000), scale = 40000, fill=c("transparent","black"))
text1 = list("sp.text", c(710000,2190000), "0", cex = 0.5)
text2 = list("sp.text", c(750000,2190000), "40 km", cex = 0.5)
arrow = list("SpatialPolygonsRescale", layout.north.arrow(),offset = c(740000,2160000), scale = 15000)

#pour le Tanimoto cumulé, la procédure pour l'aucep est identique
Auvergne@data <- BrutRes[BrutRes$NumSim == 1000,]

jpeg(filename = "C:/Users/alguttma/6 - PhD/redaction/TC-fig1.jpg",
    width = 84, height = 100, units = "mm", pointsize = 6,
    quality = 100,res = 600, antialias = "cleartype")

spplot(obj = Auvergne, zcol = c("Tsum"), sp.layout=list(echelle,text1,text2,arrow),col.regions= gray(seq(0.2,1,length = 16)),col = "white", lwd = 0.5)
dev.off()
```

2 Programme R : répartition spatiale de l'erreur de type I

```
library(foreach)
library(SpatialEpi)
library(doSNOW)

# données sources contenant les tailles de population à risque
load("C:/ Dataset1.rda")
SauvSim <- SauvSim[, c(1:4,7)]

cl <- makeCluster(8)
registerDoSNOW(cl)

#incidence de base
Rb <- 0.0226

#simulation et analyse des données
foreach(a=1:20, .inorder=FALSE, .verbose = TRUE) %do% {
   ListIDmlClust <- foreach(b=1:1000, .inorder=FALSE, .packages="SpatialEpi") %dopar% {
      ## récupération du jeu de données
      SauvSim[SauvSim$indice == b,"Observed"] <- rpois(221,Rbase*SauvSim[SauvSim$indice ==
b,"Population"])
      SauvSim[SauvSim$indice == b,"Expected"] <- expected(SauvSim[SauvSim$indice ==
b,"Population"],SauvSim[SauvSim$indice == b,"Observed"],1)
      Simx <- SauvSim[SauvSim$indice == b,]
      ## Kulldorff using Poisson likelihoods
      scanKN <- kulldorff(Simx[,c(2,3)],Simx$Observed, Simx$Population,
Simx$Expected,0.5,999,0.05,FALSE)
      ## traitement des résultats du scan de KN
      cluster1 <- scanKN$most.likely.cluster$location.IDs.included
      return(list(cluster1, scanKN$most.likely.cluster$p.value, Simx))
   }
 save(ListIDmlClust,  file = paste("C:/ results/ListIDmlClust",a,"_",Rbase,".rda", sep = ""))
  }
   stopCluster(cl)

# étape 1 : récuperer et compiler les résultats dans un objet
```

```
wd <- "C:/results"
Auvergne <- readOGR(dsn="D:/AG/carte",layer = "carte")
# liste de fichiers contenus dans le répertoire des résultats
listfichier <- list.files(wd, pattern = ("^ListIDmlClust"))
Rbase <- 0.0226
cl <- makeCluster(8)
registerDoSNOW(cl)

# caractéristiques générales des agrégats détectés à tort
WDC <- foreach(rb = Rbase, .inorder=TRUE, .combine = rbind, .verbose = TRUE)%do%{
  #fichiers cluster
  x <- listfichier[regexpr(rb, listfichier) != -1]
  Clustrb <- foreach(i = 1:length(x), .inorder=TRUE, .combine = rbind) %do% {
    load(paste(wd,"/",x[i],sep = ""))
    Clustrb.b <- foreach(j = 1:1000, .combine = rbind, .packages = c("rgeos","maptools")) %dopar% {
        Cluster <-polygons(Auvergne)[ListIDmlClust[[j]][[1]]]
        centroid <- gCentroid(Cluster)
        #il faut "merger" les polygons du cluster avant de pouvoir calculer l'aire
        km2 <-area.poly(as(unionSpatialPolygons(Cluster, rep(1, length(Cluster))), "gpc.poly"))/1000000
        data.frame(Rbase = rb, NumSim = j +((i-1)*1000), pval = ListIDmlClust[[j]][[2]], size =
length(ListIDmlClust[[j]][[1]]), km2 = km2, x = centroid$x , y = centroid$y)
    }
  }
}

#US appartenant aux agrégats détectés à tort
US <- foreach(rb = Rbase,.inorder=TRUE, .verbose = TRUE)%do%{
  #fichiers US
  x <- listfichier[regexpr(rb, listfichier) != -1]
  USrb <- foreach(i = 1:length(x), .combine = rbind) %do% {
    load(paste(wd,"/",x[i],sep = ""))
    USrb.b <- foreach(j=1:1000, .inorder=FALSE, .combine = rbind) %dopar% {
        US <- ListIDmlClust[[j]][[1]]
        data.frame(NumSim = rep(j +((i-1)*1000), length(US)), US = US, pval =
rep(ListIDmlClust[[j]][[2]], length(US)))
    }
  }
  list(rb, USrb)
}
stopCluster(cl)
```

```
# calcul des taux de participation des US à l'erreur de type I
alpha <- 0.05
# Calcul du PR pour les US des WDC avec pval < alpha
USPR <- NULL
for(i in 1:length(US)){
  USPR <- rbind(USPR, cbind(US[[i]][[2]], Rb = rep(US[[i]][[1]], nrow(US[[i]][[2]]))))
}
USPR <- USPR[USPR$pval < alpha,]
USPR.w <-  WDC[WDC$pval < 0.05,c("Rbase","NumSim","size")]

x <- NULL
for(i in 1:length(US)){
x <- c(x,rep(1/table(USPR.w$Rbase)[i], table(USPR.w$Rbase)[i]))
}
USPR.w <- cbind(USPR.w, w = x/USPR.w$size)
colnames(USPR.w)[1] <- "Rb"

USPR <- merge(USPR.w[,c(1,2,4)],USPR, by = c("NumSim", "Rb"), all.x = TRUE)
# calcul du taux de participation « PR »
USPR <- data.frame(US= c(1:221),Rb = rep(sort(unique(USPR$Rb)),each = 221),  PR =
as.vector(tapply(USPR$w, list(USPR$US, USPR$Rb), FUN = sum)))

rm(x); rm(i); rm(USPR.w)

#effet de bord : Calcul de la statistique E et test
# préparation des données et calcul du coefficient de bord ou edge coefficient « coef.lit »
US.dist <- data.frame(x = coordinates(Auvergne)[,1],y = coordinates(Auvergne)[,2])
coordinates(US.dist) = ~x+y
proj4string(US.dist) <- CRS(proj4string(Auvergne))
USPR.wide <- reshape(USPR, v.names = "PR", idvar = "US", timevar = "Rb", direction = "wide")
USPR.wide <- cbind(USPR.wide, Dist = rep(t(gDistance(US.dist, spgeom2=carte, byid=TRUE))[,1]))

Auvergne_grid <- spsample(Auvergne,n=500000,type="regular")
carte <- as(unionSpatialPolygons(Auvergne, rep(1, length(Auvergne))),"SpatialLines")
Dist.grid <- gDistance(Auvergne_grid, carte,byid=TRUE)
Dist.grid <- as.data.frame(t(Dist.grid))
colnames(Dist.grid) <- "Dist"

a <- -2/max(Dist.grid$Dist)
b <- 1
Dist.grid <- cbind(Dist.grid,Dist.scale = (a*Dist.grid$Dist) + b)
```

```
USPR.wide <- cbind(USPR.wide,Coef.lit = (a*USPR.wide$Dist) + b)
rm(a);rm(b)

# calcul de la statistique et inférence par méthode de Monte Carlo
cl <- makeCluster(4)
registerDoSNOW(cl)
replicats <- foreach(i = 2 : (1+length(US))) %do% {
  Observed <- sum(USPR.wide$Coef.lit*(USPR.wide[,i]-(1/221)))
  replicats <- foreach(j = 1:999999, .combine = c)%dopar%{
    x <- sample(USPR.wide[,i])
    sum(USPR.wide$Coef.lit*(x-(1/221)))
  }
  list(Rb = sort(unique(USPR$Rb))[i-1] ,Observed = Observed,replicats = replicats)
}
stopCluster(cl)
```