Système d'information décisionnel sur les interactions environnement-santé : cas de la Fièvre de la Vallée du Rift au Ferlo (Sénégal)

Fanta Bouba



Université Pierre et Marie Curie

Université Cheikh Anta Diop de Dakar

Ecole Doctorale Informatique, Télécommunication et Electronique

Unité de Modélisation Mathématique et Informatique des Systèmes Complexes

Système d'information décisionnel sur les interactions environnement-santé

Cas de la Fièvre de la Vallée du Rift au Ferlo (Sénégal)

Par **Fanta BOUBA**

Thèse de doctorat en Informatique

Dirigée par Christophe CAMBIER et Samba NDIAYE

Présentée et soutenue publiquement le 25 Septembre 2015

Devant le jury composé de :

Christophe Cambier	Maitre de Conférences UPMC	Directeur de Thèse Nord
Samba Ndiaye	Maitre de Conférences UCAD	Directeur de Thèse Sud
Maguelonne Teisseire	Directrice de recherche IRSTEA	Examinateur
Jacques André Ndione	Directeur de recherche CSE	Examinateur
Ousmane Sall	Maitre de Conférence UT	Examinateur
Jean-Daniel Zucker	Directeur de recherche IRD	Président
Mathieu Roche (nord)	Chercheur CIRAD	Rapporteur
Cheikh Talibouya Diop	Maitre de conférences UGB	Rapporteur

"Connaître profondément pour agir !"

Dédicaces

A ma famille, je ne suis rien sans vous, je n'existe qu'à travers vous.

A ma grand-mère et mon père qui resteront à jamais mes modèles de générosité, de pardon et de travail.

C'est l'histoire d'une fratrie de 5 filles et 3 garçons ... cette histoire a commencé en 1975 et cette histoire continue allègrement avec la grâce de DIEU.

Remerciements

Je suis particulièrement reconnaissante envers Alassane BAH qui a été à l'origine de ce travail. En effet, il n'a ménagé aucun effort pour me mettre en contact avec toute l'équipe encadrante, pour me trouver un sujet assez proche de mes aspirations mais également pour m'aider à bénéficier de ma bourse d'études.

Je remercie Jean-Daniel ZUCKER pour avoir bien voulu présider ce jury mais surtout pour ces remarques, lors de mon évaluation à mi-parcours, qui m'ont permis de mieux m'approprier mon travail de recherche.

J'exprime également mes remerciements aux rapporteurs de cette thèse, Cheikh Talibouya DIOP et Mathieu ROCHE, pour le temps accordé à lire et à commenter ce document. J'espère pouvoir mettre à profit leurs corrections et remarques en vue de m'améliorer dans mes prochains travaux de recherche.

Je remercie Ousmane SALL qui a accepté de faire partie de ce jury. Merci pour sa réactivité malgré les délais très serrés.

Je tiens à remercier mon directeur de thèse Christophe CAMBIER qui a tenu durement la laisse durant ces dernières années malgré les conditions de ma thèse qui ne correspondaient pas toujours à sa vision du doctorant.

Mes remerciements s'adressent également à mon directeur de thèse Samba NDIAYE pour la confiance accordée malgré les nombreuses périodes d'instabilité. Merci d'avoir supervisé ce travail.

Un merci particulier à Jacque-André NDIONE qui a tenu à m'accompagner aussi bien dans mes activités de recherche que dans ma vie sociale. Il a toujours pris son temps pour me mettre en rapport avec les experts métiers, pour me donner des cours à la fois sur la maladie de la FVR et les facteurs environnemenatux. Il a veillé régulièrement à ce que je sois dans les meilleurs conditions morales en particulier durant les moments de doute.

Un grand merci à Maguelonne TEISSEIRE qui a été la carte mère de cette thèse. Elle a mis à ma disposition son environnement de travail, son équipe mais également toute son expertise pour m'accompagner aussi bien dans les concepts théoriques que dans la mise en oeuvre pratique. Merci de vous être autant investie et m'avoir soutenue pendant les périodes difficiles.

Merci également à tous les experts métiers, qui ont bien voulu nous recevoir pour des entretiens, pour le recueil de données mais aussi pour la restitution de certains résultats.

Je remercie vivement toute l'équipe TETIS et particulièrement Lilia COUNIENC, Mickael FABREGUE, Guilhem MOLLA, Hugo ALATRISTA SALAS pour leur sympathie et leur précieuse aide.

Un immense merci à mes fidèles patrons Alex CORENTHIN et Mouhamed Tidiane SECK qui ont su m'accorder de nombreuses faveurs et aménagements afin que je puisse avoir le temps de me consacrer à mes travaux. Merci pour vos encouragements, votre amitié et nos nombreuses discussions enrichissantes sur tous les sujets.

A mes meilleurs amis, ma Agne team, mes frères et soeurs de sang, mes frères et soeurs de coeur, mes enfants de sang, mes enfants de coeur : merci de m'aimer, merci de me soutenir ; je vous raconterai l'histoire cette thèse.

Je remercie toutes les personnes qui de près ou de loin m'ont aidée et/ou soutenue tout au long de cette thèse.

Enfin merci au projet QWeCI qui a été le socle de nos expérimentations.

Résumé

Notre recherche se situe dans le cadre du projet QWeCI (Quantifying Weather and Climate Impacts on Health in Developing Countries, UE FP7) en partenariat avec l'UCAD, le CSE et l'IPD, autour de la thématique environnement-santé avec comme cas pratique les maladies à vecteurs au Sénégal et plus particulièrement la Fièvre de la Vallée du Rift (FVR). La santé des populations humaines et animales est souvent fortement influencée par l'environnement. D'ailleurs, la recherche sur les facteurs de propagation des maladies à transmission vectorielle, telle que la FVR, prend en compte cette problématique dans sa dimension aussi bien physique que socio-économique.

Apparue en 1912-1913 au Kenya, la FVR est une anthropo-zoonose virale répandue dans les régions tropicales qui concerne principalement les animaux mais dont les hommes peuvent aussi être touchés. Au Sénégal, la zone à risque concerne en majorité la vallée du fleuve Sénégal et la zone sylvo-pastorale du Ferlo. Bien que de climat sahélien, le Ferlo regorge de nombreuses mares qui sont des sources d'approvisionnement en eau pour les hommes et le bétail mais également les gîtes larvaires pour les vecteurs potentiels de la FVR. La maîtrise de la FVR, carrefour de trois (03) grands systèmes (agro-écologique, pathogène, économique/sanitaire/social), implique nécessairement la prise en compte de plusieurs paramètres si l'on veut d'abord comprendre les mécanismes d'émergence mais aussi envisager le travail de modélisation du risque.

Notre travail porte sur le processus décisionnel pour quantifier l'utilisation de données sanitaires et environnementales dans l'évaluation de leur impact pour le suivi de la FVR. Les équipes de recherche impliquées produisent des données lors de leurs enquêtes de terrains et des analyses de laboratoire. Ce flot de données croissant devrait être stocké et préparé à des études corrélées grâce aux nouvelles techniques de stockage que sont les entrepôts de données. A propos de l'analyse des données, il ne suffit pas de s'appuyer seulement sur les techniques classiques telles que les statistiques. En effet, la valeur ajoutée de contribution sur la question s'oriente vers une analyse prédictive combinant à la fois les techniques agrégées de stockage et des outils de traitement. Ainsi, pour la découverte d'informations, nouvelles et pertinentes à priori non évidentes, il est nécessaire de s'orienter vers la fouille de données. Par ailleurs, l'évolution de la maladie étant fortement liée à la dynamique spatio-temporelle environnementale des différents acteurs (vecteurs, virus et hôtes), cause pour laquelle nous nous appuyons sur les motifs spatio-temporels pour identifier et mesurer certaines interactions entre les paramètres environnementaux et les acteurs impliqués. Grâce au processus décisionnel, les résultats qui en découlent sont multiples :

— suivant la formalisation de la modélisation multidimensionnelle, nous avons construit

un entrepôt de données intégré qui regroupe l'ensemble des objets qui participent à la gestion du risque sanitaire - ce modèle peut être généralisé aux maladies à vecteurs ;
— malgré une très grande variété de moustiques, les *Culex* de type *neavei* et les *Aedes* de type *ochraceus* et *vexans* sont les vecteurs potentiels de la FVR les plus présents dans la zone d'étude et ce, durant la saison des pluies, période la plus sujette à des cas suspects ; la période à risque reste quand même le mois d'Octobre ;
— les mares analysées ont quasiment le même comportement, mais des variations significatives subsistent par endroits.

Ce travail de recherche démontre une fois de plus l'intérêt pour la mise en évidence des relations entre les données environnementales et la FVR à partir de méthodes de fouille de données, pour la surveillance spatio-temporelle du risque d'émergence.

Mots-clefs *Processus décisionnel, Fouille de données, Motifs spatio-temporels, Fièvre de la vallée du Rift, Entrepôt de données, Modélisation multidimensionnelle*

Abstract

Our research is in part of the QWeCI european project (Quantifying Weather and Climate Impacts on Health in Developing Countries, EU FP7) in partnership with UCAD, the CSE and the IPD, around the theme of environmental health with the practical case on vector-borne diseases in Senegal and particularly the Valley Fever (RVF). The health of human and animal populations is often strongly influenced by the environment. Moreover, research on spread factors of vector-borne diseases such as RVF, considers this issue in its dimension both physical and socio-economic.

Appeared in 1912-1913 in Kenya, RVF is a widespread viral anthropo-zoonosis in tropical regions which concerns animals but men can also be affected. In Senegal, the risk area concerns mainly the Senegal River Valley and the forestry-pastoral areas Ferlo. With a Sahelian climate, the Ferlo has several ponds that are sources of water supply for humans and livestock but also breeding sites for potential vectors of RVF. The controlling of the RVF, which is crossroads of three (03) large systems (agro-ecological, pathogen, economic/health/social), necessarily entails consideration of several parameters if one wants to first understand the mechanisms emergence but also consider the work on risk modeling.

Our work focuses on the decision making process for quantify the use of health data and environmental data in the impact assessment for the monitoring of RVF. Research teams involved produce data during their investigations periods and laboratory analyzes. The growing flood of data should be stored and prepared for correlated studies with new storage techniques such as datawarehouses. About the data analysis, it is not enough to rely only on conventional techniques such as statistics. Indeed, the contribution on the issue is moving towards a predictive analysis combining both aggregate storage techniques and processing tools. Thus, to discover information, it is necessary to move towards datamining. Furthermore, the evolution of the disease is strongly linked to environmental spatio-temporal dynamics of different actors (vectors, viruses, and hosts), cause for which we rely on spatio-temporal patterns to identify and measure interactions between environmental parameters and the actors involved. With the decision-making process, we have obtained many results :

— following the formalization of multidimensional modeling, we have built an integrated datawarehouse that includes all the objects that are involved in managing the health risk - this model can be generalized to others vector-borne diseases ;

— despite a very wide variety of mosquitoes, *Culex neavei*, *Aedes ochraceus* and *Aedes vexans* are potential vectors of FVR. They are most present in the study area and, during the rainy season period which is most prone to suspected cases ; the risk period still remains the month of October ;

— the analyzed ponds have almost the same behavior, but significant variations exist in some points.

This research shows once again the interest in the discovery of relationships between environmental data and the FVR with datamining methods for the spatio-temporal monitoring of the risk of emergence.

Keywords *Decision-making, Data mining, Spatio and temporal patterns, Rift Valley Fever, Datawarehouse, Multidimensional modeling*

Table des matières

Table des figures

Sigles et Abbréviations

CE Conductivité Electrique

CSE Centre de Suivi Ecologique

DM Datamining

DSV Direction des Services Vétérinaires

DW Datawarehouse

FVR Fièvre de la Vallée du Rift

IPD Institut Pasteur de Dakar

MES Matière En Suspension

MOLAP Multidimensional OnLine Analytical Processing

OIE World Organisation for Animal Health

OLAP OnLine Analytical Processing

OLPT OnLine Transaction Processing

OOLAP Object OnLine Analytical Processing

pH potentiel Hydrogène

PNLP Programme National de Lutte contre le Paludisme

QWeCI Quantifying Weather and Climate Impacts on Health in Developing Countries

ROLAP Relational OnLine Analytical Processing

SAD Système d'Aide à la Décision

SID Système d'Information Décisionnel

SIG Système d'Information Géographique

TDS Total dissolved solids

TIC Technologies de l'Information et de la Communication

UCAD Université Cheikh Anta DIOP

CHAPITRE 1

Introduction

Nos travaux de recherche s'inscrivent dans le cadre de l'aide à la décision dans un environnement multidisciplinaire de données complexes et de volume élevé (big data). Notre domaine d'application se situe en environnement-santé avec comme cas pratique la maladie de la Fièvre de la Vallée du Rift au Sénégal.

Un système sanitaire est assez complexe dans la compréhension de la maladie, de son émergence, de sa propagation et des soins nécessaires à apporter aux hôtes malades. Selon (OECD, 2005) (*Organisation for Economic Co-operation and Development*), « les TIC appliquées à la santé offrent d'immenses perspectives : elles ont le pouvoir de modifier notre compréhension de la maladie, de transformer les prestations de services de santé, et d'améliorer les résultats dans ce domaine. »

Dans son analyse sur l'apport de l'informatique décisionnelle dans la santé, (Allegri, 2004) conclut que les systèmes décisionnels représentent un excellent moyen de guider les acteurs du système sanitaire dans leur quête de la maitrise des maladies. Il ne s'agit pas uniquement de mettre à leur disposition de nouveaux outils (tableau de bord, rapport, etc.) mais surtout de les aider à comprendre et à définir au préalable les processus de fonctionnement et les interactions existantes entre les différentes disciplines impliquées.

La Fièvre de la Vallée du Rift (FVR) est une maladie infectieuse et virale ; elle est une zoonose qui affecte les mammifères (Pepin, 2011). Cette maladie est décrite comme une maladie au carrefour de trois (03) grands systèmes que sont les systèmes (1) agroécologique, (2) pathogène et (3) économique, sanitaire et social (Figure 1.1). Cette complexité implique la prise en compte de plusieurs paramètres pour comprendre les mécanismes d'émergence et ensuite envisager le travail de modélisation du risque (Rocque, 2004).

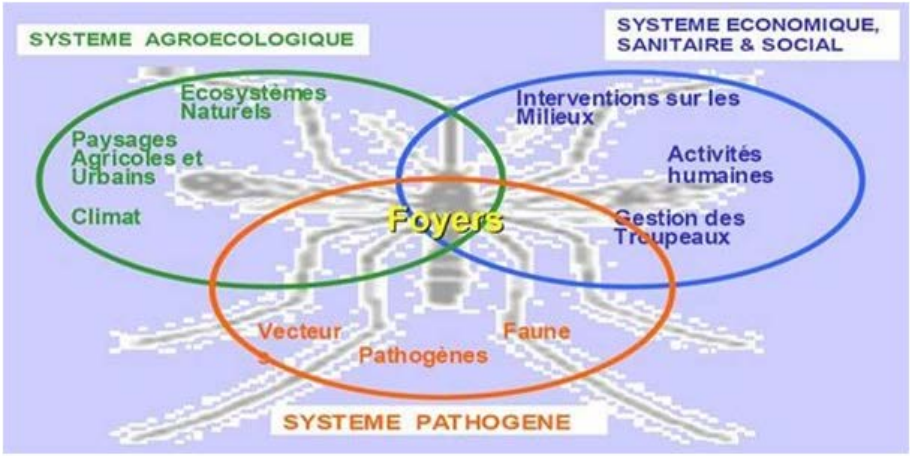

FIGURE 1.1 – Complexité de la FVR (Rocque, 2004).

Les recherches sur les maladies visent à comprendre et à maitriser plusieurs processus. Ainsi, on s'intéresse, d'une part au processus lié à l'apparition, à la persistance et à la diffusion des maladies, et d'autre part aux moyens de lutte à mettre en place. La compréhension du processus d'émergence et de propagation, telle qu'abordée actuellement, n'offre que des vues cloisonnées spécifiques à chaque discipline. Pourtant, d'importants travaux, ((Davies *et al.*, 1985) ; (Meegan et Bailey, 1988) ; (Prehaud et Bouloy, 1997) ; (Linthicum *et al.*, 1999) ; (Ndione *et al.*, 2003)), ont permis de montrer les interactions entre les facteurs environnementaux (climatiques, hydrologiques, transhumance) et l'apparition de la FVR. La complexité de l'analyse et de la corrélation de ces facteurs justifie la proposition de solutions adaptées à la problématique d'analyse multi-critères. Ceci afin de proposer des mesures de gestion du risque sanitaire en maitrisant les impacts spatiaux et temporels des facteurs environnementaux.

Les indicateurs environnementaux sont une source majeure de variation de la santé. Ainsi, les réponses aux questions que l'on se pose sur certaines maladies doivent prendre en compte les interactions entre l'environnement et les facteurs déclenchants. Afin de soutenir les équipes dans leurs recherches impliquées sur cette problématique, nous allons proposer des solutions qui répondent à la question suivante : « De quelle manière et dans quelle mesure les déplacements des vecteurs et les variables environnementales influencent-ils le déclenchement et la propagation des maladies vectorielles ? »

L'analyse de cette question nous conduit à prendre en compte deux aspects respectivement relatifs aux attentes des utilisateurs et à l'apport scientifique de nos travaux. La phase d'entretiens avec les acteurs nous a permis d'identifier des questions pour lesquelles ils souhaitaient avoir des éléments de réponse. Il s'agit entre autres des questions suivantes :

- Quelles sont les variables d'environnement impliquées ?
- Quelles sont les caractéristiques des vecteurs impliqués ?
- Quels sont les connaissances ayant un impact dans le processus de prise de décision ?

Par ailleurs, les résultats en termes de connaissances seront obtenus à partir des questionnements issus des différentes disciplines. Il s'agit, entre autres, des interrogations ci-

dessous :

— Quels sont les liens entres les localités et la taxonomie/densité vectorielles ?
— Comment évaluer l'impact de variabilité des caractéristiques des mares, des facteurs climatiques sur la densité vectorielle ?
— Quel est l'impact des facteurs climatiques sur la qualité des eaux des mares ?

Sur le plan scientifique, on cherchera à répondre aux questions suivantes :

— Comment représenter les données et leurs interactions en tenant compte de caractères fondamentaux tels que la complexité, la diversité et la quantité ?
— Quelle architecture des données à mettre en œuvre ?
— Quel algorithme utiliser pour des analyses corrélées ?

Afin de répondre à ces besoins, dans le cadre de cette thèse, nous nous orientons vers une approche décisionnelle qui apporte des solutions pour la modélisation, l'interrogation et la visualisation de données dans un objectif d'aide à la décision et de découverte de connaissances. Ainsi, il s'agit d'exploiter efficacement d'importants volumes de données provenant de diverses sources afin de fournir des informations décisionnelles. L'exploitation efficace de ces données nécessite de revoir les techniques et les méthodes existantes de stockage, de traitement de données (acquisition, fouille, découverte de données) et le cas échéant d'en proposer de nouvelles adaptées à notre contexte. L'objectif est de concevoir et de valider des solutions robustes en termes de fonctionnalités et de performances ; les données du processus de simulation doivent être modélisées, stockées, traitées et manipulées par des algorithmes robustes, performants, et adaptés aux supports décisionnels.

Le but final est de proposer un environnement décisionnel qui permettra d'avoir une vue globale sur les interactions entre les indicateurs environnementaux et la propagation des maladies à vecteurs. On doit, ainsi, proposer (1) des vues fonctionnelles sur la description et la répartition des objets manipulés (mares, troupeaux, etc.) et (2) des vues opérationnelles pour la définition de scénarii (impact de variabilité saisonnière sur la répartition des vecteurs, etc.).

Processus décisionnel

Pour la mise en place de l'environnement décisionnel, nous suivons l'approche décisionnelle que nous avons adaptée à une problématique de santé publique.

La première phase consiste à identifier les données brutes et leurs sources de stockage. Pour ce faire, il est nécessaire de définir le contexte de l'étude décisionnelle. Il s'agit de comprendre le système de santé et de maîtriser l'environnement dans lequel la FVR évolue mais aussi d'identifier ses caractéristiques directes (propres à la maladie) et indirectes (liées à l'environnement). Cette étape nous permettra de collecter les données des différentes disciplines impliquées.

Dans la deuxième phase, il est nécessaire de construire une structure de données adaptée à l'analyse décisionnelle. Ce système de stockage de données devra permettre de qualifier et d'intégrer toutes les données, sous format brute ou agrégé.

La troisième phase consistera à la sélection et à l'organisation des données permettant de fournir des vues agrégées du système. Elle nous permettra de rechercher des informations satisfaisant les besoins des utilisateurs finaux. C'est la première étape de restitution considérée comme un diagnostic. Cependant, l'évaluation des alternatives en termes de choix humains ou d'évènements naturels nécessite de mettre en place des outils d'investigation (déduction et simulation) qui puisse répondre aux interrogations des experts scientifiques. Cette deuxième étape de restitution devrait permettre d'identifier les indicateurs pertinents, de comprendre leurs liens directes/indirectes. L'évaluation des résultats a pour objectif d'adapter les décisions suivant les variations des éléments non maitrisables.

Ainsi, nous organisons le plan de cette thèse à travers le déroulement du processus décisionnel.

Plan de la thèse

Cette thèse est structurée en quatre (4) chapitres qui nous permettront d'étudier et d'appliquer le processus complet de mise en œuvre d'un environnement décisionnel.

Le premier chapitre décrit la maladie de la FVR par les acteurs impliqués et son histoire géographique. Nous présentons également la position du Sénégal par rapport à cette maladie. Dans un deuxième temps, nous cherchons à comprendre le contexte organisationnel dans lequel s'inscrit notre projet de thèse. Le caractère pluridisciplinaire de la problématique est à la fois un atout car il nous permet de comprendre la complexité de la maladie étudiée ; mais aussi un "ralentisseur" car pour des raisons pratiques, il a fallu satisfaire à la fois les experts utilisateurs "scientifiques" et les utilisateurs finaux. Nous décrivons aussi les données manipulées par les experts métiers.

Dans le deuxième chapitre, nous justifions l'approche retenue par une étude comparative, d'une part, entre les standards de modélisation épidémiologiques et d'autre part, entre les approches opérationnelles et décisionnelles. Par la suite, notre étude bibliographique est décomposée en trois (3) grandes sections :

— les définitions préliminaires sur la modélisation multidimensionnelle pour comprendre le formalisme de conception du modèle de données et sur l'entrepôt de données afin d'en maitriser les règles organisationnelles ;

— la fouille de données qui décrit les premières méthodes utilisées pour répondre à certaines des questions posées par les experts ;

— les motifs spatio-temporels pour la prise en charge d'une analyse à la fois spatiale et temporelle des données corrélées.

Le troisième chapitre expose les contributions apportées lors de notre étude. Elles sont présentées conformément au processus décisionnel. D'abord, nous décrivons le modèle de données implémenté pour la mise en place de l'entrepôt et des outils de reporting. Ensuite, nous proposons divers scenarii d'évaluation de l'impact des facteurs environnementaux (1) en analysant les variables d'un même objet et (2) en confrontant les variables de différents objets. Pour compléter notre étude sur la compréhension des interactions, nous l'orientons vers la découverte de motifs spatio-temporels présentés dans la dernière section.

En chapitre 4, les conclusions « fortes » issues de l'interprétation des résultats sont regroupées sous forme d'une discussion. Dès lors, nous proposons des mesures de gestion du risque sanitaire et des perspectives, en vue de compléter le travail réalisé.

Les dépendances entre les chapitres sont synthétisées dans la figure 1.2 ci-après :

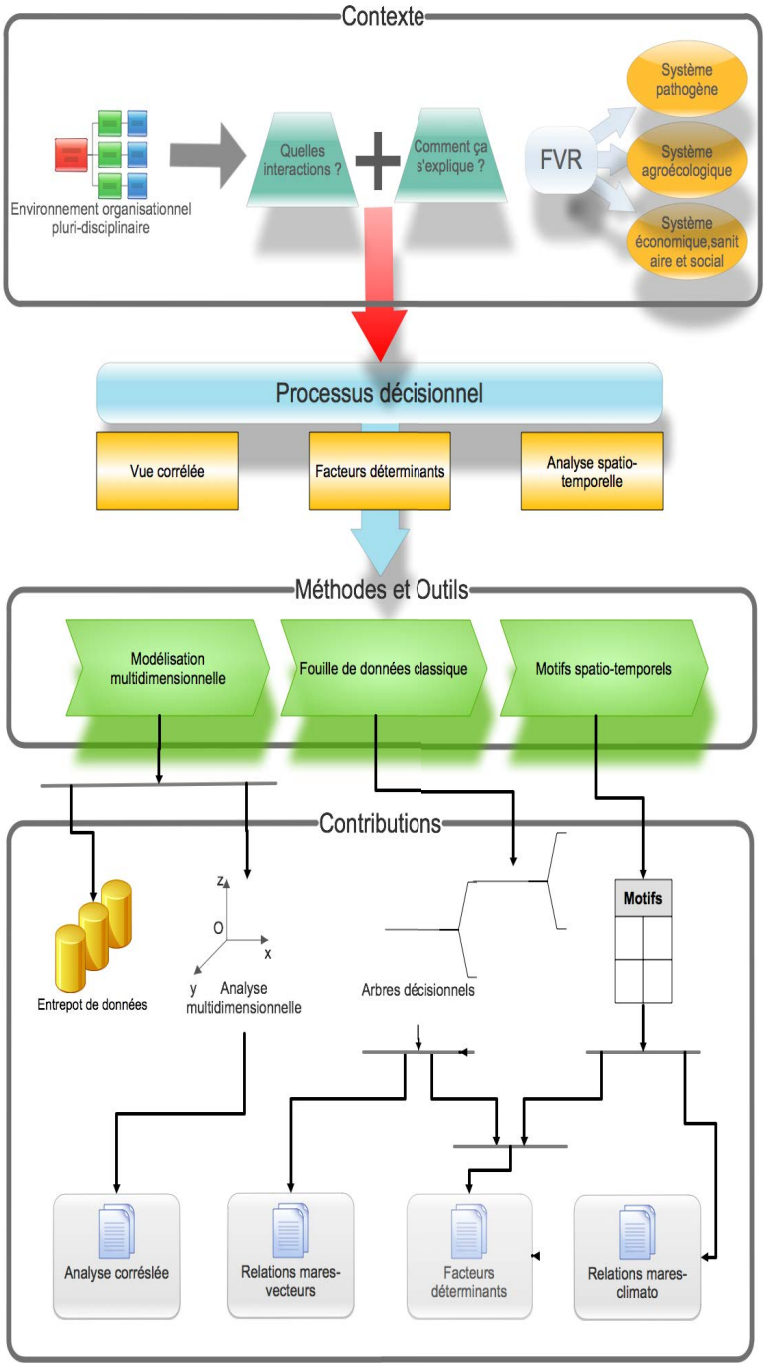

FIGURE 1.2 – Organisation du document

CHAPITRE 2

Contexte

2.1 La Fièvre de la Vallée du Rift

2.1.1 Définition

La FVR est une anthropozoonose et une arbovirose qui provoque chez les animaux des avortements des femelles gestantes, la mortalité des jeunes et chez les hommes des grippes, des affections hémorragiques et nerveuses (Cêtre-Sossah et Albina, 2009). Le nom de la maladie est originaire de la zone dans laquelle le virus a été isolé pour la première fois ; il s'agit de la Vallée du Rift au Kenya (Daubney *et al.*, 1931). Cette maladie, transmise par des arthropodes vecteurs ((Fontenille *et al.*, 1995) ; (Fontenille *et al.*, 1998) ; (Diallo *et al.*, 2000)), est causée par un virus du genre *Phlebovirus* de la famille des *Bunyaviridae*.

Bien que son taux de mortalité ne soit pas élevé, cette maladie est considérée comme une « arme biologique et économique ». D'une part, la dissémination de son virus peut se faire par aérosol et les œufs des vecteurs sont résistants à la durée et à la sécheresse ((SPIEZ, 2006)). D'autre part, une épidémie de la FVR est un facteur négatif pour le développement de l'élevage qui est une source primaire de revenu dans les zones affectées. C'est également une problématique de santé publique dans la mesure où elle touche à la fois les animaux et les hommes. C'est pourquoi l'OIE la classe dans la liste commune des maladies à plusieurs espèces ayant « un grand pouvoir de diffusion et une gravité particulière » ((Diagne, 1992) ; (OIE, 2010)). Sur le plan clinique, elle est caractérisée par des symptômes variant d'une espèce à l'autre et selon la classe d'âge ; les animaux domestiques, les femelles et les plus jeunes sont les plus touchés. Du point de vue géographique, la FVR touche principalement les régions tropicales ; notamment en Afrique et en Asie de l'Est (Moyen-Orient).

2.1.2 Description - Les acteurs de la maladie

2.1.2.1 Virus

Le virus de la FVR est stable pour un pH compris entre 6.2 et 8.0 (Lefèvre, 2003). Il peut survivre pendant 80 minutes à une température ambiante (25 à $30\,^{\circ}\mathrm{C}$), 21 jours à $37\,^{\circ}\mathrm{C}$ mais bien au-delà dans le sang ou le sérum : 1 mois à $-20\,^{\circ}\mathrm{C}$, 8 mois à $4\,^{\circ}\mathrm{C}$ et 1 an à $20\,^{\circ}\mathrm{C}$ (Balkhy et Memish, 2003). Mais ce virus peut être inactivé en 40mn à 50 ˚C ou par les solvants des lipides tels que l'éther, le chloroforme et les désinfectants usuels, tels que le formol, la bétapropiolactone ou le disoxycholate de soude. Il peut également être détruit par des rayons ultraviolets (Dampfhoffer, 2009).

2.1.2.2 Mode de transmision

Plusieurs travaux, ((Ndione *et al.*, 2003) ; (Pepin, 2011)), confirment que la FVR peut être transmise par quatre modes :

— la transmission vectorielle par la piqûre de nombreuses espèces de moustiques (*Aedes*, *Anopheles*, *Culex*, *Eretmapodites* et *Mansonia*) et d'autres insectes hématophages ;
— la transmission directe par le contact avec un hôte infecté ; les sources de virus sont les sécrétions nasales, oculaires et vaginales, les embryons, le placenta et la viande des animaux infectés. Les portes d'entrée sont soit des micro-lésions cutanées lors de la manipulation de produits souillés (notamment les avortons), soit la muqueuse nasale par inhalation d'aérosols infectieux ;
— la transmission transovarienne aux larves pour certaines espèces entretient la circulation du virus à la descendance avec amplification chez l'animal ;
— la transmission par inhalation d'aérosols de sang vireémique.

2.1.3 Historique géographique

La FVR est décrite pour la première fois en 1912-1913 près du lac Naivasha au Kenya (Prehaud et Bouloy, 1997). Dans les années 1930-1931, le virus est isolé dans la Vallée du Rift du Kenya (Daubney *et al.*, 1931). En 1948, la FVR est déclaré comme étant une arbovirose (Smithburn *et al.*, 1948).

Dès lors, plusieurs épidémies ont été notées :

— au Kenya, elles se renouvellent en 1968, en 1975 (milliers de bêtes et 7 pertes humaines), puis dans les années 1978-1979 et en 1997 (89000 personne atteintes dont 500 décès) (Gerdes *et al.*, 2004) ;
— en 1977, l'Egypte est touchée (18.000 cas humains dont 600 décès) puis en 2003 (45 personnes infectées et 17 décès) (Meegan *et al.*, 1979), (Eisa *et al.*, 2004) ;
— en 1987, une épizootie atteint l'Afrique de l'Ouest par la Mauritanie (Meegan et Bailey, 1988),(Jouan *et al.*, 1990) - en 1997, la FVR se manifeste chez des hommes dans la région du sud-est de la Mauritanie ;
— en 1998, l'Afrique Australe est fortement atteinte au Kenya, en Tanzanie et en Somalie (Woods *et al.*, 2002) ;

— en 2000, l'Asie est touchée par le Yémen (un millier de cas et 121 morts) et l'Arabie Saoudite (863 cas et 120 morts) ;
— depuis 2006, de nombreux cas sont identifiés au Kenya, en Somalie en Tanzanie (Brugere-Picoux et Kodjo, 2007).

2.1.4 Cas du Sénégal

Au Sénégal, cette maladie occupe une place importante dans la vallée du fleuve Sénégal avec une attention particulière concernant le Ferlo. D'après l'étude menée par (Diallo *et al.*, 1995), la propagation de la FVR, identifiée dans la zone du Ferlo (Figure 2.1), est liée au cycle de vie des moustiques et au cycle d'évolution des mares (Figure 2.2).

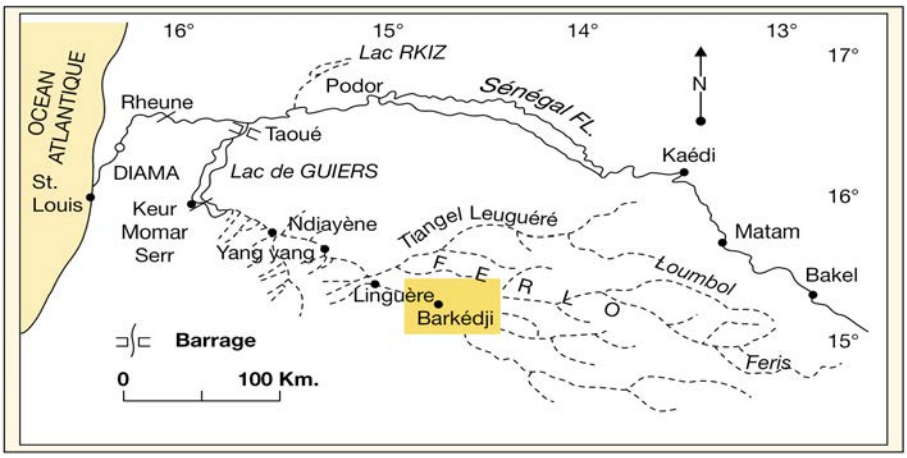

FIGURE 2.1 – Localisation de la zone d'étude, Adapté de (Coly, 1996).

En effet, les mares du Ferlo constituent à fois la source d'approvisionnement en eau pour les populations et le bétail tout en étant le gîte larvaire des vecteurs.

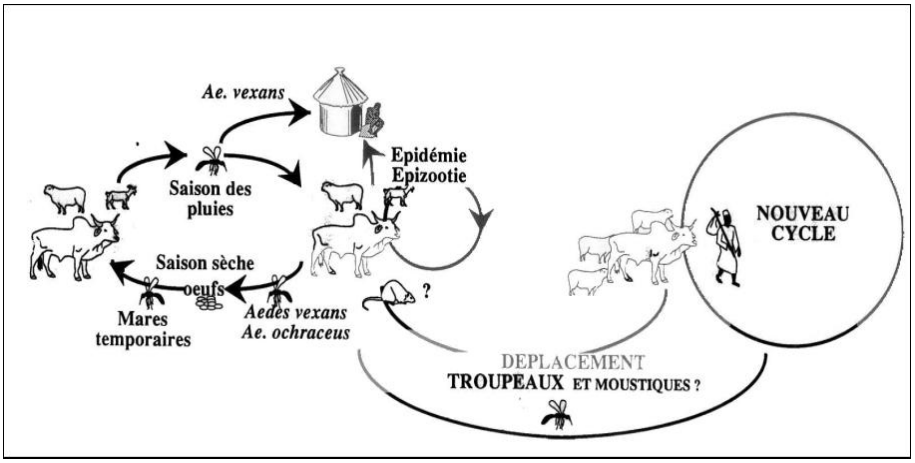

FIGURE 2.2 – Cycle épidémiologique de la FVR (Diallo *et al.*, 1995).

Dans cette zone, la variabilité intra-saisonnière de la pluviométrie, la dynamique de la végétation et la turbidité des mares temporaires, dont la taille est relativement petite, sont les facteurs principaux qui expliquent la forte concentration des moustiques ((Ndione *et al.*, 2009)).

Plusieurs équipes de diverses disciplines (climatologie, entomologie, virologie, environnement, etc.) tentent d'apporter des réponses qui permettront de mettre en œuvre des moyens de contrôle et d'éradication de cette maladie. Mais il n'existe pas, jusqu'à présent, de système d'information intégré qui leur permettrait de corréler leurs données, de confronter les résultats et de faire des simulations pour comprendre le fonctionnement des mares, la dynamique des populations de vecteurs et le déplacement des hôtes (troupeaux et populations). Pourtant, une grande variété de données est régulièrement recueillie lors des enquêtes de terrain ou des tests de laboratoire. Mais, les rapports qui en découlent se limitent à la restitution cloisonnée propre à chaque discipline. C'est dans un tel contexte que les équipes impliquées ont décidé de mettre en synergie leurs travaux. Ainsi, cette thèse, dans le but de faire cohabiter les différentes équipes, se propose de :

— développer une approche transparente basée sur la modélisation des interactions entre les différents facteurs ;
— quantifier l'utilisation des informations météorologiques et climatiques, de prévision dans l'évaluation de l'impact de la FVR.

Il s'agira, dans une première phase, d'identifier, de recueillir et de rassembler toutes les données nécessaires puis, dans une seconde phase, de produire des rapports d'analyses et d'en extraire des connaissances qui permettraient de répondre aux questions de lutte contre la maladie de la FVR.

Conclusion Avec l'évolution des technologies informatiques, plusieurs équipes ont pu développer des systèmes de surveillance et de supervision permettant d'offrir des vues statistiques en format géo référencés. Toutefois, les problèmes liés à la maîtrise du déclenchement et à la propagation des maladies vectorielles doivent êtres plus approfondis par les techniques de fouilles de données dans un but de description, de classification et de prédiction. Les algorithmes de datamining ont joué un grand rôle dans le diagnostic et le pronostic

de nombreuses maladies. En effet, les différentes techniques proposées permettent de classer les maladies, les agents ou les hôtes, d'identifier les paramètres environnementaux à fort ou faible impact, de prédire les tendances de futures périodes, etc.

2.2 Environnement organisationnel

La complexité des systèmes étudiés (environnement, climat, hydrologie, qualité de l'eau, entomologie, épidémiologie, virologie, socio-économie, pastoralisme, etc.) a rendu nécessaire la prise en compte dans un même environnement de phénomènes multiples et variés, mais aussi à différentes échelles (temporelle, spatiale et organisationnelle). Le projet QWeCI, [1] (*Quantifying Weather and Climate Impacts on Health in Developing Countries*) implique plusieurs partenaires qui sont classés en deux catégories :

— les utilisateurs scientifiques qui regroupent l'Université Cheikh Anta Diop (UCAD), le Centre de Suivi Ecologique (CSE) et l'Institut Pasteur de Dakar (IPD) ;
— les utilisateurs finaux que sont le Programme National de Lutte contre le Paludisme (PNLP) et la Direction des Services Vétérinaires (DSV).

Ces partenaires nous ont fourni des données provenant des enquêtes de terrain (données observées ou générées par des équipements de mesure) ou des tests de laboratoire (données d'analyse). Ces données se présentent sous des formats tableurs et textes, des formats « ASCII » (data) ou des données géo-spatiales (shape file). La Figure 2.3 présente les partenaires identifiés selon leurs centres d'intérêts et les formats de données utilisés.

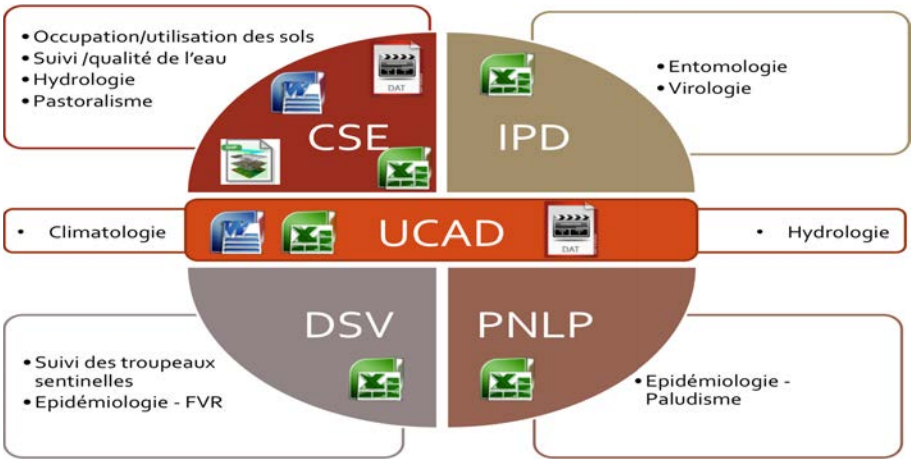

FIGURE 2.3 – Complexité organisationnelle [2]

Le défi à relever par ces équipes de recherche est la maitrise de l'impact des facteurs environnementaux sur le déclenchement et la propagation de la FVR. La compréhension d'un système pluridisciplinaire nécessite d'avoir une vision transversale afin de corréler les

1. https ://www.liv.ac.uk/qweci/
2. Le cas pratique illustré est la FVR ; le paludisme est cité car il fait partie du projet et nous l'utilisons pour confirmer notre modèle générique.

différents facteurs. L'objectif est de croiser des données qui n'ont pas forcément la même structure pour une interrogation directe. C'est pourquoi notre modèle de données se veut intégrateur des données issues des différentes disciplines concernées.

2.2.1 Données spatiales

Cette étude est conduite sur le site de la communauté rurale Barkédji de l'arrondissement, portant le même nom, de la commune Dahra-Linguère du Département Linguère de la région de Louga. Le choix de cette zone géographique est entièrement lié à la forte présence de la FVR dans cette localité. Barkédji ((Fontenille *et al.*, 1998) ; (Ndione *et al.*, 2003) ; (Mamadou *et al.*, 2010)) se caractérise par sa forte concentration d'éleveurs de bétails mais aussi son hydrologie particulière. Ainsi, on distingue, d'une part les mares et d'autres part les campements. Dans un souci de généralisation, nous avons choisi d'intégrer une plus large dimension en partant du découpage administratif du Sénégal. La mare est l'élément central de notre étude car c'est le lieu de rencontre par excellence entre les vecteurs et les hôtes.

2.2.2 Données de suivi/qualité de l'eau, climatiques et hydrologiques et

L'un des sujets fondamentaux de notre étude est la mare qui est rattachée à plusieurs éléments :

— les postes pluviométriques et les stations sont installés autour des mares : les mesures telles que la vitesse du vent, la température, la pluviométrie sont générées automatiquement et recueillies périodiquement ;
— les échelles de mesures limnométriques (3) sont placées dans les mares d'eau : elles permettent de mesurer le niveau d'eau des mares ;
— les prélèvements des eaux des mares sont identifiés par les coordonnées du point de prélèvement, appelé site, ils permettent de mesurer le PH, la température, les matières en suspension de l'eau des mares.

2.2.3 Données entomologiques et virologiques

Lors des campagnes de terrain, les moustiques sont capturés en utilisant des pièges à appât (mouton, poulet) et des pièges à CO_2. Ces enquêtes entomologiques sont faites aussi bien autour des mares que dans les campements à l'intérieur des maisons et dans les troupeaux. Après capture, un travail d'identification individuelle est mené pour déterminer le sexe du moustique. Par ailleurs, des prélèvements sont faits dans les troupeaux ; il peut s'agir de prélèvements sanguins, d'avortons, etc. La virologie des moustiques est évaluée sur la base d'un échantillon broyé. En revanche, la virologie animale est testée individuellement sur le prélèvement ; on cherche ainsi à vérifier s'il y a une présence ancienne ou récente du virus.

Conclusion La santé des populations humaines et animales est souvent fortement influencée par les indicateurs environnementaux. C'est pourquoi la recherche sur les facteurs de propagation des maladies à transmission vectorielle, telle que la Fièvre de la vallée du

Rift (FVR), prend en compte ces paramètres. Pour comprendre l'impact des facteurs environnementaux sur le déclenchement et la propagation de la FVR, il est nécessaire de maitriser tous les paramètres directs, rattachés aux acteurs de la maladie, et les paramètres indirects qui dépendent de l'environnement. Ainsi, cette première étape, nous a permis de comprendre le fonctionnement épidémiologique de la FVR et d'identifier toutes les données y afférentes. Nous avons identifié trois (3) grandes classes de données qui se positionnent dans la sphère complexe de la FVR (Figure 1.1). L'étape suivante nous permettra de justifier le choix de notre approche décisionnelle et d'identifier les méthodes et outils qui nous permettront d'apporter des réponses aux questions fondamentales soumises par les experts métiers.

CHAPITRE 3

Etat de l'art

Dans ce chapitre, nous proposons un contenu structuré en trois (3) sections. La première section sera consacrée, d'une part, au choix du système d'information, et d'autre part, à celui du formalisme de modélisation. Dans la deuxième section, nous dressons un état des lieux des travaux portant sur la fouille de données appliquée en épidémiologie. Nous nous intéressons plus particulièrement à l'application de ces techniques dans le contexte de l'analyse spatio-temporelle des maladies à vecteurs. La dernière section présentera les techniques de fouille pouvant répondre aux objectifs de recherche définis par les experts métiers. Nous étudierons particulièrement les motifs spatio-temporels.

3.1 Choix d'une approche

3.1.1 Système d'information

Un système d'information opérationnel ou système d'information supports d'opérations concerne les données relatives aux différentes fonctions de l'entreprise ; il s'agit des bases de données résultantes des sources d'information internes. Il assure le traitement de transactions, le contrôle de processus industriels, les supports d'opérations de bureau et de communication, etc.

A l'heure actuelle, l'informatique et ses technologies dérivées ont largement dépassé le cap de l'automatisation des données ; seule la gestion des informations, des connaissances est une problématique. Cette défaillance des systèmes opérationnels a d'abord été perçue comme une imperfection des systèmes d'information existants ; c'est ainsi qu'à partir des années 1970, le processus de méthodologie a fortement été reconsidéré.

La conduite de projets a certes été améliorée dans les années suivantes mais le problème persistait ; il a fallu se rendre à l'évidence : un système automatisé n'était pas conçu pour informer. Ce qui était considéré comme une limite supportable est dans le contexte actuel devenu insupportable.

Un système opérationnel présente ainsi de nombreuses limites :

— il n'est pas possible de faire ressortir des événements antérieurs aisément et encore moins de faire des prévisions car les données ne sont pas historisées nativement ;
— il n'est pas évident d'interroger plusieurs sources de données ;
— il n'est pas aisée d'obtenir des données de requêtes non prévues ;
— dans le cas de multiples bases de données, la cohérence des données n'est pas toujours garantie.

Cette architecture a été conçue à la base en exclusivité pour les applications de production. Ce qui ne concorde pas avec notre problématique. Cette faille nous renvoie donc à l'adoption de l'architecture décisionnelle mais elle ne nous enferme pas non plus dans cette approche. Car nous n'excluons pas la possibilité de recourir aux concepts de base de l'approche transactionnelle.

Dans son article sur l'intégration d'un système décisionnel dans une organisation, (Bruley, 2010) a fait noter que beaucoup d'entreprises se laissent submerger par les données internes et externes qu'elles utilisent et qu'elles ont générées pour la plupart ; « elles ne savent pas comment créer une infrastructure analytique adéquate pour convertir des données en information, des informations en opportunités et des opportunités en actions. » Le concept de Système d'Aide à la Décision, issu des Sciences de Gestion, a été initialement défini par (Gorry et Morton, 1971) de façon formelle sous l'appellation Systèmes de Décision et de Gestion (*Management Decision Systems*) comme « un système informatisé interactif aidant le décideur à manipuler des données et des modèles pour résoudre des problèmes mal structurés ». En 1978, dans le cadre de leurs travaux sur l'implication du décideur dans le management des systèmes d'information, (Keen et Morton, 1978) ont introduit la dimension cognitive humaine à travers ses limites. Il en ressort alors les objectifs des SAD de la manière suivante : « Les SAD (SAD) réunissent les ressources intellectuelles des individus avec des potentialités des ordinateurs dans le but d'améliorer les décisions prises ».

(Sprague Jr et Carlson, 1982), décrivent le système décisionnel comme un système d'information qui s'appuie sur les systèmes transactionnels et interagit avec les autres composants du système d'information global pour soutenir les activités de prise de décision des gestionnaires dans les organisations.

(Bonczek *et al*., 1981), dans leur article sur le « Fondement des Systèmes décisionnels », ont conclut qu'il est fondamental d'utiliser les outils informatiques pour améliorer les capacités de prise de décision des individus. La théorie de l'ensemble du processus de prise de décision devrait être la base pour l'introduction des technologies informatiques dans les processus décisionnels afin d'améliorer la productivité.

Selon (Navetier, 2005), les apports des systèmes décisionnels sont classés en 2 catégories :

— l'amélioration de l'efficacité de la communication et de la distribution des informations de pilotage ;
— l'amélioration du pilotage résultant de meilleures décisions, prises plus rapidement.

Un tel système devrait permettre de mieux maitriser les évènements antérieurs mais surtout de mieux appréhender le futur non pas par des intuitions et données non maitrisées mais par des informations justifiables et cohérentes.

(Gouarné, 1998) nsuite à ses travaux menés sur le projet décisionnel, nous résume les

caractéristiques fondamentales des SAD, sur lesquelles les experts semblent aujourd'hui unanimes, et qu'il est utile de souligner :

- un SAD est, par rapport aux applications de production, à la fois séparé dans sa conception et dépendant pour son alimentation ;
- l'information décisionnelle est conditionnée d'une manière intégrée et indépendante de ses sources d'alimentation. En d'autres termes, les caractéristiques techniques des applications de production et des supports externes dans lesquels le système décisionnel puise ses données n'influent pas sur les modalités selon lesquelles l'utilisateur accède à l'information ;
- l'information décisionnelle est, dans son contenu et dans sa forme, indépendante des structures et des procédures courantes de la production. Elle porte sur le métier de l'utilisateur, sans être délimitée par l'exercice de ce métier. C'est une information « orientée sujet » ;
- parmis les traitements qu'effectue un Système d'Information Décisionnel (SID), beaucoup ne sont pas déterminés par des algorithmes préétablis, ne comportent pas de transactions au sens habituel du terme, et ont pour but de permettre à l'utilisateur d'établir lui-même, entre les données, des rapprochements et des consolidations non prédéfinis. Le modèle de données de diffusion, qui est l'élément clé de la définition du système, doit être conçu dans cette perspective selon une approche multidimensionnelle ;
- l'information décisionnelle est chronologique. Elle est vouée, non pas au contrôle d'une situation instantanée, mais à l'analyse de phénomènes évoluant dans le temps. Le traitement du temps est un aspect distinctif essentiel, mais aussi un facteur de complexité ;
- les spécifications d'un Système d'Information Décisionnel sont hautement instables pour deux raisons : d'une part les objectifs stratégiques à atteindre sont des cibles mouvantes, et d'autre part le déploiement du système modifie l'expression des besoins.

Drucker (1995) souligne que « la connaissance a pris la place du capital en tant qu'élément moteur des organisations » ; ainsi, il n'est plus temps de confondre « données et connaissance, informatique et information». Pour rappel, la donnée est une caractéristique, une propriété d'un objet du monde réel ; elle se présente de façon brute. En revanche, l'information est la traduction, l'interprétation d'une ou de plusieurs données sous une forme compréhensible par l'homme. Exemple : N = 4. N est une donnée ; la traduction de cette donnée en « le nombre de moustiques identifiés est 4 » représente l'information. L'information permet d'atteindre l'objectif de toute organisation : "transmettre un message dans le but d'informer".

Comme son nom l'indique, le rôle fondamental d'un système décisionnel est de fournir des informations destinées aux gérants pour les aider à prendre des décisions stratégiques.

(Marakas, 2003) propose une architecture basée sur le processus décisionnel qui se décline en quatre (4) composants :

- le système de gestion des données : il intègre les données dans leur état brut (base de données, fichier plat, etc.) ;
- le système de gestion des modèles : il est assimilé au système de stockage des

informations – les données quantitatives le constituent ;
— le moteur de connaissances : il s'agit du système de traitement des informations et d'analyse ;
— l'interface utilisateur : qui donne une vue à l'utilisateur final sur tous les outils – tableau de bord, rapport.

Dans un contexte de suivi épidémiologique, la mise en place d'un environnement décisionnel permettrait de répondre à la principale problématique de veille sanitaire ; en effet, la collecte d'une masse de données et la compréhension de leurs interactions permettrait d'anticiper dans un but de maîtrise et/ou d'éradication de la maladie.

3.1.2 Approche de modélisation

(de Vinci, 2008) définit l'épidémiologie comme « une discipline scientifique qui étudie notamment les différents facteurs intervenant dans l'apparition des maladies ou de phénomènes de santé ainsi que leur fréquence, leur mode de distribution, leur évolution et la mise en œuvre des moyens nécessaires à la prévention ».

En épidémiologie, la modélisation permet de représenter un système en prenant en compte tous les paramètres qui le constituent, d'étudier son évolution et de définir les scenari. Les travaux de recherche sur la représentation de l'impact de l'environnement sur la santé et plus particulièrement sur les maladies vectorielles, (Tran *et al.*, 2005), ont permis de distinguer deux grandes classes de modèles :

— les modèles mathématiques qui se basent sur une représentation formelle ;
— les modèles conceptuels basés sur l'analyse des besoins des utilisateurs finaux.

Nous présentons dans cette section les variantes de ces modèles et quelques exemples d'applications utilisées dans le cadre des problématiques liées aux maladies vectorielles.

3.1.2.1 Modèles mathématiques

Les modèles mathématiques offrent des descriptions quantitatives du fonctionnement d'un système, en écrivant sous forme d'équations les lois qui le régissent (Valleron, 2000). On distingue deux catégories de modèles adaptés à notre contexte :

— les modèles géographiques, issus des modèles statistiques.
— les modèles épidémiologiques, issus des modèles théoriques.

3.1.2.1.1 Modèles géographiques

Les modèles géographiques ont pour but d'offrir une vue cartographique sur la distribution spatiale des vecteurs, des maladies en fonction de conditions environnementales. Pour être construit, on requiert les données quantifiantes sur les éléments intervenants et les données géo référencées sur les entités spatiales.

(Rakotomanana *et al.*, 2001), dans leur étude menée sur le paludisme à Madagascar, propose un SIG couplé à la télédétection qui permet d'identifier et de classer les gîtes larvaires de moustiques suivant leur taux de production. Les zones les plus à risque pourront faire l'objet d'un plan de pulvérisation. Ce modèle est applicable à d'autres maladies

à vecteurs mais l'intégration des dimensions temporelle et spatiale devra être spécifique à chaque maladie.

Une étude similaire est menée en 2010 par (Brahmi *et al.*, 2010) pour élaborer une carte de risque de nuisance des vecteurs de type *Aedes*. Cette étude, menée au Nord de la Tunisie, s'intéresse aux espèces de *Aedes* les plus fréquentes dans la zone ; il s'agit de l'*Aedes caspius* et de l' *Aedes detritus*. Les auteurs construisent une carte numérique de terrain en utilisant des images satellitaires et des cartes/photographies aériennes. Ainsi, en s'appuyant sur la technique NDVI (*Normalized Differential Vegetation Index*), ils s'intéressent aux zones humides et aux zones végétales qui sont des zones propices au développement et à la multiplication des vecteurs. Pour déterminer la carte de nuisance, ils font un croisement entre une carte d'aléa (probabilité d'occurrence des gites larvaires sur un site) et une carte de vulnérabilité (densité de la population, organisation des collectivités, etc.). En perpective les auteurs proposent une granularité plus fine de la dimension spatiale.

En utilisant les mêmes outils auxquels ils appliquent des algorithmes de classification, (Girdary et Grandchamp,) proposent un SIG sur la dengue en Guadeloupe. L'objectif est de déterminer le lien entre les contextes d'habitation, leur évolution et la propagation de la maladie. L'équipe s'appuie sur des données environnementales, cadastrales et démographiques. Ainsi, elle arrive à déterminer les contextes d'habitation dans lesquels la dengue a le plus de chance de se répandre ; leurs conclusions permettront aux autorités de proposer un plan d'intervention lié au risque de propagation. La seule limite de ces travaux est liée au format "raster" des cartes utilisées. En effet, les données socio-démographiques et physiques seraient mieux exploitables à partir des formats "vecteurs".

Limites Les modèles géograhiques permettent de visualiser la distribution spatiale de la maladie et ainsi de fournir des cartes de risque. Pour la mise en place d'un système géographique, il est nécessaire de disposer de données quantifiantes sur les éléments intervenants et des données géo-référencées sur les entités spatiales. Mais ces modèles montrent leurs limites dans la recherche de corrélations entre les éléments. Il est difficile d'expliquer l'impact de certains éléments sur d'autres et encore mois d'évaluer les variables proportionnelles.

3.1.2.1.2 Modèles épidémiologiques

Les modèles épidémiologiques, quant à eux, permettent de décrire les mécanismes de la transmission de la maladie de manière explicite, en imposant qu'une cause engendre un effet et ainsi de simuler la dynamique de la maladie à partir de conditions initiales données et d'identifier les facteurs les plus déterminants dans le cycle de la maladie. Ils imposent la maitrise du schéma de transmission de la maladie étudiée.

Les premiers travaux de modélisation mathématique de l'épidémiologie ont porté sur la variole. Afin de mesurer le risque de contagion et l'impact de cette maladie suivant la classe d'âge, (Bernoulli, 1760) fait une analyse comparée de la variolisation au moyen d'un modèle mathématique. Il a pour objectif de savoir si la variolisation (l'inoculation du pus d'une personne atteinte de variole) était plus avantageuse ou plus risquée pour les personnes ayant contracté cette maladie. Il propose d'évaluer la probabilité d'être immunisé et resté en vie ainsi :

$$\frac{dw}{da} = [1 - c(a)]\lambda(a)u(a) - \mu(a)w$$

Les variables représentent les éléments ci-dessous :

— c(a) est le taux de létalité ;
— $\lambda(a)$ est le taux selon lequel les sujets sensibles sont infectés ;
— u(a) représente la probabilité pour un nouveau-né d'être en vie et sensible à l'âge a ;
— $\mu(a)$ est le taux de mortalité causé par d'autres maladies en dehors de cette infection.

Des travaux de (Bernoulli, 1760), (Sabatier *et al.*, 2005) tirent plusieurs conclusions dont :

— la faible probabilité de l'apparition de la maladie chez l'adulte ;
— la mortalité totale causée par la variole sur une même cohorte (100 morts pour 1 300 nouveau-nés, ce qui est en accord avec certaines observations de l'époque) ;
— la probabilité de mortalité des personnes atteintes est plus fréquente pour les enfants de moins de 5 ans ;
— l'âge moyen de la cohorte imaginaire supposée exempte de variole grâce à la variolisation, d'abord supposée sans risque, serait de 29,9 ans contre 26,7 ans sans cette variolisation.

Ainsi, il démontre que le procédé de la variolisation est efficace pour prévenir la variole en termes de nombre de morts évités et de gain d'espérance de vie (Valleron, 2000).

Entre temps, d'autres travaux ont été menés. En 1911, (Ross, 1911) propose un modèle mathématique sur le paludisme en se basant sur deux (2) variables que sont la proportion des humains infectieux et celle des moustiques infectieux. Pour éradiquer cette maladie, la proportion des moustiques infectieux doit être inférieure à un certain nombre. En considérant que le système de santé d'une maladie transmissible intégre des individus saints (susceptibles d'être infectés) et des individus malades (porteurs de la maladie), il est proposé dès lors plusiers variantes de modèle SI (Susceptibes, Infectés). Ainsi, une décennie plus tard, en 1927, il est proposé un modèle sur la dynamique de la transmission des maladies infectieuses. Ce modèle sera décrit par Kermack et McKendrick en trois articles ((Kermack et Mckendrick, 1927) ; (Kermack et McKendrick, 1932) ; (Kermack et McKendrick, 1933)). Ils décrivent la dynamique de transmission comme dépendante de la fréquence et de l'intensité des interactions entre les individus susceptibles (S) et les individus infectés et infectieux (I). Les auteurs introduisent la variable R qui représente les individus rétablis.

Dans le contexte des maladies vectorielles, les modèles épidémiologiques permettent d'illustrer le potentiel des maladies pour évaluer leur propagation. (Rogers et Packer, 1993) décrivent l'élément de mesure comme étant le taux de reproduction R_0 qui représente le nombre de nouveaux cas de la maladie dans une population susceptible. Ce modèle ne s'applique que dans la première phase de déclenchement de la maladie car l'imminuté ralentit la vitesse de propagation de la maladie. Il permet également de montrer que la prévalence de la maladie et la capacité de la vaccination sont simplement lié à Ro. L'équation définie est la suivante :

($a^2 * b * c * m *$ exp [-uT]) $\div$ (u $*$ r)

Les variables sont décrites comme suit :

— a, est le taux de piqûres du vecteur ;
— b et c, sont des coefficients de transmission de vecteur aux vertébrés et inversement ;
— m, est le rapport entre le nombre de vecteurs et le nombre d'hôtes ;
— u, est le taux de mortalité des vecteurs ;
— T, est la durée du cycle de développement du parasite dans le vecteur (période d'incubation) ;
— r, est le taux de récupération de l'hôte vertébré contre l'infection.

Toutes les variables, à l'exception de r, sont liées aux vecteurs. Certaines variables sont liées à la température. Ces travaux abordent ainsi la problématique des interactions entre la propagation des vecteurs et le climat de façon indirecte.

Limites Les modèles épidémiologiques permettent de simuler de manière réaliste la dynamique d'une maladie. Les formules les plus courantes sont :

— la capacité vectorielle qui représente le nombre moyen de piqûres que les vecteurs, ayant piqué un individu infectant le jour t, infligent à la population d'hôtes pendant le reste de leur vie, une fois achevé le cycle d'incubation extrinsèque ;
— le taux de reproduction de base désigne le nombre de cas secondaires générés à partir de l'introduction d'un premier cas infecté dans une population d'hôtes sensibles ; il traduit la notion du seuil pour qu'une maladie se propage.

Pour atteindre cet objectif, il est nécessaire de disposer du schéma de transmission de la maladie étudiée. Mais cette classe de modèle présente ses limites. D'une part, les modèles mathématiques sont "spécifiques". Il ne peuvent s'appliquer qu'aux cas de figures définis dans l'hypothèse de départ. Dès lors, il est important de bien définir les hypothèses de départ et les limites de validité. D'autre part, ces modèles ne sont pas prédictifs. En effet, il n'est pas possible d'appliquer certaines mesures pour définir la dynamique de la maladie.

3.1.2.2 Modèles conceptuels

Les modèles conceptuels permettent de fournir des explications qualitatives des phénomènes. Il s'agit d'identifier les objets, de décrire leur structure (caractéristiques et actions) et de déterminer les liens d'association et les règles de gestion associées. Les approches les plus populaires sont les approches relationnelle, agent et multidimensionnelle.

Les approches relationnelles ne prennent en compte la structure des objets que du point de vue de leurs propriétés. Les actions sont traitées comme des processus indépendants. Les approches agents conçoivent l'objet comme une entité autonome « intelligente » (Elfazziki *et al.*, 2006). A titre d'exemple, nous avons étudié le modèle agent de (Cissé *et al.*, 2012). Le modèle proposé permet de représenter le cycle de transmission de la FVR et l'impact des paramètres climatiques sur l'émergence de cette maladie. Ainsi, en partant de l'hypothèse environnementale définie par (Ndione *et al.*, 2003), les auteurs utilisent le langage UML pour décrire les principaux objets :

— l'agent Vecteur duquel découle les spécifications "Culex" et "Aedes" (d'après (Fontenille *et al.*, 1998) et (Moutailler *et al.*, 2008), ce sont les principaux vecteurs responsables de la FVR) ;
— l'agent Hôte qui désigne les ovins, les caprins et les bovins ;

— l'environnement géographique traduit pas les objets "Mares" (les Culex pondent sur l'eau des mares) et "Zones Mares" (les Aedes pondent sur des substrats humides des mares).

Les fonctions qui sont représentées dans le modèle sont (i) la capacité que les vecteurs ont d'identifier les mares les plus proches de leur position (perception) ; (ii) les phases de piqûres des vecteurs sur les hôtes (interaction) ; (iii) les occurences de la maladie (émergence). Ainsi, en s'appuyant sur la plateforme de simulation GAMA, leur implémentation leur permet d'expérimenter des scénarii suivant la variation pluviométrique qui confirment les résultats obtenus par (Ndione *et al.*, 2006) sur les occurrences périodiques de la Fièvre de la Vallée du Rift dans la zone de Barkédji. La modélisation à base d'agents permet de s'affranchir des limites de modélisation analytique classique. Cependant, cette classe de modèle présente quelques limites :

— elle ne permet pas de simuler l'ensemble des dynamiques ;
— la structure des objets n'est pas conforme à une étude multi-critères ;
— elle n'est pas adaptée à une analyse descriptive ou de regroupement.

Les approches multidimensionnelles s'intéressent plutôt à l'analyse de l'objet suivant différentes perspectives.

Dans notre contexte, l'approche multidimensionnelle est la plus adaptée car la modélisation multidimensionnelle offre un formalisme dédié aux systèmes décisionnels. (Ravat *et al.*, 2001) décrivent la modélisation multidimensionnelle comme une démarche de représentation qui consiste à considérer un sujet analysé comme un point dans un espace à plusieurs dimensions. Ainsi, pour représenter un objet ou une donnée complexe, il est nécessaire de représenter, en plus des descripteurs de bas niveau, des connaissances et des métadonnées associées. Le défi est alors de trouver de nouveaux modèles, pour représenter les objets complexes, orientés analyse. Cela passe par la définition de nouvelles métriques afin d'agréger ou de comparer les objets complexes entre eux. En effet, les métriques quantitatives sont insuffisantes d'où la nécessité de définir des métriques qualitatives ou sémantiques. Un autre problème à traiter est la représentation des connaissances associées à ces objets, mais également à l'ensemble du processus.

Limites Les modèles conceptuels permettent de fournir une explication qualitative d'une maladie. Pour leur mise en œuvre, il est nécessaire d'identifier les objets, de les décrire et de maitriser les règles de gestion entre les objets. Ainsi, ils représentent les données à des fins transactionnelles (opération courante) ou décisionnelles (analyse approfondie). Cependant, ces modèles ne prennent pas en compte les dynamique spatiales et temporelles des maladies.

Conclusion Face au développement des maladies vectorielles, il serait donc intéressant de faire une étude plus poussée afin de mutualiser les différentes méthodes de représentation pour la prise en compte des objets sur tous les aspects : structure, fonctionnement et dynamique. Cette approche fédératrice pourrait nous permettre d'intégrer :

— les données issues des modèles conceptuels dans des SIG ;
— les résultats des modèles épidémiologiques comme données d'analyse.

3.1.3 Synthèse

Cette analyse comparative justifie aisément le choix d'un environnement décisionnel pour mettre en place une plateforme destinée aux gestionnaires et aux décideurs des diverses structures du projet. La technologie "décisionnelle" répond bien au besoin d'analyse approfondie des données pertinentes, issues des systèmes opérationnels. Un Système d'Information Décisionnel est un outil destiné à recueillir, à organiser, à traiter et à diffuser des données pour obtenir des informations, des connaissances. Ainsi, les données dans leur, représentation physique, doivent être structurées de façon à pouvoir faciliter l'interprétation et l'analyse. Il devient alors nécessaire d'avoir des outils puissants et fiables qui faciliteront l'aide à la prise de décision, le pilotage en s'appuyant sur un système de données. Les entrepôts de données ont été conçus de manière à répondre à ce besoin ; en effet, alimenté à partir de données d'origines diverses, il permet d'en faire ressortir les aspects informationnels qui pourront être analysés suivant divers angles suivant les besoins. Le système décisionnel est composé de trois principaux composants : les sources de données, l'entrepôt de données (base des données d'analyse) et les outils pour l'interrogation de l'ensemble de données. La mise en place de notre base de stockage (entrepôt de donnée)s nécessite un travail préalable d'organisation et de représentation de données ; le formalisme le plus approprié est celui de la modélisation multidimensionnelle. La modélisation des données est un élément fondamental dans la démarche de spécification d'un système d'information quel qu'il soit. Nous allons dans ce sens définir les concepts fondamentaux afin d'avoir une parfaite maîtrise de la modélisation multidimensionnelle.

3.2 Définitions préliminaires

3.2.1 Entrepôt de données

Dans le cadre de sa recherche sur la représentation des données de banques, (Codd, 1970), propose la structure en schéma relationnel dans lequel les données sont regroupées dans des tables ; les informations des enregistrements (lignes d'une table) sont utilisées pour identifier les liens. On parle dès lors de modèle relationnel duquel va découler la modélisation entité-relation sur laquelle les bases de données s'appuient. Une base de données est définie comme un ensemble structuré et cohérent de données. Cette méthode de stockage a vu le jour pour répondre aux problématiques liées à la sauvegarde dans des fichiers plats, (Gardarin, 2003) : redondance, non cohérence, sécurité, maintenance, accès simultanés, etc. Les applications basées sur les bases de données relationnelles étaient de types opérationnels pour répondre aux besoins basiques des utilisateurs que sont le stockage, la mise à jour et la consultation en temps réel. Les traitements sont assurés par les processus transactionnels appelées OLPT (*On Line Transactional Processing*). (Boly, 2006) dans sa synthèse, basée sur les articles de (Codd *et al.*, 1993), (**?**) et (Kimball, 1996), fait ressortir les caractéristiques des OLPT à travers trois (3) éléments :

- le traitement réduit du nombre d'enregistrements ;
- la limitation des requêtes au mise à jour ;
- l'accès en parallèle.

Contrairement, aux entrepôts de données, qui se base sur la technique OLAP (*On Line Analytical Processing*) qui est conçue pour prendre en charge :

— un nombre beaucoup plus important de données ;
— des traitements plus complexes orientés analyse sur un volume plus élevé de données ;
— un nombre limité d'utilisateurs pour la définition des modèles d'analyse et leur consultation.

Par ailleurs, (Codd *et al.*, 1993) annonce douze (12) règles qui régissent la conception des systèmes OLAP :

— le Modèle OLAP est multidimensionnel par nature ;
— l'emplacement physique du serveur OLAP est transparent pour l'utilisateur ;
— l'utilisateur OLAP dispose de l'accessibilité à toutes les données nécessaires à ses analyses ;
— la performance des rapports restent stables indépendamment du nombre de dimensions ;
— le serveur OLAP s'intègre dans une architecture client serveur ;
— le dimensionnement est générique afin de ne pas fausser les analyses ;
— le serveur OLAP assure la gestion des données clairsemées ;
— le serveur OLAP offre un support multi-utilisateurs (gestion des mises à jour, intégrité, sécurité) ;
— le serveur OLAP permet la réalisation d'opérations inter dimensions sans restriction ;
— le serveur OLAP permet une manipulation intuitive des données ;
— la flexibilité (ou souplesse) de l'édition des rapports est intrinsèque au modèle ;
— le nombre de dimensions et de niveaux d'aggrégation possibles est suffisant pour autoriser les analyses les plus poussées.

Suite à cette publication, plusieurs systèmes décisionnels adoptent l'architecture basée sur les entrepôts de données. Ce qui justifie aisément le choix de l'entrepôt de données pour mettre en place notre base de stockage destinée aux gestionnaires et aux décideurs des diverses structures. La technologie OLAP répond bien au besoin d'analyse approfondie des données pertinentes, issues des systèmes opérationnels (Wehrle, 2009).

Un entrepôt de données a pour objectif principal de centraliser des données issues de différentes sources de données ; sa structure conceptuelle répond aisément aux exigences stratégiques définies par les acteurs décisionnels. Cette définition est tirée de celle de (Inmon, 2005), père du concept qui la présente ainsi « Un datawarehouse est une collection de données thématiques, intégrées, non volatiles et historisées pour la prise de décision » [1]. Il résume également ses caractéristiques en quatre points : « Orientée sujet, intégrée, non volatile et historisée ».

— le *datawarehouse* est orienté sujets, cela signifie que les données collectées doivent être orientées « métier » et donc triées par thème : cette représentation offre un avantage sur l'analyse des données qui peut se faire sur différents sujets des sous structures fonctionnelles et organisationnelles de la structure. Toutefois, le rattachement de la donnée à une seule structure est inévitable pour pallier le problème de redondance ;

1. traduit de l'anglais par « Comment ça marche»

— le *datawarehouse* est composé de données intégrées, c'est-à-dire qu'un « nettoyage » préalable des données est nécessaire dans un souci de rationalisation et de normalisation : ce processus couramment dénommé « ETL » est déterminant sur la qualité et la quantité d'informations nécessaires pour le processus de décision. Il s'agit d'aboutir à un format unifié pour garantir la cohérence, la non redondance . . . des données. Cela nécessite la définition d'un référentiel des données prenant en compte les contraintes d'intégrité, les règles de gestion, la sémantique. Ce référentiel constitue les métadonnées ;
— les données du *datawarehouse* sont non volatiles ce qui signifie qu'une donnée entrée dans l'entrepôt l'est définitivement et ne peut être supprimée ; cette caractéristique offre la garantie de traçabilité des événements antérieurs (informations, décisions . . .) ;
— les données du *datawarehouse* doivent être historisées, donc datées : cela permet de prendre en compte la dimension temporelle pour suivre l'évolution dans le temps des données.

Dans ses travaux sur les techniques de construction des entrepôts de données multidimensionnelles, (Kimball, 1996) propose la définition suivante : « Un entrepôt de données est un espace de stockage centralisé sur lequel repose un système décisionnel, son rôle est d'intégrer et de stocker l'information utile aux décideurs et de conserver l'historique des données pour supporter les analyses effectuées lors de la prise de décision. »

Cette pièce maîtresse de l'informatique décisionnelle organise donc les données de manière à faciliter la prise de décision grâce à des outils d'analyse en ligne ou de fouille de données.

Desnos(2003), dans sa synthèse sur la présentation des entrepôts de données , classe les données des entrepôts dans différentes catégories « selon un axe de synthèse ou d'historique » :

— les données agrégées correspondent à des éléments d'analyse représentant les besoins des utilisateurs. Elles constituent déjà un résultat d'analyse et une synthèse de l'information contenue dans le système décisionnel, et peuvent être facilement accessibles et compréhensibles ;
— les données détaillées reflètent les événements les plus récents. Les intégrations régulières des données issues des systèmes de production (système opérationnel) sont généralement réalisées à ce niveau ;
— les métadonnées constituent l'ensemble des données qui décrivent des règles ou processus attachés à d'autres données. Ces dernières constituent la finalité du système d'information ;
— les données historisées - chaque nouvelle insertion de données provenant du système de production ne détruit pas les anciennes valeurs, mais créée une nouvelle occurrence de la donnée.

La modélisation des données est un élément fondamental dans la démarche de spécification d'un système d'information quelqu'il soit. La mise en place des entrepôts de données nécessite un travail préalable d'organisation et de représentation de données ; le formalisme le plus approprié est celui de la modélisation multidimensionnelle. Nous allons dans ce sens définir les concepts fondamentaux afin d'avoir une parfaite maîtrise de la modélisation des

données.

3.2.2 Modélisation multidimensionnelle

La modélisation multidimensionnelle est une approche dédiée aux systèmes décisionnels. (Ravat *et al.*, 2001) décrivent la modélisation multidimensionnelle comme une « démarche de représentation qui consiste à considérer un sujet analysé comme un point dans un espace à plusieurs dimensions ».

Les données sont organisées de manière à mettre en évidence les données quantitatives (les faits) et données qualifiantes (les dimensions).

3.2.2.1 Concept de fait

Le sujet analysé est représenté par le concept de fait. « Le fait modélise le sujet de l'analyse. Un fait est formé de mesures correspondant aux différentes valeurs de l'activité analysée » (Teste, 2000).

Le fait représente un indicateur susceptible de constituer tout ou une partie du résultat d'une requête. Les mesures d'un « fait » sont numériques et généralement valorisées de manière continue. Les mesures sont numériques pour permettre de résumer un grand nombre d'enregistrements en quelques enregistrements (addition, nombre, calcul du minimum, etc.). Les mesures sont valorisées de façon continue car il est important de ne pas valoriser le fait avec des valeurs nulles. Elles sont aussi souvent additives ou semi-additives afin de pouvoir les combiner au moyen d'opérateurs agrégatifs.

Nous pouvons considérer comme fait la « Maladie » avec comme mesure "nombre de malades souffrant de la FVR".

3.2.2.2 Concept de dimension

Le sujet analysé, c'est à dire le fait, est appréhendé suivant différentes perspectives. Elles correspondent à une catégorie utilisée pour caractériser les mesures d'activité analysées ; il s'agit des dimensions.
(Wehrle, 2009) décrit ainsi la dimension : « Chaque dimension représente un axe d'analyse qui permet d'avoir des vues suivant les paramètres de la dimension ; elle doit pouvoir intervenir comme un critère dans les traitements analytiques. » Chaque dimension est formée par un ensemble d'attributs et chaque attribut peut prendre différentes valeurs. Dans sa thèse, (Teste, 2000) donne la définition suivante : « Une dimension modélise une perspective de l'analyse. Une dimension se compose de paramètres correspondant aux informations faisant varier les mesures de l'activité. »

Les dimensions servent à enregistrer les valeurs pour lesquelles sont analysées les mesures de l'activité. Une dimension est généralement formée de paramètres (ou attributs) textuels et discrets. Les paramètres textuels sont utilisés pour restreindre la portée des requêtes afin de limiter la taille des réponses. Les paramètres sont discrets, c'est à dire que les valeurs possibles sont bien déterminées et sont des descripteurs constants.

Par exemple, dans la requête "Nombre de malades (de la FVR) dans la région de Kolda durant le mois de Janvier 2010". Nous trouvons les paramètres : lieu "Kolda" et période "Janvier 2010".

(Bellatreche, 2003) justifie la décomposition des dimensions par le niveau de granularité requis. Les dimensions possèdent en général des hiérarchies associées qui organisent les attributs à différents niveaux pour observer les données à différentes granularités. Une dimension peut avoir plusieurs hiérarchies associées, chacune spécifiant différentes relations d'ordre entre ses attributs.

3.2.2.3 Concept de hiérarchie

Lors du processus d'analyse, les données sont généralement analysées en partant d'un faible niveau de détail vers des données plus détaillées pour "forer" vers le bas. Pour définir ces différents niveaux de détail, chaque dimension est munie d'une (ou plusieurs) hiérarchie(s) des paramètres. La hiérarchie sert lors des analyses pour restreindre ou accroître les niveaux de détail de l'analyse. (Teste, 2000) propose la définition suivante : « Une hiérarchie organise les paramètres d'une dimension selon une relation « est plus fin »conformément à leur niveau de détail. »

Toujours à travers l'exemple précédent, le paramètre « lieu » pourrait même être affiné en précisant la ville, la région, le département, etc. Ainsi, une granularité plus fine (hiérarchie) est obtenue.

3.3 Fouille de données

La collecte de données est une des étapes fondamentales nécessaire à la gestion de tout système opérationnel. Aussi conscient de cette problématique, plusieurs campagnes sont régulièrement organisées dans la zone d'étude de Barkédji pour recueillir les données environnementales et sanitaires. Malgré toutes ces procédures, les équipes sont souvent limitées dans la production d'information à forte valeur ajoutée car les données sont peu exploitées. Les équipes de recherche utilisent des outils statistiques classiques qui ne sont pas adaptés à l'analyse de grand volume de données de structure complexe. Ces outils statistiques se basent plus sur des hypothèses formelles. Pourtant, une analyse approfondie et corrélée des données recueillies pourrait fournir des informations importantes qui permettraient de comprendre et de maitriser le risque sanitaire. Pour répondre à cette opportunité, la fouille de données est un domaine de recherche opérationnel qui propose des algorithmes d'analyse de données dans le but d'en extraire des connaissances. Le *datamining* propose de nombreuses techniques qui diffèrent par leurs paramètres d'entrée, de sortie suivant les objectifs en termes de résultats. Ainsi, la phase initiale de ce processus est d'identifier le type de problème pour pouvoir choisir l'algorithme le mieux adapté après une étude comparative. Il est clair que chacune des méthodes présente ses forces et ses limites. Dans un tel contexte, le consensus qui en ressort est d'adopter une approche intégrée, qui s'appuie sur divers algorithmes, afin de valoriser au maximum les données. Le *datamining* est le moteur essentiel du processus décisionnel. Aussi, il repose sur la mise

en place d'une structure de données conforme à l'analyse décisionnelle : l'entrepôt de données. Cette structure de données est basée sur le formalisme multidimensionnel.

3.3.1 Description

La fouille de données a pour objectif de faire ressortir des informations et des connaissances à priori non évidentes à partir de données brutes. Cette technologie propose deux (2) grandes techniques pour le traitement de ces données : l'apprentissage supervisé et l'apprentissage non supervisé. L'apprentissage supervisé est utilisé pour prédire des valeurs de certaines variables à partir d'autres paramètres connus. En revanche, l'apprentissage non supervisé est plus utilisé dans le cas où les valeurs des paramètres prédéfinis ne sont pas connues et les résultats recherchés non plus. Les données mises à notre disposition sont d'une grande diversité aussi bien en terme de temporalité que de la sémantique. Nous nous situons plus dans le cadre de l'apprentissage supervisé car les variables à expliquer sont connues et labelisées. Par ailleurs, la fouille de données classe ses algorithmes en quatre (4) groupes (Obenshain, 2004) :

- les algorithmes de prédiction qui permettent de prédire la valeur continue ou discrète de certaines données en fonction d'autres. Dans notre contexte, il pourrait s'agir de prédire la densité vectorielle en fonction de la pluviométrie ;
- les algorithmes de classification qui déterminent des modèles pour prédire des valeurs discrètes en fonction de données d'entrée. On peut utiliser des algorithmes pour l'apprentissage supervisé (arbre décisionnel, réseau de neurones, etc.) et d'autres pour l'apprentissage non supervisé (clustering, réseau de Konohen, etc.). On pourrait évaluer le comportement normal ou anormal des mares en fonction des conditions climatiques ;
- les algorithmes d'exploration permettent de faire des regroupements d'objet en fonction d'attributs similaires. On utilise généralement les arbres décisionnels pour l'apprentissage supervisé et l'analyse des liens ou le clustering pour l'apprentissage non supervisé. Avec ces algorithmes, on pourrait identifier des groupes de mares suivant les caractéristiques de leurs eaux ;
- les algorithmes d'analyse d'affinité qui permettent de déterminer les évènements susceptibles de se produire en conjonction avec d'autres. Cette classe d'algorithmes est utilisée pour l'apprentissage non supervisé avec les algorithmes d'associations et de séquences. On pourrait établir un lien proportionnel entre la densité vectorielle et le niveau d'eau des mares en fonction de la pluviométrie.

Chacun des algorithmes a des points forts et des points faibles. En fonction du type de résultats attendus, on utiliserait tantôt l'un, tantôt l'autre. Notre approche a pour ambition d'offrr une vision transversale afin de pouvoir extraire le maximum d'informations qui permettraient de comprendre et de prédire la maladie de la FVR. Nous nous proposons d'approfondir ces différentes approches, en analysant des contextes applicatifs dans lesquels elles ont été implémentées.

3.3.2 Quelques applications aux données de santé

Pour évaluer les épidémies dans le temps et l'espace, plusieurs tendances sont proposées. Les approches rétrospectives pour les périodes antérieures, les approches prospectives pour une épidémie en cours et les approches prédictives pour une épidémie future.

La plage temporelle des données collectées nous limitent dans une approche rétrospective qui nous permettrait de justifier et/ou d'identifier des interactions entre différents paramètres. Cette approche est également une porte d'entrée pour l'analyse prédictive.

L'équipe de (Robertson *et al.*, 2010), dans leur article sur la surveillance en santé publique, constate que la plupart des travaux sur la fouille de données en épidémiologie se base sur le couplage entre les méthodes statistiques, les outils de visualisation et les SIG pour la détection des épidémies et le suivi de l'évolution spatio-temporelle. Les travaux de (Eisen et Eisen, 2011) présentent les évolutions de la modélisation spatio-temporelle, des systèmes décisionnels et des systèmes d'information géographique (SIG) comme moyens de prévention et de lutte contre les maladies à vecteurs telles que le paludisme, la dengue, la peste. En effet, pour la gestion spatio-temporelle des maladies vectorielles, plusieurs études ont permis de démontrer que l'application des techniques de fouilles sur des données spatiales permettrait d'obtenir des informations géo localisées et de déterminer des liens de voisinage entre des objets géo référencés. Les travaux de recherche sur la représentation de l'impact de l'environnement sur la santé et plus particulièrement sur les maladies vectorielles, (Soti, 2011), ont permis de distinguer deux grandes classes de méthodes :

— les méthodes statistiques permettant de comprendre la distribution de la maladie ou celle des vecteurs impliqués en fonction des conditions environnementales. Pour être construit, il est requis des données quantifiantes sur les éléments intervenants ;
— les méthodes basées sur le processus permettant de décrire les mécanismes de la transmission de la maladie de manière explicite et ainsi de simuler la dynamique de la maladie à partir de conditions initiales données et d'identifier les facteurs les plus déterminants dans le cycle de la maladie. Ils imposent la maitrise du schéma de transmission de la maladie étudiée.

L'approche retenue dans la thèse de (Soti, 2011) prend en compte la distribution et la dynamique des gîtes larvaires des vecteurs en intégrant des paramètres environnementaux pour identifier les zones et les périodes à risque d'émergence de la FVR dans la région de Barkédji. Ainsi, en utilisant la télédétection, elle présente comme premiers travaux la cartographie des zones à risque de circulation de la FVR. Ses seconds travaux proposent un modèle de dynamique de la population des principaux vecteurs conducteurs du virus (*Ae. vexans* et *Cx. Poicilipes*) qui intègre la dynamique des mares de la zone d'étude. Ces résultats permettent de confirmer, d'une part, le fort impact de la présence des mares dans la densité vectorielle et d'autre part, que les vecteurs identifiés ont été plus présents lors des périodes de circulation du virus de la FVR. Les perspectives de ce travail de recherche se dessinent vers un modèle de diffusion spatiale des vecteurs et une approche multifactorielle de différents critères (environnementaux, climatiques . . .) sur les agents pathogènes et les hôtes. L'analyse décisionnelle de la grippe conduit (Hechmati, 2004) à proposer un système d'information intégrant un tableau de bord qui présente une vue corrélée de tous les indica-

teurs. Pour construire ce tableau de bord, il représente les deux (2) scenarii (porteurs sains et population malade) à travers les différentes acteurs et interventions qu'ils subissent (état sanitaire, transport, structure sanitaire, etc.). Par ailleurs, il établit un algorithme sous Excel pour faire une analyse prédictive des personnes infectées en période d'épidémie ou de pandémie. Cette évaluation est faite en se basant sur les données d'épidémies précédentes. Ainsi, le tableau de bord, destiné aux décideurs, leur permet de constater les évolutions de l'épidémie, d'évaluer la répartition des ressources et de mesurer l'impact de l'épidémie. Toutefois, les discussions qui en découlent font ressortir la limite de la fréquence de mise à jour des indicateurs qui joue considérablement sur l'évaluation temporelle de la maladie. En effet, la répartition temporelle des données brutes permettrait de mettre en évidence les éventuelles similarités entre des périodes suivant des paramètres environnementaux et ainsi de prédire les dynamiques des acteurs impliquées suivant les mêmes paramètres. Par ailleurs, plusieurs équipes ont développé des plateformes décisionnelles qui s'appuient sur les SIG pour le suivi des répartitions des maladies vectorielles, l'évaluation des vecteurs suivant des zones géographiques. Dans leur recherche d'outils pour la visualisation de données spatio-temporelles, (Compieta *et al.*, 2007) ont proposé un système basé sur des techniques de visualisation 3D :

— Google Earth couplé à des couches géographiques pour l'affichage des résultats ;
— Java 3D pour la détermination d'interactions entre des données non géo référencées pour évaluer la répartition spatiale de certaines variables.

Dans le cas du paludisme, (Coleman *et al.*, 2006) présente un système décisionnel qui intègre les données climatiques, sanitaires (transmission de la maladie) et sur la résistance aux insecticides. Ainsi, il propose un système sur la distribution spatiale de la résistance des vecteurs responsables et de son impact potentiel sur la transmission du paludisme au profit des programmes de lutte anti vectorielle. (Saul *et al.*, 2008) utilisent Google Earth associé à d'autres logiciels (HealthMapper développé par l'OMS et SIGEpi développé par l'OPS) pour visualiser la répartition spatiale de la dengue au Mexique. Cet outil, représentant les images satellitaires des villes de Chetumal et Merid, permet de contrôler les zones à haut risque. En se basant sur le nombres de personnes infectées dans une ville et le nombre de maisons de la même ville, on peut ainsi évaluer l'incidence annuelle de la maladie par habitation, puis par ville. La démarche proposée peut être adaptée pour d'autres maladies vectorielles. Les limites de ce système sont essentiellement liées aux fonctionnalités de Google Earth que sont :

— la nécessite de disposer d'une bonne connexion Internet pour accéder aux images - ce qui n'est pas le cas pour la plupart des zones concernées ;
— la mauvaise qualité des images des zones rurales alors qu'elles sont le plus touchées par les maladies vectorielles ;
— l'âge des images ne correspond pas toujours à la réalité - en milieu urbain, l'environnement évolue rapidement.

Par ailleurs, on note d'autres inconvénients tel que le manque (1) d'outils de gestion de données, (2) de capacité de modélisation et (3) d'analyse spatiale. (Chanda *et al.*, 2012) ont également porté une étude sur le paludisme en utilisant les systèmes décisionnels basés sur les SIG. En utilisant les données sur les indicateurs du paludisme et celles sur la résistance aux insecticides des vecteurs concernés, il serait possible de formuler des

actions de gouvernance et d'aider à l'utilisation rentable des ressources limitées pour un meilleur contrôle des vecteurs du paludisme.

En 2008, une équipe de chercheurs indiens, (Murty *et al.*, 2008), s'appuie sur les cartes de Kohonen (cartes auto organisatrices) pour classifier les zones, du quartier de Manipur, selon le niveau (haut, moyen, bas) d'endémie de la malaria. Les travaux de (Fathima *et al.*, 2011), sur les modèles de diagnostic et de pronostic des maladies vectorielles, décrivent les infections à arbovirus comme sujets à confusion. Ses recherches proposent l'approche par classification pour distinguer les infections à arbovirus. Ainsi, le principal auteur, (Fathima et Manimeglai, 2012), a ensuite présenté la technique SVM de classification pour l'identification de la maladie de la Dengue. Son étude lui a permis de mettre en place un système, qui à partir des données virales, donne un taux de prédiction assez précis de la Dengue. Il est également proposé, (Smitha et Sundaram, 2012), la technique d'arbre de décision pour prédire les occurrences d'une maladie, classer les habitants suivant leur état (susceptible, infecté, immunisé) par rapport à la maladie dans une zone. Dans leurs études sur la filariose de Bancroft, (Kumar *et al.*, 2005) regroupent des données entomologiques, climatologiques, socio-économiques auxquelles ils appliquent des algorithmes de règles de classification et de régression pour prévoir la densité mensuelle des moustiques. Ils utilisent des données entomologiques, météorologiques et socio-économiques. Leurs travaux font ressortir deux principaux résultats : (1) l'humidité relative influe fortement sur la densité vectorielle ; (2) la pluviométrie, la température et la vitesse du vent sont proportionnelles à la densité vectorielle. En ce qui concerne la prédiction, (Stevens et Pfeiffer, 2011) ont présenté une synthèse des méthodes d'analyse spatiale en épidémiologie en les classant en trois (3) catégories. On distingue ainsi les méthodes traditionnelles qui se basent sur la modélisation généralisée (GAM, Bayésion, GLM), les méthodes purement spatiales (MAXENT, GARP) et les méthodes décisionnelles d'analyse multi critère (MCDA).

3.3.3 Synthèse

La fouille de données classique offre de nombreux algorithmes soit dans un but d'analyse exploratoire soit dans un but d'analyse prédictive. Dans le cadre d'étude sur l'environnement et l'épidémiologie, de nombreuses caractéristiques sont liées à un attribut temporel et/ou un attribut spatial. Ainsi, dans le but d'analyser ce type de données, il est nécessaire d'étudier les techniques de fouille de données adaptées à l'analyse spatio-temporelle. En épidémiologie, l'analyse spatio-temporelle permettrait de suivre l'évolution des épidémies par types de région (risque faible ou élevé).
(Meliker et Sloan, 2011), dans leur revue bibliographique sur l'épidémiologie spatio-temporel, identifient des pistes de recherche suivant les sous domaines de recherche liés à cette approche. Ainsi, ils proposent :

- la prise en compte de la biologie des maladies ;
- le développement de protocoles d'échanges de données sécurisés ;
- le choix de méthodes de traitement statistiques suivant le jeu de données manipulé.

3.4 Fouille de données spatio-temporelle

La fouille de données spatio-temporelle est définie comme l'application des techniques de fouille de données classiques sur des données issues de SGBD géographiques. La dimension spatiale permet de définir la localisation d'un objet dans un espace géographique et la dimension temporelle le rattache à un pas de temps (date, période, etc.). Cette technique permet l'extraction de séquences récurrentes, appelées motifs, d'un ensemble de valeurs d'attributs pour certains objets. Ils permettent d'identifier des informations pertinentes dans un volume élevé de données. Les travaux sur les motifs spatio-temporels ont permis d'aborder les dimensions spatiales et temporelles suivant différentes approches.

3.4.1 Description

La fouille de données spatio-temporelles consiste en l'extraction de relations temporelles et/ou spatiales entre des objets qui intègrent une caratéristique sur leur position ou leur histoire. Dès lors, il devient nécessaire de stocker ces données dans une Base de Données Spatiales. Du point de vue de la structure, les SGBD spatiaux héritent des SGBDR mais les attributs spatiaux sont stockés sous forme de point, ligne ou polygone. La donnée temporelle est stockée comme une date dont le format peut différer suivant l'usage qui en est fait.

La manipulation de ces données se fait généralement sur l'attribut spatial qui les caractérise. De part la notion de coordonnées géographiques, il existe de nombreuses fonctions, spécifiques au SGBD géographique, permettant de faire ressortir les relations spatiales entre les objets (inclusion, union, etc.). Aussi, pour relever le défi de la performance algorithmique, il est également intégré un système d'indexation spécifique.

3.4.2 Panorama des cas d'applications

Les premiers travaux sur les motifs séquentiels ne prenaient en compte que la dimension temporelle (Agrawal et Srikant, 1995) ou la dimension spatiale.

Dans leur travaux, (Wang *et al.*, 2005) présentent l'ensemble des données dans des grilles séquentielles dans le temps. Chaque cellule de la grille correspond à un jeu de données dans une localisation précise. Plusieurs extractions peuvent en découler :

— un ensemble de jeu de données d'une même localisation et d'une même période temporelle ;
— un ensemble de jeu de données de la même localisation suivant l'évolution temporelle.

On peut donc extraire des données suivant des indicateurs de références (date et/ou localisation). La dimension spatiale est toutefois limitée à la grille.

(Tsoukatos et Gunopulos, 2001), dans leur article sur la granularité de l'espace, utilisent également le principe d'une grille d'évènements pour une dimension spatiale et d'un ensemble de grilles spatiales pour la dimension temporelle. Le niveau de granularité permet de fusionner certaines cellules d'une grille ; la valeur de l'élément granulé est proportionnelle au nombre de cellules fusionnées. Cela facilite la prise en charge de la hiérarchie

spatiale mais ne répond pas aux limites de l'approche.

Une méthode basée sur les graphes GenSpace est proposée par l'université de Regina ; en effet, l'équipe de (Hamilton *et al.*, 2006) cherche à identifier des anomalies dans un jeu de données minières combinant à la fois des indicateurs spatiaux et temporels. Cette technique permet de classer les séquences qui se rapprochent de faits réels et ainsi de réduire les coûts (temporel, stockage) de l'exploitation minière.

D'autres articles, tels que celui de (Huang *et al.*, 2008), proposent d'établir des relations de voisinage, suivant les dimensions spatiales et temporelles, entre les objets. Les règles d'association sont utilisées pour décrire les motifs. Toutefois, ces travaux ne permettent pas de prendre en charge les relations entre les objets spatiaux. En se basant également sur le principe de voisinage, (Salas *et al.*, 2012) intègre la proximité entre séquences. Ainsi, les séquences peuvent être regroupées (opérateur [] puis avoisiner (opérateur $\cdot$)). On note que seule la proximité spatiale est considérée. L'approche par voisinage est limitée par la prise en charge d'une seule échelle spatiale. En s'appuyant sur les algorithmes de (Salas *et al.*, 2012), (Fabrègue *et al.*, 2012) proposent de considérer à la fois les relations entre les objets spatiaux et la hiérarchie spatiale (motifs spatio-temporels reliés - MSTR). Cette approche permet d'obtenir plus de motifs en intégrant les niveaux de granularité spatiale les plus fins.

(Cao *et al.*, 2011) utilisent les motifs spatio-temporels pour la prédiction de rendement de maïs ; en effet, cette approche couplée à certaines techniques leur a permis :

— de déterminer l'impact de la distribution spatiale de la fertilité des sols à proximité de la zone d'étude en utilisant les réseaux de neurones ;
— de prédire le rendement en utilisant les régressions linéaires ;
— d'extraire des séquences en utilisant des fonctions statistiques.

(Fenzhen *et al.*, 2004) utilisent également la fouille de données spatiales pour déterminer les interactions entre les facteurs environnementaux et la distribution spatiale des poissons. La méthode proposée, STAMM (*Spatio Temporal Assignment Mining Model*), permet d'extraire les motifs constitués des indicateurs les plus influents sur la réparition des poissons. Cette approche utilise la technique de voisinage pour construire des tables de décision et les règles d'association récursives pour l'allocation spatio-temporelle.

3.4.3 Synthèse

Les missions sur les problématiques de santé épidémiologique regorgent de nombreuses données qui couvrent plusieurs périodes et les localités géographiques concernées. Ainsi, les données qui sont recueillies sont explicitement ou implicitement liées à des attributs géographique et/ou temporel. La surveillance qui en découle, les alertes et les réponses attendues nécessitent également de fournir des précisions sur la position géographique et temporel des processus épidémiologiques.

La fouille de données propose les motifs spatio-temporels pour détecter toutes les séquences d'attributs similaires ; cette approche permet d'exploiter les similarités entre des objets suivant une période ou dans un environnement géographique. La technique utilisée repose sur la recherche de données ou d'ensembles de données qui se répètent fréquemment ; c'est ainsi que l'on peut identifier les comportements fréquents et les rattacher à des dates et des positions.

Plusieurs algorithmes ont été proposés pour prendre en compte l'une des deux dimensions mais ces méthodes ont montré leurs limites dans la prise en charge simultanée des deux échelles et d'objets complexes. La méthode la plus récente et la plus adaptée à notre contexte est celle de (Salas *et al.*, 2012). L'algorithme détermine au cours du temps les occurrences des événements et leurs localisations. Ceci permet de mettre en évidence des corrélations entre ces deux dimensions. L' application a été faite sur l'évolution de la Dengue qui est une également une maladie vectorielle au même titre que la FVR.

3.5 Conclusion

Pour aider à comprendre et à maitriser les relations entre le climat, l'environnement et leur impact sur la maladie de la Fièvre de la Vallée du Rift, nous avons choisi de nous orienter vers la mise en place d'un environnement décisionnel. Notre choix se justifie par la recherche d'informations d'analyse et de prédiction pour les experts et les utilisateurs finaux afin de répondre à des questions stratégiques. Ainsi, nous nous proposons d'appliquer le processus d'extraction de connaissances qui se présente en deux (2) couches décrites dans la section de notre état de l'art.

La première couche est celle de la préparation et de l'accès aux données. Elle nécessite la définition d'une structure de données adaptée à l'analyse décisionnelle. Nous utilisons le formalisme de la modélisation multidimensionnelle pour décrire ce modèle. Cette étape est fondamentale car elle permet de construire l'entrepôt de données qui est la source de données des outils de la couche métier. La préparation des données se fera en plusieurs étapes. Après la collecte des données, il sera nécessaire de délimiter le domaine d'étude à travers la sélection des données et leur nettoyage pour ne pas prendre en compte les anomalies, les redondances, les formats variés, etc. La dernière étape sera le traitement des données pour intégrer les données issues de combinaisons et/ou de calculs préalables.

La deuxième couche prend en compte deux (2) sous couches : (i) la sous couche "présentation" chargée de la présentation des informations agrégées pour les utilisateurs finaux (DSV, PNLP) et (ii) la sous couche "métier", orientée pour l'extraction de connaissances, destinée aux utilisateurs scientifiques. Elle s'appuie sur la couche des données pour la sélection des données pertinentes et fournit les informations et les connaissances susceptibles d'apporter des réponses aux interrogations des scientifiques. Cette couche est constituée des outils d'analyse de données tels que les techniques de fouille de données. Notre état de l'art présente deux (2) approches de fouilles de données : (1) la fouille de données classique et (2) les techniques d'analyse spatio-temporelle. Les travaux présentés dans cette section s'articulent essentiellement autour de la problématique de la compréhension des épidémies. Les techniques de fouille de données sont très diversifées et s'avèrent souvent efficaces. Cependant, la principale critique qui en ressort est que les études présentées n'utilisent pas des données multi disciplinaires. Cette limite se révèle critique car elle ne permet pas de comprendre les épidémies en intégrant tous les systèmes qu'elles englobent. Le besoin d'analyse est souvent consacré à une partie des données ; il n'est pas possible de maitriser les éventuelles interactions entre tous les éléments. Ainsi, dans la suite de nos travaux, nous appliquerons les deux (2) approches présentées pour l'ana-

lyse des données. D'une part, il s'agira d'utiliser les algorithmes usuelles. Les avantages et les utilisations qui en découlent sont multiples : centralisation, corrélation des données, évaluation globale des données. D'autre part, nous nous orientons vers une analyse spatio-temporelle afin de comprendre l'impact des variations temporelles et spatiales.

CHAPITRE 4

Nos contributions

Les multiples données manipulées par les experts métiers sont traitées indépendamment les unes des autres en se limitant à la production de statistiques cloisonnées. Les utilisateurs sont ainsi noyés dans une masse de données qu'ils ont du mal à exploiter. Le défi à relever est de fournir, d'une part, aux utilisateurs finaux des données corrélées et d'autre part, aux scientifiques des informations plus précises sur les variables déterminantes et sur l'évolution spatio-temporelle de ces variables. Nous présentons dans ce chapitre les contributions apportées tout au long de nos travaux de recherche qui se situe à deux niveaux. s'agit essentiellement de :

1. la gestion des données :
   - nous proposons un modèle multidimensionnel qui intégre toutes les données manipulées des systèmes sanitaires et environnementaux ; il en découlera l'entrepôt de données ;
   - les maladies vectorielles peuvent avoir des cibles différentes mais leur point commun reste le vecteur qui "transporte" l'agent pathogène ; ainsi, nous avons proposé un modèle générique qui puisse prendre en charge d'autres maladies vectorielles telles que le palusdisme ;
2. la recherche d'informations :
   - la première étape nous conduit à l'analyse croisée et multidimensionnelle des données de l'entreprôt en nous appuyant sur un outil décisionnel ;
   - dans la deuxième étape, nous appliquons les techniques classiques de fouilles de données afin (i) d'identifier les interactions entre les données, (ii) de comprendre leur variabilité par zone géographique et/ou dans le temps ;
   - enfin, nous utilisons les motifs spatio-temporels pour identifier les facteurs déterminants et maitriser l'évolution des interactions de certaines données en intégrant à la fois les dynamiques spatiales et temporelles.

La première section de ce chapitre est consacrée à la présentation des méthodes et outils de mise en oeuvre.

4.1 Méthodes et Outils

Dans son article sur l'intégration d'un système décisionnel dans une organisation,(Bruley, 2010) fait ressortir le cycle d'évolution des données en informations, des informations en connaissances et des connaissances en plan d'actions. La donnée, décrivant un objet, se présente de façon brute. Traitée, elle permet de fournir des informations. Situer l'information dans un contexte, aboutit à une connaissance. Cette démarche est l'objectif fondamental des environnements décisionnels. Nous nous appuyons sur l'architecture basée sur le processus décisionnel proposée par (Marakas, 2003) qui se décline en quatre (4) composants (Figure 4.1) :

— le système de gestion des données (source de données) : il intègre les données dans leur état brut (base de données, fichier plat) ;
— le système de gestion des modèles (organisation des données) : il est assimilé au système de stockage des informations (les données quantitatives le constituent) ;
— le moteur de connaissances (moteurs de reporting et de fouille de données) : il s'agit du système de traitement des informations et d'analyse ;
— l'interface utilisateur : qui donne une vue à l'utilisateur final sur tous les outils (tableau de bord, rapport).

Ainsi, à partir des données brutes, issues des enquêtes de terrain et des tests de laboratoires, on construit un système de données consolidées (entrepôt de données ou *data-warehouse* - DW) qui seront analysées pour produire des informations (rapport, tableau de bord) et corrélées pour en extraire des connaissances (datamining - DM).

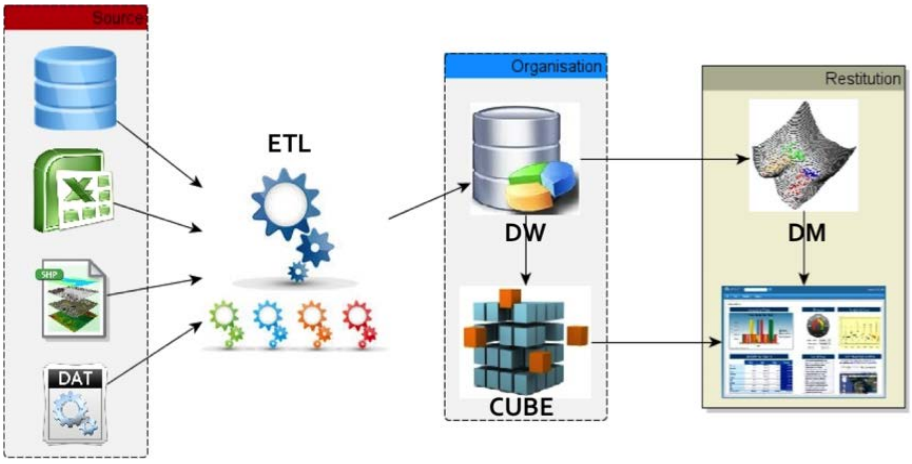

FIGURE 4.1 – Notre démarche décisionnelle.

4.1.1 Source de données

La première étape de notre travail a consisté à recueillir les données. Cette étape a demandé un temps de coordination et de communication assez important. Nous avons utilisé plusieurs méthodes dont les questionnaires soumis aux structures partenaires, les inter-

views des utilisateurs et l'analyse rétrospective des bilans et rapports périodiques. Ainsi, nous avons pu récolter des données sous des formats tableurs (excel, csv), texte (doc, pdf), géographique (shape file) et données (dat). Une synthèse est présentée en Annexe A. Dans un contexte d'analyse décisionnelle, les informations fournies doivent être agrégées, résumées et observables sur plusieurs niveaux de précision. Ainsi, il est nécessaire de les intégrer dans une structure à plusieurs dimensions, un entrepôt de données. Dès lors, un traitement des données brutes est nécessaire. A cet effet, on utilise Talend, un outil d'extraction et de traitement de données (ETL). Pour rassembler les données, il est nécessaire de les transformer pour les :

— nettoyer (suppression des données incohérentes) ;
— homogénéiser (définition de format commun) ;
— organiser (structure multidimensionnelle).

L'ETL récupère ainsi les données des différentes sources et les intègre dans l'entrepôt de données après traitements préalables.

4.1.2 Entrepôt de données

La construction de l'entrepôt de données nécessite la définition de sa structure qui est basée sur un modèle conceptuel qui s'appuie sur le formalisme multidimensionnel. Dans un premier temps, les données brutes sont analysées, par discipline, pour construire des modèles conceptuels de données basés sur le formalisme relationnel. Nous utilisons la méthode d'analyse UML et le SGBD MySQL. Dans un second temps, les modèles ainsi obtenus sont confrontés pour identifier les tables de fait et de dimension afin de construire un modèle multidimensionnel avec des axes temporel et spatial. Le modèle multidimensionnel sera traduit en modèle physique en utilisant la technologie de stockage ROLAP. Ainsi, le modèle de données est implémenté sous PostgreSQL en un entrepôt de données en utilisant le formalisme "entité-relation" des SGBD relationnels. Pour l'analyse géo-référencée, nous avons activé le *plugin* PostGIS. L'intégration des données a été faite en utilisant le logiciel Talend qui propose tous les connecteurs d'entrée et de sortie adaptés à nos besoins.

4.1.3 Reporting décisionnel

Il s'agit de présenter les indicateurs permettant d'évaluer la situation sanitaire à un moment donné ou sur une période pour une zone géographique. Le reporting décisionnel s'appuie sur la suite décisionnelle Pentaho. La première étape consite à construire le cube de données qui est une représentation abstraite de données agrégées avec plus ou moins de détails. Selon les besoins des utilisateurs en termes de reporting, on construit un cube de données qui intègre les indicateurs à produire avec tous les axes d'analyse possibles. Le module utilisée est Mondrian. Pour la génération des différents rapports et du reporting a été faite en utilisant essentiellement les modules "Report Designer" et "Saiku" qui sont accessibles en version web. Ainsi, nous fournissons des tableaux croisés et des graphiques qui permettent de visualiser les paramètres sanitaires et environnementaux à plusieurs niveaux.

4.1.4 Méthodes de fouille de données

Nos travaux s'orientent vers l'application des techniques de fouille de données dans un contexte de classification et de prédiction des maladies à vecteurs.

Afin de mieux comprendre les applications qui en découleront, nous présentons dans cette section les opérateurs de fouille sélectionnés suivant les besoins définis avec les experts métiers.

4.1.4.1 Classification

La classification est une technique de fouille de données permettant d'affecter des objets à des groupes en se basant sur les structures conditionnelles et/ou alternatives appliquées à leurs propriétés (Phyu, 2009). La fouille de données propose plusieurs algorithmes de classification dont les arbres de décision (Decision Tree). Un arbre de décision est construit de haut en bas à partir d'un nœud racine et consiste à diviser les données en sous-ensembles (également appelés nœud) qui contiennent des instances ayant des valeurs similaires (homogène) (Rakotomalala, 2005). Les nœuds sont générés sur la base des valeurs d'un attribut discriminant. Chaque nœud représente une valeur numérique ou catégorique. Cette représentation permet d'évaluer les conséquences d'une décision en considérant un nœud "parent" comme une décision et les nœuds "fils" comme les conséquences de cette décision. Dans notre contexte, nous pouvons utiliser cet algorithme pour regrouper les mares suivants leurs caractéristiques les plus pertinentes. Dès lors, certains paramètres, moins affectant, pourraient être écartés de l'étude. Les caractéristiques les plus pertinentes vont ensuite être évaluées pour ressortir leurs interactions et mesurer leurs impacts sur la dynamique vectorielle.

4.1.4.2 Clustering

Cette méthode consiste à maximiser la similarité intra-groupe et la minimiser entre groupes distincts (Berkhin, 2006) ; ces groupes sont appelés "clusters". Cette méthode est basée sur deux (2) règles essentielles : (1) Un cluster est un sous-ensemble d'occurrences qui sont "similaires ; (2) la distance entre deux occurences du cluster est inférieure à la distance entre les autres occurrences du cluster ou d'autres clusters. Par ailleurs, il est important que le jeu de données soit relativement volumineux. K-Means, un des premiers algorithmes et un des plus populaires (Jain, 2010), permet ainsi de créer des clusters selon une mesure de similarité adoptée par l'utilisateur ou spécifique au jeu de données. Le nombre de clusters est prédéfini, de façon automatique en se basant sur les valeurs des propriétés des instances d'objet. Le principe de cet algorithme est relativement simple (Kanungo *et al.*, 2002). A partir d'un ensemble de N points, représentant les données, en d-dimensions dans un espace R^d et d'un entier k, il s'agit de déterminer un ensemble de k points dans R^d, appelés centres, de manière à minimiser la moyenne distance au carré à partir de chaque point de données à son centre le plus proche.

4.1.4.3 Règles d'association

Pour l'identification de relations entre des valeurs particulières, il est proposé la dépendance "associative" (Agrawal *et al.*, 1993). Cet algorithme permet d'identifier les éventuelles corrélations entre des données. On utilise deux paramètres fondamentaux : la confiance, qui est la valeur à atteindre et le support qui représente la valeur minimale à prendre en compte. Notre approche se base sur deux (2) opérateurs exécutés séquentiellement :

— "*FPGrowth*" qui permet d'identifier les séquences les plus fréquentes qui seront injectées au prochain algorithme ;
— "*Create Association Rule*" qui détecte les règles d'association entre les attributs (conclusion des valeurs de certains attributs en se basant sur des attributs de départ).

En appliquant cet algorithme, nous visons à identifier les caractéristiques des mares à fort impact.

4.1.4.4 Motifs spatio-temporels

Nous avons choisi comme méthode d'analyse des données spatio-temporelles celle définie par (Salas *et al.*, 2011) ; elle est issue de (Agrawal et Srikant, 1995). Dans cette méthode, les auteurs se focalisent sur l'inclusion et la notion de spatialisation dans le processus d'extraction de connaissances à partir des données (ECD). Ils ont proposé une approche basée sur le pré-traitement des données afin d'inclure les caractéristiques spatiales dans les données temporelles. Grâce à cette transformation, les caractéristiques spatiales des données sont intégrées. Les séquences spatiales ainsi obtenues sont ensuite utilisées en entrée de l'étape de fouille de données afin d'en extraire des séquences spatialement fréquentes. Ceci est réalisé à l'aide d'un algorithme d'extraction de motifs séquentiels classique, décrit par (Mortazavi-Asl *et al.*, 2000), qui permet d'extraire des motifs représentant les évolutions temporelles spatialement fréquentes des zones. Cette méthode a été largement discutée dans (Salas *et al.*, 2011).

4.2 Modèle de données

4.2.1 Modélisation conceptuelle

Conceptuellement, notre modèle de données s'appuie sur le formalisme de la modélisation multidimensionnelle.

L'objectif étant de proposer un environnement orienté besoins d'analyse, nous proposons un schéma de l'entrepôt à construire qui intègre les données de toutes les disciplines impliquées.

Afin de faciliter la construction des vues corrélées des données des différentes disciplines, nous avons identifié les éléments communs que sont les dimensions spatiales et temporelles.

Nous présentons les tables de fait et de dimension suivant le type de données manipulées.

4.2.1.0.1 Données de suivi sanitaire Les données de suivi sanitaire nous ont permis de proposer les tables suivantes.

Viro-Animal

Cette table de fait permet d'établir l'état sanitaire d'un troupeau suite aux enquêtes de terrain et aux tests de laboratoire :

— nombre de prélèvement sanguin positif IgG ;
— nombre de prélèvement sanguin positif IgM ;
— nombre de prélèvement sanguin négatif IgG ;
— nombre de prélèvement sanguin négatif IgM ;
— nombre de prélèvement d'avortons positif IgG ;
— nombre de prélèvement d'avortons positif IgM ;
— nombre de prélèvement d'avortons négatif IgG ;
— nombre de prélèvement d'avortons négatif IgM ;
— nombre d'animaux malades ;
— nombre d'animaux morts.

Troupeau Cette table de dimension permet de recenser les troupeaux avec les attributs suivants :

— nombre d'animaux ;
— type : transhumant ou résident.

Taxo-Animal

Cette dimension classifie l'état sanitaire d'un troupeau suivant sa taxonomie « race » et « espece ».

Viro-Vecteur Ce fait regroupe les éléments d'analyse des vecteurs capturés lors des enquêtes de terrain :

— nombre de vecteurs mâles ;
— nombre de vecteurs femelles ;
— nombre de vecteurs broyés ;
— état de séro positivité de l'échantillon broyé.

Méthode de capture Cette dimension permet de regrouper les résultats de virologie vectorielle suivant la méthode utilisée (CO2, appât animal . . .) pour capturer les vecteurs analysés.

Piège La dimension « piège » est un axe d'analyse suivant l'emplacement et le type de piège posé.

Taxo-Vecteur

Cette dimension permet de regrouper les vecteurs suivant le genre et l'espèce.

4.2.1.0.2 Données environnementales Ces données concernent essentiellement l'environnement météo climatique et hydrologique.

Climatologie

La table de fait intègre les données générées par les stations météorologiques et les postes pluviométriques :

— humidité ;
— température ;
— vitesse du vent ;

— direction du vent ;
— ensoleillement ;
— pluviométrie.

Station Cette dimension permet d'analyser les données climatiques suivant les stations et postes installés dans les différentes localités de la zone d'étude.

Hydrologie Cette table de fait permet de gérer les mesures des niveaux d'eau des mares de la zone d'étude. Suivant les protocoles, les mesures étaient faites une à deux fois par jour ; dans un souci d'harmonisation, nous avons utilisé un seul attribut qui peut résulter de la moyenne pour les doubles mesures.

4.2.1.0.3 Données spatiales et temporelles Le découpage spatial est fait suivant le découpage administratif du Sénégal : « Communauté rurale », « Arrondissement », « Commune », « Département » et « Région ». Pour chaque zone géographique, nous identifions la position géographique pour la mise en place du SIG. Dans notre contexte, l'élément central de la dimension spatiale est la mare.

Pour la gestion de la temporalité, nous suivons l'axe temporel classique (jour, mois, année) auquel nous avons intégré la période pour la prise en compte du découpage saisonnier (saisons sèche et pluvieuse).

4.2.2 Modélisation multidimensionnelle

Pour la représentation de nos tables de fait et dimension, nous avons utilisé les trois (3) méthodes de représentation physique (Figure 4.2) :

— le modèle en étoile : nous construisons un modèle en étoile pour chaque discipline – une table de fait suivant plusieurs axes d'analyse :
  — Viro Animal ;
  — Viro Vecteur ;
  — Climatologie ;
  — Suivi/Qualité de l'eau ;
  — Hydrologie ;
— le modèle en flocon : pour affiner les résultats d'analyse, les dimensions sont hiérarchisées :
  — dimension spatiale : on suit le découpage administratif puis on affine par mare, station, site et échelle ;
  — dimension temporelle : année, période, mois et jour ;
  — dimension taxonomique : espèce, genre, race ;
— le modèle en constellation : pour rassembler les différents modèles en étoile – leurs éléments communs étant les dimensions spatiale et temporelle.

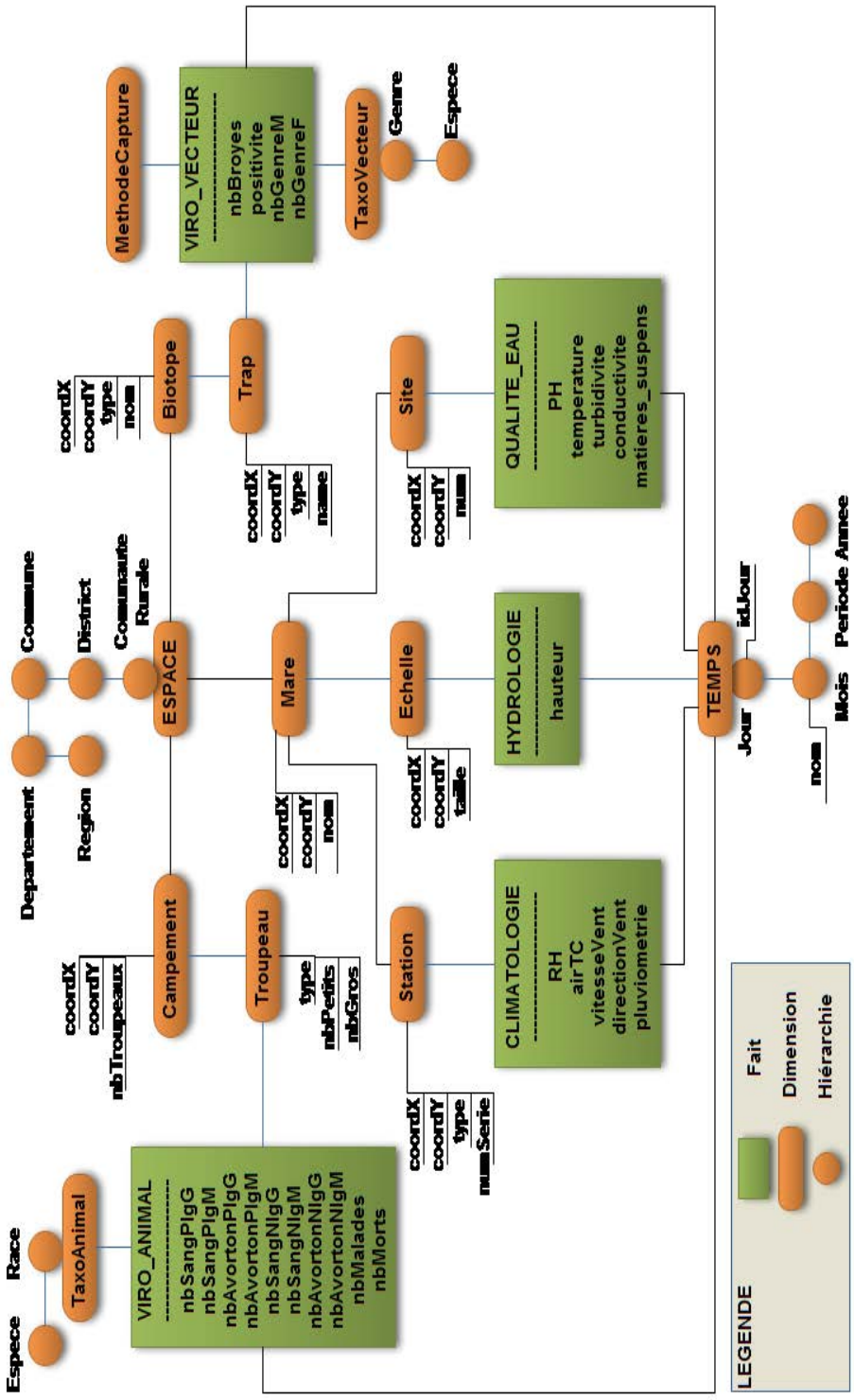

FIGURE 4.2 – Modèle multidimensionel.

4.2.3 Modélisation logique

Au niveau logique, plusieurs possibilités sont envisageables pour la modélisation multidimensioelle. L'approche la plus couramment utilisée consiste à utiliser un système de gestion de bases de données relationnelles, on parle de l'approche ROLAP ("Relational On-Line Analytical Processing"). Le modèle multidimensionnel est alors traduit de la manière suivante :

— chaque fait correspond à une table, appelée table de fait ;
— chaque dimension correspond à une table, appelée table de dimension.

Ainsi, la table de fait est constituée d'attributs représentant les mesures d'activité et les attributs clés étrangères de chacune des tables de dimension. Les tables de dimension contiennent les paramètres et une clé primaire permettant de réaliser des jointures avec la table de fait.

Plus récemment, une autre approche s'appuie sur le paradigme objet ; on parle de l'approche OOLAP ("Object On-Line Analytical Processing"). Le modèle multidimensionnel se traduit ainsi :

— chaque fait correspond à une classe, appelée classe de fait ;
— chaque dimension correspond à une classe, appelée classe de dimension.

Une alternative à ces deux approches consiste à utiliser un système multidimensionnel "pur" qui gère des structures multidimensionnelles natives ; on parle de l'approche MOLAP ("Multidimensional On-Line Analytical Processing"). Cette approche permet de stocker les données de manière multidimensionnelle. L'intérêt est que les temps d'accès sont optimisés, mais cette approche nécessite de redéfinir des opérations pour manipuler ces structures multidimensionnelles.

Notre choix s'est porté sur la technique ROLAP, qui est adaptée au SGBD géographique à mettre en place. Ainsi, conformémment aux règles définies par ce formalisme, nous avons converti chaque fait et dimension en une table relationnelle.

Les attributs des tables sont conservés ; nous utilisons un numéro séquentiel comme clé primaire pour chaque table de fait.

Pour la conversion suivant les règles de cardinalité, notre modèle ne propose pas le concept de « cardinalité » de façon explicite mais implicitement, il en ressort les règles suivantes :

— la hiérarchisation d'une dimension implique des relations de type (X,1) - (X,N) ; 1 du côté de la dimension fille et N du côté de la dimension parente :
  Fille Parent
— la relation entre une table de fait et une table de dimension implique une relation de type (X,1) - (X,N) ; 1 du côté de la table de fait et N du côté de la table de dimension ;
  Dimension Fait

Ainsi, nous obtenons le modèle logique suivant :

Dimension Taxonomie

— Espece (id, libelle, type)
— Race (id, libelle, idEspece)
— Genre (id, libelle, idEspece)

Dimension spatiale

— Region (id,libelle, coordX, coordY, idPays)
— Departement (id, libelle, coordX, coordY, idRegion)
— Commune (id, libelle, coordX, coordY, idDepartement)
— Arrondissement (id, libelle, coordX, coordY, idCommune)
— CommunauteRurale (id, libelle, coordX, coordY, idArrondissement)
— Mare (id, libelle, coordX, coordY, idCommunauteRurale)
— Station (id, libelle, coordX, coordY, idMare)
— ite (id, libelle, coordX, coordY, idMare)
— Echelle (id, libelle, coordX, coordY, idMare)
— Campement (id, libelle, coordX, coordY, idMare)
— Troupeau (id, type, nbAnimal, idCampement)

Dimension temporelle

— Annee (id, annee)
— Periode (id, libelle, idAnnee)
— Mois (id, libelle, idPeriode)
— Jour (id, idMois)

Tables de fait

— Climatologie (id, humidite, air, temperature, vent, ensoleillement, idStation)
— SuiviEau (id, conductivite, turbidite, temparature, pH, matiereSuspens, idSite)
— Hydrologie (id, hauteur, idEchelle)
— Viro-Vecteur (id, nbInsectesBroyes, positivite, nbFemelles, nbMales)
— Viro-Animal (id, nbMalades, nbMorts, nbSanguinPIgG, nbSanguinPIgM, nbSanguinNIgG, nbSanguinNIgM, nbAvortonsPIgG, nbAvortonsPIgM, nbAvortonsNIgG, nbAvortonsNIgM,)

Les tables ont été implémentées sous le SGBD PostgreSQL. L'alimentation des données a été faite en utilisant l'ETL Talend. Plusieurs sources de données ont été intégrées :

— « data » pour les données météorologiques ;
— « csv » et « excel » pour les autres.

4.2.4 Restitution des données

Les acteurs intervenant dans notre système sont de (2) types : les scientifiques (analystes) et les utilisateurs finaux (décideurs). Aussi, il est important de fournir d'une part des indicateurs clés et d'autre part des analyses multi critères. Ainsi, les données de notre modèle multidimensionnel nous permettent aisément de représenter les informations nécessaires selon les dimensions spatiales et temporelles.

4.2.4.1 Vue bi-dimensionnelle

Cette représentation est une jointure qui fait intervenir les occurrences d'une table de fait avec une occurrence d'une dimension spatiale et une d'une dimension temporelle. Le résultat peut être représenté dans un tableau de produit cartésien ou un tableau croisé.

4.2.4.2 Vue en cube

Cette représentation se fait en utilisant des axes d'un repère qui représentent chacun une dimension. Les axes d'analyse sont gradués suivant les occurrences de la table de dimension et les croisements sont des cubes qui représentent les valeurs des mesures contenues dans les tables de fait.

4.2.5 Généralisation

Le modèle de données tel que représenté actuellement présente quelques limites telles que :

— la prise en compte d'une seule maladie en l'occurrence la FVR ;
— la limitation géographique au Sénégal.

En dehors de ces limites, nous considérons que notre modèle répond aux critères de généricité que sont :

— le cadre de définition : les concepts utilisés pour la gestion de la FVR sont similaires à ceux des autres maladies vectorielles ;
— le schéma (Figure 4.3) : c'est la représentation des entités (concepts) et des relations qui en découlent ;
— l'instance : elle permet de confirmer les définitions de table de fait et de dimension avec un cas concret ; nous illustrons notre approche à travers l'instanciation de notre modèle pour la prise en compte du paludisme.

Ainsi, nous présentons une approche permettant d'étendre notre modèle, Figure 4.3, aux autres maladies à vecteur en intégrant les dimensions supplémentaires que sont :

— la typologie des maladies et des virus associés ;
— la taxonomie qui peut être décomposée suivant la cible examinée (avorton, sang, etc.), la population (humaine ou animale), l'espèce qui est "reflexive" pour pouvoir prendre en charge toutes les structures taxonomiques ;
— la dimension « pays ».

Par ailleurs, la mesure relative à la virologie est généralisée pour prendre en compte aussi bien les animaux que les êtres humains. La dimension spatiale rattachée à la virologie est également revue pour pouvoir prendre en compte différents types de hiérarchies spatiales ; pour ce faire, nous introduisons le concept de "reflexivité" issu du modèle relationnel.

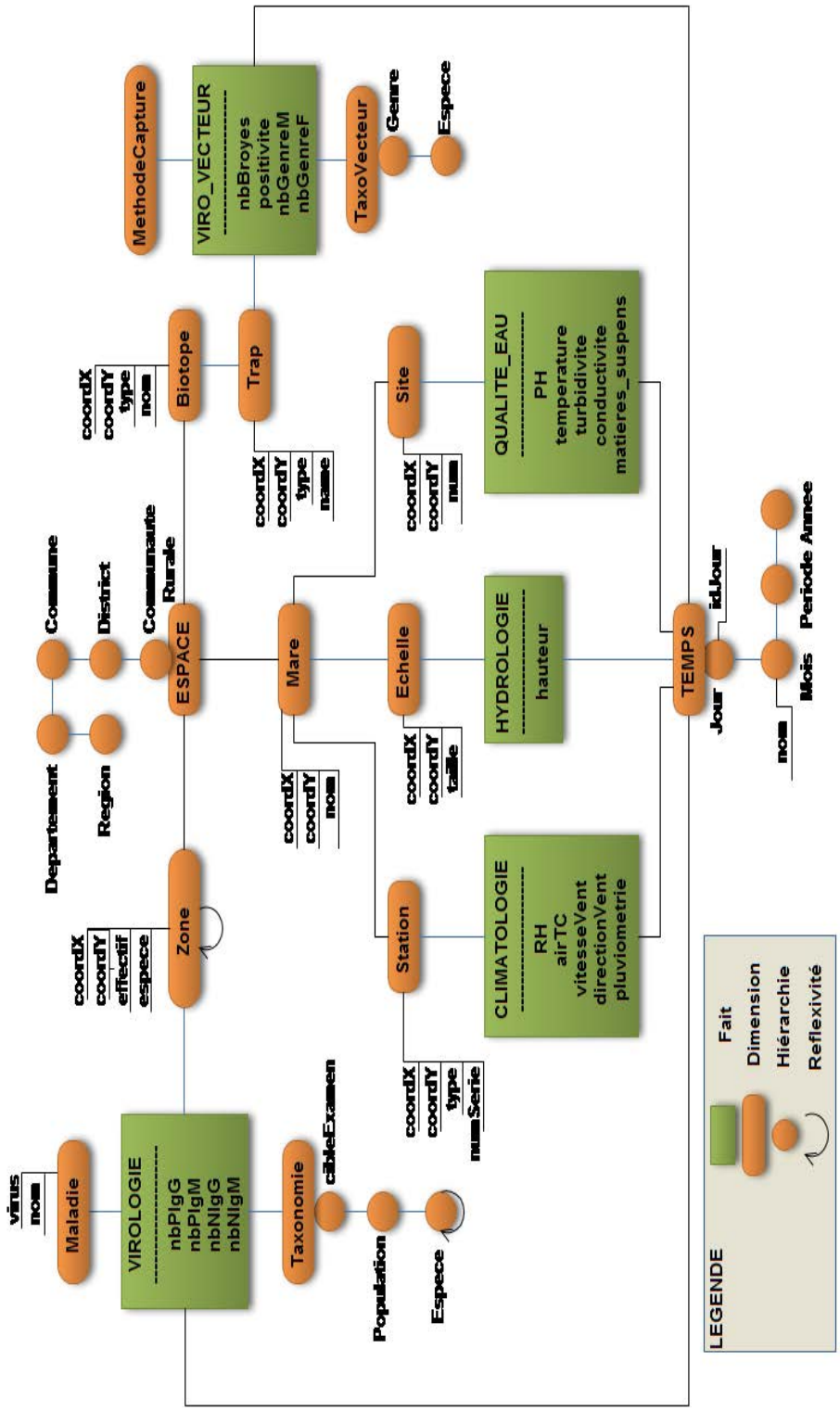

FIGURE 4.3 – Modèle générique.

4.2.6 Synthèse

Le modèle multidimensionnel proposé offre une solution permettant de répondre favorablement aux exigences de « corrélation, agrégation » fixées dans notre objectif. Pour faciliter les analyses sur de longues périodes, nous introduisons également la notion d'« historisation » empruntée au modèle entité-association. Cette historisation nous permettra de suivre l'évolution de certains attributs de table de fait. Aussi, sa valeur ajoutée s'inscrit particulièrement par son caractère générique pour la prise en charge de diverses maladies à vecteurs.

4.3 Fouille de données

4.3.1 Clustering avec K-means

4.3.1.1 Analyse sur les cinq mares

En appliquant l'algorithme de clustering avec K-means, nous avons pour objectif d'identifier les regroupements de caractéristiques des mares les plus similaires en fonction du nombre de vecteurs capturés dans une même période. Les données manipulées sont celles des enquêtes entomologiques et d'analyses de laboratoire des eaux des mares de Niakha, Kangalédji, Beli Boda, Furdu et Ngao pour les périodes de Août, Septembre et Décembre 2010. Il s'agit essentiellement des éléments suivants :

- pH : indique si l'eau présente une acidité ou une basicité ;
- CE : indique son caractère plus ou moins salin - mesure la capacité de l'eau à conduire le courant ;
- TDS : désigne la teneur d'une eau en matières (organiques, inorganiques) suspendues qui la troublent ;
- MES : représente les matières en suspension - particules présentes dans l'eau ;
- T $^{\circ}\mathrm{C}$: température ;
- nombre : nombre de vecteurs capturés.

Le choix du nombre de clusters se justifie, suivant le nombre de données, pour valider la variabilité des caractéristiques de la qualité des eaux des mares Niakha, Kangalédji, Beli Boda, Furdu et Ngao. Les figures ci-dessous représentent les valeurs des différents éléments de mesures de qualité de l'eau des principales mares de la zone d'étude confrontés au nombre de moustiques capturés dans la même période. Chaque courbe représente un des clusters identifiés. Pour d'importants volumes de données, nous pouvons construire trois (3) à quatre (4) clusters (figures 4.4 à 4.7). En fin de saison, seuls deux clusters (figure 4.8) ont pu être générés.

Dans notre contexte, les clusters générés peuvent être décrits comme suit :

- les caractéristiques de la qualité de l'eau des mares, telles que le pH et la tempréteaure sont relativement stables quelque soit le cluster d'une même période (figures 4.4 à 4.8) ;
- en ce qui concerne le CE, on constate une nette variabilité entre les clusters mais on note également une stabilité en pleine saison pluvieuse (figures 4.4 à 4.7) et un

niveau très élevé en fin de saison (figure 4.8) ;
— le TDS est assez varié quelque soit le cluster (figures 4.4 à 4.7) mais très élevé en décembre (figure 4.8) (la plupart des mares dans la zone d'étude s'assèchent entre fin octobre et décembre ; quasiment, seul Niakha dispose d'eau au-delà du mois de décembre).
— pour la période d'août à octobre, les espèces sont quasiment les mêmes, s'y ajoutent que le nombre de vecteurs capturés est particullièrement élevé pour un TDS assez faible. En revanche, pour le mois de décembre (TDS élevé), très peu de moustiques ont été capturés : il y a lieu alors de s'interroger sur l'impact du TDS sur la densité vectorielle ;
— pour cette même période, le MES est proportionnel au nombre de vecteurs – ceci ce justifie aisément par l'impact des matières en suspension sur la turbidité des mares.
— par ailleurs, on constate que toutes ces caractéristiques de la qualité de l'eau des mares augmentent considérablement en décembre contre de faibles densités de vecteurs capturés.

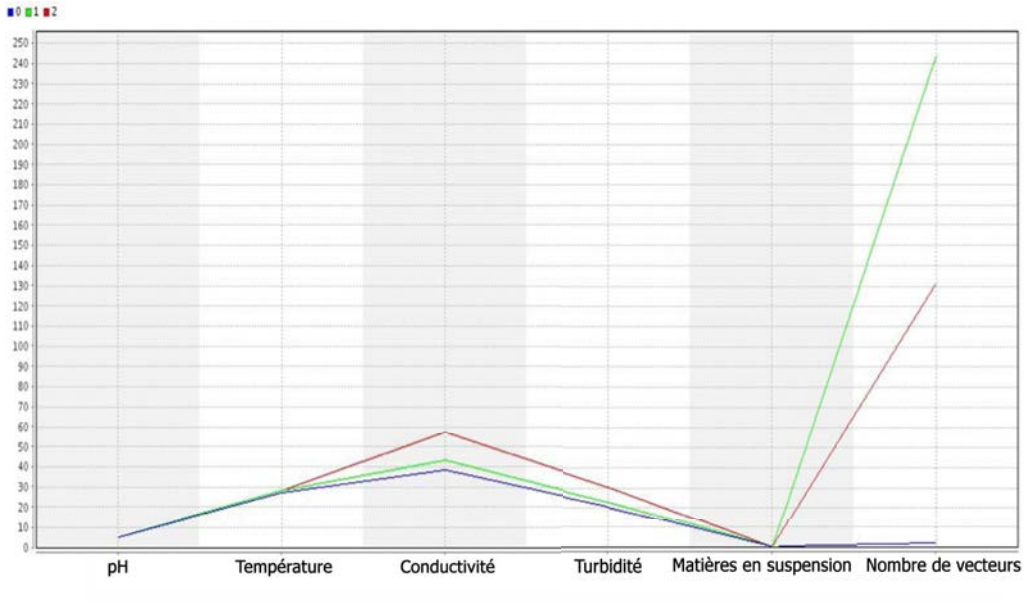

FIGURE 4.4 – Trois (3) clusters - Aout 2010.

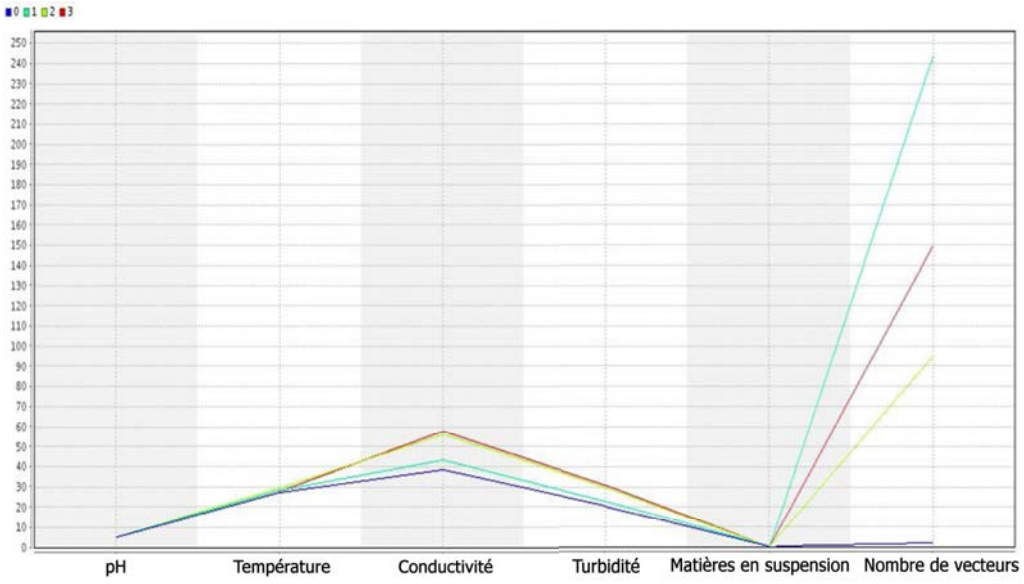

FIGURE 4.5 – Quatre (4) clusters - Aout 2010.

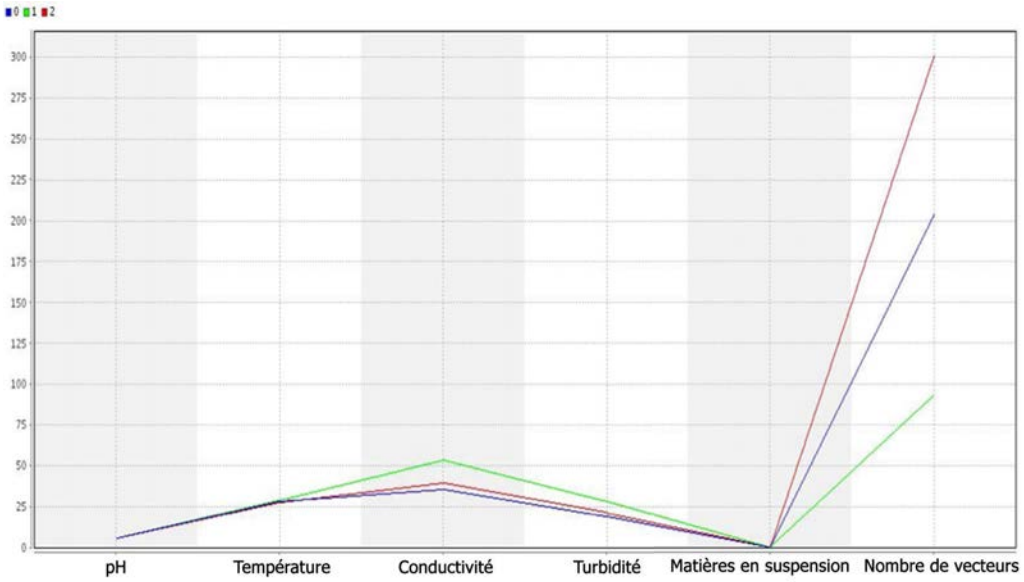

FIGURE 4.6 – Trois (3) clusters - Septembre 2010.

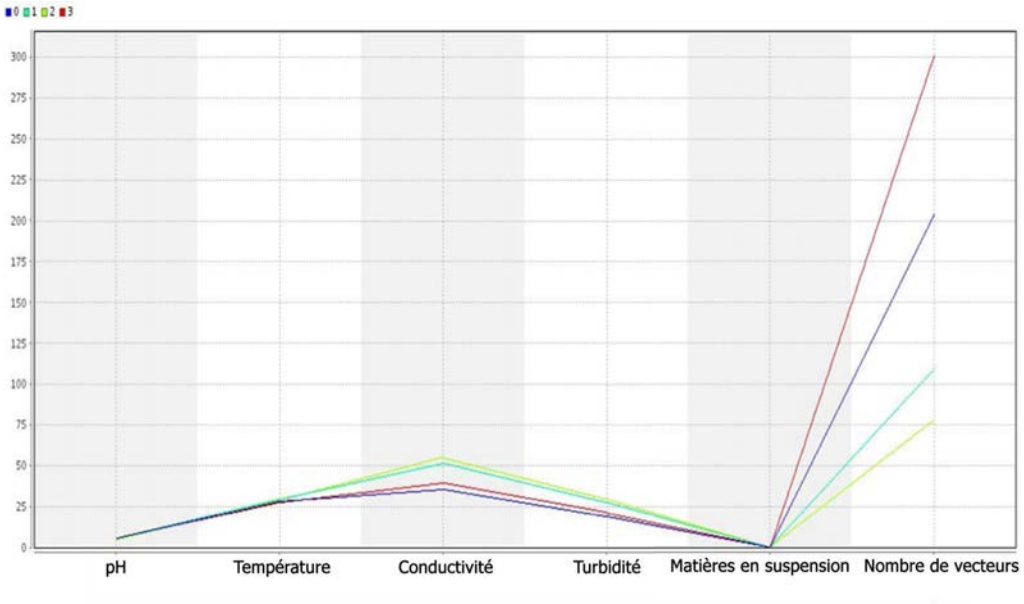

FIGURE 4.7 – Quatre (4) clusters - Septembre 2010.

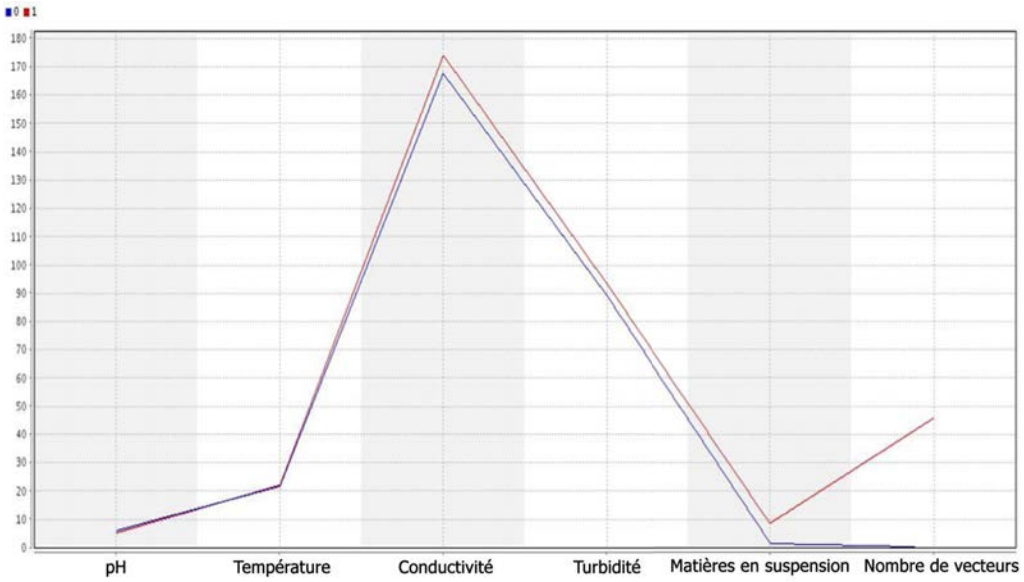

FIGURE 4.8 – Deux (2) clusters - Décembre 2010.

Il est important de noter que durant cette période, les contrôles sérologiques (recherche d'anticorps IgG et IgM, spécifiques du virus de la FVR) se sont tous avérés négatifs. D'une part, cette analyse nous pousse à nous interroger sur la caractérisation de chacune des mares identifiées dans la zone d'étude en faisant une analyse comparative entre les diffé-

rentes variables de qualité de l'eau (Voir Section 2). D'autre part, afin d'évaluer avec plus de précision l'impact des paramètres environnementaux sur ces données entomologiques (Pouémé, 2009), nous allons les confronter aux données générées par les stations météorologiques de la zone d'étude. Les données de la même période n'étant disponible que pour la mare de Niakha, nous avons mené notre analyse sur cet échantillon (Voir Section 1.2).

4.3.1.2 Analyse sur la mare de Niakha

Les données utilisées concernent les enquêtes pastoralisme, de suivi de qualité des eaux des mares et de la station météo automatique de Niakha d'Août, Octobre et Décembre 2010.

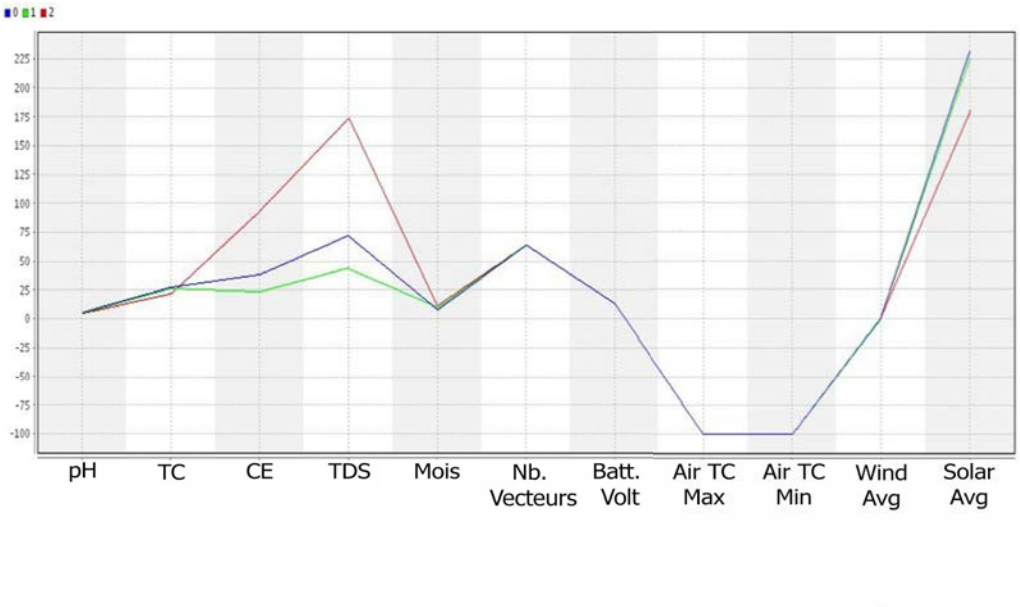

FIGURE 4.9 – Clustering sur la mare de Niakha. [1]

Pour la zone de Niakha, les périodes choisies (Août, Octobre et Décembre) se font remarquer par leur différence sur les caractéristiques telles que le CE, le TDS et l'ensoleillement. On ne peut se pronconcer sur leur impact sur la population des moustiques car les captures révèlent le même nombre d'individus.

Par ailleurs, les périodes choisies étant basées sur la même année, il n'est pas possible de faire ressortir avec précision les éventuels effets des facteurs environnementaux sur les caractéristiques de qualité de l'eau des mares.

1. Cluster Vert : Aout, Cluster Bleu : Octobre et Cluster Rouge : Décembre.

4.3.2 Classification avec Random Tree

L'algorithme « Random Tree » est appliqué pour comparer les caractéristiques des eaux des mares pour la période de août à décembre 2010.

Durant cette période, il en ressort les analyses suivantes :

— les températures des mares étudiées varient entre 25 et $28\,^{\circ}\mathrm{C}$; Niakha a la température la plus faible (figure 4.10) ;

— la turbidité est fonction des matières en suspension (MES) : plus on a de MES (figure 4.12) plus la turbidité est élevée, (figure 4.11) mais on constate une légère anomalie pour Furdu (la proportionnalité n'est pas respecté) et Niakha (considérablement envahie par des matières organiques/inorganiques) ;

— la conductivité (figure 4.13) est certes très élevée à Niakha mais il ressort aussi que toutes ces mares sont pauvres en ions – ces paramètres ne sont pas conformes à la littérature dans laquelle la conductivité est liée à la température. En effet, les variations de température n'étant pas très notables, il devrait en être de même pour la conductivité. Ce qui nous emmène à revoir l'impact de la température sur la conductivité dans de tels milieux. Il serait alors intéressant de se pencher sur des analyses plus approfondies d'autres caractéristiques telles que la qualité du sol de la mare ;

— le PH est assez stable (figure 4.14), compris entre 5,30 et 5,70 quelle que soit la mare ; donc toutes les mares ont une eau acide ;

— la mare de Ngao est beaucoup plus sujette à la fréquentation de vecteurs (figure 4.15) : plus de 500 vecteurs capturés contre 175 à 275 dans les autres mares.

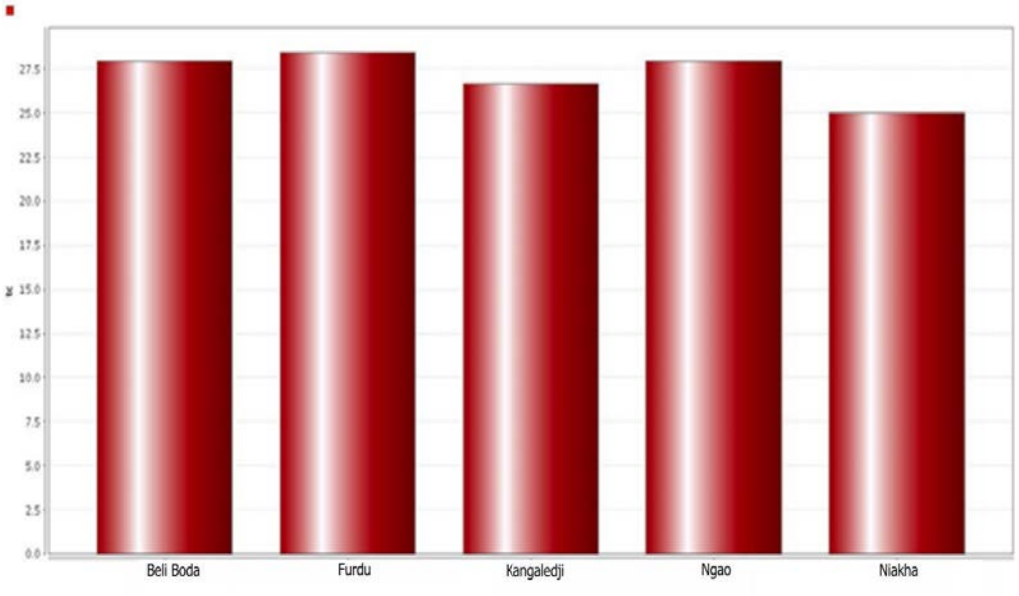

FIGURE 4.10 – Random Tree sur la température.

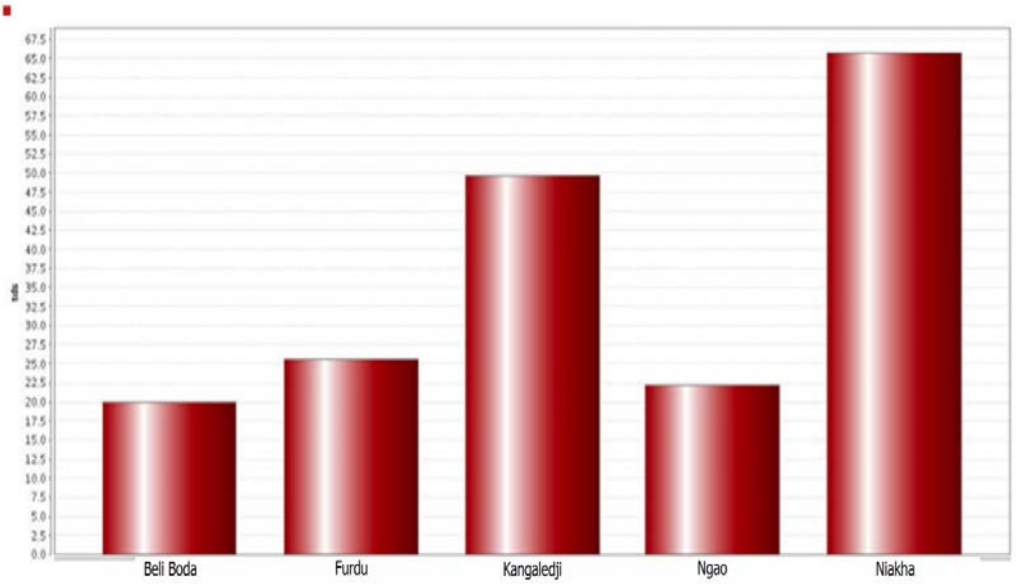

FIGURE 4.11 – Random Tree sur la turbidité.

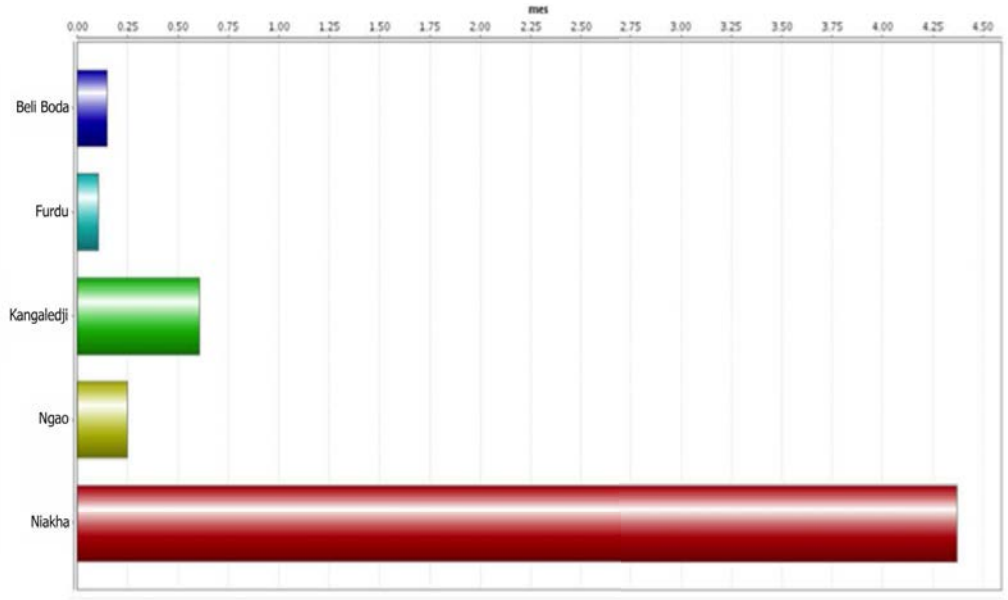

FIGURE 4.12 – Random Tree sur les matières en suspension.

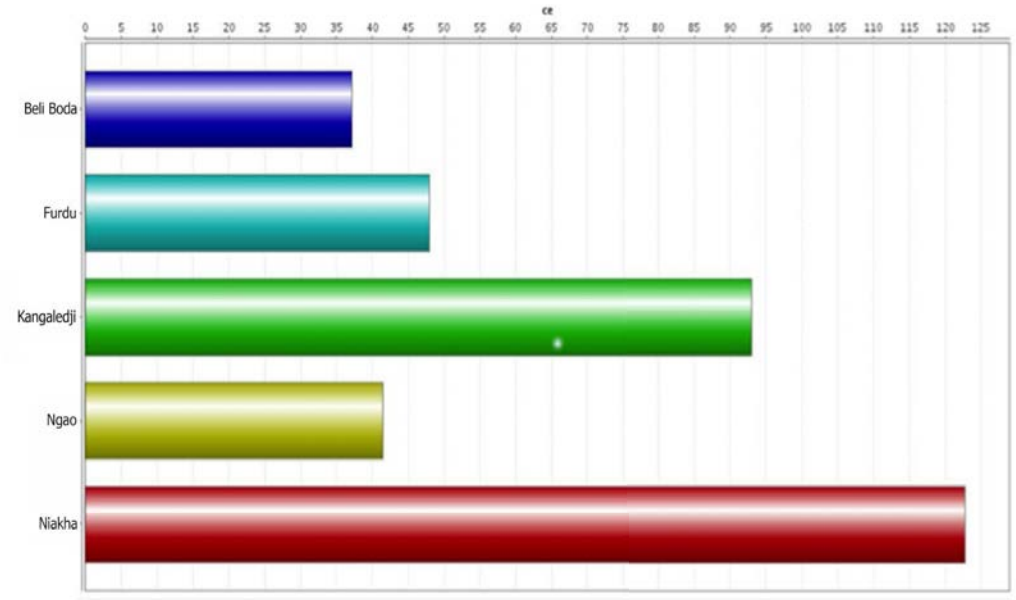

FIGURE 4.13 – Random Tree sur la conductivité.

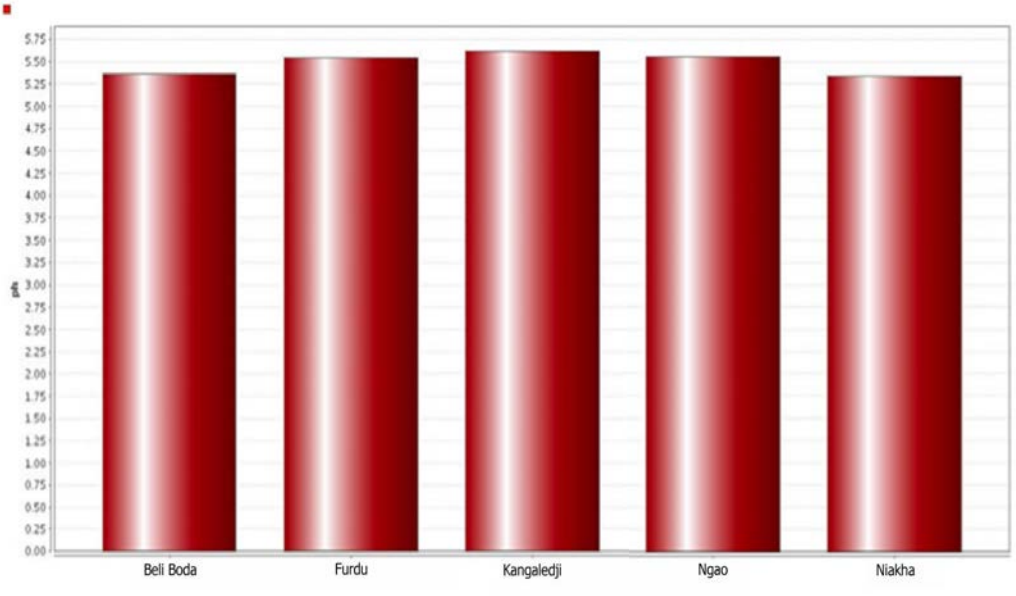

FIGURE 4.14 – Random Tree sur le pH.

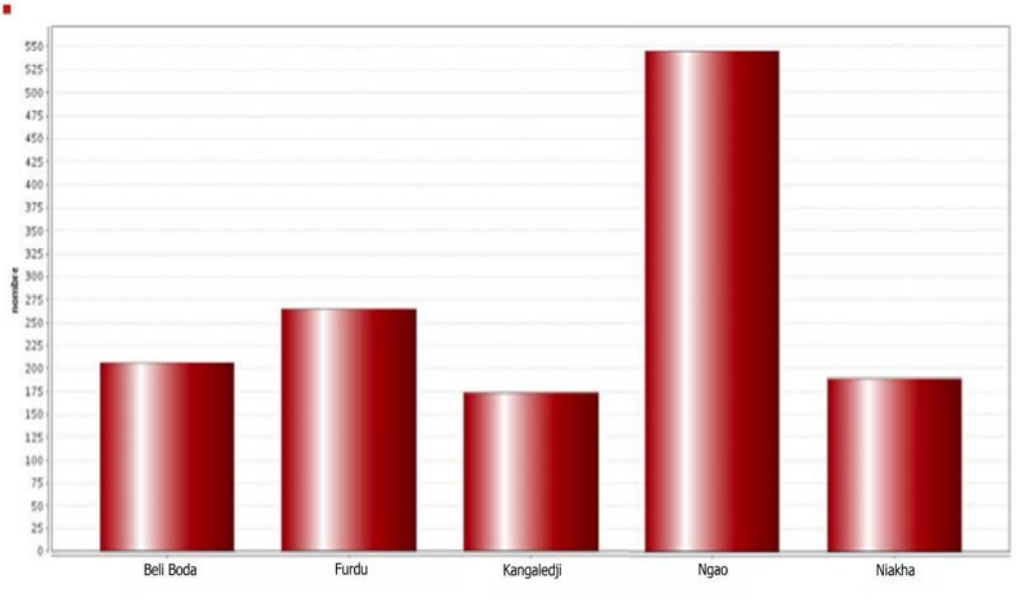

FIGURE 4.15 – Random Tree sur le nombre de vecteurs capturés.

4.3.3 Classification avec Decision Tree

Cet algorithme est appliqué pour prédire la présence des espèces vectorielles les plus fréquentes dans les localités de la zone d'étude (figure 4.16).

Ainsi, nous obtenons les résultats suivants :

— en juillet, sur les 11 localités que compte la zone d'étude, 9 enregistrent la présence d'Aedes vexans contre 2 pour celle de Culex (2 espèces) ;
— en août, une forte variété d'Aedes (7 espèces) dans 16 villages, des Culex poicilipes dans 1 village (Furdu) et des Mansonia à Kangalédji ;
— en septembre, les Aedes sont toujours aussi présents (2 variantes d'Aedes présents dans 9 villages), mais on note 2 espèces de Culex dans 7 villages et des Mansonia à Kangaldéji ;
— en octobre, les Culex se développent (3 variantes dans 10 localités), les Aedes sont beaucoup moins présents (1 espèce dans 2 villages), les Mansonia sont toujours aussi discrets (1 espèce dans 1 village) et les Anopheles se déclarent (1 espèce, 1 région) ;
— en novembre, on n'a plus d'Aedes, seules 2 espèces de Culex sont présentes dans 6 villages, les Anopheles sont toujours en expansion (2 espèces dans 5 villages) et les Mansonia ne sont relevées qu'à Niakha ;
— en décembre, 2 variétés de Culex sont présentes dans 4 localités, on a des Aedes aegypti à Ngao et des Anopheles à Ngao et Barkédji.

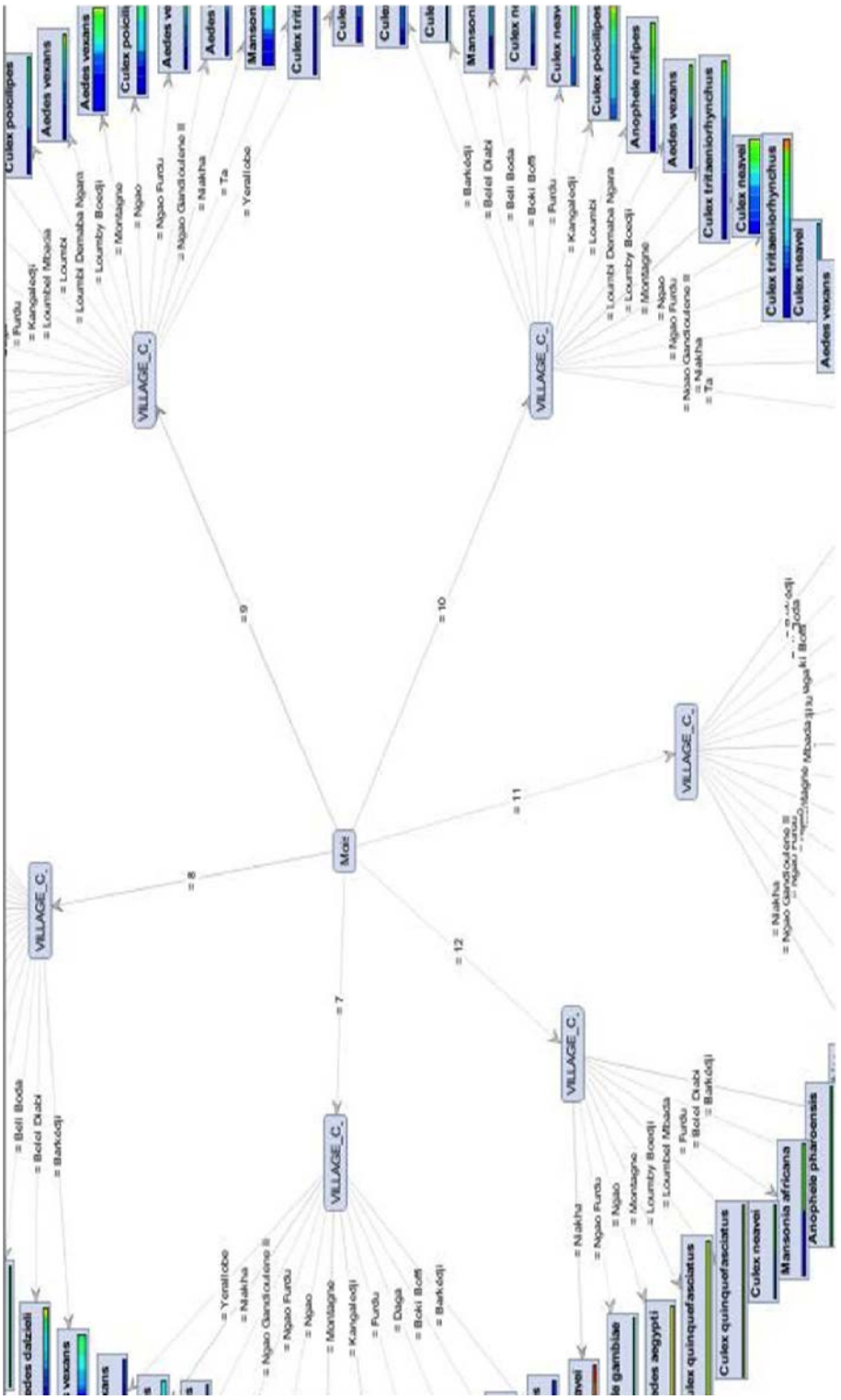

FIGURE 4.16 – Répartition mensuelle de la diversité vectorielle.

Ces résultats nous ont conduit à une analyse de regroupement des localités suivant la taxonomie vectorielle (Section 4).

4.3.4 Classification avec W-Random Tree

Cet algorithme est appliqué pour décrire la diversité vectorielle des localités rattachées aux mares de la zone d'étude. Nous nous sommes intéressés aux vecteurs potentiels de la FVR (Pépin, 2009). Cette analyse de la diversité vectorielle par localité, effectuée sur les principales mares, nous confirme essentiellement :

- la forte présence d'Aedes vexans dans les localités de Barkédji (figure 4.17), Ngao (figure 4.18), Niakha (figure 4.19) et Beli Boda (figure 4.20) ;
- la forte présence de Culex poicilipes dans les localités de Ngao (figure 4.18), Beli Boda (figure 4.20) et Furdu (figure 4.22) ;
- très peu d'Anopheles à Furdu (figure 4.22) ;

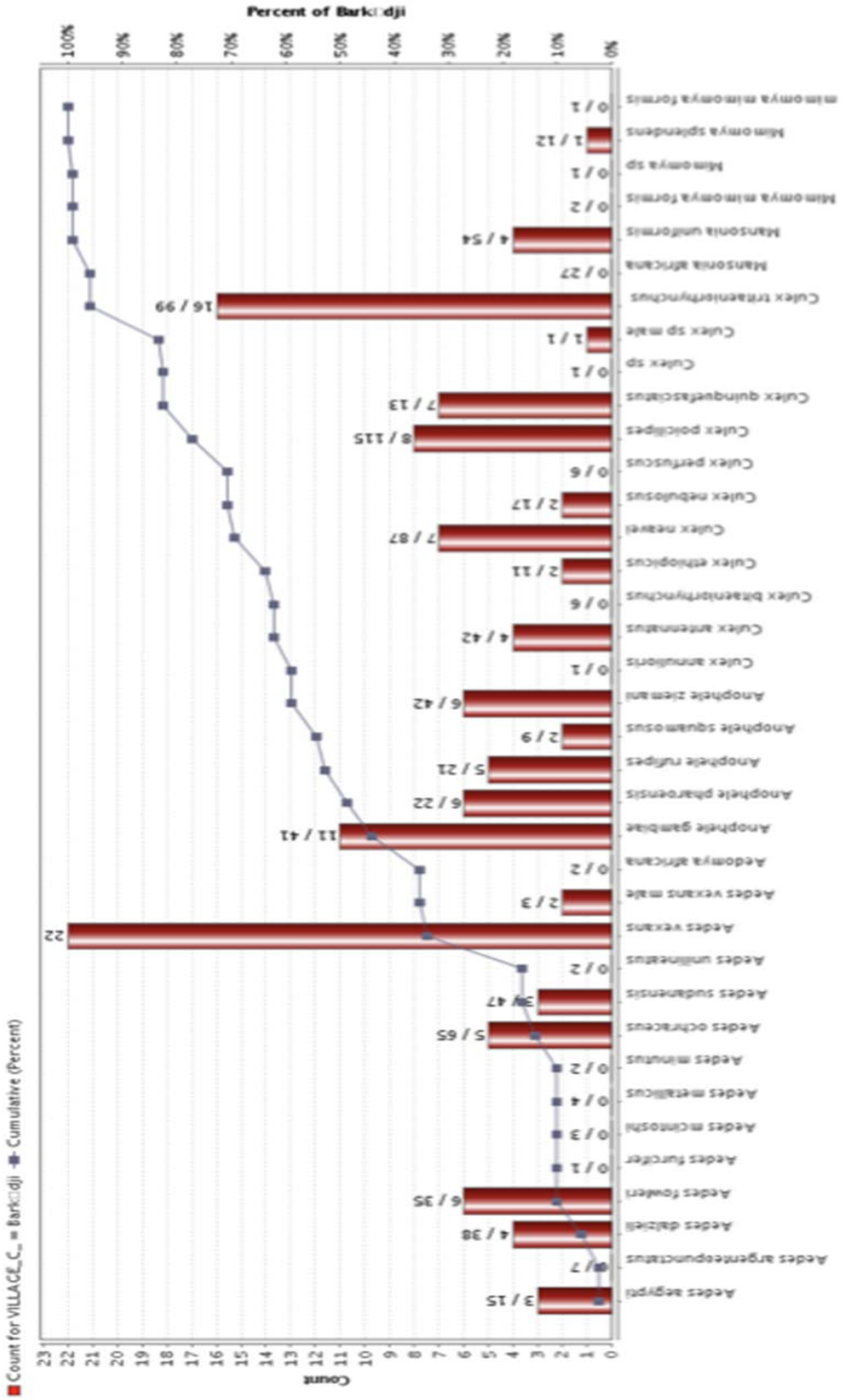

FIGURE 4.17 – Diversité vectorielle à Barkédji.

FIGURE 4.18 – Diversité vectorielle à Ngao.

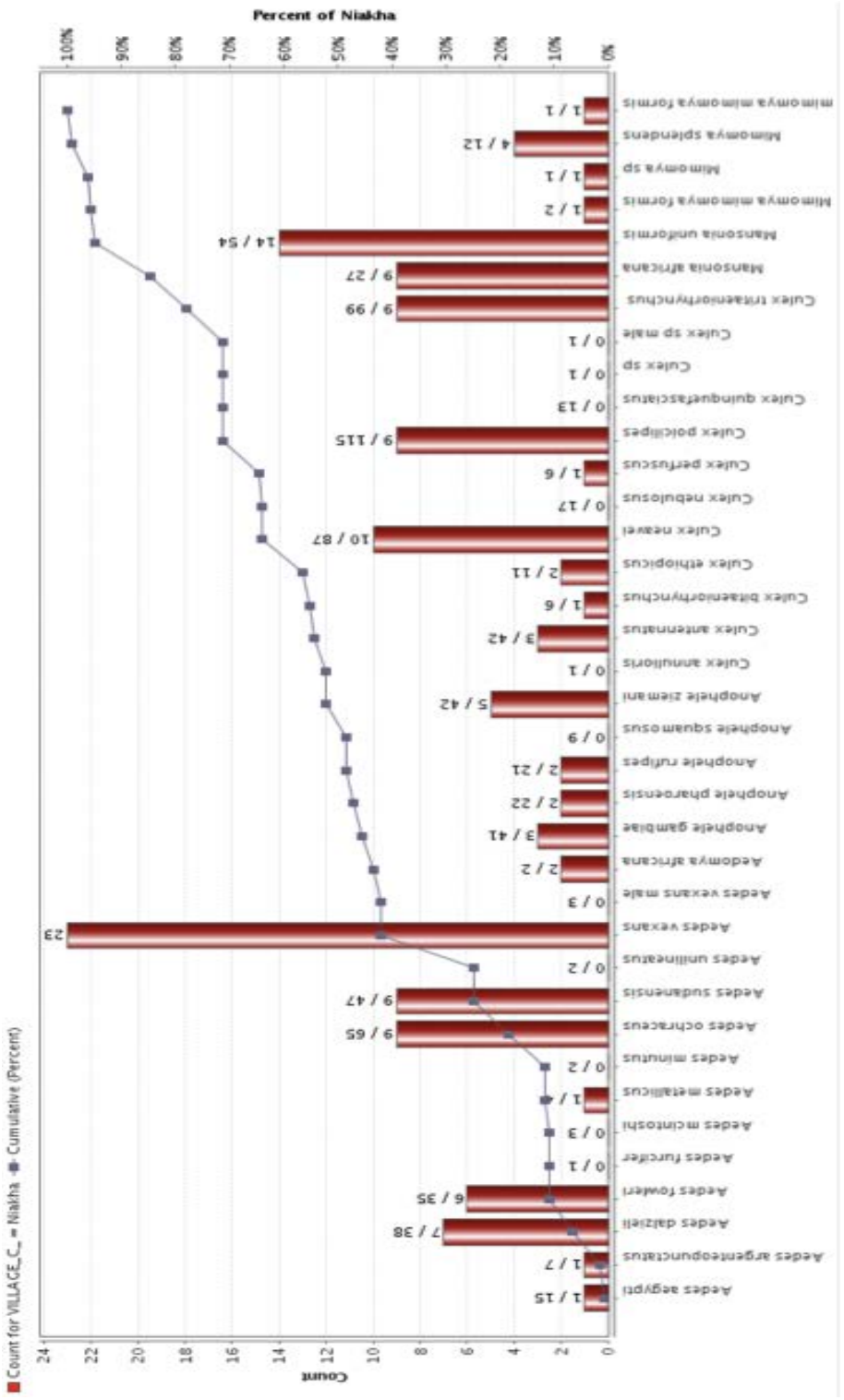

FIGURE 4.19 – Diversité vectorielle à Niakha.

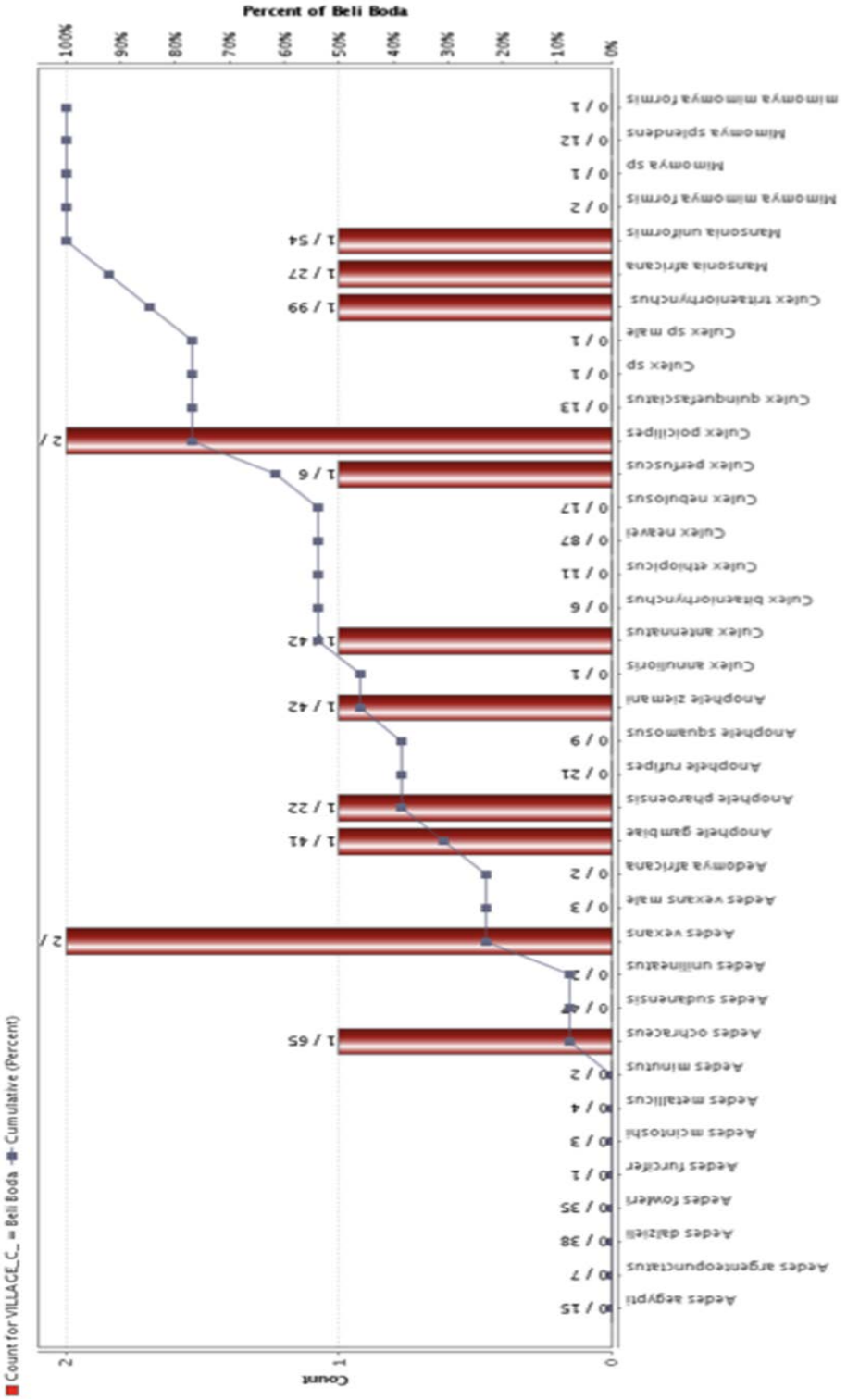

FIGURE 4.20 – Diversité vectorielle à Beli Boda.

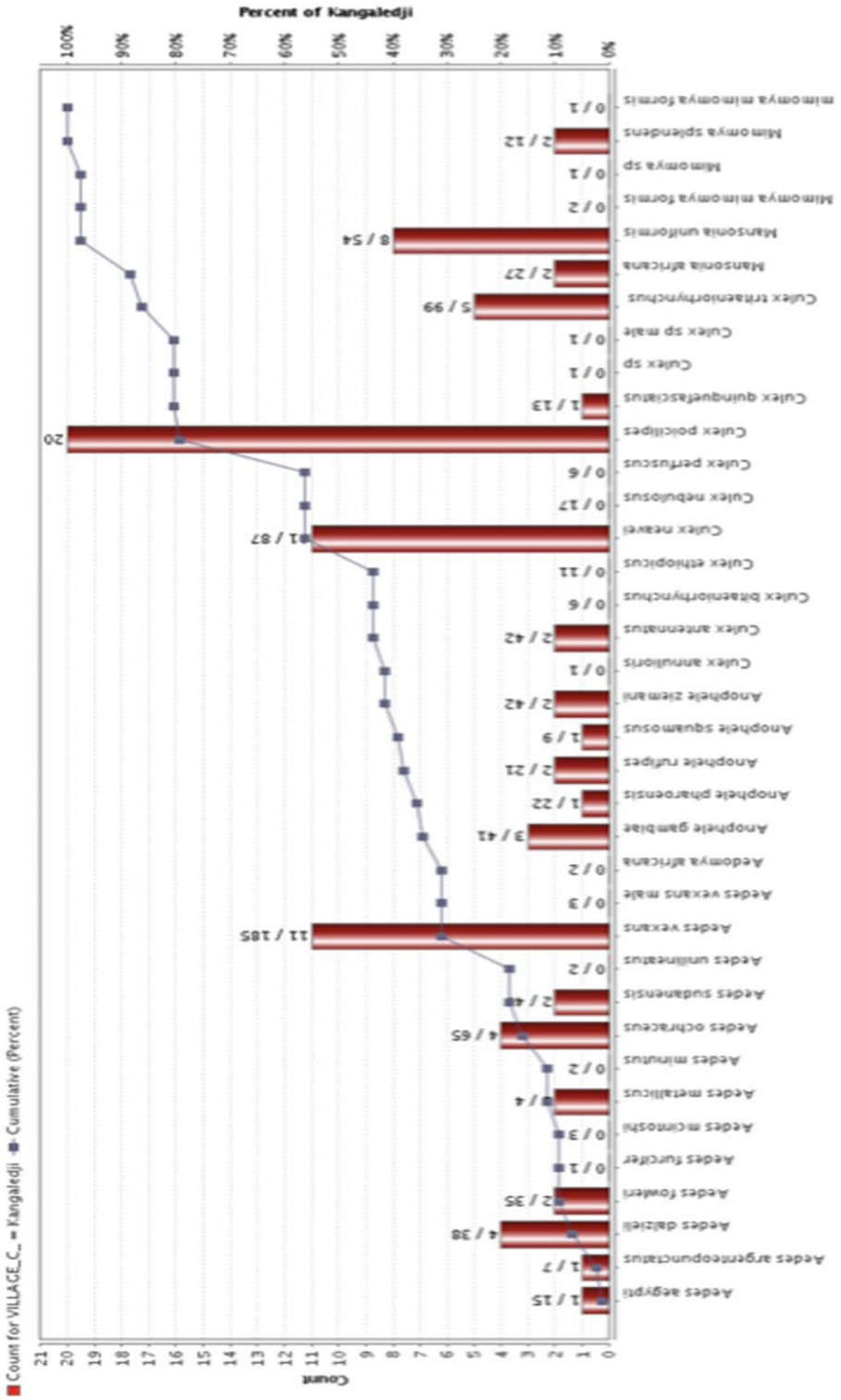

FIGURE 4.21 – Diversité vectorielle à Kangalédji.

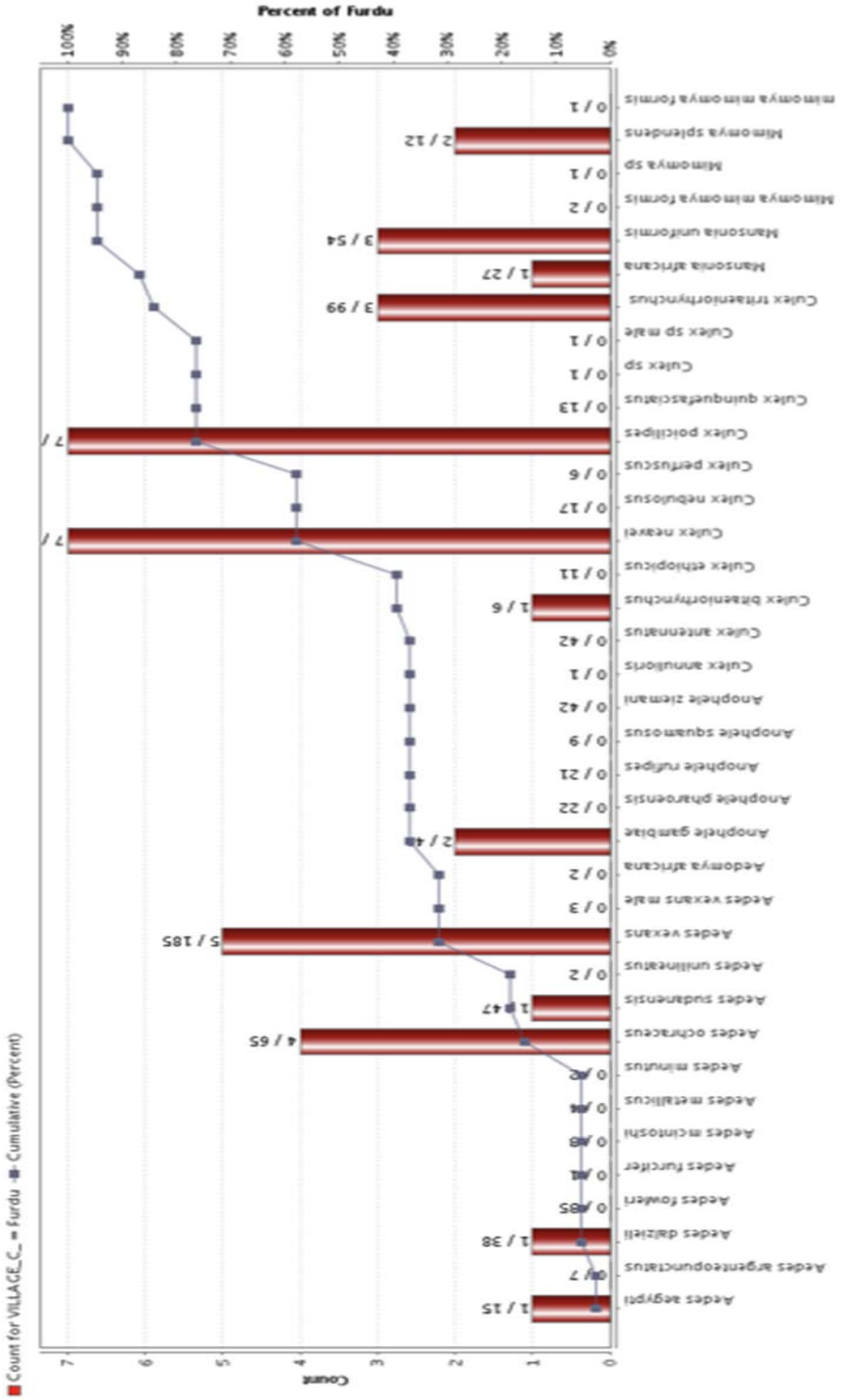

FIGURE 4.22 – Diversité vectorielle à Furdu.

Il convient donc d'identifier les points de similarité entre ces zones géographiques. Diverses études ont montré que le climat sahélien caractérise toutes ces localités. Ainsi, nous nous sommes penchés sur leurs positions cardinales. Suivant la description cartographique, il en ressort le découpage suivant :

— « Nord ouest » : Niakha, Barkédji ;
— « Nord est » : Kangalédji ;
— « Sud est » : Furdu, Ngao et Beli boda.

De cette première analyse, on peut donc conclure que la position géographique pourrait être un des facteurs d'affluence indépendamment des éléments climatologiques. En effet, la forte concentration des espèces de type Culex poicilipes se fait ressentir plus au Sud que dans les autres zones.

4.4 Extraction de motifs spatio-temporels

Dans notre approche, nous utilisons l'extraction de motifs spatio-temporels afin de recueillir le maximum de séquences des attributs des objets manipulés (mare, campement, vecteurs, etc.) pour nous permettre d'établir des relations entre ces objets. Les définitions utilisées, Annexe B, sont issues de (Salas *et al.*, 2012) et de (Fabrègue *et al.*, 2012).

La démarche suivie se résume en quatre (4) étapes :

— la recollection des données ;
— la transformation pour la construction de séquences ;
— la fouille de motifs ;
— la visualisation/validation.

Les données en entrée sont issues des données descriptives (MES, température, conductivité, turbidité, pH) des cinq (5) mares de la zone d'étude. Le jeu de données est consolidé via des jointures en une table dans laquelle nous intégrons une caractéristique spatiale de type "geometry".

Les expérimentations ont été faites en prenant en compte la mare comme "élément central". Ainsi, nous nous sommes focalisés sur les caractéristiques des eaux des mares.

4.4.1 Interactions intra hydrologiques

Une première analyse a porté sur les éventuels impacts de certaines caractéristiques des eaux sur d'autres. Les données sont regroupées par mare. Pour ce premier jeu de données (mares), nous avons une grande variété temporelle ; ainsi, avec un support minimal de 0,9, on obtient 21.131 séquences fréquentes.

Motif	Support	Equivalence
L1		
(ph:(5.36;5.54] temp:>29.05)	1	5/5
(ph:(5.54;5.66])	1	5/5
(ph:(5.54;5.66] temp:(28.35;29.05])	1	5/5
(ph:>5.66)	1	5/5
(temp:(28.35;29.05])	1	5/5
(tds:<=22.65)	1	5/5
(tds:(22.65;27.45])	1	5/5
(temp:>29.05)	1	5/5
(ce:(42.45;51.30])	1	5/5
(ce:(42.45;51.30] tds:(22.65;27.45])	1	5/5
L2		
(ph:(5.36;5.54])(ph:(5.36;5.54])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54] temp:>29.05)	1	5/5
(ph:(5.36;5.54])(temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ce:<=42.45 tds:<=22.65)	1	5/5
(ph:(5.36;5.54])(ph:(5.54;5.66] temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(tds:<=22.65)	1	5/5
(ph:(5.36;5.54])(ce:(42.45;51.30])	1	5/5
(ph:(5.36;5.54])(ce:(42.45;51.30] tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54] temp:>29.05)(temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54] temp:>29.05)(ph:(5.54;5.66] temp:(28.35;29.05])	1	5/5
L3		
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(ph:(5.54;5.66] temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(ce:(42.45;51.30] tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54] temp:>29.05)(temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54] temp:>29.05)(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54] temp:>29.05)(ph:(5.54;5.66] temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54] temp:>29.05)(ce:(42.45;51.30] tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54] temp:>29.05)(tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54] temp:>29.05)(ph:>5.66)	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(tds:(22.65;27.45])	1	5/5
L4		
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ph:(5.54;5.66] temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30] tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:(28.35;29.05])(ph:(5.54;5.66] temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:(28.35;29.05])(ce:(42.45;51.30])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:(28.35;29.05])(ce:(42.45;51.30] tds:(22.65;27.45])	1	5/5

Motif	Support	Equivalence
L5		
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(temp:(28.35;29.05])(temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(temp:(28.35;29.05])(ph:(5.54;5.66] temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(temp:(28.35;29.05])(ce:(42.45;51.30])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(temp:(28.35;29.05])(ce:(42.45;51.30] tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(temp:(28.35;29.05])(tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ph:(5.54;5.66])(temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ph:(5.54;5.66])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ph:(5.54;5.66])(ph:(5.54;5.66] temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ph:(5.54;5.66] temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
L6		
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(temp:(28.35;29.05])(temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(temp:(28.35;29.05])(temp:(28.35;29.05])(ce:(42.45;51.30])	1	5/5
(ce:(42.45;51.30])(temp:(28.35;29.05])(temp:(28.35;29.05])(ce:(42.45;51.30])(temp:(28.35;29.05])(ph:(5.54;5.66] temp:(28.35;29.05])	1	5/5
(ce:(42.45;51.30])(temp:(28.35;29.05])(temp:(28.35;29.05])(ce:(42.45;51.30])(ph:(5.54;5.66])(temp:(28.35;29.05])	1	5/5
(ce:(42.45;51.30])(temp:(28.35;29.05])(temp:(28.35;29.05])(ce:(42.45;51.30])(ph:(5.54;5.66])(ph:(5.54;5.66])	1	5/5
(ce:(42.45;51.30])(temp:(28.35;29.05])(temp:(28.35;29.05])(ce:(42.45;51.30])(ph:(5.54;5.66])(ph:(5.54;5.66] temp:(28.35;29.05])	1	5/5
(ce:(42.45;51.30])(temp:(28.35;29.05])(temp:(28.35;29.05])(ce:(42.45;51.30])(ph:(5.54;5.66] temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ce:(42.45;51.30])(temp:(28.35;29.05])(temp:(28.35;29.05])(tds:(22.65;27.45])(temp:(28.35;29.05])(temp:(28.35;29.05])	1	5/5
(ce:(42.45;51.30])(temp:(28.35;29.05])(temp:(28.35;29.05])(tds:(22.65;27.45])(temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ce:(42.45;51.30])(temp:(28.35;29.05])(ph:(5.54;5.66])(ph:(5.54;5.66])(temp:(28.35;29.05])(temp:(28.35;29.05])	1	5/5

Motif	Support	Equivalence
L7		
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30] tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30] tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(ph:(5.54;5.66])(temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30] tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(ph:(5.54;5.66])(ph:(5.54;5.66] temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30] tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(ph:(5.54;5.66] temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30] tds:(22.65;27.45])(ce:(42.45;51.30])(temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30] tds:(22.65;27.45])(tds:(22.65;27.45])(ph:(5.54;5.66])(temp:(28.35;29.05])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30] tds:(22.65;27.45])(tds:(22.65;27.45])(ph:(5.54;5.66])(ph:(5.54;5.66])	1	5/5
(ph:>5.66)(ce:(42.45;51.30] tds:(22.65;27.45])(ce:<=42.45)(ce:(42.45;51.30])(ph:(5.54;5.66])(ph:(5.54;5.66] temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ph:>5.66)(ce:(42.45;51.30] tds:(22.65;27.45])(ce:<=42.45)(tds:(22.65;27.45])(ph:(5.54;5.66])(temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ph:>5.66)(ce:(42.45;51.30] tds:(22.65;27.45])(ce:<=42.45)(tds:(22.65;27.45])(ph:(5.54;5.66])(ph:(5.54;5.66])(ph:(5.54;5.66])	1	5/5
L8		
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30],tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(ph:(5.54;5.66])(temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30],tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(ph:(5.54;5.66])(ph:(5.54;5.66])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30],tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(ph:(5.54;5.66])(ph:(5.54;5.66] temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30] tds:(22.65;27.45])(ce:(42.45;51.30])(ph:(5.54;5.66])(temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30],tds:(22.65;27.45])(ce:(42.45;51.30])(ph:(5.54;5.66])(ph:(5.54;5.66])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30] tds:(22.65;27.45])(ce:(42.45;51.30])(ph:(5.54;5.66])(ph:(5.54;5.66] temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(temp:>29.05)(ce:(42.45;51.30] tds:(22.65;27.45])(tds:(22.65;27.45])(ph:(5.54;5.66])(temp:(28.35;29.05])(ph:(5.54;5.66])	1	5/5
(tds:(22.65;27.45])(ce:(42.45;51.30])(ce:(42.45;51.30])(temp:(28.35;29.05])(temp:(28.35;29.05])(ph:(5.36;5.54])(temp:(28.35;29.05])(temp:(28.35;29.05])	1	5/5
(tds:(22.65;27.45])(ce:(42.45;51.30])(ce:(42.45;51.30])(temp:(28.35;29.05])(temp:(28.35;29.05])(ph:(5.36;5.54])(temp:(28.35;29.05])(ce:(42.45;51.30])	1	5/5
(tds:(22.65;27.45])(ce:(42.45;51.30])(ce:(42.45;51.30])(temp:(28.35;29.05])(temp:(28.35;29.05])(ph:(5.36;5.54])(temp:(28.35;29.05])(ce:(42.45;51.30] tds:(22.65;27.45])	1	5/5

Motif	Support	Equivalence
L9		
(ph:(5.36;5.54])(ph:(5.36;5.54])(ce:(42.45;51.30] tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(temp:>29.05)(temp:(28.35;29.05])(temp:(28.35;29.05])(ce:(42.45;51.30])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(ce:(42.45;51.30] tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(temp:>29.05)(temp:(28.35;29.05])(temp:(28.35;29.05])(ce:(42.45;51.30] tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(ce:(42.45;51.30] tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(temp:>29.05)(temp:(28.35;29.05])(temp:(28.35;29.05])(tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(tds:(22.65;27.45])(ce:(42.45;51.30])(tds:(22.65;27.45])(temp:>29.05)(temp:(28.35;29.05])(temp:(28.35;29.05])(tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(tds:(22.65;27.45])(tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(temp:>29.05)(temp:(28.35;29.05])(temp:(28.35;29.05])(ce:(42.45;51.30])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(tds:(22.65;27.45])(tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(temp:>29.05)(temp:(28.35;29.05])(temp:(28.35;29.05])(ce:(42.45;51.30] tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(tds:(22.65;27.45])(tds:(22.65;27.45])(ce:(42.45;51.30] tds:(22.65;27.45])(temp:>29.05)(temp:(28.35;29.05])(temp:(28.35;29.05])(tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(tds:(22.65;27.45])(tds:(22.65;27.45])(ce:(42.45;51.30])(temp:>29.05)(temp:(28.35;29.05])(temp:(28.35;29.05])(ce:(42.45;51.30])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(tds:(22.65;27.45])(tds:(22.65;27.45])(ce:(42.45;51.30])(temp:>29.05)(temp:(28.35;29.05])(temp:(28.35;29.05])(ce:(42.45;51.30] tds:(22.65;27.45])	1	5/5
(ph:(5.36;5.54])(ph:(5.36;5.54])(tds:(22.65;27.45])(tds:(22.65;27.45])(ce:(42.45;51.30])(temp:>29.05)(temp:(28.35;29.05])(temp:(28.35;29.05])(tds:(22.65;27.45])	1	5/5

Avec un support de 1 pour tous les motifs générés, nous concluons donc que toutes les mares analysées ont exactement le même comportement.

Ces résultats combinés aux précédents résultats sur les arbres décisionnels appliqués aux caractéristiques des mares, permettent de confirmer que la diversité vectorielle est étroitement liée à la qualité des eaux environnantes. En effet, pour des caractéristiques des mares similaires durant une même période, on constate que l'on retrouve les mêmes espèces de vecteurs ; même s'il faut noter que la densité vectorielle est très variable. Dans un tel contexte, il conviendrait de faire une analyse comparative en intégrant à la fois des sites de recueil en profondeur et aux alentours mais également en évaluant à la fois la densité des larves et celle des vecteurs adultes.

Avec les données mises à notre disposition, des précisions peuvent être apportées sur le domaine des valeurs possibles pour chacune des caractéristiques en comparaison aux données générées par les stations météorologiques.

4.4.2 Interactions climato-hydrologiques

La deuxième étude a permis d'analyser les éventuels interactions entre les eaux des mares et les paramètres météorologiques. L'objectif de cette approche est de regrouper les données en sites. Par exemple, on pourra grouper les stations par communauté rurale (pour Mare) et par zone de pointe (pour les données entomologiques).

Motif	Support	Equivalence
L1		
(temp:<=28.35 ce:(42.4551.30] tds:(22.6527.45])	0,44	7/16
(ph:>5.62 temp:(28.3529.05] ce:<=42.45 tds:<=22.65)	0,44	7/16
(ph:>5.62 temp:>29.05 tds:>27.45)	0,44	7/16
(ph:>5.62 temp:>29.05 ce:>51.30)	0,44	7/16
(ph:>5.62 temp:>29.05 ce:>51.30 tds:>27.45)	0,44	7/16
(temp:<=28.35 ce:<=42.45 tds:<=22.65)	0,5	8/16
(ph:<=5.45 tds:>27.45)	0,5	8/16
(ph:<=5.45 ce:>51.30 tds:>27.45)	0,5	8/16
(ph:>5.62 ce:(42.4551.30])	0,5	8/16
(ph:>5.62 ce:(42.4551.30] tds:(22.6527.45])	0,5	8/16
(temp:(28.3529.05] ce:>51.30 tds:>27.45)	0,56	9/16
(ph:(5.455.62] ce:(42.4551.30])	0,56	9/16
(ph:(5.455.62] ce:(42.4551.30] tds:(22.6527.45])	0,56	9/16
(ph:>5.62 ce:<=42.45)	0,56	9/16
(ph:>5.62 ce:<=42.45 tds:<=22.65)	0,56	9/16
(temp:(28.3529.05] ce:(42.4551.30])	0,62	10/16
(temp:(28.3529.05] ce:(42.4551.30] tds:(22.6527.45])	0,62	10/16
(temp:(28.3529.05] tds:(22.6527.45])	0,62	10/16
(ph:<=5.45 temp:<=28.35)	0,62	10/16
(temp:<=28.35 tds:>27.45)	0,69	11/16
(temp:<=28.35 ce:>51.30 tds:>27.45)	0,69	11/16
(temp:>29.05 ce:>51.30)	0,69	11/16
(temp:>29.05 ce:>51.30 tds:>27.45)	0,69	11/16
(ph:<=5.45 ce:(42.4551.30])	0,75	12/16
(ph:<=5.45 ce:(42.4551.30] tds:(22.6527.45])	0,75	12/16
(ph:<=5.45 tds:(22.6527.45])	0,75	12/16
(ph:(5.455.62] temp:(28.3529.05])	0,75	12/16
(temp:<=28.35)	0,81	13/16
(temp:(28.3529.05] tds:<=22.65)	0,81	13/16
(temp:(28.3529.05] ce:<=42.45)	0,81	13/16
(temp:(28.3529.05] ce:<=42.45 tds:<=22.65)	0,81	13/16
(temp:>29.05)	0,81	13/16
(tds:(22.6527.45])	0,88	14/16
(ce:(42.4551.30])	0,88	14/16
(ce:(42.4551.30] tds:(22.6527.45])	0,88	14/16
(tds:>27.45)	0,88	14/16
(temp:(28.3529.05])	0,94	15/16
(ph:<=5.45)	0,94	15/16
(ph:>5.62)	0,94	15/16
(tds:<=22.65)	1	16/16
(ce:<=42.45 tds:<=22.65)	1	16/16
L2		
(tds:(22.6527.45])(temp:>29.05 ce:(42.4551.30])	0,44	7/16
(temp:<=28.35 ce:>51.30)(temp:>29.05 ce:>51.30 tds:>27.45)	0,44	7/16
(temp:<=28.35 ce:>51.30 tds:>27.45)(ph:(5.455.62] temp:<=28.35)	0,44	7/16
(ce:<=42.45)(ph:<=5.45 temp:<=28.35)	0,44	7/16
(tds:(22.6527.45])(ph:(5.455.62] temp:<=28.35)	0,5	8/16
(tds:(22.6527.45])(ph:(5.455.62] tds:>27.45)	0,5	8/16

Motif	Support	Equivalence
(ce:<=42.45)(temp:(28.3529.05] tds:<=22.65)	0,5	8/16
(ce:<=42.45)(temp:(28.3529.05] ce:<=42.45 tds:<=22.65)	0,5	8/16
(tds:(22.6527.45])(ph:(5.455.62] temp:(28.3529.05])	0,56	9/16
(tds:(22.6527.45])(ph:>5.62 tds:>27.45)	0,56	9/16
(temp:>29.05 tds:>27.45)(ph:>5.62)	0,56	9/16
(temp:>29.05 tds:>27.45)(ce:>51.30)	0,56	9/16
(temp:<=28.35)(ph:(5.455.62] temp:(28.3529.05])	0,62	10/16
(temp:<=28.35 ce:>51.30)(ce:<=42.45)	0,62	10/16
(temp:<=28.35 ce:>51.30)(ce:<=42.45 tds:<=22.65)	0,62	10/16
(ce:<=42.45)(tds:<=22.65)	0,62	10/16
(ce:<=42.45)(temp:(28.3529.05])	0,62	10/16
(tds:(22.6527.45])(temp:(28.3529.05] tds:<=22.65)	0,69	11/16
(tds:(22.6527.45])(temp:(28.3529.05] ce:<=42.45)	0,69	11/16
(ph:>5.62)(ce:>51.30 tds:>27.45)	0,69	11/16
(tds:>27.45)(temp:>29.05)	0,69	11/16
(tds:(22.6527.45])(tds:(22.6527.45])	0,75	12/16
(tds:(22.6527.45])(ce:(42.4551.30])	0,75	12/16
(tds:>27.45)(temp:(28.3529.05] ce:<=42.45)	0,75	12/16
(tds:>27.45)(temp:(28.3529.05] ce:<=42.45 tds:<=22.65)	0,75	12/16
(tds:>27.45)(ce:<=42.45 tds:<=22.65)	0,81	13/16
(tds:>27.45)(ph:>5.62)	0,81	13/16
(tds:>27.45)(tds:<=22.65)	0,81	13/16
(tds:>27.45)(ph:<=5.45)	0,81	13/16
(tds:(22.6527.45])(ph:(5.455.62])	0,88	14/16
(tds:(22.6527.45])(ce:<=42.45)	0,88	14/16
(ce:(42.4551.30])(ph:<=5.45)	0,88	14/16
(tds:>27.45)(temp:(28.3529.05])	0,88	14/16
(ph:<=5.45)(ce:<=42.45)	0,94	15/16
(ph:<=5.45)(ce:<=42.45 tds:<=22.65)	0,94	15/16
(ph:<=5.45)(temp:(28.3529.05])	0,94	15/16
L3		
(tds:(22.6527.45])(ph:(5.455.62] tds:>27.45)(ce:<=42.45)	0,44	7/16
(ce:<=42.45)(ph:<=5.45)(ce:>51.30)	0,44	7/16
(ce:<=42.45)(ph:<=5.45)(ce:>51.30 tds:>27.45)	0,44	7/16
(ce:<=42.45)(ph:<=5.45)(temp:(28.3529.05])	0,44	7/16
(tds:(22.6527.45])(ph:(5.455.62])(temp:>29.05)	0,5	8/16
(tds:(22.6527.45])(ph:(5.455.62])(ph:>5.62 temp:>29.05)	0,5	8/16
(ce:<=42.45)(temp:(28.3529.05])(ce:>51.30)	0,5	8/16
(ce:<=42.45)(temp:(28.3529.05])(ce:>51.30 tds:>27.45)	0,5	8/16
(ce:<=42.45)(temp:(28.3529.05])(ph:(5.455.62])	0,5	8/16
(tds:>27.45)(ph:<=5.45)(temp:<=28.35)	0,56	9/16
(tds:>27.45)(ph:<=5.45)(temp:(28.3529.05] tds:>27.45)	0,56	9/16
(tds:>27.45)(ph:<=5.45)(temp:(28.3529.05] ce:>51.30)	0,56	9/16
(tds:>27.45)(ph:<=5.45)(temp:(28.3529.05] ce:>51.30 tds:>27.45)	0,56	9/16
(tds:(22.6527.45])(ph:(5.455.62])(tds:(22.6527.45])	0,62	10/16
(tds:(22.6527.45])(ph:(5.455.62])(ce:(42.4551.30])	0,62	10/16
(tds:(22.6527.45])(ph:(5.455.62])(ce:(42.4551.30] tds:(22.6527.45])	0,62	10/16
(tds:>27.45)(ph:<=5.45)(temp:(28.3529.05] ce:<=42.45 tds:<=22.65)	0,62	10/16
(tds:>27.45)(ph:<=5.45)(ph:<=5.45)	0,62	10/16

Motif	Support	Equivalence
(tds:>27.45)(ph:<=5.45)(ce:(42.4551.30])	0,69	11/16
(tds:>27.45)(ph:<=5.45)(ce:(42.4551.30] tds:(22.6527.45])	0,69	11/16
(tds:>27.45)(ph:<=5.45)(ph:>5.62)	0,69	11/16
(tds:>27.45)(ph:<=5.45)(ph:(5.455.62] temp:(28.3529.05])	0,69	11/16
(tds:(22.6527.45])(tds:>27.45)(ph:>5.62)	0,75	12/16
(tds:(22.6527.45])(ce:>51.30 tds:>27.45)(ph:>5.62)	0,75	12/16
(tds:>27.45)(ph:<=5.45)(tds:<=22.65)	0,75	12/16
(tds:>27.45)(ph:<=5.45)(ce:>51.30 tds:>27.45)	0,75	12/16
(tds:(22.6527.45])(ph:<=5.45)(temp:(28.3529.05])	0,81	13/16
(ce:>51.30 tds:>27.45)(ph:<=5.45)(ph:(5.455.62])	0,81	13/16
(tds:>27.45)(ph:<=5.45)(ph:(5.455.62])	0,81	13/16
(tds:>27.45)(ph:<=5.45)(temp:(28.3529.05])	0,81	13/16
(ph:<=5.45)(tds:(22.6527.45])(ph:(5.455.62])	0,88	14/16
(ph:<=5.45)(ce:(42.4551.30])(ph:(5.455.62])	0,88	14/16
(ph:<=5.45)(ce:(42.4551.30] tds:(22.6527.45])(ph:(5.455.62])	0,88	14/16
L4		
(tds:(22.6527.45])(ph:(5.455.62])(tds:(22.6527.45])(ph:(5.455.62])	0,44	7/16
(temp:<=28.35 ce:>51.30 tds:>27.45)(ph:>5.62)(ph:(5.455.62] ce:(42.4551.30])(ph:(5.455.62])	0,44	7/16
(tds:>27.45)(ce:>51.30 tds:>27.45)(ce:(42.4551.30] tds:(22.6527.45])(ce:>51.30 tds:>27.45)	0,44	7/16
(ce:<=42.45)(ph:<=5.45)(ce:>51.30)(temp:(28.3529.05])	0,44	7/16
(tds:(22.6527.45])(ph:(5.455.62])(tds:(22.6527.45])(ce:<=42.45)	0,5	8/16
(temp:<=28.35)(ce:(42.4551.30] tds:(22.6527.45])(ph:>5.62)(ce:<=42.45)	0,5	8/16
(temp:<=28.35)(temp:(28.3529.05])(temp:<=28.35)(ce:>51.30)	0,5	8/16
(tds:>27.45)(ph:<=5.45)(ph:<=5.45)(ph:(5.455.62])	0,5	8/16
(tds:(22.6527.45])(ph:(5.455.62])(ph:>5.62)(ce:<=42.45)	0,56	9/16
(tds:>27.45)(ph:<=5.45)(temp:(28.3529.05])(ph:<=5.45)	0,56	9/16
(tds:>27.45)(ph:<=5.45)(ph:<=5.45)(ph:>5.62)	0,56	9/16
(tds:>27.45)(ph:<=5.45)(ph:<=5.45)(temp:(28.3529.05])	0,56	9/16
(tds:>27.45)(ph:<=5.45)(ce:>51.30)(temp:(28.3529.05])	0,62	10/16
(tds:>27.45)(ph:<=5.45)(ph:(5.455.62])(ce:<=42.45)	0,62	10/16
(tds:>27.45)(ph:<=5.45)(ph:(5.455.62])(ce:<=42.45 tds:<=22.65)	0,62	10/16
(tds:>27.45)(ph:<=5.45)(ph:(5.455.62])(tds:<=22.65)	0,62	10/16
(tds:>27.45)(ph:<=5.45)(ph:(5.455.62])(ph:(5.455.62])	0,62	10/16
(tds:(22.6527.45])(temp:(28.3529.05])(ph:<=5.45)(temp:(28.3529.05])	0,69	11/16
(temp:<=28.35)(ph:>5.62)(tds:(22.6527.45])(ph:<=5.45)	0,69	11/16
(tds:>27.45)(ph:<=5.45)(tds:(22.6527.45])(ph:(5.455.62])	0,69	11/16
(tds:>27.45)(ph:<=5.45)(ce:(42.4551.30])(ph:(5.455.62])	0,69	11/16
(tds:>27.45)(ph:<=5.45)(ce:(42.4551.30] tds:(22.6527.45])(ph:(5.455.62])	0,69	11/16
(temp:<=28.35)(temp:(28.3529.05])(ph:<=5.45)(tds:(22.6527.45])	0,75	12/16
(temp:<=28.35)(temp:(28.3529.05])(ph:<=5.45)(ce:(42.4551.30])	0,75	12/16
(temp:<=28.35)(ph:<=5.45)(ce:(42.4551.30])(ph:(5.455.62])	0,75	12/16
(temp:<=28.35)(ph:<=5.45)(ce:(42.4551.30] tds:(22.6527.45])(ph:(5.455.62])	0,75	12/16

Motif	Support	Equivalence
L5		
(temp:<=28.35)(temp:>29.05)(ph:>5.62)(tds:(22.6527.45])(temp:(28.3529.05])	0,44	7/16
(temp:<=28.35 tds:>27.45)(temp:>29.05)(ph:>5.62)(ce:(42.4551.30])(temp:(28.3529.05])	0,44	7/16
(temp:>29.05)(temp:>29.05)(ce:(42.4551.30] tds:(22.6527.45])(ce:(42.4551.30])(temp:(28.3529.05])	0,44	7/16
(tds:>27.45)(ph:<=5.45)(ph:<=5.45)(ph:>5.62)(tds:<=22.65)	0,44	7/16
(temp:<=28.35)(ph:(5.455.62])(ph:>5.62)(tds:(22.6527.45])(ph:(5.455.62])	0,5	8/16
(tds:>27.45)(ph:<=5.45)(ce:>51.30 tds:>27.45)(temp:(28.3529.05])(tds:<=22.65)	0,5	8/16
(tds:>27.45)(ph:<=5.45)(ce:>51.30)(temp:(28.3529.05])(ce:<=42.45)	0,5	8/16
(tds:>27.45)(ph:<=5.45)(ph:(5.455.62])(ce:(42.4551.30] tds:(22.6527.45])(tds:<=22.65)	0,5	8/16
(temp:<=28.35)(tds:(22.6527.45])(tds:(22.6527.45])(tds:>27.45)(ph:>5.62)	0,56	9/16
(temp:<=28.35)(tds:(22.6527.45])(tds:(22.6527.45])(ph:>5.62)(ph:>5.62)	0,56	9/16
(tds:>27.45)(temp:(28.3529.05])(ph:<=5.45)(ce:(42.4551.30] tds:(22.6527.45])(ce:<=42.45 tds:<=22.65)	0,56	9/16
(tds:>27.45)(temp:(28.3529.05])(ph:<=5.45)(ce:(42.4551.30] tds:(22.6527.45])(tds:<=22.65)	0,56	9/16
(temp:<=28.35)(tds:(22.6527.45])(ph:<=5.45)(tds:>27.45)(ph:>5.62)	0,62	10/16
(temp:<=28.35)(tds:(22.6527.45])(ph:<=5.45)(ce:>51.30 tds:>27.45)(ph:>5.62)	0,62	10/16
(tds:>27.45)(ce:>51.30)(temp:(28.3529.05])(ce:(42.4551.30] tds:(22.6527.45])(ce:<=42.45)	0,62	10/16
(tds:>27.45)(ce:>51.30)(temp:(28.3529.05])(ce:(42.4551.30] tds:(22.6527.45])(ce:<=42.45 tds:<=22.65)	0,62	10/16
(tds:>27.45)(ce:>51.30)(temp:(28.3529.05])(ce:(42.4551.30] tds:(22.6527.45])(tds:<=22.65)	0,62	10/16
L6		
(temp:<=28.35)(tds:(22.6527.45])(ph:(5.455.62])(ph:>5.62)(ph:>5.62)(ce:<=42.45)	0,44	7/16
(temp:<=28.35)(tds:(22.6527.45])(ph:(5.455.62])(ph:>5.62)(ph:>5.62)(ce:<=42.45 tds:<=22.65)	0,44	7/16
(tds:>27.45)(ce:(42.4551.30])(ph:<=5.45)(ce:>51.30)(ph:>5.62)(tds:<=22.65)	0,44	7/16
(tds:>27.45)(ph:>5.62)(temp:(28.3529.05])(temp:(28.3529.05])(tds:(22.6527.45])(ph:(5.455.62])	0,44	7/16
(temp:<=28.35)(tds:(22.6527.45])(tds:(22.6527.45])(ph:<=5.45)(tds:>27.45)(ph:>5.62)	0,5	8/16
(temp:<=28.35)(ce:(42.4551.30])(ph:<=5.45)(ce:>51.30)(ph:>5.62)(ce:<=42.45)	0,5	8/16
(temp:<=28.35)(temp:(28.3529.05])(ph:<=5.45)(ph:>5.62)(ph:>5.62)(tds:<=22.65)	0,5	8/16
(temp:<=28.35)(temp:(28.3529.05])(ph:<=5.45)(ph:>5.62)(ph:>5.62)(temp:<=28.35)	0,5	8/16

Les résultats sont constitués de 67.912 séquences fréquentes avec un support variant entre 0,4 et 1. Plus les séquences comportent d'attributs moins on retrouve d'objets similaires. En analysant les données les plus significatives, on pourrait regrouper les supports en trois (3) classes :

— la classe des faibles valeurs < 0,5 contenant 1/4 des séquences :
— l'intervalle des valeurs moyennes variant de 0,5 à 0,69 qui inclut également 1/4 des séquences ; le pH, la CE et la TDS sont très liés ;
— l'intervalle des valeurs se rapprochant de 1 (>0,7) qui comporte quasi 30.000 séquences ; les niveaux maximales et minimales des valeurs des attributs restent les mêmes pour tous les sites.

La faible variablité temporelle se fait ressentir dans ce jeu de données ; elle s'explique par les écarts entre les missions des différentes équipes. Ce facteur est très limitant car il ne permet pas de mettre en corrélation les dimensions temporelles et spatiales de façon fiable ; nous nous ne pouvo appréhender la dynamique spatiale.

Cela nous pousse a évaluer les sites de recueil d'eau et ceux de positionnement des stations/postes suivant la position des mares géographiques auxquelles ils font référence. Le regroupement par mares, nous permet donc de distinguer trois (3) groupes dont les attributs se rapprochent. Cette analyse correspond également à la proximité géographique des mares sur la carte de la zone d'étude qui peut être traduite suivant le découpage ci-dessous :

— nord-ouest : on retrouve la mare de Niakha isolée ;
— centre : Barkédji et Kangaledji ;
— sud : Furdu, Beli Boda et Ngao.

4.5 Article Acta Biotheorica

Acta Biotheor
DOI 10.1007/s10441-014-9235-7

REGULAR ARTICLE

Decision Making Environment on Rift Valley Fever in Ferlo (Senegal)

Fanta Bouba · Alassane Bah · Christophe Cambier · Samba Ndiaye · Jacques-André Ndione · Maguelonne Teisseire

Received: 22 December 2013 / Accepted: 11 July 2014

Abstract The Rift Valley fever (RVF), which first appeared in Kenya in 1912, is an anthropozoonosis widespread in tropical areas. In Senegal, it is particularly felt in the Ferlo area where a strong presence of ponds shared by humans, cattle and vectors is noted. As part of the studies carried out on the environmental factors which favour its start and propagation, the focus of this paper is put on the decision making process to evaluate the impacts, the interactions and to make RVF monitoring easier. The present paper proposes a model based on data mining techniques and dedicated to trade experts. This model integrates all the involved data and the results of the analyses

F. Bouba (✉) · A. Bah · S. Ndiaye
UMI 209, UMMISCO-UCAD, Dakar, Senegal
e-mail: boubafanta@gmail.com

A. Bah
e-mail: alassane.bah@gmail.com

S. Ndiaye
e-mail: samba.ndiaye@ucad.edu.sn

F. Bouba · C. Cambier
UPMC, Paris, France
e-mail: christophe.cambier@upmc.fr

A. Bah
ESP/UCAD, Dakar, Senegal

C. Cambier
UMI 209, UMMISCO-IRD, Bondy, France

J.-A. Ndione
CSE and LPA/UCAD, Dakar, Senegal
e-mail: jacques-andre.ndione@cse.sn

M. Teisseire
TETIS - Irstea, Montpellier, France
e-mail: maguelonne.teisseire@teledetection.fr

Published online: 09 August 2014

Springer

made on the characteristics of the surrounding ponds. This approach presents some advantage in revealing the relationship between environmental factors and RVF transmission vectors for space–time epidemiology monitoring purpose.

Keywords Data mining · Decision-making system · Multidimensional modelling · Rift Valley fever · Spatio-temporal patterns

1 Introduction

The Rift Valley fever (RVF) is an arbovirosis which is transmitted by vectors. Despite a low death rate, this disease, seen as an "economic and biological weapon, has a great power of diffusion and is particularly serious" (Diagne 1992). As a matter of fact, its virus diffusion can be made by aerosol and the vectors eggs are resistant to a long duration period and dryness (Flick and Bouloy 2005). That is why it appears in the common list of several species diseases of the International Organization for Animal Health (Information 2013). In the current approach, the understanding of emergence and spreading process offers a compartmental cellular view. However, significant works (Ndione et al. 2003; Préhaud and Bouloy 1997), have demonstrated the impact of factors such as climate, livestock movement, hydrology on the emergence and persistence of RVF. The complexity of this analysis requires the involvement of all the actors concerned by multidisciplinary dynamics. This approach, based on the multi-criteria and multidimensional analysis, makes it possible to propose health risk management measures by controlling the impacts of the environmental factors and their interactions. It is therefore proposed that a decision making approach be adopted in order to provide trade experts and decision makers with analysis tools, related to the dynamics of animal and human populations. So, to find out new and relevant information, and in order to understand the interactions between the different objects and make predictions, it is necessary to resort to data mining techniques for classification and prediction purposes. These techniques are applied to assess the relevance of the parameters retained for the RVF space–time monitoring. Further in this paper, Sect. 2 is devoted to the description of the studied disease and to the data search applied in epidemiology. Section 3 describes the manipulated data and the experimented data mining algorithms. Section 4 presents (1) the data model gathering the environment and health parameters integrating the time-space dimensions and (2) extracts of correlated analysis on the characterization of temporary pools in the studied area. Section 5 concludes this approach by discussing the choices made and by proposing study tracks for the continuation of this work.

2 Thematic and Scientific Context

2.1 Rift Valley Fever

RVF is an infectious viral disease which affects animals and humans (OMS 2010). This disease, conveyed by arthropod vectors (Diallo et al. 2000), is caused by a

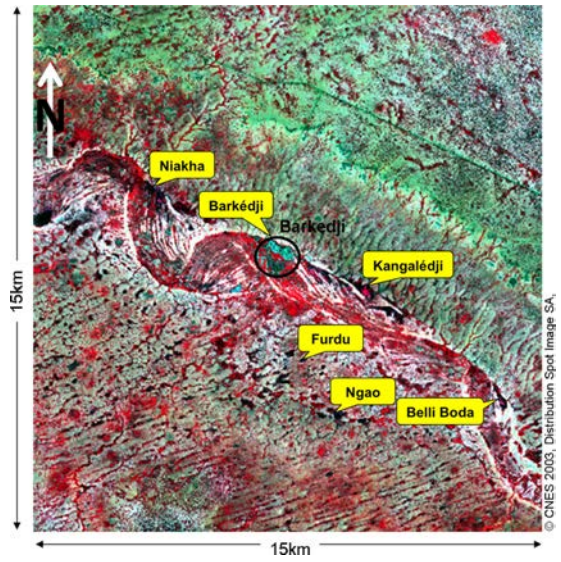

Fig. 1 Ponds of the study area, adapted from Ndione et al. (2009) and Tourre et al. (2008)

virus of the Bunyaviridae family Phlebovirus kind. The RVF can be transmitted by two main modes: the vectorial transmission mode by some blood-sucking insects and the direct transmission mode by personal contact with an infected host (Greboval 2003). In this article, virological data issued from vectors are used. In Senegal, this disease holds an important place in Ferlo. According to Diallo (1995), RVF spread identified in the Ferlo area (Fig. 1) is linked to the mosquito life cycle and ponds evolutionary cycle.

Indeed, Ferlo's ponds constitute both the water supply source for the population and the livestock while being the larva living place of the vectors. In this area, the rainfall intra-seasonal variability, the vegetation dynamics and the turbidity of temporary ponds (the size of which is comparatively small) are the main factors explaining the strong concentration of mosquitoes (Ndione et al. 2009). To understand the vectorial diseases process, it is necessary to take into account the multiple and varied phenomena in the same environment, but also at different scales (time, space and organization).

The partners[1] involved in the QWeCI project (http://www.liv.ac.uk/qweci) in Senegal (Fig. 2) provide data from field investigations (data observed or generated by measurement equipment) and laboratory tests (analysis data).

[1] UCAD (Université Cheikh Anta Diop)—IPD (Institut Pasteur de Dakar)—CSE (Centre de Suivi Ecologique)—DSV (Direction des Services Vétérinaires)—PNLP (Programme National de Lutte contre le Paludisme).

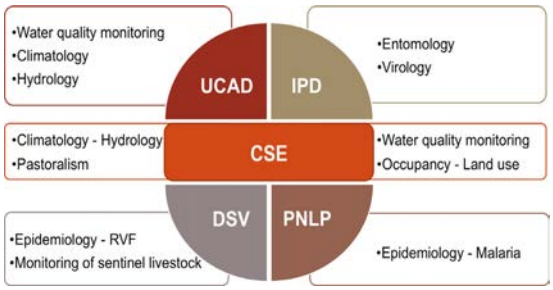

Fig. 2 Organizational complexity

The challenge to be taken up by the research teams is the impact control of the environmental factors on the triggering and spreading of RVF. It is therefore fundamental to take into account the whole set of the different disciplines data in the proposed data model.

2.2 Decision Making Environment and Data Mining in Epidemiology

Data mining is an outstanding stage in the decision-making process allowing to review information, not a priori obvious, in a high data volume. As part of this work, the analysis method of the space time data is defined by Agrawal and Srikant (1995) and Alatrista et al. (2011). The authors of this method focus on including the spatialization notion in the knowledge extraction process. This approach is based on pre-processing the data to include space characteristics into time data. The space sequences thus obtained are then used as inputs for the data mining stage in order to extract spatially frequent sequences. This is achieved with the help of a classical algorithm (Mortazavi-Asl et al. 2000) to extract the sequential motives representing the spatially frequent time evolutions of the studied area.

3 Materials and Methods

This section presents data identified during the analysis phase (field surveys or laboratory results) and the formalisms used to process these data.

3.1 Data Description

3.1.1 Space Data

The present study is carried out with the data of Barkédji rural community (Fig. 1). Thus, a distinction can be made between ponds and encampment (Fig. 3). For possible generalizations, the present paper has chosen to integrate a larger

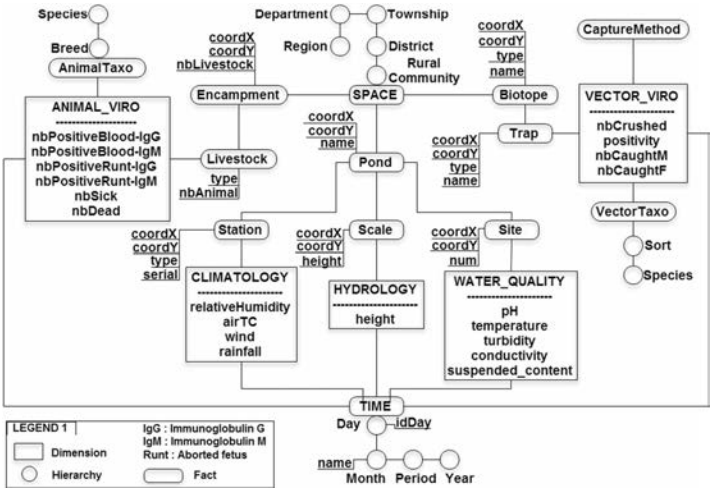

Fig. 3 Data model

dimension by starting from the Senegal administrative division. The pond, however, is the key element of the present study as it is the junction place between vectors and hosts.

3.1.2 Climate, Hydrology and Water Quality Data

Ponds of this geographical area are linked to several elements:

- rainfall stations and automatized meteo stations are installed around ponds: measurements such as wind speed, temperature, rainfall are automatically generated and periodically collected;
- gauging scales are set on water ponds to record the water level;
- water samples are identified by their collection point coordinates: they make it possible to measure pH, temperature and matters in suspension.

3.1.3 Entomological and Virological Data

We also use data from field surveys and laboratory health teams. During the field campaigns, mosquitoes are captured by using bait (sheep, chicken) and CO_2 traps. Then, individual identification follows to determine age, sex, etc. Besides, blood samples are taken up among sentry flocks. The mosquito's virology is assessed

based on a crushed sample. On the other hand, animal virology is tested individually on the sampling.

3.2 Methodology

3.2.1 Multidimensional Modelling

The research work on the environment impact on health and more particularly on vector diseases (Tran et al. 2005) made it possible to distinguish two big classes of models: (1) mathematical models (geographical and epidemiological) which are based on a formal representation to provide quantitative descriptions and (2) conceptual models (relational, object and multidimensional) based on the need of the final users. In the present context, the multidimensional approach is best adapted as it provides some formalism dedicated to decision making systems. Multidimensional modelling is described as a representation approach, which considers the analysed subject as a point (the fact) in a several dimension space. Dimensions can be decomposed following a hierarchy linked to the required granularity level Wehrle (2009). In their article on the OLAP algebra Ravat et al. (2010), Ravat et al. propose the following definitions:

- a dimension is defined by its attributes, hierarchies and the set of instances;
- a hierarchy is defined by an ordered set describing the attributes hierarchy (each attributes, called parameter, corresponds to an analysis granularity level);
- a fact is defined by a set of measurements, a set of instances and a function associating each instance to the instances of the dimensions linked to the fact.

3.2.2 Spatial and Temporal Patterns

The spatial dimension is used to define the location of an object in a geographical space and the temporal dimension relates to the time step (date, time, etc.). To better understand this data processing technique, we present the main concepts used, based on the work of Fabrègue et al. (2012):

- item: an item I is a literal value of dimension D_i, $I \in \text{dom}(D_i)$;
- itemset: an itemset, $IS = \langle I_1 I_2 \ldots I_n \rangle$ is a non-empty set of items. All items of an itemset are associated with different dimensions of analysis;
- itemsets sequence: a sequence S is a non empty ordered list of itemsets noted $\langle s_1 s_2 \ldots s_p \rangle$ where s_j is a itemset;
- absolute support: the absolute support of a sequence S is the number of data warehouse areas which satisfy S;
- relative support: the relative support is defined as the ratio between absolute support and the number of areas;
- pattern: let S be the sequence and θ the minimal support defined, if the relative support $\geq \theta$ then S is a spatio-temporal pattern.

4 Results

In this section, we introduce the representation model and the first data mining results obtained with the decision-making environment described by Marakas (2003).

Based on the formalism of multidimensional modelling (Sect. 3.2.1), we built the data model (Sect. 4.1), which is one of the activities of the first step (data management) of this methodology. This model allowed us to implement our data warehouse in which we integrate all handled data (Sect. 3.1).

For the second phase (knowledge management), we use a decision algorithm to identify the species of vectors more present in the different localities of the study area (Sect. 4.2). Then, we apply a spatio-temporal algorithm (Sect. 3.2.2) to get similar parameters of these localities ponds (Sect. 4.3). This work is a technique we proposed to identify and classify risk areas.

4.1 Data Model

In accordance with the multidimensional formalism, based on Golfarelli's proposal (Golfarelli et al. 1998), a constellation model (Fig. 3) has been built, including two (2) fact tables for animal and vector virology and three (3) fact tables for environmental data (climatology, hydrology and water quality). Dimensions have been identified in function of different analysis criteria:

- the vector (mosquitoes) dimensions and the livestock;
- the space dimension hierarchized following the administrative division;
- the time dimension taking into account study periods;
- environment analysis dimensions: scale (measurement ruler used for gauging), station (meteorology station, rainfall station) and site (pond water collection point).

The implemented model is used as a data source for selected algorithms and tools.[2] Yet, it is worth noting that several problems have been encountered:

- irregular temporality: investigations for data production are irregularly carried out, while collection missions are not jointly led between the different teams;
- the data volume: the available data sets are too small and do not guarantee right interpretation;
- the data format: following the project adopted protocols, data are not collected data with the same format. Thus the low quality of this data constitutes a strong constraint in the choice of the algorithms.

[2] This data model is experienced through the implementation of a PostgreSQL data warehouse fed from different sources using Talend ETL (Extract Transform Load). ETL is used to retrieve data from a source, to process those data (cleaning, formatting or structural change) and to load data in another file or database.

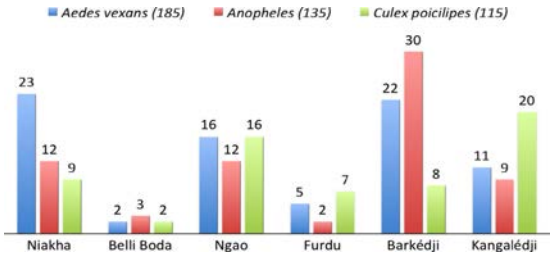

Fig. 4 Distribution of vectors in studied area

The reporting tests made, with the Pentaho tools, on our data warehouse confirm the completeness of the model.

4.2 Decision Tree

The "W-Random Tree" algorithm[3] is applied to describe the vector diversity of the locality related to the studied area ponds.[4] The focus has particularly been put on the potential vectors of RVF (*Aedes vexans*—blue bar, *Anopheles*—red bar, *Culex poicilipes*—green bar). This analysis (Fig. 4)[5] of the vector diversity by locality essentially confirms:

- the high presence of *Aedes vexans* in Niakha, Ngao and Barkédji;
- the high presence of *Culex poicilipes* in Ngao and Kangalédji;
- very few *Anopheles* in Belli Boda and Furdu.
  It is therefore worth identifying the similarity points among these geographical areas. Various studies have shown that the Sahelian climate characterizes all these localities. Then, it became interesting to focus on their water parameters using the space–time patterns.

4.3 Spatio-Temporal Patterns

The data warehouse is transformed in a database of sequences with the algorithm proposed by Alatrista et al. (2011).

This method first retrieves the list of all frequent occurrences (or frequent item) in the database based on the minimum support. A frequent item means that a pattern of greater length was found. For our experimentation, we use data sets for climate elements (temperature, wind, rainfall…) and water quality parameters of the ponds

[3] Tests were performed with RapidMiner to which we integrated the Weka algorithms.

[4] This graphic format was chosen to facilitate interpretation by trade experts.

[5] The number in parenthesis represents the total number of captured vectors in the studied area.

Table 1 Patterns "Ponds"

Patterns	Support
(EC:[42.45;51.30])(TDS:[22.65;27.45])(Temp:>29.05)(pH:[5.36;5.54])	1 (5/5)
(pH:[5.54;5.66])(Temp:[28.35;29.05])	1 (5/5)
(pH:[5.36;5.54])(TDS:[22.65;27.45])(EC:[42.45;51.30])(Temp:[28.35;29.05])	1 (5/5)

(acidity content, dissolved solids, temperature…).[6] Two analyses have been carried out: pond gathered data and site gathered data. For the first data set (ponds), the time dimension is a great deal varied, with a 0.9 minimal support (at least 90 % of the ponds meet the supplied sequences requirements), 21,131 frequent sequences are obtained (Table 1).

These obtained patterns make it possible to certify that the water quality of the five (5) study ponds of the Ferlo zone (Fig. 1) is very similar. Indeed, these patterns have a 100 % support. This means that all the ponds analysed during similar periods have exactly the same behaviour. It is so equally deemed more relevant to confront the variants of the ponds with those of the meteorology-climate environment to control the impact of the latter on the pond water quality. Thus, studying the same data on the basis of the rainfall stations and posts, we obtained very diversified supports, varying between 44 % (temperature, pH, dissolved solids) and 94 % (temperature; Table 2). This data set includes more geographical areas (localities) but fewer dates. The results consist of 67,912 for a 0.4 minimal support (at least 40 % of the localities meet the sequence requirements).

These latest results provide no coherent interpretation. Indeed, this analysis brings no answer based on possible correlations among the pond water characteristics and those supplied by meteorological automatized stations. The low time density could justify this. The meteorological data confirm that the rural community climate in these study areas is very similar. It should then be fit to focus on other parameters such as soil quality to be able to justify the study area vector diversity.

5 Conclusion and Perspectives

As a reminder, the present research work aims at proposing a decision environment describing the interactions between the environment indicators and the spreading of the RVF. This paper has presented the data model and a few extracts of the patterns obtained on the pond water quality. The proposed data model makes it possible to (1) identify the different analysis axes; (2) correlate the different quality measurement indicators (Trujillo et al. 2003); (3) propose views corresponding to the expectations of the different involved disciplines (Golfarelli et al. 2002). Moreover, the data warehouse is also used as a data source for data mining tools and the algorithm of spatio-temporal patterns. These experimentations have made it

[6] TDS: Total Dissolved Solids, EC: Electrical Conductivity, Temp: Temperature, pH : potential Hydrogen.

Table 2 Patterns "Sites"

Patterns	Support
(Temp:[28.35;29.05])	0.94 (15/16)
(pH:≤5.45)(EC:[42.45;51.30])(TDS:[22.65;27.45])	0.75 (12/16)
(pH:[5.45;5.62])(Temp:[28.35;29.05])(TDS:≤22.65)	0.44 (7/16)

possible to confirm that the Ferlo area ponds present the same water quality characteristics. In spite of the weak data set, the realized experiments confirm this integration approach of the spatio-temporal attributes to understand and have control over the RVF health risk. For the continuation of this research work, it is important to be able to confront the achieved results with virological data. In addition, we are working on identifying neighbourhoods of geo-referenced objects following polar and/or Euclidean approach. Thus, the problem will be to identify all correlations between the geographical environmental parameters and the entomological and health data taking into account basic characters such as time and space. Furthermore, observing the livestock transhumance should provide additional elements for space–time projections of the RVF appearance and spreading. In data mining, data make up the information raw material upon which the decisions should be made. So, issues related to the data quality are the cause of the main failures in setting up a decision system. To go thoroughly into the present approach, it would be appropriate to propose a method for rebuilding the missing data, taking into account their proportion and their type.

Acknowledgments Hugo Alatrista Salas, Lilia Beharrou, Centre National d'Etudes Spatiales (CNES), Institut Pasteur de Dakar (IPD), Centre de Suivi Ecologique (CSE).

References

Agrawal R, Srikant R (1995) Mining sequential patterns. In: Yu PS, Chen ALP (eds) Proceedings of the eleventh international conference on data engineering (ICDE), Taipei, Taiwan, pp 3–14. IEEE Computer Society

Alatrista H, Cernesson F, Bringay S, Azé J, Flouvat F, Semaloui N, Teisseire M (2011) Recherche de séquences spatio-temporelles peu contredites dans des données hydrologiques. Revue des Nouvelles Technologies de l'Information (RNTI), numéro spécial Qualité des Données et des Connaissances/ Evaluation des Méthodes d'Extraction des Connaissances dans les Données, vol RNTI-E-22, pp 165–188, ISBN: 978-2-70568-286-6

Diagne FF (1992) In: Etude de la Fièvre de la Vallée du Rift chez les ruminants domestiques au Sénégal: enquêtes sérologiques dans la vallée du Fleuve. le Ferlo et la Casamance, PhD, FMPOS/UCAD, pp 29–34

Diallo M (1995) Dynamique comparée des populations de Culicidae à Kédougou (zone soudano-guinéenne) et à Barkédji (zone de savane sahélienne): conséquences dans la transmission des arbovirus. Université Cheikh Anta Diop de Dakar, DEA en Biologie Animale

Diallo M, Lochouarn L, Ba K, Sall A, Mondo M, Girault L, Mathiot C (2000) First isolation of the Rift Valley fever virus from *Culex poicilipes* (Diptera: Culicidae) in nature. Am J Trop Med Hyg 6(62):702–704

Fabrègue M, Braud A, Bringay S, Le Ber F, Teisseire M (2012) Extraction de motifs spatio-temporels à différentes échelles avec gestion de relations spatiales qualitatives, Infosid 2012, Montpellier

Flick R, Bouloy M (2005) Rift valley fever virus. Curr Mol Med 5:827–834

Golfarelli M, Maio D, Rizzi S (1998) The dimensional fact model: a conceptual model for data warehouses. Int J Coop Inf Syst 7(2–3):215–247

Golfarelli M, Rizzi S, Saltarelli E (2002) WAND: a case tool for workload based design of a data mart. 10th National convention on systems evolution for data bases, pp 422–426

Greboval M (2003) Facteurs environnementaux influençant la dynamique des vecteurs du virus de la Fièvre de la Vallée du Rift: conséquences pour la modélisation de la maladie. Thèse de Doctorat Vétérinaire, ENVL - Ecole Nationale Vétérinaire de Lyon

Information on http://www.oie.int/fr/sante-animale-dans-le-monde/maladies-de-la-liste-de-loie-2013 (2013)

Marakas GM (2003) Decision support systems in the 21st century. Prentice Hall, Upper Saddle River

Mortazavi-Asl B, Pinto H, Dayal U (2000) PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. In: Proceedings 17th international conference on data engineering, pp 215–224

Ndione JA, Besancenot J, Lacaux J, Sabatier P (2003) Environnement et épidémiologie de la Fièvre de la Vallée du Rift (FVR) dans le bassin inférieur du fleuve Sénégal. Environnement, Risques et Santé 3(2):176–182

Ndione JA, Lacaux JP, Tourre Y, Vignolles C, Fontanaz D, Lafaye M (2009) In: Mares temporaires et risques sanitaires au Ferlo : contribution de la télédétection pour l'étude de la fièvre de la vallée du Rift entre août 2003 et janvier 2004. Secheresse 20(1):153-160

OMS, La fièvre de la Vallée du Rift. Aide-Mémoire 207, Nations Unies (2010). http://www.who.int/mediacentre/factsheets/fs207/fr/index.html

Préhaud C, Bouloy M (1997) La fièvre de la vallée du rift, un modèle d'étude des fièvres hémorragiques virales. Technical report, Institut Pasteur

Ravat F, Teste O, Zurfluh G (2010) Algèbre OLAP et langage graphique. CoRR abs/1005.0213

Tourre YM, Lacaux JP, Vignolles C, Ndione JA, Lafaye M (2008) Mapping of zones potentially occupied by Aedes vexans and Culex poicilipes mosquitoes, the main vectors of Rift Valley fever in Senegal. Geospatial Health 3(1):69–79

Tran A, Biteau-Coroller F, Guis H, Roger F (2005) Modélisation des maladies vectorielles. Epidémiol et santé anim 47:35–51

Trujillo JC, Luján-Mora S, et Song I (2003) Applying UML for designing multidimensional databases and OLAP applications. In: Siau K (ed) Advanced topics in database research, vol 2. Idea Group Publishing, Hershey, PA, pp 13–36

Wehrle P (2009) Modèle multidimensionnel et OLAP sur architecture de grille. PhD thesis, Institut national des sciences appliquées de Lyon

CHAPITRE 5

Conclusion et perspectives

Les maladies vectorielles sont fortement dépendantes de l'environnement. De ce fait, les recherches sur ces maladies s'intéressent particulièrement aux interactions entre les facteurs environnementaux, leur apparition et leur propagation. Nos travaux de recherche se placent au du processus décisionnel en épidémiologie vectorielle avec comme cas pratique la maladie de la Fièvre de la Vallée du Rift. Cette maladie est une zoonose causée par un virus et qui affecte les animaux et les hommes. Notre choix portant sur la mise en place d'un environnement décisionnel se justifie par la recherche d'informations, d'analyse et de prédiction pour les experts et utilisateurs finaux afin de répondre à des questions stratégiques.

Ainsi, l'objectif est de fournir aux analystes et aux décideurs qui interviennent dans ce domaine :

— des vues fonctionnelles : description et traitement des caractéristiques des objets manipulés ;
— des vues opérationnelles : description de scénarii – les décisions sont liées à un contexte suivant des objectifs.

Avec l'évolution des technologies informatiques, des systèmes de surveillance et de supervision ont été développés afin d'offrir des vues prévisionnelles et décisives aux acteurs de la santé. Toutefois, les problèmes liés à la maîtrise du déclenchement et à la propagation des maladies vectorielles doivent être plus approfondis par des techniques de fouilles de données dans un but de compréhension et de prévision.

Dans son article sur l'intégration d'un système décisionnel dans une organisation, (Bruley, 2010) fait ressortir le cycle d'évolution des données en information, des informations en connaissances et des connaissances en plan d'actions. Ce processus se décline dans nos travaux à travers l'utilisation de plusieurs techniques et outils dont (1) la définition du modèle de données basée sur la modélisation multidimensionnelle, (2) la mise en place d'une entrepôt de données, (3) le déploiement d'un outil de *reporting*, (4) le traitement des données par des outils de fouille et enfin (4) l'interprétation des résultats obtenus.

Le modèle de données est conçu en utilisant la modélisation multidimensionnelle qui est un formalisme adapté à la représentation des données destinées à une analyse décision-

nelle. Ce formalisme considère un sujet analysé, le fait, comme un point dans un espace à plusieurs dimensions qui peuvent être décomposées suivant une hiérarchie liée au niveau de granularité requis (Bellatreche, 2003),(Ravat *et al.*, 2001), (Teste, 2000), (Wehrle, 2009). Suivant le formalisme multidimensionnel, basé sur la proposition de (Golfarelli *et al.*, 1998), nous avons identifié deux (2) tables de fait pour les virologies animale et vectorielle et trois (3) tables de fait pour les données environnementales telles que la climatologie, l'hydrologie et le suivi de la qualité de l'eau. Les dimensions ont été identifiées en fonction des différents critères d'analyse. On retrouve ainsi :

- les dimensions vecteurs (moustiques) et les troupeaux ;
- la dimension spatiale hiérarchisée suivant le découpage administratif ;
- la dimension temporelle (date) avec la prise en compte des périodes d'étude ;
- les dimensions d'analyse environnementales : Echelle (règle de mesure utilisée en limnimétrie), Station (stations météorologiques ou postes pluviométriques) et Site (point de recueil d'eau dans les mares).

Ce modèle a été évalué et validé par la mise en place d'outils de reporting correspondants aux attentes des utilisateurs finaux. Ainsi, pour faciliter la lecture analytique suivant plusieurs angles, nous avons construit plusieurs cubes suivant les objectifs visés. Dans notre contexte, deux types de partenaires sont impliqués : les utilisateurs finaux et les scientifiques. Les utilisateurs finaux sont plus intéressés par les outils de mesure que sont les rapports et les tableaux de bord., visualisés via l'outil de visualisation de Pentaho. Ainsi, nous leur proposons une interface de génération de tableaux et graphiques dynamiques suivant une analyse multi-critères.

L'appropriation de la méthodologie utilisée pour le recueil et le traitement des données par les experts nous a permis de proposer un modèle de données générique stable pour la prise en compte d'autres maladies vectorielles. De part la prise en compte de la dimension "sanitaire" liée à la maladie et de la dimension liée à la cible (humains ou animaux), ce modèle peut ainsi être utilisé pour l'étude d'autres maladies vectorielles.

Les attentes des utilisateurs suscitées par cette première phase ont été très importantes et nous ont permis de définir des questions de recherche scientifique auxquelles les techniques de fouilles de données peuvent répondre. La fouille de données est une des étapes du processus décisionnel qui permet d'extraire des informations, à priori non évidentes, dans un volume élevé de données. La fouille de données suit un processus de sélection, d'exploration et de modélisation qui permet de générer des modèles inconnus et d'identifier les relations pouvant exister entre les données analysées (Giudici, 2003). Les différentes expérimentations ont été réalisées en utilisant des algorithmes (k-means, decision tree, generate prediction) de Weka intégrés à l'outil RapidMiner. Les résultats obtenus permettent aux scientifiques de classer les zones géographiques, d'identifier les paramètres environnementaux à fort ou faible impact, de prédire les tendances de futures périodes, etc. Ces résultats nous ont permis d'aboutir, entre autres, aux conclusions suivantes :

- les *Culex* de type *Neavei* et les *Aedes* de types *Ochraceus* et *Vexans* sont moins populaires que les autres types de Moustiques dans la zone d'étude ;
- les pièges à appâts sont moins attractifs que les pièges à CO2 ;
- lors de précipitations similaires, on note la présence du même statut taxonomique des vecteurs

— une forte présence de vecteurs de type *Aedes Vexans* durant la saison pluvieuse ;
— certaines espèces sont fortement ou faiblement plus présentes que dans la période de très faibles précipitations.

Cependant, nous avons pu identifier les limites de ces algorithmes pour la compréhension des impacts temporels et/ou spatiaux sur certaines données. En conséquence, il nous a fallu faire évoluer nos travaux vers des outils de fouille adaptés à l'analyse spatio-temporelle. Pour ce faire, nous nous sommes appuyés sur la méthode d'analyse des données spatio-temporelles définie par (Salas *et al.*, 2012) et (Agrawal et Srikant, 1995) ; ceci en nous basant sur les travaux de (Fabrègue *et al.*, 2012) .

Des résultats obtenus, nous avons réalisé les interprétations suivantes :

— toutes les mares analysées pendant des périodes similaires ont exactement le même comportement ;
— les données météorologiques confirment que, durant la saison pluvieuse, le climat des communautés rurales de cette zone d'étude est très similaire.

Limites

La mise en place d'outils de fouille de données repose fondamentalement sur le volume des données mais la variabilité des données du point de vue spatial et temporel est également un facteur à fort impact de réussite pour un tel projet. Aussi, vu le caractère interdisciplinaire de notre projet de recherche, il est important d'avoir des données qui concordent dans leur lieu de recueil mais aussi dans la période de recueil.

Cependant, tout au long de notre travail de recueil de données, nous avons été confrontés à de nombreux freins. Il s'agit entre autres :

1. des délais entre le recueil des données et leurs analyses par les laboratoire (ex : lors d'un entretien en 2012, les résultats des prélèvements de la campagne de 2011 n'étaient toujours pas disponibles) ;
2. le protocole de recueil des données varie d'un projet à un autre ; cela impacte fortement sur :
   — le format et la structure des données (ex : la limnimétrie a été relevée deux (2) fois par jour dans certains cas et une (1) fois par jour dans d'autres cas) ;
   — le niveau de détails des zones géographiques ;
   — la période de recueil (ex : dans certains cas, les mesures varient de Juillet à Décembre et dans d'autres, de Aout à Octobre ...) ;
   — la disponibilité des données (les données sont la propriété du projet dans lequel elles ont été récoltées) ;

Notre diagnostic aurait été plus rapide et sûrement plus précis et complet n'euent été ces problèmes.

Perspectives

Au terme de nos travaux de recherche, les perspectives qui se dessinent sont liées d'une part aux problèmes auxquels nous avons été confrontés et d'autre part aux améliorations qui pourraient être apportées à nos résultats.

Ainsi, les problématiques liées aux données pourraient être résorbées suivant les recommandations ci-dessous :

1. la définition d'un protocole d'accord national auquel tous les projets pourraient adhérer ;
2. la programmation de campagnes de recueil communes à toutes les disciplines ;
3. la mise en place d'une base de données commune à tous les acteurs pour garantir l'archivage des données de campagne.

Mais pour combler les intervalles de valeurs manquantes, il serait approprié d'utiliser un algorithme pour imputer les données manquantes.

Dans le cadre de la modélisation, deux pistes peuvent être explorées :

— du point de vue "application", le modèle générique pourrait être implémenté pour prendre en charge des jeux de données liés à différentes maladies vectorielles et ainsi faire une classification des vecteurs par cible ou maladie, identifier les éventuels croisements dans le cycle des maladies vectorielles, etc. ;
— du point de vue "évolution", il pourrait être intéressant d'intégrer (i) la gestion du cycle de vie des vecteurs et des agents pathogène qui pourrait permettre d'évaluer leurs concordances (ii) la notion de "paramétrage de modèle" pour pouvoir l'adapter plus aisément à d'autres maladies non vectorielles, d'autres hiérarchies spatiales mais surtout pour la prise en compte de nouveaux paramètres environnementaux.

En ce qui concerne l'analyse décisionnelle, des environnements personnalisés pourraient être développés. Ainsi, des interfaces adaptées (structure et contenu) à chaque besoin (par utilisateur/par maladie) pourraient être générées. Pour ce faire, les algorithme de classification doivent être plus approfondis.

L'analyse spatio-temporelle quant à elle pourrait être valorisée à travers :

— l'adaptation de l'algorithme pour réduire les motifs en se concentrant sur les facteurs les plus pertinents (une approche récursive pourraient permettre d'éliminer les facteurs "non parlants" au cours de l'exécution de l'algorithme) ;
— le traitement automatisée des résultats de l'algorithme pour les intégrer à un outil de visualisation graphique.

ANNEXE A

Données manipulées

N°	Désignation	Période	Nb. Champs	Nb. Enregis.	Observations
Structure		Centre de Suivi Ecologique - CSE	**Discipline**		Climatologie
1	Données Station Niakha	22/07/2009 au 04/06/2011	8	15359	Donnée horaire
2	Données Station Niakha	29/07/2009 au 04/06/2011	17	632	Donnée journalière
3	Pluies journalières	01/06 au 30/10/2011	2	1989	153 lignes par site (13 sites)
4	Observatoire Barkédji	01/06 au 30/10/2011	2	153	1 feuille par site
5	Pluviométrie Barkédji	01/06 au 30/10 – 1998 à 2011	2	153	
Structure		Centre de Suivi Ecologique - CSE	**Discipline**		Occupation des sols
1	Données Station Niakha	/	6	80	1973, 1987, 2009
Structure		Institut Pasteur de Dakar	**Discipline**		Entomologie/Virologie
1	Données entomologiques	29/07 au 31/07/1991	11	52	
2	Données virologiques	06/04/1993	7	57	
3	Données entomologiques – Base IPD	15/07 au 29/12/2010	8	1322	

N°	Désignation	Période	Nb. Champs	Nb. Enregis.	Observations
Structure		Centre de Suivi Ecologique - CSE	**Discipline**		Hydrologie
1	Barkédji Hydro Mare 2007	01/08 au 30/10/2007	3	122	Limnimétrie de Niakha
2	Barkédji Hydro Mare 2008	01/07 au 30/11/2008	4	153	Limnimétrie de Niakha
		01/07 au 30/11/2008	2	153	Limnimétrie de Kangalédji
3	Barkédji Hydro Mare 2009	01/07 au 30/11/2009	2	153	Limnimétrie de Niakha
		01/07 au 30/11/2009	2 à 3	153	Limnimétrie de Kangalédji
		01/09 au 30/10/2009	2	60	Limnimétrie de Ngao
4	Barkédji Hydro Mare 2010	01/07 au 31/12/2010	2 à 3	183	Limnimétrie de Niakha
		01/07 au 31/12/2010	2 à 3	183	Limnimétrie de Kangalédji
		01/07 au 30/11/2010	2	153	Limnimétrie de Ngao
		01/07 au 31/12/2010	2	183	Limnimétrie de Furdu
5	Barkédji Hydro Mare 2011	01/06 au 30/06/2011	2	30	Limnimétrie de Niakha
		01/06 au 30/06/2011	2	30	Limnimétrie de Kangalédji
		01/06 au 10/07/2011	2	40	Limnimétrie de Furdu
Structure		Centre de Suivi Ecologique - CSE	**Discipline**		Qualité des eaux
1	QUAEMA Niakha juil 08	17/07 et 18/07/2008	9	14	
2	QUAEMA Niakha août 08	04/08 et 07/08/2008	9	18	
3	QUAEMA Ngao juil 08	17/07/2008	9	9	
4	QUAEMA Ngao août 08	06/08 et 09/08/2008	9	18	
5	QUAEMA Kangalédji juil 08	16/07/2008	9	9	
6	QUAEMA Kangalédji août 08	05/08 et 08/08/2008	9	18	

ANNEXE B

Motifs spatio-temporels : Définitions préliminaires

Definition B.0.0.1. ***Base de données spatio-temporelles.***
Une base de données spatio-temporelle est un ensemble de données structurées contenant des composantes géographiques (mares, campements, etc.), des composantes temporelles (température, vent, etc.) et des données décrivant les composantes géographiques à un temps donné. Plus formellement, une base de données spatio-temporelle est définie comme un triplet DB = (D_T,D_S,D_A) où D_T est la dimension temporelle, DS est la dimension spatiale et D_A=$\{ D_{A_1}, D_{A_2}, \ldots, D_{A_p} \}$ est un ensemble des dimensions d'analyse associées aux attributs. Le tableau suivant illustre la structure d'une base de données spatio-temporelles.

TABLE B.1 – Structure d'une base de données spatio-temporelle

Zone	*Date*	*Température*	*Précipitations*	*Vent*	*Moustiques*
Z_1	D_1	T_1	P_1	V_1	M_1
Z_1	D_2	T_2	P_2	V_2	M_2
Z_1	D_3	T_3	P_3	V_3	M_3
Z_2	D_4	T_4	P_4	V_4	M_4
Z_2	D_5	T_5	P_5	V_5	M_5
Z_2	D_6	T_6	P_6	V_6	M_6

La dimension temporelle est associée à un domaine de valeurs noté dom (D_T)= T_1, T_2, ..., T_t où $\forall$ i $\in$ [1..t], T_i est souvent appelé estampille temporelle et $T_1 < T_2 < \ldots < T_t$. Chaque dimension D_{A_i} ($\forall$ i $\in$ [1..p]) appartenant à la dimension d'analyse est associée à un domaine de valeurs noté par dom(A_i). Dans ce domaine, les valeurs peuvent être ordonnées ou non. La dimension spatiale est associée à un domaine de valeurs noté dom(D_S)=Z_1,Z_2,...,Z_l où $\forall$ i $\in$ [1..l], Z_i est une zone.

Definition B.0.0.2. ***Item et Itemset.***
Soit un item I, une valeur littérale pour la dimension D_{A_i}, I $\in$ dom(D_{A_i}). Un itemset, I_S = (I_1 I_2 ... I_n) avec n $\leqslant$ p est un ensemble non vide d' items tel que $\forall$ i,j $\in$ [1..n], $\forall$ k,k' $\in$ [1 ..p], $I_i \in$ dom(D_{A_k}), $I_j \in$ dom($D_{A'_k}$)) et k $\neq$ k'. Tous les items appartenant à un itemset sont associés à différentes dimensions d'analyse. Un itemset avec k items est appelé un k-itemset.

Definition B.0.0.3. ***Séquence d'itemsets.***
Une séquence d'itemsets S est une liste ordonnée, non vide, d'itemsets notée <s_1 s_2 ... s_p> où s_j est un itemset. Une n-séquence est une séquence composée de n items. Par exemple, considérons les événements produits dans la zone de T_1 à T_3 selon la séquence S = <(T_b P_m V_m)(T_m P_m V_b)(T_b P_m V_m M)> indiquée dans le Tableau B.1. Ceci signifie qu'hormis les événements T_b, P_m et V_m se sont produits ensemble, i.e. lors de la même transaction, les autres événements de la séquence se sont produits dans deux autres dates. Dans notre exemple, S est une 10-séquence. Une séquence <s_1 s_2 ... s_p> est une sous-séquence d'une autre séquence <s'_1 s'_2 ... s'_m> s'il existe des entiers <i_1 i_2 ... i_j ... i_p> tels que $s_1 \subseteq s'_{i_1}$,$s_2 \subseteq s'_{i_2}$,..., $s_p \subseteq s'_{i_p}$. Par exemple, la séquence S' = <(T_m P_m)(T_b P_m V_m M)> est une sous-séquence de S car (T_m P_m) $\subseteq$ (T_m P_m V_b) et (T_b P_m V_m M) $\subseteq$ (T_b P_m V_m M). Toutefois, S' = <(T_m P_h)(T_b P_m V_m)> n'est pas une sous-séquence de S car les deux itemsets de S' ne sont pas inclus dans deux itemsets de S. Tous les événements produits dans une même zone sont regroupés et triés par date. Ils constituent la séquence de données de la zone. Une zone supporte une séquence S si S est incluse dans la séquence de données de cette zone (S est une sous-séquence de la séquence de données).

Definition B.0.0.4. ***Itemset spatial - inclusion.***
La relation dans entre une zone Z et un itemsets IS est l'occurrence de l'itemset IS dans la zone Z au temps t dans la base de données DB. Plus formellement : (dans(IS,Z,t)=vrai si IS apparait dans DB pour la zone Z au temps t dans(IS,Z,t)=false sinon) Maintenant, nous définissons la notion de voisin entre zones. Plusieurs zones peuvent avoir une relation de voisinage. Deux zones sont voisines si : (voisin(Z_i,Z_j)=vrai si Z_i et Z_j sont voisin voisin(Z_i,Z_j)=false sinon)

Definition B.0.0.5. ***Item spatial - voisinage.***
Soient IS_i et IS_j deux itemsets, IS_i et IS_j sont spatialement proches si $\exists$ Z_i,$Z_j \subseteq$ dom(D_S) et $\exists$ t $\subseteq$ dom(D_T) tel que dans(IS_i,Z_i,t) $\wedge$ dans(IS_j,Z_j,t) $\wedge$ voisin(Z_i,Z_j) est vrai. Deux itemsets IS_i et IS_j qui sont spatialement proches, forment un itemset spatial noté I_ST= $IS_i \cdot IS_j$. Pour alléger les notations, nous introduisons un opérateur de groupement d'itemsets associé à l'opérateur $\cdot$ (voisin) et noté []. Le symbole θ représente l'absence d'itemsets dans une zone. La Figure XXX montre les trois types d'itemsets spatiaux que nous pouvons construire en utilisant ces opérateurs. Les lignes pointillées représentent le voisinage spatial.

Definition B.0.0.6. ***Association zone, itemset spatial et temps.***
Soit I_ST = $IS_i \cdot IS_i$ un itemset spatial, Z$\in$ dom(D_S) une zone et t $\in$ dom(D_T) une estampille temporelle, nous définissons la relation vérifier qui représente la présence de l'itemset spatial I_ST dans Z au temps t comme suit : (vérifier(I_ST,Z,t)= vrai si dans(IS_i,Z,t)= vrai et Z' $\in$ dom(D_S) tel que voisin(Z,Z')= vrai et dans(IS_j,Z',t)= vrai) vérifier(I_ST,Z,t)=false sinon)

Definition B.0.0.7. ***Inclusion d'itemsets spatiaux.***
Un itemset spatial I_ST = $IS_i \cdot IS_i$ est inclus, noté par $\subseteq$, dans autre itemset spatial I'_ST = $IS'_k \cdot IS'_l$, si et seulement si $IS_i \subseteq IS'_k$ et $IS_j \subseteq IS'_l$. Nous modélisons la notion d'évolution d'événements dans les zones en prenant en compte leur relation de voisinage via la notion de séquence spatiale.

Definition B.0.0.8. ***Séquence spatiale.***
Une séquence spatiale ou simplement S2 est une liste ordonnée d'itemsets spatiaux notée par s = < $I_(ST_1)$ $I_(ST_2)$... $I_(ST_m)$> où $I_(ST_i)$,$I_(ST_(i+1))$) satisfaisant la contrainte de séquentialité temporelle, i.e. i ∈ [1 ..m-1]. La figure suivante illustre la dynamique temporelle d'une séquence spatial.

Definition B.0.0.9. ***Inclusion de séquences spatiales (2S).***
Une 2S notée s = < I_{ST_1} I_{ST_2} ... I_{ST_m} > est plus spécifique qu'une autre 2S s'= < I'_{ST_1} I'_{ST_2} ... I'_{ST_n} notée par s $\preccurlyeq$ s', s'il existe $j_1 \leqslant \ldots \leqslant j_m$ tel que $I_{ST_1} \subseteq I'_{ST_{j_1}}$, $I_{ST_2} \subseteq I'_{ST_{j_2}}$,..., $I_{ST_m} \subseteq I'_{ST_{jm}}$.

Definition B.0.0.10. ***Motif spatio-séquentiel.***
Soient la 2S s et σ le support minimum spécifié par l'utilisateur, si $supp_rel$ (s,DB) $\geqslant \sigma$ alors s est une 2S fréquente appelée motif spatio-séquentiel où $supp_rel$ est le support relatif d'un séquence spatial . Concernant les mesures d'élagage, il est également défini un nouveau support absolu pour des séquences spatiales défini comme le nombre de zones contenant la séquence étudiée et satisfaisant les contraintes de proximité des itemsets spatiaux.

Definition B.0.0.11. ***Support absolu d'un motif spatio-séquentiel.***
Soit la 2S s = < I_{ST_1} I_{ST_2} ... I_{ST_m}>, le support absolu de s représenté par suppabs(s, DB) est défini comme le nombre de zones de la base de données DB qui vérifient s, autrement dit : $supp_abs$(s,BD)=|(Z∈ dom(D_S) tel que ∀ i∈[1..n], ∃ t_i ∈ dom (D_T) @ et vérifier(I_{ST_i},Z,t_i)=vrai)| De la même manière, nous définissons le support relatif dénoté par $supp_rel$(s,DB) ou simplement supp(s) pour une 2S s comme le ratio entre le support absolu et le nombre de zones total.

Definition B.0.0.12. ***Support relatif d'un motif spatio-séquentiel.***
Soient la 2S s = < I_{ST_1} I_{ST_2} ... I_{ST_m}>, suppabs(s, DB) le support absolu de s et |dom(D_S) | le nombre total de zones de la base de données spatio-temporelle DB, le support relatif de s est défini par :

$supp_rel$ (s,BD)=supp(s)=($supp_abs$ (s,BD))/dom(D_S)

La problématique d'extraction de motifs spatio-séquentiels à partir d'une base de données spatio-temporelles BD consiste à retrouver toutes les séquences spatiales dont le support relatif est supérieur à un seuil spécifié par l'utilisateur minSupp. Chacune de ces séquences fréquentes est communément appelée motif spatio-séquentiel.

ANNEXE C

Communications

C.1 Articles

— Bouba, F., Bah, A., Cambier, C., Ndiaye, S., Ndione, J. A., Teisseire, M. (2014). Decision Making Environment on Rift Valley Fever in Ferlo (Senegal).Acta biotheoretica, 62(3), 405-415.
— Bouba F., Bah A., Ndione J-A., Ndiaye S., Cambier C., 2012. Modèle de Données multidimensionnel sur les interactions environnement-santé, CNRIA'12 (Colloque National sur la Recherche en Informatique et ses Applications- Edition 2012), Volume 1 - 2012, pages 1 à 9 - ARIMA
— Bouba F., Bah A., Ndione J-A., Ndiaye S., Kebe C.M.F., 2011. Intelligence Information System on the health-environment interactions : Case of Rift Valley Fever (RVF) in Ferlo (Senegal), MSE 2011 - Advances in Artificial Intelligence, ISSN 2160-147X

C.2 Communications orales

— PPZS – 7ème Conseil Scientifique – Avril 2014 : Présentation d'un extrait du modèle multidimensionnel et des résultats de reporting
— SFBT – Mai 2013 : Présentation orale « Vers un environnement décisionnel sur la FVR au Ferlo (Sénégal) »
— PPZS – 6ème Conseil Scientifique – Avril 2012 : Présentation d'un extrait du modèle multidimensionnel et des résultats de reporting
— Atelier QWeCI 2011 – 14 au 16 Novembre 2011 : Présentation du projet de thèse et de la plateforme de saisie des données

C.3 Posters

— 11-12 Mai 2010, Premières Journées Régionales de l'IRD, Dakar, Salle de conférence de l'Ucad II « Contribution à l'étude d'un système d'aide à la décision intégrant des données complexes : Modèles informatiques intégrés des interactions Climat-Santé dans le Ferlo au Sénégal »
— 15 Décembre 2011, Conference sur les Systèmes Complexes, Paris, Salle de conférence de l'IRD Bondy « Système d'Information décisionnel sur les interactions environnement-santé : Cas de la Fièvre de la Vallée du Rift au Sénégal »

Bibliographie

AGRAWAL R., IMIELINSKI T. et SWAMI A. (1993). Mining association rules betweeen sets of items in large databases. *In Proceedings of the Eleventh International Conference*, pages 207–216.

AGRAWAL R. et SRIKANT R. (1995). Mining sequential patterns. *In Data Engineering, Proceedings of the Eleventh International Conference*, pages 3–14.

ALLEGRI B. (2004). L'informatique décisionnelle dans le milieu de la santé. ITBM-RBM News, volume 5, pages 10–13.

BALKHY H. H. et MEMISH Z. A. (2003). Rift valley fever : an uninvited zoonosis in the arabian peninsula. International journal of antimicrobial agents, Elsevier, volume 21, pages 153–157.

BELLATRECHE L. (2003). Techniques doptimisation des requêtes dans les datawarehouses. *In 6th International Symposium on Programming and Systems*, pages 81–98.

BERKHIN P. (2006). A survey of clustering data mining techniques. Grouping multidimensional data, Springer, pages 25–71.

BERNOULLI D. (1760). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. Histoire de l'Acad. Roy. Sci.(Paris) avec Mém. des Mathématiques et Physique, pages 1–45.

BOLY A. (2006). *Fonctions d'oubli et résumés dans les entrepôts de données*. Thèse de doctorat, Ecole Nationale Supérieure des Télécommunications.

BONCZEK R. H., HOLSAPPLE C. W. et WHINSTON A. B. (1981). A generalized decision support system using predicate calculus and network data base management. Operations Research, INFORMS, volume 29, pages 263–281.

BRAHMI N., HATIRA A. et RABIA M.-C. (2010). Contribution de la télédétection et des systèmes d'information géographique à la prise en compte du risque de prolifération des aedes dans les zones humides de bizerte (nord de la tunisie). Physio-Géo. Géographie, physique, et environnement, Martin, Claude, volume 4, pages 151–168.

BRUGERE-PICOUX J. et KODJO A. (2007). Actualités sur les zoonoses émergentes et résurgentes. Bull Acad Vét France, volume 160, pages 279–288.

BRULEY M. (2010). Propos sur le développement d'un système d'information décisionnel. URL : `http://www.decideo.fr/bruley/`

Propos-sur-le-developpement-d-un-systeme-d-information-decisionnel_a24.html/. Consulté le 02/09/2013.

CAO L., XIAOHUI S., YUELING Z. et CHEN G. (2011). The application of the spatio-temporal data mining algorithm in maize yield prediction. Mathematical and Computer Modelling, volume 58, pages 507–513.

CÊTRE-SOSSAH C. et ALBINA E. (2009). Fievre de la vallee du rift : aspects veterinaires et impacts sur la sante humaine. Medecine Tropicale, volume 69, page 358.

CHANDA E., MUKONKA V. M., MTHEMBU D., KAMULIWO M., COETZER S. et SHINONDO C. J. (2012). Using a geographical-information-system-based decision support to enhance malaria vector control in zambia. Journal of tropical medicine, Hindawi Publishing Corporation, volume 2012. 10 pages.

CISSÉ A., BAH A., DROGOUL A., CISSÉ A. T., NDIONE J. A., KÉBÉ C. M. et TAILLANDIER P. (2012). Un modèle à base d'agents sur la transmission et la diffusion de la fièvre de la vallée du rift à barkédji (ferlo, sénégal). Stud. Inform. Univ., volume 10, pages 77–97.

CODD E. F. (1970). A relational model of data for large shared data banks. Communications of the ACM, ACM, volume 13, pages 377–387.

CODD E. F., CODD S. B. et SALLEY C. T. (1993). Providing olap (on-line analytical processing) to user-analysts : An it mandate. Codd and Date, Codd & Date, Inc., volume 32.

COLEMAN M., SHARP B., SEOCHARAN I. et HEMINGWAY J. (2006). Developing an evidence-based decision support system for rational insecticide choice in the control of african malaria vectors. Journal of medical entomology, BioOne, volume 43, pages 663–668.

COLY A. (1996). *Le système fluviolacustre du Guiers : étude hydrologique et gestion quantitative intégrée*. Thèse de doctorat, Université Cheikh Anta Diop de Dakar.

COMPIETA P., DI MARTINO S., BERTOLOTTO M., FERRUCCI F. et KECHADI T. (2007). Exploratory spatio-temporal data mining and visualization. Journal of Visual Languages & Computing, Elsevier, volume 18, pages 255–279.

DAMPFHOFFER M. (2009). Elaboration d'un plan de surveillance et d'urgence fièvre de la vallée du rift pour la france métropolitaine. Mémoire de D.E.A., Ecole des Hautes Etudes en Santé Publique.

DAUBNEY R., HUDSON J. et GARNHAM P. (1931). Enzootic hepatitis or rift valley fever. an undescribed virus disease of sheep cattle and man from east africa. The Journal of Pathology and Bacteriology, Wiley Online Library, volume 34, pages 545–579.

DAVIES G. G., LINTHICUM K. J. et JAMES A. D. (1985). Rainfall and epizootic rift valley fever. Bulletin of the World Health Organisation, volume 5, pages 941–943.

de VINCI L. (2008). Recommandations de déontologie et bonnes pratiques en épidémiologie (version france-2007). Revue d'Épidémiologie et de Santé Publique, volume 56, pages 121–148.

DIAGNE F. (1992). *Etude de la Fièvre de la Vallée du Rift chez les ruminants domestiques au Sénégal : enquêtes sérologiques dans la vallée du Fleuve, le Ferlo et la Casamance*. Thèse de doctorat, FMPOS/UCAD.

DIALLO M., LOCHOUARN L., BA K., SALL A., MONDO M., GIRAULT L. et MATHIOT C. (1995). Dynamique comparée des populations de culicidae à kédougou (zone soudano-guinéenne) et à barkédji (zone de savane sahélienne) : conséquences dans la transmission des arbovirus. Mémoire de D.E.A., UCAD.

DIALLO M., LOCHOUARN L., BA K., SALL A. A., MONDO M., GIRAULT L. et MATHIOT C. (2000). First isolation of the rift valley fever virus from culex poicilipes (diptera : Culicidae) in nature. The American journal of tropical medicine and hygiene, ASTMH, volume 62, pages 702–704.

EISA M., Kheir el SID E., SHOMEIN A. et MEEGAN J. (2004). An outbreak of rift valley fever in the sudan - 1976. Trans R Soc Trop Med Hyg, volume 74, pages 417–426.

EISEN L. et EISEN R. J. (2011). Using geographic information systems and decision support systems for the prediction, prevention, and control of vector-borne diseases. Annual review of entomology, Annual Reviews, volume 56, pages 41–61.

ELFAZZIKI A., NEJEOUI A. et SADGAL M. (2006). Une approche multi-agents pour la modélisation et l'optimisation des systèmes de gestion de transport maritime. Revue d'Information Scientifique et Technique - RIST, volume 15, pages 173–200.

FABRÈGUE M., BRAUD A., BRINGAY S., LE BER F., TEISSEIRE M. *et al.* (2012). Extraction de motifs spatio-temporels à différentes échelles avec gestion de relations spatiales qualitatives. *In 30ème édition, Inforsid 2012*, pages 123–138.

FATHIMA A. S., MANIMEGALAI D. et HUNDEWALE N. (2011). A review of data mining classification techniques applied for diagnosis and prognosis of the arbovirus-dengue. IJCSI International Journal of Computer Science Issues, Citeseer, volume 8, pages 322–328.

FATHIMA A. S. et MANIMEGLAI D. (2012). Predictive analysis for the arbovirus-dengue using svm classification. International Journal of Engineering and Technology, volume 2, pages 521–527.

FENZHEN S., CHENGHU Z., LYNE V., YUNYAN D. et WENZHONG S. (2004). A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution. Ecological modelling, Elsevier, volume 174, pages 421–431.

FONTENILLE D., TRAORE-LAMIZANA M., DIALLO J., THONNON J., DIGOUTTE J. et ZELLER H. (1998). Nouveaux vecteurs de la fièvre de la vallée du rift en afrique de l'ouest. Emerging Infectious Diseases, volume 4, pages 289–293.

FONTENILLE D., TRAORE-LAMIZANA M., MONDO M., DIALLO M. et DIGOUTTE J. (1995). Short report : Rift valley fever in western africa : isolations from aedes mosquitoes during an interepizootic period. Am. J. Trop. Med. Hyg, volume 52, pages 403–404.

GARDARIN G. (2003). Bases de données. Editions Eyrolles. 832 pages.

GERDES G. *et al.* (2004). Rift valley fever. Revue scientifique et technique-Office International des Epizooties, Paris : L'Office, 1982-, volume 23, pages 613–624.

GIRDARY L. et GRANDCHAMP E. Contributions de la télédétection et des systèmes d'information géographiques pour l'étude de la transmission de la dengue. URL : `http://www.selperbrasil.org.br/selper2012/PDF/FP_SELPER-161.pdf`. Consulté le 10/02/2015.

GIUDICI P. (2003). *Applied Data Mining Statistical Methods for Business and Industry*. Wiley And Sons. 376 pages.

GOLFARELLI M., MAIO D. et RIZZI S. (1998). The dimensional fact model : a conceptual model for data warehouses. International Journal of Cooperative Information Systems, World Scientific, volume 7, pages 215–247.

GORRY G. A. et MORTON M. S. S. (1971). *A framework for management information systems*, volume 13. Massachusetts Institute of Technology.

GOUARNÉ J.-M. (1998). *Le projet décisionnel : enjeux, modèles et architectures du Data Warehouse*. Eyrolles.

HAMILTON H. J., GENG L., FINDLATER L. et RANDALL D. J. (2006). Efficient spatio-temporal data mining with genspace graphs. Journal of Applied Logic, Elsevier, volume 4, pages 192–214.

HECHMATI G. (2004). *Epidémies de grippe : système d'information pour la prise de décision en santé publique*. Thèse de doctorat, Univ. Genève.

HUANG Y., ZHANG L. et ZHANG P. (2008). A framework for mining sequential patterns from spatio-temporal event data sets. Knowledge and Data Engineering, IEEE Transactions on, IEEE, volume 20, pages 433–448.

INMON W. H. (2005). *Building the data warehouse*. John wiley & sons, quatrième édition. 576 pages.

JAIN A. K. (2010). Data clustering : 50 years beyond k-means. Pattern Recognition Letters, Elsevier, volume 31, pages 651–666.

JOUAN A., ADAM F., RIOU O. et AL (1990). Evolution des indicateurs de santé dans la région du trarza lors de l'épidémie de la fièvre de la vallée du rift en 1987. Bull Soc Pathol Exot, volume 83, pages 621–628.

KANUNGO T., MOUNT D. M., NETANYAHU N. S., PIATKO C. D., SILVERMAN R. et WU A. Y. (2002). An efficient k-means clustering algorithm : Analysis and implementation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, IEEE, volume 24, pages 881–892.

KEEN P. G. et MORTON S. (1978). Decision support systems : An organizational perspective. volume 35.

KERMACK M. et MCKENDRICK A. (1927). Contributions to the mathematical theory of epidemics. part i. *In Proc. R. Soc. A*, volume 115, pages 700–721.

KERMACK W. O. et MCKENDRICK A. G. (1932). Contributions to the mathematical theory of epidemics. ii. the problem of endemicity. Proceedings of the Royal society of London. Series A, The Royal Society, volume 138, pages 55–83.

KERMACK W. O. et MCKENDRICK A. G. (1933). Contributions to the mathematical theory of epidemics. iii. further studies of the problem of endemicity. Proceedings of the Royal Society of London. Series A, The Royal Society, volume 141, pages 94–122.

KIMBALL R. (1996). The data warehouse toolkit : practical techniques for building dimensional data warehouse. John Willey & Sons, volume 248, page 4.

KUMAR D. V. R. S., SRIRAM K., RAO K. M. et MURTY U. S. (2005). Management of filariasis using prediction rules derived from data mining. Bioinformation, Biomedical Informatics Publishing Group, volume 1, pages 8–11.

LEFÈVRE P. (2003). Fièvre de la vallée du rift. Principales maladies infectieuses et parasitaires du bétail, Europe et régions chaudes I : Généralités Maladies Virales, Lavoisier, pages 643–657.

LINTHICUM K., ASSAF A., COMPTON J., KELLEY P., MYERS M. et PETER C. (1999). Climate and satellite indicators to forecast rift valley fever epidemics in kenya. Science, volume 285, pages 397–400.

MAMADOU B., SOUSSOU S., FADEL K. C. M. et JACQUES-ANDRE N. (2010). Modélisation du fonctionnement hydrologique d'un bassin endoréique pour une application à l'étude de la fièvre de la vallée du rift (fvr). IAHS-AISH publication, International Association of Hydrological Sciences, pages 305–313.

MARAKAS G. M. (2003). *Decision support systems in the 21st century*, volume 134. Prentice Hall Upper Saddle River, NJ.

MEEGAN J., HOOGSTRAAL H. et MOUSSA M. (1979). An epizootic of rift valley fever in egypt in 1977. Vet Rec, volume 105, pages 124–129.

MEEGAN J. M. et BAILEY C. H. (1988). Climate and satellite indicators to forecast rift valley fever epidemics in kenya. Monath, T. P., editor. The arboviruses : epidemiology and ecology, volume IV, page 61–76.

MELIKER J. R. et SLOAN C. D. (2011). Spatio-temporal epidemiology : principles and opportunities. Spatial and spatio-temporal epidemiology, Elsevier, volume 2, pages 1–9.

MORTAZAVI-ASL B., PINTO H. et DAYAL U. (2000). Prefixspan : Mining sequential patterns efficiently by prefix-projected pattern growth. *In 17th International Conference on Data Engineering (ICDE)*, IEEE Computer Society, pages 215–224.

MOUTAILLER S., KRIDA G., SCHAFFNER F., VAZEILLE M. et FAILLOUX A.-B. (2008). Potential vectors of rift valley fever virus in the mediterranean region. Vector-borne and zoonotic Diseases, Mary Ann Liebert, Inc. 2 Madison Avenue Larchmont, NY 10538 USA, volume 8, pages 749–754.

MURTY U. S., SRINIVASA RAO M. et MISRA S. (2008). Prioritization of malaria endemic zones using self-organizing maps in the manipur state of india. Informatics for Health and Social Care, Informa UK Ltd UK, volume 33, pages 170–178.

NAVETIER L. (2005). *Que peut on espérer des systèmes décisionnels ?* Valutis.

NDIONE J., BESANCENOT J., LACAUX J. et SABATIER P. (2003). Environnement et épidémiologie de la fièvre de la vallée du rift (fvr) dans le bassin inférieur du fleuve sénégal. Environnement, Risques et Santé 3, volume 2, page 176–182.

NDIONE J., DIOP M., MONDET B., DIOP C. et DACOSTA H. (2006). Emergence de la fièvre de la vallée du rift au sénégal et variabilité intra-saisonnière de la pluviométrie. pages 596–589.

NDIONE J.-A., LACAUX J.-P., TOURRE Y., VIGNOLLES C., FONTANAZ D. et LAFAYE M. (2009). Mares temporaires et risques sanitaires au ferlo : contribution de la télédétection pour l'étude de la fièvre de la vallée du rift entre août 2003 et janvier 2004. Science et changements planétaires/Sécheresse, volume 20, pages 153–160.

OBENSHAIN M. K. (2004). Application of data mining techniques to healthcare data. Infection Control and Hospital Epidemiology, JSTOR, volume 25, pages 690–695.

OECD (2005). *Technologies De La Santé et Prise De Décision*. Éditions OCDE. 176 pages.

OIE (2010). Maladies de la liste de l'oie. URL : `http://web.oie.int/fr/maladies/fr_classification2010.htm`. Consulté le 04/01/2012.

PEPIN M. (2011). Fièvre de la vallée du rift. Medecine et maladies infectieuses, Elsevier, volume 41, pages 322–329.

PHYU T. N. (2009). Survey of classification techniques in data mining. *In Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, pages 18–20.

PREHAUD C. et BOULOY M. (1997). La fièvre de la vallée du rift-un modéle d'étude des fièvres hémorragiques virales. *In Annales de l'Institut Pasteur/Actualites*, Elsevier, volume 8, pages 233–244.

RAKOTOMALALA R. (2005). Arbres de décision. Revue Modulad, volume 33, pages 163–187.

RAKOTOMANANA F., JEANNE I., DUCHEMIN J., PIETRA V., RAHARIMALALA L., TOMBO M. et ARIEY F. (2001). Approche géographique dans la lutte contre le paludisme dans la région des hautes terres centrales à madagascar. Archives de l'Institut Pasteur de Madagascar, Institut Pasteur de Madagascar, volume 67, pages 27–30.

RAVAT F., TESTE O. et ZURFLUH G. (2001). Modélisation multidimensionnelle des systèmes décisionnels. *In EGC*, pages 201–212.

ROBERTSON C., NELSON T. A., MACNAB Y. C. et LAWSON A. B. (2010). Review of methods for space-time disease surveillance. Spatial and Spatio-temporal Epidemiology, Elsevier, volume 1, pages 105–116.

ROCQUE S. D. L. (2004). Des maladies qui entrent en scène : Homme, plante, animal : tous exposés. *In Journée du développement durable : le changement climatique, risques ou opportunités ?*

ROGERS D. et PACKER M. (1993). Vector-borne diseases, models, and global change. The Lancet, Elsevier, volume 342, pages 1282–1284.

ROSS R. (1911). The prevention of malaria, with addendum on the theory of happenings. Murray, London.

SABATIER P., BICOUT D. J., DURAND B. et DUBOIS M. A. (2005). Le recours à la modélisation en épidémiologie animale. Epidemiol Santé Anim, volume 47, pages 15–33.

SALAS H. A., AZÉ J., BRINGAY S., CERNESSON F., FLOUVAT F., SELMAOUI-FOLCHER N. et TEISSEIRE M. (2011). Recherche de séquences spatio-temporelles peu contredites

dans des données hydrologiques. *In Revue des Nouvelles Technologies de l'Information , RNTI-E-22*, pages 165–188.

SALAS H. A., BRINGAY S., FLOUVAT F., SELMAOUI-FOLCHER N. et TEISSEIRE M. (2012). The pattern next door : Towards spatio-sequential pattern discovery. *In Advances in Knowledge Discovery and Data Mining*, Springer, pages 157–168.

SAUL L.-F., DARWIN E.-Q., ARTURO F.-A. J., ALBA L.-P. M., JULIAN G.-R., SALVADOR G.-C., VICTOR L.-Z., ROSARIO N.-V., ILDEFONSO F.-S., JOAQUIN C.-M. *et al.* (2008). Use of google earthtm to strengthen public health capacity and facilitate management of vector-borne diseases in resource-poor environments. Bulletin of the World Health Organization, SciELO Public Health, volume 86, pages 718–725.

SMITHA T. et SUNDARAM V. (2012). Classification rules by decision tree for disease prediction. International Journal of Computer Applications, volume 43.

SMITHBURN K., HADDOW A. et GILLETT J. (1948). Rift valley fever. isolation of the virus from wild mosquitoes. British Journal of Experimental Pathology, Blackwell Publishing, volume 29, page 107.

SOTI V. (2011). *Caractérisation des zones et périodes à risque de la Fièvre de la Vallée du Rift au Sénégal par télédétection et modélisation éco-épidémiologique*. Thèse de doctorat, AgroParisTech.

SPIEZ L. (2006). Fièvre de la vallée du rift. URL : `http://www.labor-spiez.ch/fr/dok/fa/pdf_f/Rift-Valley_Fieber_f.pdf`. Consulté le 07/09/2013.

SPRAGUE JR R. H. et CARLSON E. D. (1982). *Building effective decision support systems*. Prentice Hall Professional Technical Reference. 329 pages.

STEVENS K. B. et PFEIFFER D. U. (2011). Spatial modelling of disease using data-and knowledge-driven approaches. Spatial and spatio-temporal epidemiology, Elsevier, volume 2, pages 125–133.

TESTE O. (2000). *Modélisation et manipulation d'entrepôts de données complexes et historisées*. Thèse de doctorat, Université Paul Sabatier de Toulouse (France).

TRAN A., BITEAU-COROLLER F., GUIS H. et ROGER F. (2005). Modélisation des maladies vectorielles. Epidémiol et santé anim, volume 47, pages 35–51.

TSOUKATOS I. et GUNOPULOS D. (2001). Efficient mining of spatiotemporal patterns. Springer, volume 2121, pages 425–442.

VALLERON A. (2000). The roles for modeling in epidemiology. Comptes rendus de l'Academie des sciences. Serie III, Sciences de la vie, volume 323, pages 429–433.

WANG J., HSU W. et LEE M. L. (2005). Mining generalized spatio-temporal patterns. *In Database Systems for Advanced Applications*, Springer, pages 649–661.

WEHRLE P. (2009). *Modèle multidimensionnel et OLAP sur architecture de grille*. Thèse de doctorat, Institut national des sciences appliquées de Lyon.

WOODS C., KARPATI A., GREIN T. et AL (2002). An outbreak of rift valley fever in northern kenya, 1997-1998. Emerg Infect Dis, volume 8, pages 138–144.