

FSEA/Année 2020

THÈSE

En vue de l'obtention du diplôme
de Doctorat en sciences

Spécialité : INFORMATIQUE

Présentée par : **Farah Amina ZEMRI**

Vers une Nouvelle Formulation SOLAP pour Supporter les Objets Spatiaux Complexes dans les Entrepôts de Données

Soutenue publiquement le :

devant le jury composé de :

Yahya LEBBAH	Professeur	Univ. Oran1, Ahmed Ben Bella	Président
Lougmiri ZEKRI	MCA	Univ. Oran1, Ahmed Ben Bella	Examineur
Abdelmalek Amine	Professeur	Univ. Tahar Moulay, Saida	Examineur
Zakaria ELBERRICHI	Professeur	Univ. Djilali Liabess, Sidi Belabbes	Examineur
Nacira GHOUALMI	Professeur	Univ. Baji Mokhtar, Annaba	Examineur
Djamila HAMDADOU	Professeur	Univ. Oran 1, Ahmed Ben Bella	Directrice de Thèse
Karine Zeitouni	Professeur	Univ. de Parsi Saclay (ex UVSQ) (France)	Invité

Résumé : L'objectif visé dans cette thèse est de proposer une nouvelle formulation SOLAP (Spatial On Line Analysis Processing) intégrant les techniques de la FDS (Fouille de Données Spatiales) et les MADMC (Méthodes d'Aide à la Décision Multicritères). A travers la présente étude, nous évaluons d'une part, les avantages du SOLAP dans notre système pour la surveillance épidémiologique « EPISOLAP », et d'autre part nous exploitons les avantages des techniques de la FDS et des MADMC dans l'évaluation du risque sanitaire. Ce risque encouru est mesuré par la présence simultanée de l'aléa et de la vulnérabilité de la population menacée. L'outil EPISOLAP quant à lui, servira pour la détection et la localisation des foyers d'épidémies. Cette proposition d'intégration de ces outils d'aide à la décision puissants sera réalisée dans le cadre d'une seule plate forme décisionnelle intégrée d'une manière cohérente et transparente. Notre préoccupation majeure est de fournir plus d'explications sur les facteurs réels et responsables de la propagation des épidémies ainsi que de son émergence ou réémergence. Enfin, nous pensons que l'exploitation des capacités de la FDS qu'offrent les méthodes et techniques d'extraction de connaissances à partir des données nous permettra d'avoir dans une perspective à court terme une meilleure précision de la prédiction et ainsi des résultats plus convaincants.

Mots clé : SOLAP (Spatial On Line Analysis Processing), Fouille de données spatiales, système d'aide à la décision spatiale, Aide à la décision multicritères, Surveillance épidémiologique.

Abstract: The objective of this thesis is to propose a new SOLAP formulation (Spatial On Line Analysis Processing) integrating the techniques of SDM (Spatial Data Mining) and MCDSM (Multicriterion Decision Support Methods). Through this study, we evaluate the benefits of SOLAP in our epidemiological surveillance system "EPISOLAP", and we are exploiting the benefits of SDM and MCDSM techniques in the evaluation of health risk. This risk is measured by the simultaneous presence of the hazard and the vulnerability of the threatened population. As for the EPISOLAP tool, it will be used for the detection and localization of epidemic outbreaks. This proposal to integrate these powerful decision support tools will be made in a single integrated decision-making platform in a consistent and transparent way. Our main concern is to provide more explanations on the real factors responsible for the spread of epidemics and its emergence or re-emergence. Finally, we believe that exploiting the capabilities of the SDM offered by knowledge extraction methods and techniques from the data will allow us to have in a short-term perspective a better precision of the prediction and thus of the results. more convincing.

Key words: SOLAP (Spatial On Line Analysis Processing), Multicriteria decision support, Spatial data mining, Spatial decision support system, Epidemiological surveillance.

ملخص: الهدف من هذه الأطروحة هو اقتراح صيغة SOLAP جديدة (معالجة تحليل المكاني) تدمج تقنيات SDM (بحث البيانات المكانية) و MADSM (أساليب دعم اتخاذ القرار متعدد المعايير). من خلال هذه الدراسة، نقوم بتقييم فوائد SOLAP في نظام المراقبة الوبائية «EPISOLAP»، و نستغل فوائد تقنيات FDS و MADMC في تقييم المخاطر الصحية. يتم قياس هذا الخطر من خلال الوجود المتزامن للخطر وعرضة السكان المهددين. بالنسبة لأداة EPISOLAP، سيتم استخدامها للكشف عن تفشي الوباء وتوطينه. سيتم تقديم هذا الاقتراح لدمج أدوات دعم القرار القوية هذه في منصة واحدة متكاملة لاتخاذ القرارات بطريقة متسقة وشفافة. شاغلنا الرئيسي هو تقديم المزيد من التوضيحات حول العوامل الحقيقية المسؤولة عن انتشار الأوبئة وظهورها أو ظهورها مرة أخرى. أخيرًا، نعتقد أن استغلال إمكانات SDM لطرق وتقنيات استخراج المعرفة من البيانات سيتيح لنا الحصول على دقة أفضل للتنبؤ ومن ثم النتائج من منظور قصير الأجل أكثر إقناعاً.

الكلمات المفتاحية: SOLAP، دعم اتخاذ القرار متعدد المعايير، استخراج البيانات المكانية، نظام دعم القرار المكاني، المراقبة الوبائية.

Remerciements

En tout premier lieu, je remercie le bon DIEU le tout puissant de m'avoir guidée vers le chemin de la réussite de ce modeste travail de thèse.

Mes remerciements les plus sincères s'adressent à madame **HAMDADOU Djamila**, ma directrice de thèse, Professeur à l'université Oran 1 Ahmed BENBELLA pour avoir cru en moi en me proposant ce sujet de thèse, d'avoir dirigé mes travaux avec beaucoup de patience ainsi que pour ces orientations, ses conseils précieux et encouragements durant le cheminement de ce travail de thèse.

Je veux exprimer ma sincère reconnaissance et ma gratitude à madame **ZEITOUNI Karine** Professeur et chef d'équipe ADAM du laboratoire DAVID, pour son accueil, ses orientations, et son aide précieuse durant mon stage à l'université Paris-Saclay (ex UVSQ).

Mon profond respect et mes remerciements vont à monsieur **LEBBAH Yahya**, Professeur à l'université Oran 1 Ahmed BENBELLA, pour l'honneur qu'il me fait de présider ce jury.

Mes vifs remerciements s'adressent, également à Monsieur **ZEKRI Logmiri**, Maître de conférences à l'université Oran 1 Ahmed BENBELLA, et ma reconnaissance envers la précieuse formation qu'il nous a donnée durant notre cursus et d'avoir accepté d'examiner mon travail de thèse.

J'adresse mes sincères remerciements à madame **GHOUALMI Nacira** Professeur à l'université Badji Mokhtar, Annaba, à Monsieur **AMINE Abdelmalek** Professeur à l'université Tahar Moulay, Saida et Monsieur **ELBERRICHI Zakaria** Professeur à l'université Djilali Liabbes, Sidi Belabbes, pour avoir accepté d'examiner et de juger mon travail de thèse.

Dédicaces

A la mémoire de mon père, A ma mère, mon frère et ma sœur

A mon neveu adoré Ziad Fazil

A mes aimables grands parents

A mes tantes, mes oncles, mes cousins et cousines

A toutes mes amies sœurs, Chahéra, Hasna, Nesrine, Farah,

A toute personne qui a contribué de loin ou du pré dans l'aboutissement de ce travail

Farah Zemri

« The greatest value of a picture is when it forces us to notice
what we never expected to see »

(Tukey, 1977).

Table des matières

Introduction générale.....	1
1 Contexte de l'étude.....	1
1 Problématique	2
2 Objectifs de l'étude et méthodologie adoptée.....	2
3 Notre contribution	3
5 Acheminement de la thèse.....	5
6 Organisation de la thèse.....	5
PREMIERE PARTIE : Synthèse de l'état de l'art	8
Chapitre I : Les Systèmes d'Aide à la décision Spatio-temporelles (STDSS)	7
1 Introduction.....	10
2 Les systèmes d'Aide à la Décision	10
2.1 Définition	10
2.2 Rôle d'un SAD.....	10
2.2 Les modèles des SAD	11
3 Les systèmes Interactifs d'Aide à la Décision (SIAD)	13
3.1 Définition	13
3.2 Caractéristiques des SIAD	13
3.3 Composants d'un SIAD	14
3.4 Applications d'aide à la décision	14
3.5 Les problématiques décisionnelles	14
4 Les systèmes d'Aide à la Décision spatiale (SDSS)	15
5 Les systèmes d'Aide à la Décision Spatiotemporelle (STDSS).....	16
5.1 Définition et exemples d'application	16
5.2 Les objets mobiles	17
5.3 Les séquences.....	17
5.4 Méthodes d'exploration des séries temporelles.....	18
6 Le système EPISOLAP	20
7 Travaux connexes.....	22
8 Conclusion	23
Chapitre II : L'aide à la décision multicritère	25
1 Introduction.....	26
2 Aide à la décision	26
3 Les acteurs d'aide à la décision	26
4 Décision et processus de décision	27

5 Concepts de base de l'aide à la décision multicritères	28
5.1 Le Processus de la décision	28
5.2 Les Eléments du processus décisionnel	30
5.3 Le Du monocritère au multicritères	33
5.4 Typologie des problématiques décisionnelle	34
5.5 Les problématiques de l'aide à la décision multicritères	34
6 Procédure d'agrégation.....	36
7 Les méthodes d'analyse multicritères	37
8 Travaux connexes	41
11Exemple d'application.....	41
12Conclusion	42
Chapitre III : La fouille de données spatiales	43
1 Introduction.....	44
2 Méthodologie de fouille de données spatiales	44
3 Particularité des données à référence spatiale	46
4 Méthodes utilisées dans la fouille de données spatiales	47
4.1 Phase descriptive	47
4.2 Phase exploratoire	48
4.3 Phase explicative (décisionnelle).....	55
5 Classification des méthodes de fouille de données spatiales	59
5.1 Méthodes monothématiques.....	59
5.2 Méthodes multithématiques.....	60
6 Travaux connexes	60
6.1 Approches basées sur le prétraitement de données	60
6.2 Approches bas »e sur le traitement dynamique de l'information spatiale	61
7 Conclusion	61
DEUXIEME PARTIE : Contributions	62
Chapitre VI : Le système SOLAP-MINING.....	64
1 Motivation	65
2 Introduction.....	65
3 Objectif visé	66
4 Approche décisionnelle proposée	66
5 Description du module SOLAM.....	68
6 Description du processus de la généralisation spatiale	69
6.1 La méthode d'Induction Orientée Attribut (AOI)	72
6.2 Principe de l'algorithme de comparaison	72
7 Expérimentation du système EPISOLAM.....	74
7.1 Délimitation de la zone d'étude	74
7.2 Formulation du problème abordé	74
7.3 Expérimentation	75
7.4 Arbre de généralisation.....	83

7.5 Discussion.....	83
8 Travaux connexes	84
9 Conclusion	85
Chapitre V : Le système SDSS-MCA-DM.....	86
1 Introduction.....	87
2 Motivation	87
3 Objectif visé	88
4 Outils d'investigation	89
5 Description du modèle décisionnel proposé.....	95
5.1 Le sous-système d'analyse multicritères	95
5.2 Le sous-système fouille de données	96
5.3 Le sous-système de visualisation.....	96
6 Processus décisionnel adopté par SDSS-MCA-DM	96
6.1 Phase de structuration.....	97
6.2 Phase d'exploitation	98
6.3 Phase d'exécution.....	98
7 Modélisation UML du système décisionnel proposé.....	98
8 Expérimentation du système SDSS_MCA_DM	101
8.1 Formulation du problème décisionnel.....	101
8.2 Processus d'exploration de données	108
9 Travaux connexes	110
9.1 Travaux connexes sur la surveillance épidémiologique.....	110
9.2 Travaux connexes sur les méthodes d'analyse multicritères.....	111
9.3 Travaux connexes sur SOLAP.....	112
9.4 Travaux connexes sur le datamining.....	112
9.5 Synthèse	113
10 Conclusion	114
Chapitre VI : Proposition de l'indice de LISTA.....	107
1 Introduction.....	117
2 Motivation	117
3 Objectif visé	118
4 Mesure d'autocorrélation spatiotemporelle.....	119
4.1 Indice d'autocorrélation temporelle (ITA)	119
4.2 Indice d'autocorrélation spatiale (LISA)	120
4.3 Indice croisé de corrélation temporelle (CITC).....	121
4.4 Indice d'autocorrélation spatiotemporelle (LISTA)	121
5 Matrice des indices d'autocorrélation spatiotemporelle.....	122
6 Expérimentation de l'indice de LISTA.....	123
6.1 Analyse de l'autocorrélation spatiale par indice de LISA	124
6.2 Analyse de l'autocorrélation spatiotemporelle par indice de LISTA.....	125
6.3 Calcul de la matrice des indices croisés de corrélation temporelle.....	126
6.4 Calcul de la matrice des indices d'autocorrélation spatiotemporelle	129

Table de matières

6.5 Discussion.....	130
7 Travaux connexes	132
7.1 Analyse d'autocorrélation spatiale	132
7.2 Analyse temporelle et spatiotemporelle	133
8 Conclusion	133
Conclusion générale	135
1 Bilans et contributions	135
2 Perspectives	136
TROISIEME PARTIE : Annexes	138
Annexe A : Epidémiologie	139
1 Introduction.....	139
2 Définition de l'épidémiologie	139
3 Etapes d'une étude épidémiologique	139
4 Pratique de l'épidémiologie	140
4.1 Epidémiologie descriptive	140
4.2 Epidémiologie analytique.....	141
5 Niveau de preuve.....	144
6 Lien de causalité	144
7 Indicateurs.....	145
8 Conclusion	146
Annexe B : Analyse multicritères dans EPISOLAP	147
1 Introduction.....	147
2 Conception multicritères.....	147
3 La méthode multicritères PROMETHEE	149
4 Expérimentation	152
4.1 Identification de la zone d'étude.....	152
4.2 Identification des actions	152
4.3 Identification des critères	152
4.4 Génération du tableau de performances.....	153

Table de matières

4.5 Poids des critères.....	153
4.5 Résultat du rangement	153
4.6 Discussion.....	153
5 Conclusion	154
Annexe C : Autocorrélation spatiale.....	155
1 Introduction.....	155
2 Mesures des statistiques descriptives	155
2.1 Mesures de centralité	155
2.2 Mesures de dispersion.....	156
3 Mesures de l'autocorrélation spatiale	156
3.1 Mesures de l'autocorrélation spatiale globale.....	157
3.2 Mesures de l'autocorrélation spatiale locale.....	160
3 Mesures de l'autocorrélation spatiotemporelle locale	168
4 Mesures de classification automatique	169

Liste des figures

Figure 0.1. Diagramme d'activité des taches réalisées dans cette thèse	5
Figure 1.1. Principe des SIAD	14
Figure 1.2. Modèle décisionnel du système EPISOLAP	21
Figure 2.1 : Processus de SIMON	29
Figure 2.2 Processus de Pictet	30
Figure 2.3 La problématique du choix	35
Figure 2.4 La problématique du tri	35
Figure 2.5 La problématique du rangement	36
Figure 2.6 Carte de risque tuberculeux	42
Figure 3.1 Démarche systémique de la fouille de données spatiale	45
Figure 3.2 Méthode de la généralisation spatiale	53
Figure 4.1 Grille de mesure Analyse / Prédiction	68
Figure 4.2 Le module SOLAM du système EPISOLAP_MINING.....	68
Figure 4.3 Les étapes 1 et 2 du processus de généralisation	71
Figure 4.4 Les étapes 3 et 4 du processus de généralisation	72
Figure 4.5 Requête DMQL	76
Figure 4.6 Requêtes SQL	77
Figure 4.7 Concepts hiérarchiques	78
Figure 4.8 Visualisation du processus de généralisation spatiale.....	83
Figure 4.9 Arbre de généralisation spatiale	84
Figure 5.1 Le système d'aide à la décision propose SDSS_MCA_DM	89
Figure 5.2 La démarche de la méthode ELECTRE TRI	93
Figure 5.3 Processus de fouille de données	96
Figure 5.4 Architecture du système décisionnel proposé SDSS-MCA-DM.....	97
Figure 5.5 Processus décisionnel adopté par le système SDSS_MCA_DM	97
Figure 5.6 Modèle décisionnel proposé	98
Figure 5.7.a Digramme UML du cas d'utilisation du système SDSS_MCA_DM.....	99
Figure 5.7.b Digramme de séquence UML du système SDDD_MCA_DM.....	100
Figure 5.8 Zone d'étude	101
Figure 5.9 Actions de reference	105
Figure 5.10 Format du fichier sad	106
Figure 5.11 Format du fichier arff.....	106
Figure 5.12 Matrice de concordance	107
Figure 5.13 Matrice de discordance	107
Figure 5.14 Matrice de crédibilité	107
Figure 5.15 Résultat du tri	108
Figure 5.16 Résultat du tri	108
Figure 5.17 l'affichage de l'arbre de décision.....	109
Figure 5.18 Résultat de la prédiction	109
Figure 5.19 Résultat de la prédiction.....	110
Figure 6.1Démarche systémique de la fouille de données spatiales	118
Figure 6.2 Matrice des indices de l'autocorrélation spatiotemporelle.....	122

Liste des figures

Figure 6.3 Représentation trimestrielle du TBC à Oran	124
Figure 6.4 Autocorrélation spatiale locale par indice de LISA	124
Figure 6.5 Illustration de LISTA définie sur la matrice des indices d'autocorrélation spatio-temporelle.....	126
Figure 6.6 Calcul de l'indice de CITC entre T1 et T2.....	127
Figure 6.7 Calcul de l'indice de CITC entre T2 et T3.....	128
Figure 6.8 Illustration de la matrice des indices CITC	129
Figure 6.9 Illustration de la matrice des indices de LISTA.....	119
Figure 6.10 Représentation cartographique de l'autocorrélation spatio-temporelle de la tuberculose à Oran par trimestre calculée par indice de LISTA	130
Figure A.1 Types d'étude en épidémiologie.....	144
Figure B.1 Modèle décisionnel multicritères proposé.....	148
Figure B.2 Diagramme d'utilisation du model décisionnel proposé.....	149
Figure B.3 Le principe général de fonctionnement par PROMETHEE (I et II).....	150
Figure B.4 Le principe général de fonctionnement par PROMETHEE (II).....	151
Figure B.4 Carte de risque tuberculeux avec Map Info Professionnel 11.0	153
Figure C.1 Centre moyen des facteurs de la tuberculose à Oran en 2009.....	155
Figure C.2 Ecart type de la distance de facteurs de la tuberculose à Oran en 2009.....	156
Figure C.3 Ellipse de dispersion des facteurs de la tuberculose à Oran en 2009	156
Figure C.4 Représentation de l'incidence de la tuberculose à Oran en 2009	157
Figure C.5 Autocorrélation spatiale locale de la tuberculose à Oran par mois synthétisé par trimestre calculée par l'indice de lisa(li)	160
Figure C.6 Analyse de clusters par autocorrélation locale calculée par l'indice de lisa (li).....	161
Figure C.7 Autocorrélation spatiale locale de la tuberculose à Oran en 2009 par mois synthétisé par trimestre	162
Figure C.8 Autocorrélation locale des accidents de la route en France par année	163
Figure C.9 Autocorrélation locale des accidents de la route en France en 2009 par mois	163
Figure C.10 Analyse de points chauds avec la statistique G_i^* de Gétis-Ord	164
Figure C.11 Analyse de points chauds de la tuberculose à Oran par mois synthétisés par trimestre	165
Figure C.12 Analyse de points chauds des accidents de la route en France par année.....	166
Figure C.13 Analyse de points chauds des accidents de la route en France en 2009.....	166
Figure C.13 Analyse de points chauds des accidents de la route en France par mois en 2009	167
Figure C.13 Analyse de points chauds de la criminalité en France par année	167
Figure C.13 Analyse de regroupements (agrégats) en 2, 3, 4 groupes distincts	169

Liste des abréviations

AHP	Analytique Hierarchy Process
AD	Aide à la Décision
ADM	Aide à la Décision Multicriteres
ADS	Arbre de Decision Spatial
AMC	Analyse Multicriteres
AOI	Attribute Oriented Induction
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CITC	Cross Indice of Temporal Correlation
CLARANS	Clustering Large Applications based on RANdomised Search
DMQL	Data Mining Query Language
DSP	Direction de la Santé et de la Population
DSS	Decision Support System
ELECTRE	Elimination ET Choix Traduisant la REalité
EPISOLAP	EPIdemiological Spatial On Line Analytical Processing
FDD	Fouille de Données
FDS	Fouille de Données Spatial
GAIA	Geometrical Analysis for Interactive Assistance
GAM	Geographical Analysis Machine
GDNS	Generalisation à dominante Non spatial
GDS	Genralisation à dominante Spatial
ITA	Indice of Temporal Autocorrelation
LIO	Laboratoire d'Informatique d'Oran
LISA	Local Indice of Spatial Autocorrelation
LISTA	Indice of Spatial Temporal Autocorrelation
MAMC	Methode d'Aide Multicriteres
MAUT	Multi Attribut Utility Theory
OLAP	ON Line Analytical Processing
ONM	Office National de la météoologie
ONS	Office National des statistiques
PROMETHEE	Preference Ranking Organisation Mehod fo Enrichment Evaluations
SAD	Système d'Aide à la Décision
SAD	Système d'Aide à la Décision
SDM	Spatial Data Mining
SDPS	Spatial Data Processing System
SDSS	Spatial Decision Support Système
SESS	Spatial Expert Support System
SES-SKBS	Spatial Expert System-Spatial Knowledge Based System
SIAD	Système d'Information d'Aide ç la Décision
SIAD	Système Interactif d'Aide à la Décision

Liste des abréviations

SIG	Système d'Information Géographique
SOLAM	Spatial On Line Analytical Mining
SOLAP	Spatial On Line Analytical Processing
SQL	Structured Query Language
STDSS	Spatial Temporal Decision Support Système
STING	STatistical Information Grid
UML	Unified Modeling Language

Introduction Générale

1. Contexte de l'étude

Dans notre vie quotidienne, nous pouvons observer de nombreux phénomènes spatiotemporels. Les événements météorologiques en est un exemple classique. Dans la littérature, ces phénomènes spatiaux sont traités par les systèmes d'information géographique qui sont des outils puissants permettant de faire des analyses spatiales sur les données géographiques. Ces outils, étant insuffisants devant la complexité croissante des analyses des données à référence spatiale, d'autres solutions ont été proposées par la communauté scientifique du monde de la géomatique tels que les systèmes SOLAP (Spatial On Line Analytical Processing) ou OLAP Spatial (Bédard, 1997) et OLAP Géographique GeOlap (Bimonte, 2007) permettant d'intégrer les données spatiales dans l'OLAP et de lier l'exploration et la visualisation cartographique.

D'une part, plusieurs travaux se sont intéressés aux méthodes d'analyse multicritères d'aide à la décision (MAMC) et leur intégration aux systèmes d'information géographique (Joerin, 1997), (C.Boulemia et al, 2000) (Mottier, 1999). Ces méthodes formelles ont prouvé leur efficacité dans le domaine spatial et ont démontré leur capacité à cerner des problèmes spatiaux. Ces méthodes ont été appliquées dans différentes études menées dans notre équipe depuis une décennie et dans des domaines différents (transport, aménagement du territoire, agroforesterie, etc.).

D'une autre part, la fouille de données spatiales qui est un domaine de recherche en pleine expansion, offre de nouvelles perspectives pour beaucoup d'applications à caractère décisionnel comme l'analyse du risque d'accident routier (Chelghoum, 2004), d'épidémies (Guernier, 2006), de la pollution (Tchounikine et al, 2005) de la criminalité (Ouattara, 2010) et des risques naturels tels que les inondations, les feux, les sécheresses, les tremblements de terre ou encore l'évolution spatio-temporelle d'un orage dans différentes régions (Alatrasta, 2013).

Les SDSS « Spatial Decision Support System » ou SADS « Systèmes d'Aide à la Décision Spatiale » quant à eux, sont explicitement conçus pour répondre aux lacunes des systèmes d'information géographique (SIG) qui manquent de capacités de modélisation analytique. D'ici provient notre motivation pour intégrer les capacités des méthodes d'analyse multi critères avec les techniques de fouille de données spatiale dans le SIG conventionnel afin de produire un SDSS capable de soutenir la prise de la décision spatiale.

2. Problématique

A l'inverse de la fouille de données (ou la fouille de données spatiales), le système OLAP (ou SOLAP) ne permet pas d'extraire automatiquement des connaissances à partir des données géographiques (Chaudhuri, 1998). Les Systèmes OLAP et les systèmes SOLAP en particulier, ne présentent aucun instrument pour la description des données spatiales et l'explication des phénomènes spatiaux, l'interprétation des résultats d'analyse ou la prédiction de nouvelles connaissances.

La vision d'intégrer OLAP et la fouille de données n'est pas nouvelle. Des solutions, connues sous le terme OLAM ou OLAP sémantique (Han, 1997), (Guran et al, 2009) qui visent à intégrer les méthodes de la fouille de données dans l'OLAP, ont déjà été développées.

L'intégration des méthodes de la fouille de données spatiale dans les systèmes SOLAP est un enjeu majeur. La modélisation des entrepôts de données géographiques tout comme l'adaptation des techniques de la fouille de données classique pour les données géographiques, sont encore à l'heure actuel, des problématiques ouvertes.

3. Objectifs de l'étude et méthodologie adoptée

Notre travail de thèse contribue dans la surveillance et la gestion du risque sanitaire relative à la propagation de l'épidémie de la tuberculose dans la région d'Oran en Algérie.

La réalisation de cet objectif a été conduite en cinq principales phases.

1. Amélioration du système « EPISOLAP » (Zemri, 2011) et adaptation de ses outils de visualisation cartographique pour détecter et localiser les foyers d'épidémies actuels puis prévenir l'apparition de nouveaux foyers. Ceci est réalisé par l'implémentation d'une méthode de fouille de données prédictive.
2. Enrichissement du système « EPISOLAP » par l'intégration des méthodes d'analyse multicritères afin de classer les foyers d'épidémies (par utilisation de la méthode PROMETEE) puis les classer dans des catégories (par utilisation de la méthode ELECTRE TRI) selon le degré du risque épidémiologique {Fort, Moyen, Faible} et l'élaboration, par conséquent, de la carte de risque épidémiologique.
3. Enrichissement du système « EPISOLAP » par l'intégration des techniques de fouille de données spatiales principalement par des méthodes de généralisation spatiale permettant de produire le système « SOLAM » suggéré.
4. Intégration de l'aspect temporel des données de la surveillance et étude de l'auto corrélation spatiale et spatio-temporelle en vue d'enrichir le système SOLAM.
5. Intégration des trois sous systèmes, à savoir, le sous système « EPISOLAP », la carte de risque épidémiologique produite par le module MAMC et le sous système « SOLAM » issu du module FDS dans le système « EPISOLAP –MINING ».

L'outil produit est un système d'aide à la décision spatial qui intègre des solutions puissantes du domaine de l'informatique décisionnelle (Business Intelligence, entrepôts de données, outils de datamining spatial, tableau de bords, système d'information géographique et méthodes d'aide à la décision multi critères).

Ce système fournit des indicateurs et offre des interprétations aux résultats d'analyse des données de la surveillance, il est destiné aux épidémiologistes et aux décideurs en santé publique pour leur fournir une aide dans la prise des décisions stratégiques, en vue d'améliorer la situation épidémiologique rencontrée et orienter les efforts d'éradication et de la prévention.

4. Contributions

Ce travail de recherche doctorale s'inscrit dans le prolongement d'une étude antérieure dont l'objectif était de mettre en œuvre une solution SOLAP pour la surveillance épidémiologique baptisée « EPISOLAP » (Zemri, 2011).

Toutefois, pour aller plus loin dans l'analyse à la fois explicative et prédictive, le SOLAP seul n'est pas suffisant. Notre présente contribution dans cette thèse est la proposition d'une combinaison de la technologie SOLAP avec les Méthodes d'Analyse Multi Critères (MAMC) et les techniques de Fouille de Données Spatiales (FDS) offrant une analyse des données plus riche. Les considérations théoriques et pratiques d'une telle intégration font l'objet de la présente étude.

Notre intervention dans la présente étude se focalise sur les deux axes scientifiques et techniques suivants :

4.1 Couplage SOLAP/MAMC : Classement (rangement) et tri (affectation)

Le système « EPISOLAP » avait comme objectif majeur la détection et la localisation des foyers d'épidémie ; il reste cependant à savoir quel sont les foyers les plus à risque ; nous avons donc proposé de faire un classement de ces foyers en utilisant les différents critères qui rentrent dans l'identification du risque sanitaire. Nous étudions la possibilité de faire une intégration des méthodes d'analyse multicritères. L'idée est de concevoir un système d'aide à la décision spatial basé sur l'intégration de la technologie SOLAP et l'analyse multicritères dont l'objectif est de cerner le phénomène de la propagation épidémiologique et rendre ce problème contrôlable et par conséquent une surveillance épidémiologique accrue et efficace.

- Le rôle du SOLAP est la localisation des foyers d'épidémies et l'analyse spatiale.
- Le rôle des MAMC est le classement des foyers d'épidémie pour faciliter l'intervention par ordre de priorité (Zemri et al, 2014).

Le premier défi était de modéliser le problème de la surveillance épidémiologique en un problème multicritères en tenant compte des différents facteurs influant la propagation de la maladie puis intégrer des scénarios de l'impact environnemental et socio-économique dans un entrepôt de données spatio-temporel en appliquant des techniques de datamining spatiale afin d'expliquer les scénarios prospectifs et donner des interprétations précises aux phénomènes spatiaux.

L'analyse multicritères qui a été utilisée dans le cadre de cette première approche a débouché sur la génération d'une carte de risque épidémiologique qui range les foyers d'épidémies ou susceptibles de devenir des foyers d'épidémies (de la zone la plus favorable à la zone la moins favorable). Ainsi, les spécialistes et épidémiologistes peuvent intervenir et agir précocement sur ces régions par ordre de priorité tout en prenant les dispositifs nécessaires pour minimiser le taux de mortalité et de morbidité. Les décideurs en santé peuvent savoir donc quand intervenir ? Ou intervenir ? Tout en ayant une idée claire et précise sur le potentiel du risque sanitaire encourue et de la gravité de la situation épidémiologique rencontrée.

Une autre méthode multicritères a été intégrée par la suite pour affecter les actions potentielles (les foyers d'épidémie) dans trois catégories distinctes {Fort, Moyen, Faible} en vue de préparer l'étude prédictive utilisant la méthode de l'arbre de décision.

4.2 Intégration SOLAP/FDS : Description, Explication, Interprétation, Prédiction

Nous pensons que l'intégration de la FDS dans notre système « EPISOLAP » aidera dans la découverte des relations et corrélations entre les phénomènes spatiaux et donnera une description précise des scénarios épidémiologiques ce qui permettra d'établir de bonnes interprétations des résultats d'analyse. Notre idée était de combiner SOLAP à la fouille de données spatiales (FDS) pour enrichir les possibilités d'analyse. Nous avons choisi ces deux systèmes complémentaires (SOLAP et FDS) pour les appliquer dans notre application dans la surveillance épidémiologique pour tirer profit de leurs avantages :

- Le SOLAP permettra de représenter l'aspect spatial et cartographique de notre étude grâce aux outils qu'ils combinent pour la détection et la localisation des foyers d'épidémies
- La FDS permettra de mettre en évidence certains aspects socio environnementaux et transformations socio économiques qui influencent la dynamique spatiale de la tuberculose et contribuent dans les différents scénarios de la propagation et la transmission de la maladie dans la région d'Oran, ceux qui n'avaient pas été jusqu'à présent formalisés, ou du moins modélisés.

Ce travail d'intégration SOLAP /FDS débouchera sur l'élaboration d'un nouveau système décisionnel EPISOLAP-MINING qui se base sur le système SOLAM (SOLAP Mining) que nous essayons de concevoir à travers cette étude.

Le système SOLAM s'inspire du modèle OLAM (OLAP Mining) conçu par Jiawei Han (Han, 1997) où il a proposé un OLAP sémantique intégrant le datamining dans le data cube de son prototype. Nous nous sommes largement inspiré de ce système pour proposer le système SOLAM mais cette fois ci en tenant en compte du caractère spatial et temporel en même temps.

Pour cela, nous étudierons l'autocorrélation spatiotemporelle des données de la surveillance et nous adopterons la méthode de la généralisation spatiale basée sur la méthode d'induction orienté attribut (Han et Fu, 1995). Cette méthode permettra d'enrichir SOLAP pour ne plus se limiter à une simple fonction d'agrégation selon le découpage administratif de la zone d'étude mais qui peut être basée sur les caractéristiques des régions géographiques. Le nouvel outil permettra de classer les facteurs démographiques et socio-environnementaux selon leurs degrés de pertinences. Ainsi, les décideurs peuvent orienter leurs actions d'éradication et leurs stratégies de prévention à long terme en éliminant les facteurs par ordre de pertinence. Le système EPISOLAM disposera d'une nouvelle fonctionnalité de sectorisation permettant d'augmenter d'un niveau de généralisation des régions géographiques, chaque fois qu'un attribut non pertinent est supprimé de la base.

5. Acheminement de la thèse

Notre travail de thèse a poursuivi des étapes parfois séquentielles et parfois en parallèle avec d'autres tâches, pour répondre aux objectifs tracés. Ces différentes étapes sont illustrées à travers un diagramme d'activités qui met en relief nos principales contributions ainsi que nos travaux menés tout au long de cette thèse.

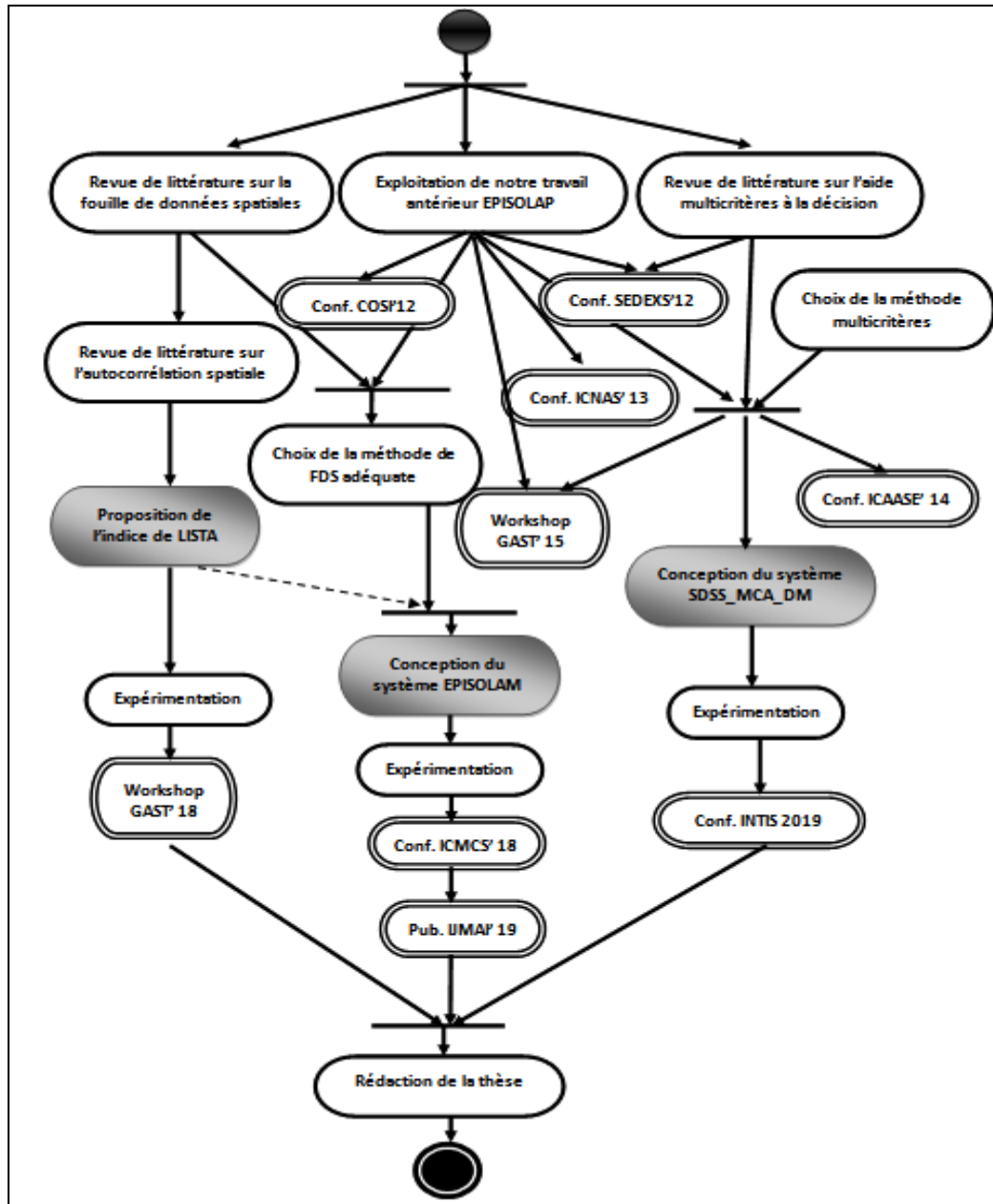


Figure 0.1 Diagramme d'activités des tâches réalisées dans cette thèse

6. Organisation de la thèse

Pour présenter nos travaux, nous avons retenu pour ce mémoire de thèse une organisation en six chapitres structurés en deux parties et une troisième partie consacrée aux annexes.

La première partie est dédiée à **une synthèse de l'état de l'art** qui s'articule autour de trois chapitres.

Chapitre 1 : Les systèmes d'aide à la décision spatiales et spatiotemporels (SDSS)

Ce chapitre constitue une présentation générale des concepts de base structurant l'aide à la décision et l'aide à la décision spatiale ainsi que les systèmes SOLAP.

Chapitre 2 : L'aide à la décision multicritères

Ce chapitre constitue une présentation générale des concepts de base structurant l'aide à la décision multicritères. Il introduit, également, les différents aspects et théories des méthodes d'aide à la décision multicritères.

Chapitre 3 : La fouille de données spatiales

Ce chapitre présente, d'une manière détaillée, les principales techniques de la fouille de données et de la fouille de données spatiales.

La deuxième partie est dédiée à la présentation de différentes **contributions** proposées dans cette thèse. Cette partie se compose de trois chapitres :

Chapitre 4 : Le système SOLAP-MINING

Ce chapitre présente une description détaillée du système SOLAM (Spatial On Line Analytical Mining), SOLAP Mining ou encore SOLAP sémantique qui est une extension du système SOLAP Spatial On Line Analytical Processing) avec les techniques de la fouille de données spatiales (FDS).

Chapitre 5 : Le système SDSS-MCA-DM

Ce chapitre présente une description détaillée du système SDSS-MCA-DM proposé pour la surveillance épidémiologique de la tuberculose à Oran. Il représente une application directe des méthodes multicritères, des méthodes de fouille de données ainsi que les systèmes d'information géographique pour répondre adéquatement aux besoins des décideurs en santé publique.

Chapitre 6 : Proposition de l'indice de LISTA

Dans ce chapitre, nous introduisons la formule de l'indice de l'autocorrélation spatiotemporelle en suivant les différentes étapes qui ont mené au développement de cette formule en passant par la formule de l'autocorrélation temporelle (ITA) puis l'indice croisé de corrélation temporelle (CITC) et enfin l'indice de LISTA suggéré. Nous introduisons plus en détail la matrice des indices de LISTA correspondante.

La troisième partie comprend trois annexes qui présentent des résultats complémentaires.

- **Annexe A** : présente le domaine de l'**épidémiologie**.
- **Annexe B** : présente une **analyse multicritères dans EPISOLAP** utilisant la méthode Prométhée.
- **Annexe C** : présente des expérimentations pour calculer l'**autocorrélation spatiale** sur différentes études de cas

Le manuscrit est achevé par une conclusion générale qui synthétise l'étude abordée dans cette thèse résumant les contributions les plus importantes ainsi qu'un ensemble d'orientations pour les travaux futurs.

PREMIERE PARTIE

SYNTHÈSE DE L'ÉTAT DE L'ART

Chapitre 1

Les Systèmes d'Aide à la Décision Spatio-temporelle(STDSS)

À travers ce chapitre nous survolons les différents types des systèmes d'aide à la décision et leur évolution en systèmes interactifs d'aide à la décision, système d'aide à la décision spatiale en intégrant la composante spatiale et système d'aide à la décision spatiotemporelle en intégrant l'aspect temporel.

Plan du chapitre

1 Introduction	10
2 Les systèmes d'Aide à la Décision.....	10
2.1 Définition.....	10
2.2 Rôle d'un SAD	10
2.2 Les modèles des SAD	11
3 Les systèmes Interactifs d'Aide à la Décision (SIAD)	13
3.1 Définition.....	13
3.2 Caractéristiques des SIAD	13
3.3 Composants d'un SIAD.....	14
3.4 Applications d'aide à la décision	14
3.5 Les problématiques décisionnelles.....	14
4 Les systèmes d'Aide à la Décision spatiale (SDSS).....	15
5 Les systèmes d'Aide à la Décision Spatiotemporelle (STDSS)	16
5.1 Définition et exemples d'application.....	16
5.2 Les objets mobiles	17
5.3 Les séquences	17
5.4 Méthodes d'exploration des séries temporelles.....	18
6 Le système EPISOLAP	20
7 Travaux connexes	22
8 Conclusion.....	23

1.1 Introduction

Dans ce chapitre nous introduisons les différents types des systèmes décisionnels existants et leurs particularités chacun par rapport aux besoins de leurs applications. Nous commençons par les systèmes d'aide à la décision en général en spécifiant leurs rôles et leurs modèles existants puis nous entamons les systèmes interactifs d'aide à la décision en citant leurs caractéristiques, leurs composants et leurs différentes applications. Par la suite, une attention est donnée aux systèmes d'aide à la décision spatiale (SDSS) car ce type de systèmes décisionnels nous intéresse dans le développement de notre application dans les prochains chapitres de cette thèse. Nous donnons pour cela une taxonomie de ce type de systèmes décisionnels. En fin, nous élargissons notre étude en introduisant l'aspect temporel aux systèmes d'aide à la décision spatiale conventionnels pour aboutir à des systèmes spatiotemporels plus complets. Nous citons également, un exemple de système spatiotemporel « EPISOLAP » que nous avons développé dans un travail antérieur. A la fin du chapitre, une étude d'état de l'art sur les technologies SOLAP et les systèmes qui ont en découlé est relatée en donnant quelques exemples de travaux connexes.

1.2 Les Systèmes d'aide à la décision (SAD)

Les systèmes d'aide à la décision (SAD) ou DSS (Decision Support System) sont devenus ces dernières années des systèmes très connus vu leur démocratisation sur les différents secteurs socioéconomiques pour fournir l'aide et répondre aux besoins des décideurs. Aujourd'hui, l'application des SAD dans des organismes privés ou publics (de santé, de transport, d'aménagement de territoire et bien d'autre) est plus que nécessaire pour répondre aux normes d'une gestion efficace des biens et des services.

1.2.1 Définition

Un SAD est un outil permettant d'assister le décideur dans sa prise de décision. En effet, un processus de décision ne peut pas être entièrement automatisable (Duvallet, 2001). Un SAD est un Système de Traitement de l'Information (STI) (Lévine and al, 1989) qui permet d'extraire et de donner au décideur l'information nécessaire au processus de prise de décision. D'une manière plus précise, dans le monde anglo-saxon, un DSS, « décision support system » est défini comme "un système d'information informatique, interactif, flexible et adaptable, spécialement développé pour aider à trouver une solution à un problème de gestion non structuré, en vue d'une meilleure prise de décision. Il utilise des données, offre une interface conviviale et permet au décideur d'apporter ses propres idées.

1.2.2 Rôle d'un SAD

Un SAD, en favorisant l'association d'un système informatique et d'un décideur, permet la formation d'un système de décision complet. Le décideur, de par sa connaissance pratique, possède un méta-modèle du processus de décision, et le SAD, par sa capacité de traitement de l'information, l'aide à structurer le modèle (Lévine and al, 1989). En d'autres termes, le décideur contrôle le processus de décision et le SAD l'assiste en effectuant les calculs standards et répétitifs sur les données (Duvallet, 2001). Dans une telle perspective, le processus de décision s'identifie à une recherche heuristique menée par le décideur ; le système jalonne le processus de

recherche à l'aide d'indicateurs et d'informations. Le décideur, en fonction de ces informations produites à l'issue du traitement des données, continue l'exploration heuristique des actions possibles, ou arrête si tout lui indique que la solution construite rencontre ses buts de façon satisfaisante (Duvallat, 2001). De ce fait, la coopération entre le décideur et le système informatique ne peut être fructueuse que dans le cadre d'un système interactif (Duvallat, 2001).

1.2.3 Les modèles des systèmes d'aide à la décision

Les modèles pour l'aide à la décision reposent généralement sur des hypothèses de rationalité des décideurs et de solutions possibles (Attonaty et al, 1991). On distingue deux types de modèles :

- a. Des modèles fondés sur la rationalité complète et la recherche de solutions optimales telle que la programmation linéaire.
- b. Des modèles non fondés sur cette rationalité complète (appelée rationalité limitée) et ayant des solutions considérées comme appropriées, tels que les systèmes interactifs d'aide à la décision (SIAD) et les systèmes multi agents (SMA).

Dans cette seconde approche, le processus de conduite de raisonnement est au centre de l'aide à la décision, alors que dans les approches plus normatives, la solution optimale est au centre de l'aide à la décision (Bourgine et al, 1990).

1.2.3.1 La rationalité complète

Les choix sont explicités comme étant le résultat de l'activité d'un décideur qui cherche à maximiser/minimiser un certain objectif. La décision est bien orientée vers un but. Une décision effectuée selon un tel modèle consiste à choisir la solution dont les conséquences possibles permettent d'atteindre le mieux les objectifs établis au départ.

La théorie de la décision définit les hypothèses de la rationalité complète (Russel, 1999) comme suit :

- Les objectifs du décideur sont clairement exprimés dès le départ et peuvent être rangés selon un ordre de préférence.
- L'ensemble des solutions possibles dans une décision est connu et peut être évoqué.
- Les conséquences rattachées à chacune des solutions sont précisées.
- Le choix de la solution s'effectue en retenant la solution optimale.

Ces hypothèses sont réalisables dans le cas d'un problème bien structuré. Dans ce courant, la solution existe en dehors des acteurs. Que l'on soit en statique, en incertain, on se ramène à un critère unique de synthèse que l'on optimise.

1.2.3.2 La rationalité limitée

Porte sur l'étude du comportement d'un individu (ici appeler acteur) face à un choix (l'achat d'un produit,...). Il suppose que l'acteur économique a un comportement rationnel, mais que sa rationalité est limitée en termes de capacité cognitive et d'information disponible. Dès lors, l'acteur va généralement s'arrêter au premier choix qu'il jugera satisfaisant.

Ce principe a été proposé par Herbert Simon afin de montrer que l'être humain est capable de justifier son comportement (ou sa décision) mais à certaines limites, car ses capacités mentales sont limitées.

En d'autres termes, l'acteur est rationnel (s'il préfère A à B et B à C, on peut en déduire qu'il préfère A à C). Cependant, lors d'un choix complexe, il cherche moins à étudier l'ensemble des possibilités qu'à trouver une solution raisonnable dans une situation d'incertitude. Il va s'arrêter généralement à la première option qui satisfera à la situation concrète, tout en évitant de consommer trop de temps à effectuer son choix.

La rationalité limitée est définie comme étant « *la capacité de générer un comportement maximal en fonction des informations disponibles et des ressources de calcul* » (Russel, 1999). Le terme d'optimisation limitée a été défini par Horvitz (Horvitz, 1989) comme étant « *l'optimisation d'une utilité calculable donnée par un ensemble d'hypothèses sur des problèmes et des contraintes de ressources* ».

Les hypothèses de la rationalité limitée sont les suivantes (Horvitz, 1989):

- Le décideur perçoit un problème auquel il réagit, ce problème représente une version simplifiée de la réalité.
- Le décideur doit d'abord identifier des actions possibles (sans toutefois toutes les découvrir).
- Il doit recueillir l'information nécessaire pour déterminer et évaluer les conséquences des actions.
- Les objectifs poursuivis par le décideur sont précisés tout au long du processus sous la forme de « niveaux d'aspiration ».
- Le décideur arrête son choix sur la première option analysée qui satisfait chacun des buts fixés.

D'après ces auteurs, les profondes modifications imposées au modèle de la rationalité complète permettent de considérer que les décisions sont prises sur la base d'une information incomplète, souvent imprécise, que l'élaboration de toutes les possibilités est un processus beaucoup trop exigeant et aussi que les objectifs tracés sont souvent conflictuels. Le modèle de la rationalité limitée correspond ainsi un peu plus aux conditions réelles de la prise de décision par rapport à la rationalité complète.

1.2.3.3 La rationalité procédurale

A la suite de la rationalité limitée, (Simon, 1982) introduit le concept de la rationalité procédurale qui met en avant le processus qui conduit l'acteur à prendre ses décisions.

1.2.3.4 La rationalité adaptative

Développée par Cyert et March (Cyert et March, 1963), elle possède les caractéristiques suivantes :

- L'adaptation de l'organisation à son environnement provient d'enseignements découlant d'expériences antérieures (*learning by doing*).

- L'apprentissage débouche sur : La création d'indicateurs qui conditionnent la perception des situations rencontrées, l'élaboration des règles d'actions pour poursuivre des buts dédiés à des problèmes identifiés à travers des situations mémorisées et le développement de compétences particulières.

Cette dernière hypothèse pose un problème dans un environnement changeant, il peut y avoir un décalage entre les règles apprises dans un environnement donné par rapport à un environnement courant pouvant conduire à des comportements erronés (Hall, 1984) et (Argyris et al, 1978).

1.3 Les systèmes interactifs d'aide à la décision (SIAD)

Suite à l'apparition des SAD, les systèmes d'Aide interactifs d'Aide à la décision (SIAD) ont vu le jour pour combler l'aspect interactif en introduisant l'interface qui gère l'interaction entre le SAD et le décideur.

1.3.1 Définition

Un SIAD est un système d'information interactif, flexible, adaptable et spécifiquement développé pour aider à la résolution d'un problème de décision en améliorant la prise de décision. Il utilise des données, fournit une interface utilisateur simple et autorise l'utilisateur à développer ses propres idées ou points de vue.

Ces systèmes n'ont pas pour objectif d'identifier une décision optimale qu'il « suffirait » ensuite de mettre en application. Leur finalité consiste à « *amplifier le raisonnement du décideur sur la base de ses propres représentations* » (Courbon, 1993). Ces systèmes ont une place privilégiée lorsque les problèmes sont peu, voire non structurés (Lévine et al, 1989).

1.3.2 Caractéristiques des SIAD

Keen et Scott-Morton (Keen et Scott-Morton, 1978) présentent les Systèmes Interactifs d'Aide à la Décision (SIAD) comme des systèmes conçus pour résoudre des problèmes de décision peu ou mal structurés. Ces problèmes possèdent les ou l'une des caractéristiques suivantes (Klen, 1997) :

- Les préférences, jugements, intuitions et l'expérience du décideur sont essentiels.
- La recherche d'une solution implique un mélange de recherche d'information, de formalisation ou définition et structuration du problème, du calcul et de la manipulation de données.
- La séquence des opérations de recherche d'une solution n'est pas connue à l'avance parce qu'elle peut être fonction des données, être modifiée, peut ne donner que des résultats partiels, ou encore peut être fonction des préférences de l'utilisateur.
- Les critères pour la décision sont nombreux, en conflit et fortement dépendant de la perception de l'utilisateur.
- La solution doit être obtenue en un temps limité.
- Le problème évolue rapidement.

1.3.3 Composants d'un SIAD

Un SIAD se compose d'au moins trois composants (Sprague, 87):

- **Le module de dialogue** représente le module charnière du système permettant d'établir une collaboration entre le décideur et la machine. Pour (Courbon et al, 1986), le module de dialogue est au centre du SIAD et sa réalisation est primordiale.
- **La base d'informations** assure la fonction de mémoire, elle stocke non seulement les données, de façon permanente ou passagère, mais elle gère aussi l'enregistrement de données volatiles ainsi que l'effacement de ces mêmes données selon le souhait de l'utilisateur. Ces données volatiles correspondent aux résultats obtenus lors de traitements de données. Les données que nous avons qualifiées de permanentes sont les informations statistiques ou autres données qui décrivent les situations courantes et passées.
- **La base de modèle** se compose d'un ensemble de modèles et d'un système de gestion de ceux-ci. Les modèles peuvent être : des outils de recherche opérationnelle, des modèles statistiques ou autres. Pour avoir davantage de flexibilité, un SIAD doit posséder plusieurs modèles (Chakhar, 1997). Dans cette optique, le SIAD organise les liens et le passage de paramètres entre les différents modèles, de même qu'il gère le module de dialogue (Lévine et al, 90).

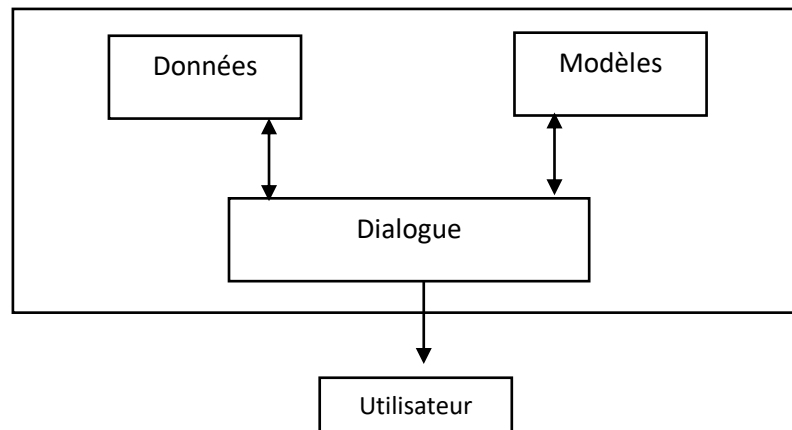


Figure 1.1 Principe des SIAD selon Sprague (Sprague, 87)

1.3.4 Applications d'aide à la décision

Les systèmes d'aide à la décision sont appliqués dans une large gamme d'activités où chacun des utilisateurs les adaptent selon l'objet de son étude et selon son domaine d'application. Sans toutefois dresser une liste exhaustive, les domaines suivants peuvent faire l'objet d'application des systèmes d'aide à la décision ; la gestion forestière, l'administration publique, la gestion du personnel, les affaires, la santé, l'aménagement de territoire, le transport...etc.

1.3.5 Les problématiques décisionnelles

Les différents types des problématiques décisionnelles proposées par Simon et reprises par [Lev et al, 89] sont :

1. **Problème structuré** : le problème est posé d'une façon claire, les données sont fiables et numériques. Les modèles associés sont clairement définis, ils permettent une résolution algorithmique programmable du processus décisionnel.
2. **Problème semi structuré** : le problème peut ne pas être clairement posé, les données sont souvent qualitatives et la résolution est difficilement exprimable sous forme de programme
3. **Problème non structuré** : le problème n'est pas clairement posé, le principe de la rationalité limitée s'applique à toutes les étapes du processus décisionnel. La décision prise est difficilement justifiable

1.4. Les systèmes d'aide à la décision spatiale

Les systèmes d'aide à la décision spatiale regroupent l'ensemble des modèles qui font appel aux systèmes d'information géographique pour traiter et analyser les données à référence spatiale alimentant les bases d'informations géographiques du SDSS. Néanmoins, ces systèmes ne se contentent pas de faire de simples analyses spatiales comme dans les SIG, mais se généralisent à d'autres techniques et méthodes poussées pour l'exploration des données spatiales.

La fouille de données spatiales (FDS) découvre des modèles et des connaissances à partir des données spatiales. Les données spatiales se réfèrent aux données géo spatiales stockées dans des référentiels de données (entrepôts de données) géo spatiales. Ces données peuvent être en format "vectoriel" ou "raster", ou sous forme d'images multimédia géo référencées. Récemment, de grands entrepôts de données géographiques ont été construits en intégrant des données thématiques et géographiquement référencées provenant de sources multiples à partir desquels nous pouvons construire des cubes de données spatiales contenant des dimensions et des mesures spatiales, et qui supportent l'OLAP spatial (SOLAP) pour faire l'analyse multidimensionnelle des données spatiales. L'exploration de données spatiales peut être effectuée sur des entrepôts de données spatiales, des bases de données spatiales ou des entrepôts de données géo spatiales

Les sujets d'actualité traités dans cette catégorie des systèmes d'aide la décision spatiale sont : la découverte des schémas des associations et des co-localisations spatiales, la classification spatiale, l'analyse des regroupements spatiaux ou le clustering spatial, l'analyse de modèles spatiaux et l'analyse des tendances et des valeurs aberrantes spatiales. Nous verrons avec plus de détails, la majorité de ses techniques dans le chapitre 5.

Tableau 1.1 Tableaux des types des SDSS

	Spatial Data Processing Systems (SDPS)	Spatial Decision Support Systems(SDSS)	Spatial Expert Systems Spatial Knowledge Based System (SES-SKBS)	Spatial Expert Support Systems (SESS)
Principal objectif	incorpore tous les éléments du problème décisionnel et analyse le modèle représentant le problème	fournit l'expérience, de la collecte et de l'exécution de données aussi bien que le conseil aux utilisateurs	fournit le conseil et l'explication intelligente	
Nature du problème traité	structuré	semi structuré mal structuré	semi structuré mal structuré	semi structuré mal structuré
Avantages	Traitement des problèmes décisionnels via des algorithmes	interface puissante et facile, capacité de combiner les modèles analytiques avec les données et offre une variété de modèles de prise de décision.	Capable de résoudre des problèmes spatiaux, emploie la connaissance experte sous forme de règles et agit l'un sur l'autre avec des décideurs.	Accessibilité et flexibilité
Inconvénients	Pas de flexibilité		Beaucoup de processus géographiques et de problèmes liés à la décision spatiale ne sont pas bien compris. L'expertise humaine complexe et différents niveaux de connaissance.	

1.5 Les systèmes d'aide à la décision spatiotemporelle STDSS

Suite aux succès grandissant des SDSS ces dernières années et leurs applications dans les différents domaines, nous assistons à l'émergence des systèmes d'aide à la décision spatiotemporelle vu la multiplicité des données temporelles et spatiales relatives à la même étude et ceci dans les différents domaines d'application.

1.5.1 Définition et exemples d'application

Les données spatiotemporelles sont des données relatives à l'espace et au temps. L'exploration des données spatiotemporelles fait référence au processus de découverte des modèles et des connaissances à partir des données spatio-temporelles (Han et al, 2012). Des exemples typiques d'exploration de données spatio-temporelles incluent :

- la découverte de l'évolution historique des villes et des terres
- la découverte des conditions météorologiques
- prévoir les tremblements de terre et les ouragans

- déterminer les tendances globales du réchauffement climatique.

L'exploration des données spatiotemporelles est devenue de plus en plus importante et a de profondes implications, étant donné la popularité des téléphones mobiles, des appareils GPS, des services de cartographie sur internet, des services météorologiques et la terre numérique, ainsi que les technologies satellite, RFID, les capteurs sans fil et la vidéo.

1.5.2 Les objets mobiles

Nous citons, parmi les nombreux types de données spatio-temporelles, *les données des objets mobiles* (c'est-à-dire les données sur les objets en déplacement) sont particulièrement importantes. Par exemple, les chercheurs en biologie animale attachent des équipements de télémétrie sur la faune pour analyser le comportement écologique, les gestionnaires de la mobilité intègrent le GPS dans les voitures pour mieux surveiller et guider les véhicules, et les météorologues utilisent des satellites météorologiques et des radars pour observer les ouragans.

Les données d'objets mobiles à grande échelle sont en train de devenir riches, complexes et omniprésents. Les exemples d'extraction des données à partir des objets mobiles incluent les modèles des mouvements des objets multiples (c'est-à-dire la découverte des relations entre des objets mobiles multiples tels que les regroupements en mouvement, les leaders et les followers, les modèles de fusion, analyse des convois, des essaims, ainsi que d'autres schémas de mouvements collectifs). Parmi les autres exemples d'exploration des données des objets mobiles, on peut citer l'exploration des modèles périodiques pour un ou plusieurs ensembles d'objets, et l'exploration des schémas des trajectoires, des clusters, des tendances et des valeurs aberrantes.

Ils existent plusieurs manières pour modéliser les données spatiotemporelles. Ce ci dépend de la nature de ces données, ainsi du type du phénomène étudié. S'agissant des problématiques liées à la surveillance et le suivi environnemental tels que :

- la détection de changements abrupts (catastrophes naturelles, etc.),
- le suivi de phénomènes évolutifs (érosion côtière, désertification, etc.),
- le suivi de phénomènes qui se propagent (épidémies, pollution de rivières, etc.)
- la mise au point de modèles (hydrologie, activité agricole, etc.)

L'analyse des séquences restent l'une des meilleures solutions utilisées pour modéliser, comprendre et prédire de tels scénarios complexes et d'origine pluridisciplinaire (données climatiques, géologiques, médicales, etc.).

1.5.3 Les séquences

Une séquence est une liste ordonnée d'événements. Les séquences peuvent être classées en trois groupes, sur la base des caractéristiques des événements décrits (Han et al, 2012) : (1) données chronologiques, (2) données de séquence symboliques, et (3) séquences biologiques. Considérons dans ce qui suit chacun des trois types (Han et al, 2012) :

1.5.3.1 Les séries temporelles ou les séries chronologiques, les données consistent en de longues séquences de données numériques, enregistrées à des intervalles de temps égaux (par exemple, par minute, par heure ou par jour). Nous pouvons citer dans cette catégorie des données spatiotemporelles par exemple des données générées par de nombreux processus

naturels ou économiques tels que l'étude des marchés, les observations scientifiques, médicales ou naturelles.

1.5.3.2 Les données de séquence symboliques, sont constituées de longues séquences d'événements ou de données nominales, qui ne sont généralement pas observés à des intervalles de temps égaux. Pour beaucoup de ces séquences, des lacunes (c'est-à-dire laps de temps entre les événements enregistrés) ne sont pas très importantes. Les exemples incluent les séquences du shopping client, les flux des clics Web, ainsi que des séquences d'événements scientifiques et de l'ingénierie et dans le domaine du développement naturel et social.

1.5.3.3 Les séquences biologiques comprennent les séquences d'ADN et de protéines. De telles séquences sont typiquement très longues, et portent une signification sémantique importante, compliquée, mais cachée. Ici, les lacunes sont généralement importantes.

Examinons l'exploration des données pour chacun de ces types de données de séquence.

1.5.4 Méthodes d'exploration des séries temporelles

Dans le cadre de cette thèse, nous nous focalisons sur le type de données des séries temporelles car nous faisons appel à ce type de données plus tard pour répondre aux besoins relatifs à notre domaine d'application en épidémiologie. Nous citons ci-dessous quelques techniques d'exploration des séries temporelles trouvées dans (Han et al, 2012) :

1.5.4.1 Recherche de similarité dans les données de séries chronologiques

Un ensemble de données chronologiques se compose de séquences de valeurs numériques obtenues dans par des mesures du temps répétées. Les valeurs sont généralement mesurées à intervalles de temps égaux (par exemple, chaque minute, heure ou jour). Les bases de données chronologiques sont populaires dans de nombreuses applications comme l'analyse des marchés, prévisions économiques, prévision des ventes, analyse budgétaire, études des services publics, études d'inventaire, projections de rendement, projections de charge de travail, processus et contrôle de qualité. Ils sont également utiles pour l'étude des phénomènes naturels (atmosphère, température, vent, séisme), des expériences scientifiques et d'ingénierie et des études des traitements médicaux (Han et al, 2012).

Contrairement aux requêtes de base de données normales, qui trouvent des données correspondant exactement à une requête, une recherche de similarité trouve des séquences de données qui ne diffèrent que légèrement d'une requête séquence donnée. De nombreuses requêtes de similarité de séries temporelles nécessitent une correspondance des sous-séquences, ce qui signifie de rechercher un ensemble de séquences contenant des sous-séquences similaires à une requête séquence donnée.

Pour la recherche de similarité, il est souvent nécessaire d'effectuer d'abord une réduction des données ou des dimensions et une transformation des données chronologiques. Les techniques typiques de réduction de dimensionnalité incluent (Han et al, 2012) :

- La transformée de Fourier discrète (DFT : discrete Fourier transform)
- La transformée en ondelettes discrète (DWT : discrete wavelet transforms)
- La décomposition en valeurs singulières (SVD : singular value decomposition) basée sur l'analyse en composantes principales (PCA : principle components analysis).

Dans de telles techniques, les données (ou les signaux) sont mappés à un autre signal dans un espace transformé. Un sous-ensemble des coefficients transformés les plus forts sont enregistrés en tant qu'entités.

Ces caractéristiques forment un *espace de fonctions*, qui est une projection de l'espace transformé. Les indices peuvent être construits sur les données de série chronologique originales ou transformées pour accélérer la recherche.

Pour une recherche de similarité basée sur une requête, les techniques incluent la transformation normalisée, appariement atomique (c'est-à-dire trouver des paires de fenêtres sans faille d'une petite longueur qui sont similaires), couture de fenêtre (c'est-à-dire assembler des fenêtres similaires pour former des paires de grandes sous-séquences similaires, laissant des espaces entre les correspondances atomiques), et ordonner la sous-séquence (c'est-à-dire, ordonner linéairement les correspondances des sous-séquences pour déterminer si suffisamment de pièces similaires existent). De nombreux logiciels existent pour une recherche de similarité dans les séries chronologiques.

Des chercheurs ont récemment proposé de transformer les séries chronologiques en approximations agrégées afin que les données puissent être visualisées comme une séquence de représentations symboliques. Le problème de la recherche de similarité est ensuite transformé en un problème de correspondance des sous-séquences dans les données de séquence symboliques.

Nous pouvons identifier des motifs (c'est-à-dire des occurrences de modèles séquentiels fréquents) et construire des mécanismes d'index ou de hachage pour une recherche efficace en se basant sur de tels motifs. Les expérimentations montrent que cette approche est simple et rapide et d'une bonne qualité de recherche comparable à celle de DFT, DWT et autres méthodes de réductions de dimensionnalité (Han et al, 2012).

1.5.4.2 Analyse de régression et de tendance dans les données de séries chronologiques

L'analyse de régression des séries chronologiques a fait l'objet d'une étude approfondie dans le domaine de la statistique et de l'analyse des signaux. Cependant, il peut souvent être nécessaire d'aller au-delà de l'analyse de la régression pure et d'effectuer une analyse de tendance pour de nombreuses applications pratiques. L'analyse des tendances crée un modèle intégré utilisant les quatre composantes ou mouvements principaux suivants pour caractériser les séries chronologiques :

1. **Tendance des mouvements à long terme** : indique la direction générale dans laquelle un graphique de série temporelle évolue dans le temps, par exemple, en utilisant la méthode de la moyenne mobile pondérée et les méthodes des moindres carrés pour trouver des courbes de tendance.
2. **Mouvements cycliques** : ce sont les oscillations à long terme autour d'une ligne de tendance ou d'une courbe.
3. **Variations saisonnières** : ce sont les modèles presque identiques qu'une série chronologique semble suivre au cours des saisons correspondantes des années successives. Pour une analyse efficace des tendances, les données doivent souvent être «désaisonnalisées» sur la base d'un indice saisonnier calculé par l'autocorrélation.
4. **Mouvements aléatoires** : ils caractérisent les changements occasionnels dus à des événements fortuits.

L'analyse des tendances peut également être utilisée pour la prévision des séries chronologiques, c'est-à-dire pour trouver une fonction mathématique qui générera approximativement les tendances historiques d'une série chronologique et l'utiliser pour établir des prévisions à court ou à long terme des valeurs futures. Nous citons le modèle ARIMA (auto-regressive integrated moving average), la modélisation des séries temporelles à long terme et l'autoregression sont des méthodes populaires pour une telle analyse (Han et al, 2012).

L'un des technologies puissantes permettant de lier l'aspect temporel et spatial des données spatiotemporelles est le système SOLAP. Celui là permet une représentation et une analyse en ligne de ce type de données ainsi un affichage rapide et une visualisation graphique, tabulaire et cartographique des résultats de l'analyse multidimensionnelle.

1.6 Le système EPISOLAP

Les systèmes SOLAP sont des systèmes qui intègrent des composantes spatiales et temporelle dans les modèles multidimensionnels pour les applications géo-décisionnelles. L'OLAP spatial (SOLAP) a été défini par Yvan Bedard comme "une plateforme visuelle conçue spécialement pour supporter une analyse spatiotemporelle rapide et efficace à travers une approche multidimensionnelle qui comprend des niveaux d'agrégation cartographiques, graphiques et tabulaire" (Bédard, 2004). Les systèmes SOLAP permettent de représenter l'information géographique au niveau des dimensions.

Le système EPISOLAP suggéré dans (Zemri, 2011) pour détecter et localiser les foyers d'épidémies permet de fournir une visualisation simple et rapide des données de la surveillance grâce à une vue multidimensionnelle, une analyse en ligne et une analyse spatiotemporelle sur une carte géographique.

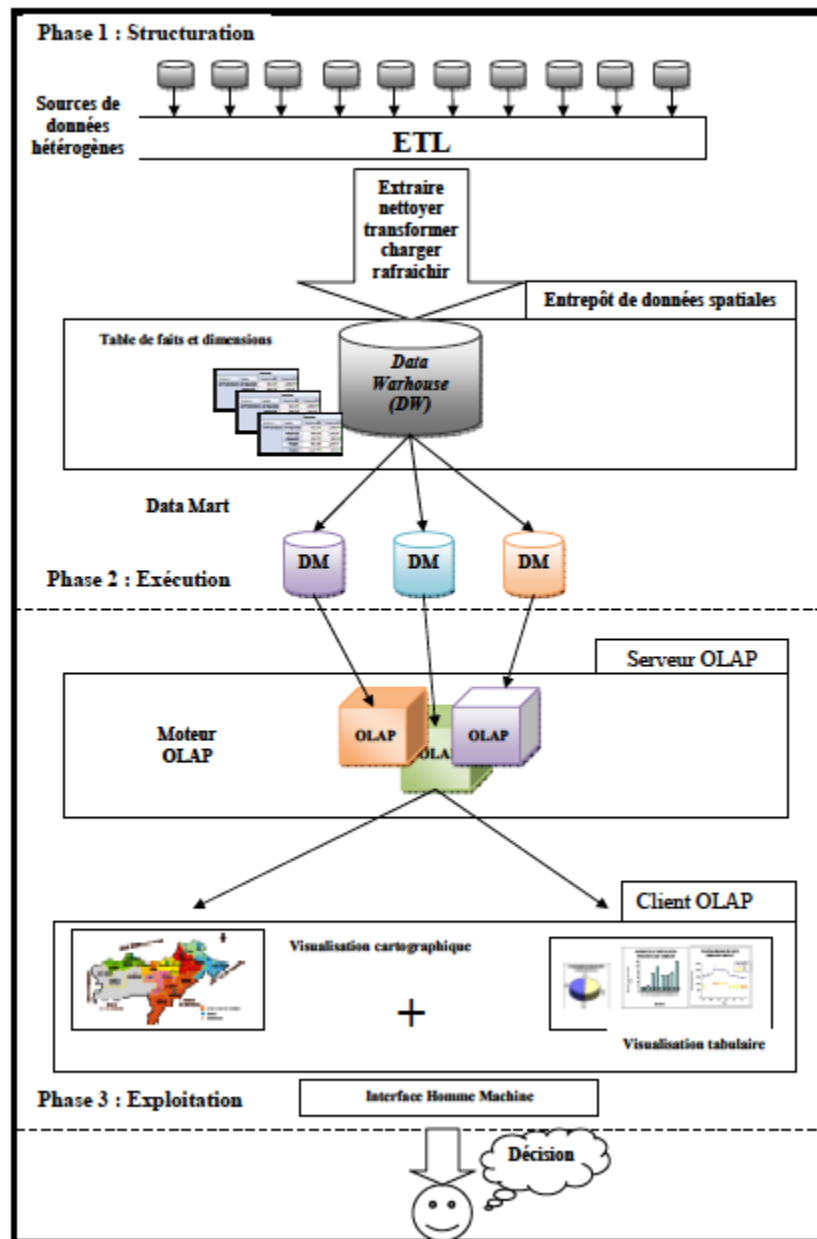


Figure 1.2 Modèle décisionnel du système EPISOLAP

Cependant, ces simples analyses ne donnent aucune interprétation des résultats issus du système SOLAP. De plus, ce dernier ne tient pas en compte des facteurs socio économiques qui contribuent aux divers scénarios de propagation ou de la transmission de la maladie et ne permet pas de découvrir les transformations socio-économiques qui influencent la dynamique spatiale de l'épidémie.

La FDS quant à elle permet d'interpréter les résultats d'analyse à travers une panoplie des méthodes et outils exploratoires et prédictives non offertes par les systèmes SOLAP seuls. La FDS offre une grande variété de techniques avec différents objectifs analytiques. Les deux approches ont été donc proposées séparément et n'étaient pas suffisamment intégrées. Cela rend le couplage entre ces deux outils plus que nécessaire pour atteindre cette complémentarité.

En se basant sur les limites que nous avons détectées dans le système EPISOLAP, nous avons pensé à son extension à d'autres fonctionnalités.

Au début de notre étude, nous nous sommes appuyés sur des travaux de la littérature qui avaient comme objectif l'extension des systèmes OLAP qui sont devenus soudainement eux aussi insuffisants pour représenter les données complexes. Nous citons dans la section suivante un ensemble de ces travaux connexes.

1.7 Travaux connexes

Depuis l'introduction du système SOLAP par le professeur Yvan Bedard, les travaux d'amélioration et de perfectionnement ne cessent d'émerger pour présenter de nouvelles propositions en vue de rendre les applications SOLAP plus complètes. Nous citons à titre d'exemple l'étude de (Rivest et al, 2001) où plusieurs applications SOLAP ont été présentées (dans les domaines de santé environnementale, le transport...), et des lacunes ont été soulevées pour être ensuite remédiées. D'autre part, et au court de cette dernière décennie, plusieurs efforts ont été faits pour enrichir les systèmes OLAP afin de pouvoir supporter les objets spatiaux complexes. Nous citons ici, les travaux de (Viswanathan et Schneider, 2011) qui ont proposé le modèle C^3 présenté comme une nouvelle stratégie de modélisation pour supporter les objets spatiaux complexes dans les cubes OLAP. Ce modèle est l'extension du modèle **"BigCube"** proposé dans (Viswanathan et al, 2010). Le modèle C^3 regroupe trois constructeurs : le constructeur de catégorisation **"categorization construct"**, le constructeur de confinement **"containment construct"** et le constructeur de combinaison **"combination or cubing construct"** permettant tous les trois de structurer des régions complexes et multi structurées et construire le cube de données multidimensionnelles. Ils ont proposé également un autre constructeur géographique appelé **"geo-construct"** pour manipuler et interroger les hiérarchies spatiales des objets géographiques. Ces objets sont subdivisés en deux types : Objets structurés (SO : Structured Object) et Objets de base (BO : Base Object). Sur la base de ces deux types d'objets géographiques ils ont proposé l'hiérarchie des régions appelée **"regH : region hierarchy"** organisée comme ceci :

$$< region : SO > := < regionLabel : BO > < face : SO >$$

Dans (Daminani et Spaccapietra, 2006) les chercheurs ont proposé **"MuSD"** (multigranular spatial data warehouse) qui se base sur le fait spatial, la dimension spatiale et la mesure spatiale multi-niveaux pour gérer et supporter les mesures spatiales dans des niveaux de granularité géométrique multiple. Outre les constructions de représentation, le modèle comprend un ensemble d'opérateurs SOLAP pour naviguer non seulement à travers les niveaux dimensionnels mais également à travers les niveaux des mesures spatiales. Récemment, le travail de (Amanzougarene, 2014) s'est intéressé à la gestion des mesures qualitatives en plus des mesures quantitatives dans les systèmes OLAP ainsi qu'à la proposition de nouveaux opérateurs de forage pour naviguer dans les différents niveaux du cube multidimensionnel. L'objectif de la recherche était l'adaptation des opérations de forage en introduisant la notion de la logique floue. Cette dernière permettait de calculer des mesures probabilistes par agrégation des dimensions qualitatives. D'autres travaux ont entamé la problématique des hiérarchies spatiales non strictes dans les entrepôts de données spatiales et ont proposé de nouvelles modèles logiques dans

(Malinowski et Zimanyi, 2004) en introduisant des schémas logiques avec **"Bridge Table"**. Plus récemment dans (Ferrahi et al. 2016) les auteurs ont proposé une autre solution d'implémentation de ce type spécial d'hierarchies qui existent dans certaines applications spatiales en introduisant l'index bitmap (Index-NN). Dans le cadre de l'amélioration de la capacité de traitement des données à référence spatiale, les chercheurs dans (Xu et Sheu, 2012) ont introduit la notion des dimensions spatiales auxiliaires et ont adapté les opérations SOLAP pour gérer ce genre de dimensions spatiales qui ont pour rôle de faciliter le traitement et l'exécution des requêtes spatiales parfois complexes et assurer un meilleur temps de réponse. Un autre travail de recherche de (Pedersen et al, 2009) s'est intéressé à la gestion des niveaux de détails complexes des hierarchies dans les systèmes OLAP. Ils étaient les premiers à proposer un "multi-model" qui utilise la sémantique dans l'agrégation des données pour éviter l'introduction des requêtes incorrectes et le non sens des réponses. Ils ont introduit également dans ce même travail, un langage de haut niveau pour les bases de données multidimensionnelles appelé **"SumQL"** qui a été étendu par la suite au langage **"SumQL++"** pour supporter les bases de données objets distribuées. Dans le même sens, un travail récent dans (Besalloua et Hamdadou, 2018) a fait l'objet d'une nouvelle approche qui définit des contraintes d'intégrités dans les systèmes SOLAP pour améliorer la précision, la cohérence et la complétude des bases de données. Ensuite, une classification des contraintes concernant les utilisateurs SOLAP était appliquée sur un schéma en constellation composé de trois tables de faits partageant les dimensions spatiales, temporelles et thématiques. L'approche était appliquée dans le domaine de l'agroforesterie dans la région de Mostaganem en Algérie.

Certaines questions de recherche ouvertes qui sont en cours d'investigation sont énumérées sous forme de nouvelles tendances du SOLAP, dans (Taher, 2008). Ces tendances incluent : SOLAP continu, SOLAP en temps réel, SOLAP pour les situations d'urgence et applications Web SOLAP.

Dans le domaine du web mapping, dans (Bernier et al, 2007) les chercheurs ont mis en place la plateforme **"UMapIT"** (Unrestricted Mapping Interactive Tool) pour une gestion efficace des données à référence spatiales sur le web. Cette plateforme web était introduite comme un nouveau paradigme qui intègre l'approche multidimensionnelle (datacube) et l'approche orienté occurrence pour supporter les demandes des utilisateurs sur le web. Cette approche combine la géométrie, la sémantique et la représentation graphique comme tables de dimensions dans un schéma en étoile et sont connecté à la table de fait spatial définit comme **"Vuel"** (Vue Element). Cette approche permet une meilleure gestion des niveaux de détails des hierarchies spatiales sur le web. Dans (Bimonte et al. 2007) GeWOLAP un autre prototype pour SOLAP basé sur le Web est présenté avec quelques nouveaux opérateurs SOLAP adaptés aux applications web.

1.8 Conclusion

Dans ce chapitre, nous avons détaillé les différentes définitions relatives aux systèmes d'aide à la décision (DSS), les systèmes interactifs d'aide à la décision ainsi que les systèmes d'aide à la décision spatiale (SDSS). Nous avons décrit les modèles conceptuels existantes et proposés dans la littérature. Puis nous avons présenté une extension des SDSS sur l'aspect temporel des données spatiales et nous avons définit les systèmes d'aide à la décision spatiotemporelles

(STDSS) vu leurs utilité dans la résolution de certains problèmes à caractère spatial et temporel en même temps (ex : épidémiologie).

A la fin de ce chapitre, nous avons présenté un exemple des systèmes SOLAP, à savoir le système EPISOLAP suggéré dans un travail antérieur. Celui là révèle que les solutions qui intègrent les fonctionnalités OLAP et SIG sont les plus adéquates à l'analyse spatiale en ligne. Les systèmes SOLAP sont très variés, et peuvent aller de la représentation sur cartes géographiques, au logiciel avec carte interactive interrogeant directement la base de données, et ce qu'elle soit spatiale ou spatiotemporelle.

Cependant, quelques limites ont été détectées relatives au manque des outils de datamining spatiale en vue de donner des interprétations exactes aux résultats d'analyses en ligne. Pour remédier à ces insuffisances, des efforts pour intégrer les techniques de fouille de données spatiales ont été fait et l'approche SOLAM a été proposée dans le chapitre 4.

Chapitre 2

L'Aide à la Décision Multicritères

Ce chapitre présente les fondements de base structurant l'analyse multicritères de la décision. Nous procédons à l'introduction des différentes notions relatives à la décision, l'aide à la décision, le processus décisionnel et ses éléments. Les différents types des problématiques décisionnelles sont aussi détaillés en présentant les méthodes les plus utilisées dans leur résolution.

Plan du chapitre

1 Introduction	26
2 Aide à la décision.....	26
3 Les acteurs d'aide à la décision	26
4 Décision et processus de décision.....	27
5 Concepts de base de l'aide à la décision multicritères	28
5.1 Le Processus de la décision	28
5.2 Le Eléments du processus décisionnel	30
5.3 Le Du monocritère au multicritères	33
5.4 Typologie des problématiques décisionnelle	34
5.5 Les problématiques de l'aide à la décision multicritères	34
6 Procédure d'agrégation	36
7 Les méthodes d'analyse multicritères.....	37
8 Travaux connexes	41
11Exemple d'application	41
12Conclusion.....	42

2.1 Introduction

Les méthodes multicritères sont des outils d'aide à la décision, leur développement a débuté dans le contexte militaire depuis les années 1960 pour deux essentielles raisons (Joe, 1998). L'amélioration de la gestion et la fourniture des moyens nécessaires pour les soldats. Les méthodes utilisées à l'époque sont issues du domaine de recherche opérationnelle. Ces méthodes permettaient d'optimiser une fonction tout en considérant un ensemble de contraintes prédéfinies. Par la suite, ces méthodes ont investi d'autres problématiques décisionnelles où le facteur humain a pris une dimension importante. Malheureusement, lorsque la décision concernait un système ouvert, qui intègre des dimensions de natures différentes telles qu'économique (optimisation de coût, de production) et sociale (acceptation d'un groupe, impact sur la santé, etc.), les méthodes de recherche opérationnelle ont montré certaines faiblesses auxquelles les méthodes multicritères semblent pallier.

Dans ce qui suit, nous détaillerons les différents concepts de l'aide à la décision multicritères (AMC) ainsi que les différentes méthodes d'analyse multicritères.

2.2 L'aide à la décision

L'aide à la décision est une approche constructiviste visant à fournir des outils permettant de progresser dans la résolution d'un problème où plusieurs points de vue, souvent contradictoires, doivent être pris en compte.

Vincke en 1989 (Vincke, 1989) définit l'aide multicritères à la décision de la façon suivante: « *L'aide multicritère à la décision vise, comme son nom l'indique, à fournir à un décideur des outils lui permettant de progresser dans la résolution du problème de décision ou plusieurs points de vue, souvent contradictoires, doivent être pris en compte* ». Cette analyse vise la résolution de problèmes avec plusieurs critères de décision simultanément. Ces derniers sont souvent conflictuels et d'importance inégale.

Selon (Roy et al, 1993), « *L'aide à la décision est l'activité de celui (homme d'étude) qui, prenant appui sur des modèles clairement explicités mais non nécessairement complètement formalisés, aide à obtenir des éléments de réponses aux questions que se pose un intervenant (décideur) dans un processus de décision, éléments concourant à éclairer la décision et normalement à recommander, ou simplement à favoriser, un comportement de nature à accroître la cohérence entre l'évolution du processus d'une part, les objectifs et le système de valeurs au service à partir desquels cet intervenant se trouve placé d'autre part* ».

Constatons que cette définition reste modeste quant à son objectif et qu'il ne s'agit pas d'une recherche de la "réalité". L'aide à la décision est donc un processus qui utilise un ensemble d'informations disponibles à un instant donné, afin de formuler un problème et aboutir à une décision sur un objet précis.

Selon (Hamdadou, 2016) « l'aide à la décision consiste à assister les décideurs et les aider à mieux exprimer leurs choix et préférences vis-à-vis une situation donnée ».

2.3 Les acteurs de l'aide à la décision.

« *Un individu ou un groupe d'individus est acteur d'un processus de décision si, par son système de valeurs, que se soit au premier degré du fait des intentions de cet individu ou groupe d'individus ou au second degré par la*

manière dont il fait intervenir ceux d'autres individus, il influence directement ou indirectement la décision » (Roy, 1985).

Pour qu'un groupe d'individus soit identifié comme un acteur unique, il faut que *«relativement à ce processus, les systèmes de valeurs des membres de ce groupe soient identiques* » (Roy, 1985).

Pour pouvoir garantir une aide convenable à la décision, on fait appel aux différents acteurs de la décision.

- **le décideur** : est la personne à qui s'adresse l'aide à la décision, il occupe une place centrale dans le processus d'étude dont les caractéristiques dépendent de ses attentes. Il s'agit parfois d'une entité un peu floue, mais son identification est primordiale.
- **L'intervenant** : c'est une personne qui cherche à influencer le décideur dans une phase du processus décisionnel, en raison de la nature de ses valeurs et donc en fonction de son système de préférences.
- **Les agis** : ils sont concernés par les conséquences de la décision, ils interviennent indirectement dans le processus par l'image que d'autres acteurs se font de leurs valeurs et plus concrètement de leurs système de préférences.
- **Les demandeurs** : ils demandent l'étude et allouent les moyens.
- **L'homme d'étude** : est un individu ou un groupe d'individus, qui a pour rôle d'établir un système de préférences, de définir le modèle d'aide à la décision, de l'exploiter afin d'obtenir des réponses et d'établir des recommandations pour conseiller le décideur sur les solutions envisageables. L'homme d'étude est à distinguer du négociateur et du médiateur.
- **Le négociateur** : mandaté par un décideur en vue de faire la position de celui-ci dans une négociation et de rechercher une action compromise.
- **Le médiateur** : intervient en vue d'aider les décideurs (ou les négociateurs) à rechercher une action compromise.
- **L'arbitre (Juge)** : intervient en se substituant aux acteurs dans la recherche d'une action compromise.

2.4 Décision et processus de décision

Nous décrivons dans cette partie, les concepts relatifs à la décision ainsi que ses niveaux et ses éléments ontologiques.

2.4.1 Décision

La décision est, avant tout, un choix que l'on doit réaliser devant différentes possibilités offertes et qui ne sont pas directement comparables ou qui présentent des aspects contradictoires. Selon plusieurs chercheurs, la décision est définie comme *« un choix entre plusieurs alternatives »*. Pour d'autres, la décision concerne aussi le processus de sélection de but et d'alternatives (Pictet, 1996).

Selon Laaribi (Laaribi, 1999), *« La décision est un choix entre actions, solutions, ou alternatives »*. Selon Marchais (Marchais,) *« la décision est une sélection d'une intention d'actes par un acteur »*. D'une manière

générale, « *la décision est l'action de l'esprit qui décide quelque chose ou se décide après délibération individuelle ou collective* ».

2.4.2 Les éléments ontologiques de la décision

La décision a comme propriétés, les éléments ontologiques suivants (Hamdadou, 2017) :

- **Objet de la décision** : but, programme, opération, instrument,.....
- **Organe de la décision** : organisation, groupe, individu,.....
- **Type de prise de décision** : routine, créatif, application d'un programme,....
- **Porté de décision** : stratégique, tactique, opérationnelle,...
- **Contrôle des éléments de la décision** : bonne, moyenne,.....

2.4.3 Les niveaux de la décision

Igor Ansoff (Ansoff et Anthonio, 2004), dans son livre « STRATEGIC MANAGMENT », a proposé une classification des décisions en fonction de leurs objectifs :

- **Les décisions stratégiques** (ou niveau stratégique) concernent les relations de l'entreprise avec son environnement (ex : décision de produire un nouveau type de voiture à destination des pays émergents).
- **Les décisions tactiques** (ou niveau tactique) sont relatives à la gestion des ressources (ex : décision d'acquérir un brevet, d'organiser une formation pour les salariés).
- **Les décisions opérationnelles** (ou niveau opérationnelle) portent sur l'exploitation courante (ex : établissement des plannings, décision de réapprovisionnement en pièces détachées).

2.5 Concepts de base de l'aide multicritères à la décision

Nous introduisons dans les sections suivantes, la notion du processus de la décision ainsi que ses éléments : les actions, les critères, la matrice de performance et les paramètres subjectifs.

2.5.1 Le processus de la décision

La décision ne peut être définie indépendamment de la notion de processus décisionnel. C'est pourquoi nous nous attacherons particulièrement à l'étude du processus de décision dans sa globalité. Cette optique permet en outre de mettre en évidence l'importance relative des différentes étapes composant ce processus.

2.5.1.1 Processus de SIMON

De manière générale, un processus de décision consiste à restreindre un ensemble de possibilités à un sous ensemble stricte et à évaluer cette restriction. SIMON (Simon, 1982).détaille ce processus en quatre étapes (Figure 2.1), non nécessairement séquentielles.

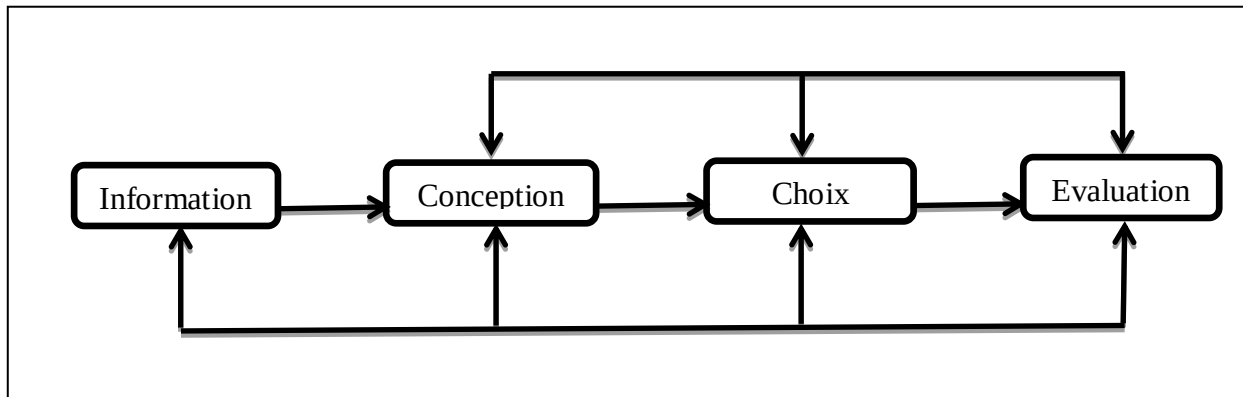


Figure 2.1 : Processus de SIMON

- **L'information** : c'est la phase qui détermine l'ensemble des données nécessaires mais pas forcément suffisantes qui seront utilisées lors des phases suivantes
- **La conception** : cette phase génère les différentes alternatives qui forment l'ensemble des possibilités. Les différentes solutions sont donc élaborées.
- **Le choix** : c'est la phase consistant à restreindre l'ensemble des possibilités aux sous-ensembles des possibilités sélectionnées.
- **L'évaluation** : en regard des trois phases précédentes, de la solution provisoirement retenue comme satisfaisante, cette phase peut amener à la réactivation de l'une des trois phases précédentes ou, au contraire, à la validation de la solution.

2.5.1.2 Approche décisionnelle de Pictet

Cette approche comprend trois phases principales (Pictet, 1996) : la structuration du modèle, son exploitation et la concrétisation des résultats. .

- **Phase de structuration du modèle** : elle a pour objectif l'identification du problème et les choix fondamentaux sur la manière de l'aborder.
- **Phase d'exploitation du modèle** : Sert à effectuer certains types d'analyses sur le modèle conçu telle que l'analyse spatiale (The, 96).
- **Phase de concrétisation** : vise essentiellement l'acceptation sociale du résultat, cependant elle comprend aussi la mise en œuvre de la décision et son contrôle.

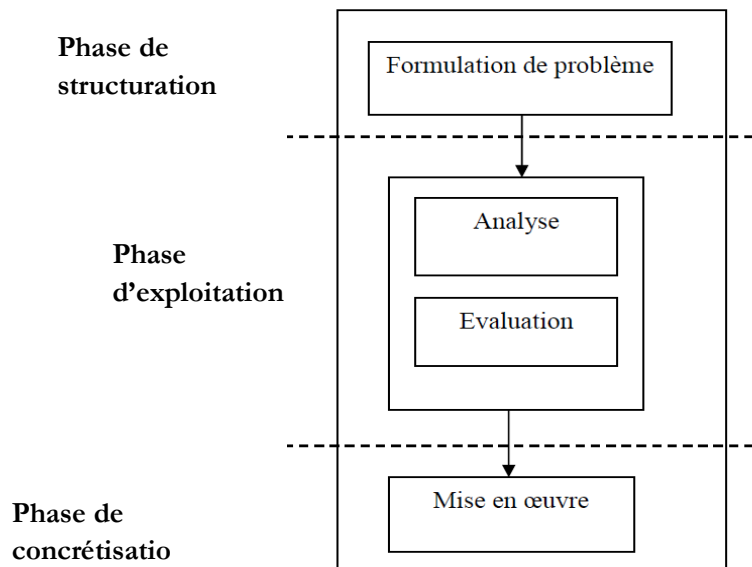


Figure 2.2 Processus de Pictet

2.5.2 Eléments du processus décisionnel

L'analyse multicritères nécessite de définir une certaine terminologie résumée comme suit :

- Un ensemble d'actions ou variantes en compétition.
- Une famille de critères, permettant de juger ces actions.
- Un tableau de performances des actions par critère.
- Une agrégation des résultats aboutissant à un classement par préférence.

L'analyse multicritères peut se faire à partir d'informations à la fois objectives et subjectives. Lors d'un exposé à l'école polytechnique Fédérale de Lausanne, le professeur Slowinsky a parfaitement exprimé le contenu subjectif de toute décision (Joerin, 1997) : « *un problème multicritère n'a pas de solution, si on n'apporte pas une information supplémentaire qui est la préférence du décideur* ». Cette sensibilité de la décision à des facteurs extérieurs explique sa composante subjective.

2.5.2.1 Les actions (les alternatives)

Une action est une représentation des éléments de la solution qui contribue à la décision par exemple (une zone industrielle, un site, etc.). C'est une éventuelle contribution à la décision globale (Roy, 1985) nommée action potentielle ou alternative une action réalisable, c.-à-d. une action dont la mise en œuvre en pratique est envisageable. Une action potentielle est une action réelle ou fictive provisoirement jugée réaliste par un acteur au moins susceptible d'être prise dans le processus décisionnel ; l'ensemble des actions potentielles sur lequel l'aide à la décision prend appui au cours d'une phase d'étude est noté $A = \{a_1, a_2, \dots, a_i, \dots, a_m\}$.

2.5.2.2 La famille de critères

Dans tout problème décisionnel multicritères, il convient de considérer un ensemble de critères que l'on nomme famille de critères et que l'on notera $F = \{g_1, \dots, g_m\}$ (on trouvera aussi la notation $F = \{1, \dots, m\}$). Pour que la famille F constitue une représentation appropriée des points de vue à prendre en compte dans la modélisation des préférences.

H.Laurent (Laurent, 2000) note que Roy définit la notion de famille Cohérente de critère à l'aide des trois propriétés suivantes. Ainsi une famille de critère sera dite cohérente si elle respecte :

- **L'exhaustivité** : il s'agit de ne pas oublier un critère pertinent pour la résolution de la problématique décisionnelle posée. C'est-à-dire quand les conséquences de deux actions sont identiques pour l'ensemble des critères en présence, il doit exister une relation d'indifférence entre ces deux actions.
- **La cohésion** : il doit y avoir une cohérence entre les préférences locales de chaque critère et les préférences globales. si une action a est égale à une action b pour tous les critères sauf un où elle lui est supérieure, ceci signifie que l'action a est globalement supérieure à l'action b.
- **L'indépendance** : il ne doit pas y avoir une redondance entre les critères. Leur nombre doit être tel que la suppression d'un des critères ne permet plus de satisfaire les deux conditions précédentes.

2.5.2.3 Les Critères

Face à un problème de décision, le décideur et l'homme d'étude sont amenés à juger et à évaluer les différentes actions potentielles. Pour rendre compte de tels jugements, on fait appel à la notion de critère qui peut être un facteur favorisant la prise de décisions ou une contrainte à éliminer. Les critères vont être le moyen utilisé pour décrire les actions par une expression quantitative ou qualitative qui permet d'examiner l'action. Formellement, on représentera les critères par des fonctions à valeurs réelles. On peut citer Vincke qui donne une définition concise des critères (Ginting, 2000).

Philippe Vincke (Vincke, 1989) a défini un critère ainsi : « *un critère est une fonction g , définie sur l'ensemble A des actions, qui prend ses valeurs dans un ensemble F totalement ordonné, et qui représente les préférences du décideur selon un point de vue* ». Lorsque le problème repose sur la considération de plusieurs critères (m critères), nous les notons $g_1...g_m$. Le problème posé est de la forme $G = \{g_1(x), g_2(x), ..., g_n(x)\} \times A = \{a_1(x), a_2(x), a_3(x), ..., a_m(x)\}$.

On peut conclure de cette définition formelle qu'il existe une action meilleure ou optimale par rapport à toutes les autres simultanément mais cette conclusion est subjective. L'optimisation dans ce cas éloigne le problème de son contexte.

Dans le contexte multicritères l'approche d'optimisation n'a pas de sens il ne s'agit pas de chercher une vérité cachée, mais plutôt d'aider le décideur à progresser vers une action de compromis (Chakhar et al, 2005).

2.5.2.4 La matrice de performances

Dans la terminologie introduite par (Vansnick, 1990), la formulation multicritère d'un problème décisionnel est définie par le modèle "A, A/F, E" où :

La formulation multicritères d'un problème décisionnel est définie par le modèle (A, F, E, PS) :

A : désigne l'ensemble des actions potentielles (envisageables, admissibles,...). Cet ensemble peut être défini :

- **Explicitement** (ensemble fini), les contraintes étant implicites.
- **Implicitement** (en général ensemble infini), les contraintes étant explicites. Dans ce deuxième cas, on a recours à la programmation mathématique à objectifs multiples (PMOM).

On désigne souvent l'ensemble des actions admissibles par le symbole X ;

- F : est l'ensemble fini des attributs ou critères, généralement conflictuels, à partir desquels les actions seront évaluées.
- E : est l'ensemble des évaluations de performances des actions selon chacun des attributs ou critères, c'est-à-dire l'ensemble des vecteurs de performances, un vecteur par action.
- PS : L'ensemble des paramètres subjectifs exprimés par le décideur.

La plupart des méthodes multicritères s'appuient sur la formulation précédente et comportent plus qu'une étape. Si l'ensemble des actions est implicite et défini par un ensemble de contraintes explicites, la méthode multicritère est généralement de type PMOM. Une partie importante des travaux en PMOM cherche à déterminer l'ensemble efficace. La détermination de cet ensemble ne fait pas intervenir les préférences du décideur, n'apporte pas une réponse définitive au problème de décision et n'est pas nécessairement une étape obligée. En effet, en général, l'ensemble efficace contient une infinité d'actions de sorte que l'on doit avoir recours à une procédure d'agrégation (Evans, 1984) pour dégager la ou les meilleures actions.

L'ensemble des évaluations de chaque action par rapport à chaque critère forme la matrice de performance illustré sur le tableau 3.1.

2.5.2.5 Les paramètres subjectifs

Vincke (Vincke, 1989) rappelle que la modélisation des préférences constitue une étape indispensable en aide à la décision. Lorsqu'un décideur est confronté à la comparaison de deux actions a_1 et a_2 il aura alors l'une des trois réactions (Joerin, 1997) :

- **Préférence** : pour l'une des deux actions notées a_1 p a_2 : si a_1 est préférée à a_2 cela signifie que l'une des deux actions est meilleure que l'autre.
- **Indifférence** : entre les deux actions notées a_1 I a_2 : signifie que les deux actions sont tellement proches qu'ils difficile de dire que est meilleure que l'autre.
- **Incomparabilité** : entre les deux actions notées a_1 R a_2 : signifie que les deux actions sont tellement différentes l'une de l'autre de qu'il est difficile de pouvoir les comparer.

Pour chaque action considérée, et pour chaque critère un seuil de préférence p , d'indifférence q et un seuil de veto v sont estimés. Chaque critère se voit attribué un poids k traduisant sa contribution dans la décision finale (Hamdadou, 2017).

- **Poids (k)** : qualifie l'importance relative d'un critère g_j donné vis à vis des autres critères de sa famille. Il est noté par $K = \{k_1, k_2, \dots, k_m\}$.
- **Seuil d'indifférence (q)** : Il s'agit de la plus petite différence qui est significative. En dessous de ce seuil, il est impossible de départager les deux actions. Le seuil d'indifférence noté par « q_j » représente la plus grande différence préservant une indifférence entre deux actions « a » et « b » selon le critère g_j .

- **Seuil de préférence (p)** : Il s'agit du seuil à partir duquel la différence entre deux actions est perceptible et fait préférer l'une à l'autre. Le seuil de préférence noté par « p_j » représente la plus petite différence compatible entre deux actions « a » et « b » avec une préférence en faveur de « a » selon le critère g_j .
- **Seuil de veto (v)** : Permet de fixer une notion supplémentaire. Si ce seuil est dépassé sur un critère, alors l'action ne peut être prise en considération. Il définit donc une situation intolérable pour un des décideurs. Il s'exprime par l'écart maximum acceptable autour de la valeur de l'évaluation. Le seuil de veto noté par « v_j » représente donc pour deux actions « a » et « b » la plus petite différence $g_j(b) - g_j(a)$ incompatible avec aSb où « S » est la relation de surclassement.

Le résultat de l'analyse des conséquences est présenté dans un tableau de performances illustré sur le tableau 2.1.

Tableau 2.1 Tableau des paramètres subjectifs

Paramètres subjectifs	g_1	g_2		G_j		g_n
Poids des criteres(k)	K1	K2		K_j		K_n
Seuil de préférence(p)	P1	P2		P_j		P_n
Seuil d'indifférence(q)	Q1	Q2		Q_j		Q_n
Seuil de Veto (v)	V1	V2		V_j		V_n
Actions (a_i)	Ensemble des évaluations E					
A1	$G_1(a_1)$	$G_2(a_1)$		$G_1(a_j)$		$G_1(a_1)$
A2	$G_2(a_2)$	$G_2(a_2)$		$G_2(a_j)$		$G_n(a_2)$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
A_i	$G_i(a_i)$	$G_i(a_i)$		$G_j(a_i)$		$G_n(a_i)$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
A_m	$G_1(a_m)$	$G_2(a_m)$		$G_j(a_m)$		$G_n(a_i)$

2.5.3 Du monocritère au multicritère

La solution d'un problème monocritères est facile à modéliser. Un problème de décision monocritère est de type $\text{Opt } \{g(x) : x \in A\}$ ou A : Ensemble des actions admissibles, G : la fonction critère à optimiser. Dans cette définition la comparaison de plusieurs actions se fait suivant un seul critère alors que d'après (Mayster et al, 1994) :

- La comparaison de plusieurs actions se fait rarement suivant un seul critère.
- Les préférences sur un critère sont dans bien de cas, difficilement modélisable par une fonction.
- Lorsqu'il y a plusieurs objectifs, il est impossible de les atteindre tous à la fois.

La recherche opérationnelle pour laquelle on fixe un objectif ambitieux qui est la désignation en toutes circonstances de la meilleure décision (L'optimum) mène à une certaine exclusivité des décisions alors que le domaine nécessite une certaine complémentarité. Ben mena (BenMena, 2000) parle de la misère de l'optimisation par contre au bonheur du multicritères et cite l'exemple suivant : Si pour acheter une voiture, un individu ne considère que l'aspect financier tout le monde roule en 2CV. Le normal pour cet individu c'est d'énumérer les critères suivants : cout, confort, sportivité, sécurité, économie à l'usage, environnement. Certains de ces critères sont contradictoires.

2.5.4 Typologie des problématiques décisionnelles

Les problématiques décisionnelles sont distinguées en trois types, à savoir :

- **Problème structuré** : le problème est posé d'une façon claire, les données sont fiables et numériques. Les modèles associés sont clairement définis, ils permettent une résolution algorithmique programmable du processus décisionnel.
- **Problème semi structuré** : le problème peut ne pas être clairement posé, les données sont souvent qualitatives et la résolution est difficilement exprimable sous forme de programme.
- **Problème non structuré** : le problème n'est pas clairement posé, le principe de la rationalité limitée s'applique à toutes les étapes du processus décisionnel. La décision prise est difficilement justifiable.

2.5.5 Les problématiques de l'aide à la décision multicritère

Roy (Roy, 1997) définit quatre problématiques de référence, le choix, le tri, le rangement et la description.

2.5.5.1 Problématique du Choix

Il s'agit de la problématique la plus classique en aide à la décision. Elle consiste à sélectionner un sous-ensemble aussi restreint que possible d'actions A dans le cas le plus favorable à un singleton d'un ensemble A qui justifie l'élimination des autres actions. Cette problématique généralise la problématique de la recherche opérationnelle. Elle aboutit à la mise au point d'une procédure de sélection.

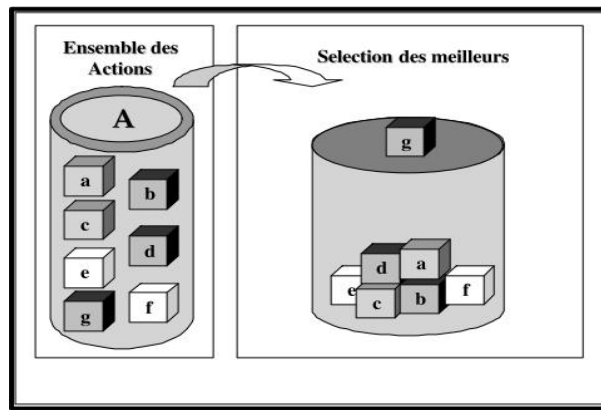


Figure 2.3 La problématique du choix. (Ginting, 2000)

2.5.5.2 Problématique du Tri

Cette problématique consiste à affecter les actions de A à des catégories prédéfinies (caractérisées par exemple par des actions de référence). Contrairement aux autres problématiques, on ne compare pas les actions de A entre elles mais on se fonde uniquement sur les comparaisons des actions de A aux actions de référence, il s'agit d'une évaluation intrinsèque des actions. Ici, on parlera de procédure d'affectation à des catégories.

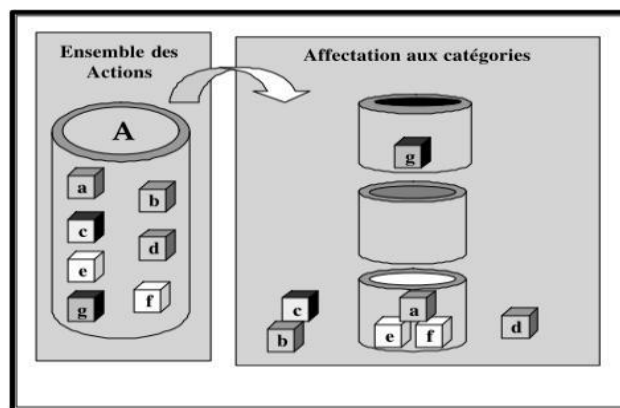


Figure 2.4 La problématique du tri. (Ginting, 2000)

2.5.5.3 Problématique de Rangement

Il s'agit sans doute de la problématique la plus ambitieuse. Elle a pour objectif d'ordonner les actions de A. Cependant, on ne recherche pas nécessairement un ordre complet sur les actions au sens d'une relation de préférence globale ; on cherche plutôt à regrouper les actions en classes d'équivalence, celles-ci étant totalement ou partiellement ordonnées. La procédure recherchée est une procédure de classement.

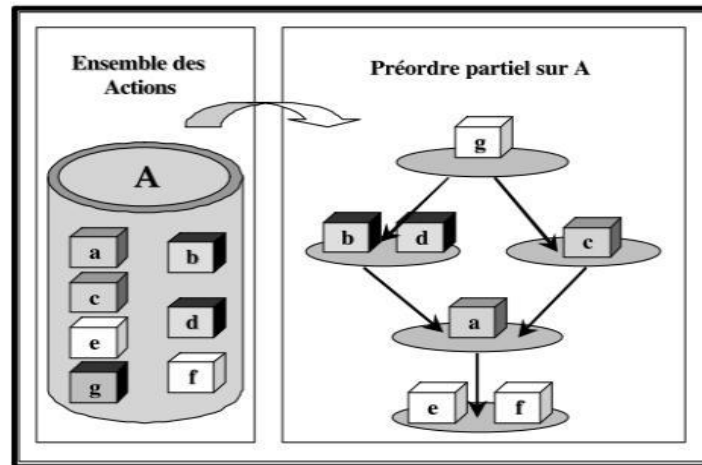


Figure 2.5 La problématique du rangement. (Ginting, 2000)

2.5.5.4 Problématique de la Description

Cette problématique fait généralement partie de la première phase d'analyse des problématiques précédentes pour améliorer la compréhension d'ensemble de différentes actions afin de mettre en contexte les conséquences possibles des interventions. Elle consiste juste à éclairer l'analyse des actions en aidant le décideur à appréhender celles-ci. Cela passe par la définition des conséquences élémentaires et des critères, ainsi que par le choix ultérieur d'une autre problématique.

2.6 Procédures d'agrégation

Il s'agit d'établir un modèle des préférences globales, c'est-à-dire une représentation formalisée de telles préférences relativement à un ensemble A d'actions potentielles, que l'homme d'étude juge appropriée au problème d'aide à la décision. Selon la méthode d'agrégation choisie, on choisit la méthode multicritère appropriée. Les spécialistes de l'aide multicritères à la décision distinguent les méthodes multicritères en trois classes (Roy, 1985) :

- **Agrégation totale** : Roy (Roy, 1985) et les auteurs dans (Mayster et al., 1994) l'appellent encore « approche du critère unique de synthèse évacuant toute incomparabilité » et que Vincke (Vincke, 1989) nomme cette approche « théorie de l'utilité multi attribut ». La technique consiste à introduire toutes les performances dans une seule fonction en leur attribuant des poids et suppose que tous les jugements sont commensurables alors qu'une des justifications de l'approche multicritère est le non commensurabilité de ces jugements.
- **Agrégation partielle** : D'autre (Mayster et al, 1994) l'appellent encore « approche du sur classement de synthèse acceptant l'incomparabilité » ou « méthodes de sur classement » (Vincke, 1989). repose sur la comparaison des actions deux à deux puis une synthèse des résultats de comparaison. Cette technique ne permet pas d'obtenir un résultat indiscutable comme on s'y attend en général, pour tout ce qui est à base de mathématique (Scharling, 1985). Cette méthode correspond aux situations où il existe un critère en moins n'est pas quantitatif ou dans le cas où les unités des critères sont très hétérogènes (différentes échelles) (Mousseau, 1993). Des indices de concordance, de discordance ou de crédibilité sont utilisés.

- **Agrégation locale** : Les deux types d'agrégation totale et partielle supposent que A soit de dimensions raisonnables. Or A peut être très grand, voire infini lorsque les actions varient en continu. Cette méthode d'agrégation correspond donc aux situations où il existe un nombre quasi infini de variantes (Scharling, 1985). La technique consiste alors à partir d'une solution de départ procède à explorer localement et itérativement s'il existe de meilleure variante par rapport à celle sélectionnée. On pratique donc une exploration locale et répétitive de A, d'où son nom technique « agrégation locale itérative avec itérations essai-erreur » (Vincke, 1989).

Le tableau 2.2 résume les différentes méthodes existantes et correspondantes à chaque procédure d'agrégation.

Tableau 2.2 Tableau de procédures d'agrégation

Procédure d'agrégation	Définition	Méthodes
Agrégation complète transitive	Approche du critère unique	– Goal-Programming – MAUT (Multiple Attribute Utility Theory) – UTA (Utilité Additives) – AHP (Analytic Hierarchy Process) – WPM (Weight Product Method) – WSM (Weight Sum Method)
Agrégation partielle	Approche du surclassement de synthèse	– ELECTRE (Elimination Et Choix Traduisant la REalité) – promethee (Preference Ranking Organisation METHod for Enrichment Evanuation)
Agrégation locale	Approche du jugement local interactif avec itération essai-erreur	– PLM (Programmation Lineaire Multicritere) – Stem (Pop) – UTA interactive – Prefcale

2.7 Les méthodes d'analyse multicritères

Toutes les méthodes multicritères ne permettent pas obligatoirement de résoudre n'importe quel problème d'aide à la décision. Chacune propose des modalités particulières (choix, tri, rangement). Nous pouvons recenser un très grand nombre de méthodes multicritères. La plupart de ces méthodes appartiennent à l'une ou l'autre des trois approches opérationnelles citées ci-dessus.

2.7.1 La famille PROMETHEE

Les méthodes PROMETHEE (Preference Ranking Organisation Method for Enrichment Evaluation) méthode d'organisation de classement de préférence pour évaluation d'enrichissement, sont des méthodes de surclassement basées sur les théories de (Roy, 1993). L'objectif principal revendiqué par la méthode PROMETHEE est d'être facilement compréhensible pour le décideur.

2.7.2 AHP (Analytical Hierarchy Process)

Elle traite la problématique du choix. Le point de départ de cette méthode est de définir une arborescence hiérarchique de critères et sous-critères. La représentation de l'arborescence se fait sous la forme de couples (pères, fils). La somme des poids de tous les critères fils d'un même père égale à 1. On compare ensuite deux à deux les différentes branches de même niveau, en attribuant une note chiffrée sur une échelle prédéfinie à la façon dont on ressent la différence entre les deux critères. Mais l'échelle trop restreinte des valeurs autorisées peut conduire à des erreurs (Saaty et al, 1984).

2.7.3 MAUT (Multi Attribute Unity Theory)

Elle traite les problèmes de choix ou de tri et utilise les méthodes de construction des fonctions d'utilité partielle pour chaque critère, la meilleure action sur un critère aura une valeur partielle égale à 1 et la plus mauvaise aura 0, et en posant des questions au décideur sur les valeurs entre 0 et 1, nous construisons les fonctions de valeurs partielles. Ensuite nous construisons la fonction d'utilité, et ce en fonction des préférences du décideur qui doit associer une utilité à chacune des actions considérées (AnandR, 1995).

2.7.4 GAIA (Geometrical Analysis for Interactive Assistance)

Cette méthode cherche à décrire la structure du problème par le passage des évaluations aux flux uni critère. Chaque action est un point dans l'espace des critères en fin elle utilise l'ACP pour projeter les points de façon optimale (Brans et Mareschal, 2001).

2.7.5 La famille ELECTRE

Les méthodes ELECTRE (ELimination Et Choix Traduisant la REalité), ont été développées par Bernard Roy au début des années 1970. Il a ainsi initié toute une série de méthodes, dites de surclassement, basées sur des comparaisons d'actions deux à deux. Celles-ci demandent peu d'information pour pouvoir être implémentées ; de plus cette information est facilement accessible au décideur (Hamdadou, 2008).

2.7.5.1 ELECTRE I

ELECTRE I (Roy, 1968) est la plus ancienne des méthodes ELECTRE, elle révèle de la problématique de *choix*, dite alpha (α). Celle-ci consiste à chercher dans l'ensemble des actions envisagées, un sous ensemble contenant les "meilleures" actions ou à défaut, les actions "les plus satisfaisantes". On soumet le surclassement à une condition de *concordance*, condition imposant

qu'une certaine majorité de critères se dégage en faveur de l'action surclassante ; et on le soumet en même temps à une condition de *non-discordance*, condition imposant qu'il n'existe pas une trop forte pression, dans un des critères de la minorité, en faveur du sur classement inverse.

2.7.5.2 ELECTRE II

Avec l'aide de P. BERTIER, Bernard Roy a ensuite développé la méthode ELECTRE II (Roy et Bertier 1971). La méthode ELECTRE II relève de la problématique gama (γ) (procédure de *classement*). Les résultats obtenus dans ELECTRE II sont plus tranchés que dans la méthode précédente, car il y a presque toujours une meilleure action qui se dégage du classement obtenu. ELECTRE II a pour objectif, en utilisant les relations d'ordre sur chacun des critères, de classer les variantes depuis les meilleurs jusqu'aux moins bonnes, tout en tolérant les ex-aequo (Mayster et al, 1994). Les actions ne sont pas qualifiées en fonction de leur valeur propre mais en valeur relative par rapport aux autres actions. Cette méthode utilise, tout comme la méthode ELECTRE I, une vérification de la concordance et de la non-discordance avec l'hypothèse de sur classement aiSak. Une distinction est néanmoins réalisée, afin de mieux respecter les nuances du réel, en introduisant une notion de respect de cette hypothèse de surclassement qui est plus floue, entre la rigidité d' ELECTRE I et le flou total d' ELECTRE III. On trouve ainsi un sur classement fort aiSFak, qui repose sur des bases solides (certitude forte sur l'hypothèse), et un sur classement faible aiSfak, qui est plus sujet à débattre (certitude faible sur l'hypothèse). Les critères utilisés dans la méthode ELECTRE II sont francs et ne comportent pas de seuil de veto.

2.7.5.3 ELECTRE III

Tout comme ELECTRE II, ELECTRE III, répond à la problématique de rangement. Cependant, ELECTRE III ne cherche pas seulement deux espèces de sur classement, les forts et les faibles, mais examine toute une famille qui va du totalement fort au totalement faible ou inexistant, en passant par toutes les nuances que permet l'échelle continue entre ces deux extrêmes. De plus, ELECTRE III étudie les sur classements tout en se demandant quel est le degré de crédibilité (Pha, 1990). Cette méthode utilise une relation de surclassement évaluée et une procédure d'exploitation du graphe évalué utilisant des techniques proches de celles de manipulation des nombres flous pour modéliser les pseudo-critères. ELECTRE III a perdu de la simplicité des méthodes originelles et elle ne rencontre plus guère de succès vu sa grande complexité et un recours important aux valeurs numériques qui la rapproche des méthodes d'utilité.

2.7.5.4 ELECTRE IV

La méthode ELECTRE IV, (Roy et al, 1982) relève aussi de la problématique γ . Sa démarche d'utilisation constitue en une simplification de la méthode ELECTRE III par l'abandon de la pondération des critères. Elle comporte deux relations de surclassement comme ELECTRE II mais un seul jeu de seuils de veto et la notion de concordance est traduite par une notion de majorité de critères en l'absence de toute pondération. Son originalité réside dans le fait qu'il n'y a plus de poids attribués aux critères, seule une hypothèse de disparité limitée étant fixée. Ce changement fondamental est accompagné d'une autre grande nouveauté : l'abandon de l'hypothèse de surclassement, qui rend inutiles les notions de concordance et de discordance.

ELECTRE IV utilise comme ELECTRE III, des pseudo-critères (critères flous) de même importance, c'est à dire des critères associés à un seuil de préférence strict et à un seuil d'indifférence. Cette méthode admet plusieurs versions des types de sur classement entre deux actions. ELECTRE IV se base sur les nuances développées dans ELECTRE III mais simplifie fortement le processus d'étude qui s'ensuit.

2.7.5.5 MELCHIOR

Cette dernière méthode est une méthode proche de ELECTRE II mais purement ordinale : elle utilise des pseudo-critères et utilise une relation d'ordre au lieu de poids pour refléter l'importance relative des critères. ELECTRE IV apparaît comme un cas particulier de MELCHIOR lorsque la relation d'ordre sur les critères est une équivalence (JP Leclercq, 1984).

2.7.5.6 ELECTRE IS

La méthode ELECTRE IS (Roy et al, 1985) est une adaptation d' ELECTRE I (problématique α) à la logique floue, permettant d'utiliser des pseudo-critères (le S veut dire seuil). Comme dans ELECTRE III, ELECTRE IS utilise les seuils d'indifférence Si, de préférence Sp et de veto Sv.

2.7.5.7 ELECTRE TRI

La méthode ELECTRE TRI (Roy et al, 1985) relève de la problématique β (procédure d'affectation) : le problème est posé en termes d'attribution de chaque action à une catégorie prédéfinie. Des actions de préférences sont utilisées pour segmenter l'espace des critères en catégories, chaque catégorie est bornée inférieurement et supérieurement par deux actions de référence et chaque action de référence sert donc à deux catégories, l'une supérieure, l'autre inférieure. Cette méthode présente trois intérêts principaux permettant :

- De juger une action potentielle pour elle-même, indépendamment des autres actions potentielles ; en ce sens, cette méthode juge chaque action potentielle sur sa valeur absolue (bien que relativement aux actions de référence prédéfinies) ;
- De fixer une ou plusieurs valeurs de référence, par exemple des normes légales ou des résultats minimaux pour l'acceptation de candidats ;
- De considérer un nombre d'actions potentielles plus important que les autres méthodes ELECTRE.

Le tableau 2.3 résume les problématiques de références et leurs méthodes correspondantes.

Tableau 2.3 Problématiques de référence

Problématique	Objectif	Résultat	Procédure	Méthode
α (alpha)	Choix d'un sous ensemble contenant les actions « les meilleures » ou, à défaut. « satisfaisantes »	Choix	Sélection	ELECTRE I ELECTRE IS
β (Béta)	Tri par affectation des actions à	Tri	Affectation	ELECTRE TRI

	des catégories prédéfinies			
γ (Gama)	Rangement de classes d'équivalence, composées d'actions, ces classes étant ordonnées de façon complète ou partielle	Rangement	Classement	ELECTRE II ELECTRE III ELECTRE IV PROMETHEE
δ (Delta)	Description, dans un langage approprié, des actions et de leurs conséquences	Description	Cognition	

2.8 Travaux connexes

Dans cette section, nous citons les travaux réalisés dans le domaine de l'aide multicritères à la décision réalisés dans différents domaines. Dans (Hamdadou et al, 2007), un modèle d'aide à la décision est conçu pour optimiser la qualité de la décision dans le contexte de l'aménagement du territoire basé sur une nouvelle stratégie de décision autorisant les interactions entre critères par intégrale de Choquet. Dans (Younsi et Hamdadou, 2009), un processus décisionnel a été proposé pour l'aménagement du territoire en exploitant les SIGs et les techniques de télédétection. Ces dernières ont été agrégées et analysées par utilisation des méthodes d'aide multicritère à la décision (AMCD) à savoir la méthode ELECTRE III. Un autre système d'aide à la décision multicritère a été développé dans le domaine industriel, par (Hamdadou et Bouamrane, 2011) pour aider les décideurs en maintenance, selon leurs préférences souvent conflictuelles, à adopter une méthode de diagnostic rapide et efficace. Le système d'aide au diagnostic de groupe (DIAG-GDSS) utilise un protocole de négociation qui se base sur la méthode ELECTRE III. Dans une autre étude proposée par (Bouamrane et al, 2012), un système de gestion de risque basé sur l'analyse multicritères a été conçu pour identifier le risque du transport routier de matières dangereuses en définissant les voies les moins risquées dans la zones urbaine d'Oran. Dans (Hamdadou et Bouamrane, 2016), un système d'aide à la décision multicritère de groupe dans le domaine de l'aménagement de territoire appelé SIGMAS, a été proposé afin de faciliter le choix de terrain adéquat pour une construction. Les auteurs ont fait appel à deux méthodes d'ELECTRE à savoir, ELECTRE III et ELECTRE TRI.

2.9 Exemples d'application

Le système « EPISOLAP » décrit dans le chapitre 1 avait comme objectif la détection et la localisation des foyers d'épidémie. Or, dans le cas où ces foyers sont nombreux, il faut pouvoir établir des priorités afin de traiter les foyers les plus à risque. Nous avons donc proposé dans (Zemri et Hamdadou, 2014) d'effectuer un classement de ces foyers en utilisant les différents critères qui rentrent dans l'identification du risque sanitaire. La conception d'un système d'aide à la décision spatial basé sur une telle intégration permettrait de cerner le phénomène de la propagation épidémiologique, d'améliorer la surveillance et enfin mieux contrôler les actions des décideurs. Nous cherchons à travers cette étude, à donner plus d'explications sur les facteurs réels et responsables de la propagation des épidémies ainsi que de son émergence ou réémergence. En fin, cette étude aboutit à la génération automatique d'une cartographie de risque sanitaire qui donne un classement des foyers d'épidémies pour faciliter l'intervention des

épidémiologistes par ordre de priorité. Les étapes détaillées de conception et de développement de cette application multicritère sont illustrées dans l'annexe B de cette thèse.

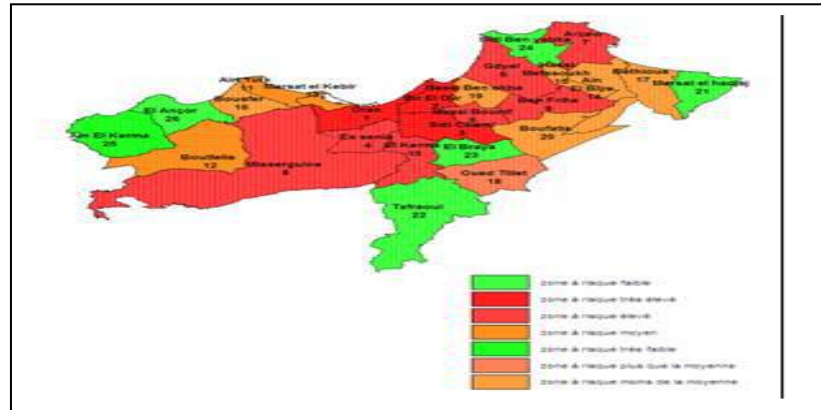


Figure 2.6 Carte de risque tuberculeux avec Map Info Professional 11.0

2.10 Conclusion

Les méthodes d'analyse multicritères sont des méthodes formelles qui ont prouvé leur efficacité dans beaucoup de domaines principalement le domaine spatial et ont démontré leur capacité à cerner, par excellence, des problèmes spatiaux. Les méthodes multicritères représentent une partie intégrante de la théorie d'aide à la décision, elles ont fait leurs preuves et ont déjà exigé leur indispensabilité dans le domaine de la décision. Ces méthodes sont très efficaces pour résoudre les problèmes multicritères réelles et permettent d'avoir des résultats fiables.

Ces méthodes ont été appliquées dans les différentes études menées dans notre équipe depuis une décennie et dans des domaines différents (Transport, Aménagement du territoire, Gestion de la production et dans le domaine industriel). Ceci nous a incités à concevoir un nouveau système d'aide à la décision spatial basé sur l'intégration de la technologie SOLAP et l'analyse multicritères d'une part et les techniques d'exploration de données d'autre part dans le système SDSS_MCA_DM dont l'objectif est de cerner le phénomène de la propagation épidémiologique et rendre ce problème contrôlable et, par conséquent, assurer une surveillance épidémiologique accrue et efficace. Le développement de cette idée ainsi que la formulation du problème multicritères en santé environnementale seront détaillés dans le chapitre 5.

Chapitre 3

Fouille de Données Spatiales

L'objectif du chapitre est de présenter les différents concepts relatifs à la fouille de données en générale et la fouille de données spatiale en particulier selon une démarche structurée et systématique. Nous présentons les différentes méthodes existantes à travers les différentes phases de cette démarche générique et fondamentale destinée pour être appliquée dans toutes études qui visent à analyser l'information spatiale dans les domaines de recherche fondamentale et appliquée.

Plan du chapitre

1 Introduction	44
2 Méthodologie de fouille de données spatiales.....	44
3 Particularité des données à référence spatiale	46
4 Méthodes utilisées dans la fouille de données spatiales	47
4.1 Phase descriptive	47
4.2 Phase exploratoire	48
4.3 Phase explicative (décisionnelle)	55
5 Classification des méthodes de fouille de données spatiales.....	59
5.1 Méthodes monothématiques.....	59
5.2 Méthodes multithématiques.....	60
6 Travaux connexes.....	60
6.1 Approches basées sur le prétraitement de données	60
6.2 Approches basées sur le traitement dynamique de l'information spatiale.....	61
7 Conclusion	61

3.1 Introduction

L'analyse spatiale combine des approches statistiques exploratoires et la fouille de données suit une démarche d'abord **descriptive** (en appliquant des méthodes de statistiques descriptives) puis **exploratoire** (en analysant les données monothématiques) et en fin explicative ou prédictive (sur des données multithématiques) dans un but d'aide à la décision. L'analyse spatiale exploratoire s'articule autour de trois composantes qui sont (Morency, 2006) : i) **la visualisation** qui permet de représenter les données spatiales à travers un système d'information géographique, ii) **l'analyse spatiale exploratoire** de données qui permet d'explorer et de synthétiser les données afin de révéler des configurations spatiales particulières (patrons, relations spatiales) et iii) **la modélisation spatiale** qui tente d'expliquer les configurations par la spécification d'un modèle statistique et l'estimation des paramètres.

L'exemple le plus connu d'analyse spatiale exploratoire est celui du Dr Snow, lequel a découvert la cause de la maladie du choléra à Londres en 1854, grâce à une cartographie et une analyse visuelle de la relation entre les cas de décès et les puits d'eau. À travers cet exemple (Zeitouni, 2006) distinguent la différence entre l'analyse spatiale et la fouille de données spatiales.

Tableau 3.1 Comparaison entre l'analyse spatiale et la Fouille de données spatiale

Analyse cartographique	Fouille de données spatiales
découverte visuelle	découverte automatique
inapplicable sur de gros volumes de données	opère sur de gros volumes de données
Confirmatoire	Exploratoire (génère des hypothèses)

La fouille de données spatiales se distingue alors de l'analyse spatiale par le fait qu'elle peut générer automatiquement à partir de gros volumes de données, des hypothèses par des règles caractéristiques ou bien des règles de classement ou encore des règles d'association.

Nous constatons que les étapes d'analyse spatiale exploratoire et celles de la fouille de données spatiales (FDS) se chevauchent pour conduire à une démarche systémique et décisionnelle.

3.2 Méthodologie de fouille de données spatiales

La démarche systémique que nous avons adoptée est largement inspirée des travaux de (Zeitouni et al, 1999). La figure 3.1 illustre cette démarche.

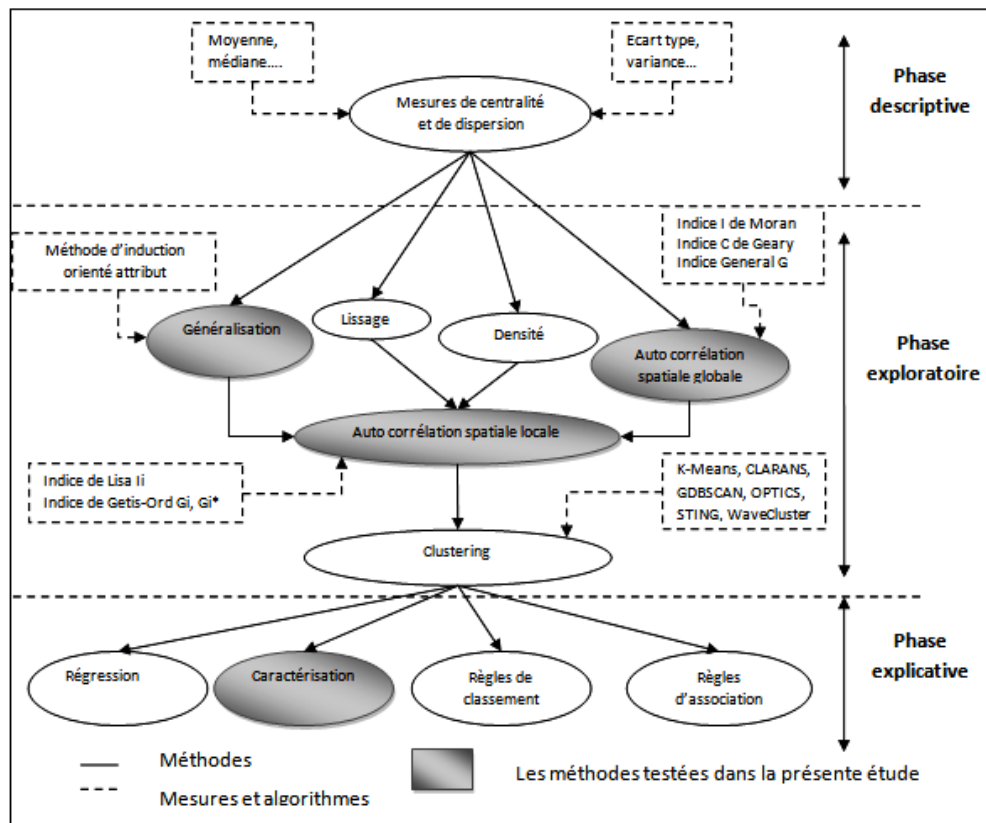


Figure 3.1 Démarche systémique de la fouille de données spatiale

La figure 3.1 représente une méthodologie générique pour toute étude spatiale qui vise à analyser l'information géographique dans différents domaines de recherche fondamentale et appliquée. La démarche est composée de trois phases principales qui sont : une phase descriptive, une autre exploratoire et une dernière explicative, voir prédictive.

3.2.1 Phase descriptive

Cette phase utilise les mesures de statistiques descriptives usuelles qui sont la moyenne et la variance. Les mesures de centralité (centre moyen, centre de gravité) et de dispersion (écart type, écart type de la distance, ellipse de dispersion) permettent de comparer rapidement des nuages de points propres à des segments clés de la population. L'effet discriminant ou non de l'espace peut être observé.

3.2.2 Phase exploratoire

Cette phase opère sur des données monorhèmes et se décompose en trois sous étapes :

Analyse globale : donne une description synthétique pour faire apparaître une tendance générale de la distribution spatiale en appliquant l'autocorrélation globale qui peut être mesurée par : l'indice de Moran I (Moran, 1950), l'indice de Geary C (Geary, 1954), l'indice de Getis-Ord General G (Getis et Ord, 1992) ou par analyse basée sur la densité, analyse multidimensionnelle lissée ou encore par généralisation.

Analyse locale : découvre les spécificités locales des sous ensembles d'objets par les mesures de l'autocorrélation locales (Indice de LISA (Anselin, 1995) ou indice de Getis-Ord Gi (Getis et Ord, 1992), Gi* (Ord et al, 1995)) ou analyse factorielle.

Clustering: permet de rassembler des agrégats homogènes partageant des caractéristiques similaires (K-Means (Anil, 2010), K-medoids (Kaufman, 1987) CLARANS (Han, 1994), GDBSCAN (Sander et al, 1998), OPTICS (Ankers, 1999), STING (Wang, 1997), WaveCluster (Gholamhosein, 1998)).

3.2.3 Phase explicative

Cette phase opère sur plusieurs couches thématiques et permet d'expliquer la particularité des agrégats découverts dans la phase exploratoire par des règles caractéristiques (Ester et al, 1998), des règles de classement (Ester et al, 1997) ou règles d'association (Koperski et Han, 1995) qui permettent éventuellement la prédiction de nouvelles données et la prise de décision.

3.3 Particularité des données à référence spatiale

Les données à référence spatiale sont de nature particulière et présentent certaines spécificités et certains problèmes relatifs à leur gestion et leur manipulation. Nous citons :

L'hétérogénéité : les données à référence spatiale sont de nature hétérogène. Cela signifie qu'il faut considérer les caractéristiques particulières de chaque zone géographique. Le problème de l'hétérogénéité spatiale traite le phénomène d'instabilité structurelle qui se traduit de deux façons :

- **Des coefficients différents :** l'instabilité spatiale des paramètres se réfère à des comportements ou à des relations hétérogènes dans l'espace.
- **Des variances différentes des termes d'erreurs** (problème d'hétéroscédasticité). (Jayet, 2001) distingue différents types d'hétérogénéité : *i) de taille* (ordre de grandeur des agrégats), *ii) de forme ou de position* (morphologie spatiale, urbain/périurbain) et *iii) de structure* (socio-démographie de la population, niveau de scolarité, revenus).

L'interaction spatiale : les relations spatiales jouent un rôle important en géographie car elles mettent en évidence l'influence du voisinage qui est la principale spécificité des données géographiques. La notion de dépendance spatiale est endémique aux données spatiales. En effet *ce qui se passe dans une localité particulière dépend de ce qui se passe dans d'autres localités et ces interactions sont d'autant plus fortes que les localités concernées sont plus proches* (Tobler, 1970).

Sophisme écologique : l'action d'inférer que les relations à l'échelle individuelle sont les mêmes que celles observées à un niveau agrégé est appelée erreur d'inférence due à l'absence de distinction entre différents niveaux d'organisation (Guide, 2004).

Le problème de la zone modifiable : le problème de la zone modifiable ou MAUP (Modifiable Areal Unit Problem) est un phénomène par lequel des résultats différents sont obtenus à partir de l'analyse des mêmes données, groupés dans différents ensembles d'unités surfaciques. Le MAUP

peut être subdivisé en deux effets distincts qui se produisent habituellement simultanément lors de l'analyse des données agrégées :

- L'effet de différentes échelles ce qui engendre différents niveaux d'agrégation.
- L'effet de regroupement des données selon différentes configurations.

3.4 Méthodes utilisées dans la fouille de données spatiales

Dans cette section nous analysons en détails, les différentes étapes de la méthodologie de la fouille de données adoptée ci dessus. Nous donnons des exemples de mesures et d'algorithmes utilisés dans chacune des étapes.

3.4.1 Phase descriptive

Pour mieux comprendre les données, plusieurs mesures sont appliquées afin d'identifier la tendance centrale, les variations et la dispersion des données géographiques. Nous distinguons les mesures de centralité et de dispersion (Morency, 2006) :

3.4.1.1 Les mesures de centralité

- **Centre moyen** : c'est l'équivalent de la moyenne, en deux dimensions.

$$x_{cm} = \frac{\sum_{i=1}^n x_i}{n} \quad (5); \quad y_{cm} = \frac{\sum_{i=1}^n y_i}{n} \quad (6)$$

- **Centre de gravité** : C'est le centre moyen avec une intensité observée en un point particulier de l'espace (le poids).

$$x_{cg} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad (7); \quad y_{cg} = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} \quad (8)$$

3.4.1.2 Les mesures de dispersion

- **L'écart type** : La première mesure de dispersion consiste à mesurer l'écart type (Standard Déviation) dans chacun des axes (x, y), de façon distincte :

$$s_x = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}} \quad (9); \quad s_y = \sqrt{\sum_{i=1}^n \frac{(y_i - \bar{y})^2}{n-1}} \quad (10)$$

- **L'écart type de la distance** : Cette statistique mesure l'écart type de la distance de chaque point par rapport au centre moyen. C'est donc l'équivalent, en deux dimensions, de l'écart type classique d'une distribution.

$$s_{xy} = \sqrt{\sum_{i=1}^n \frac{(d_{icm})^2}{n-2}} \quad (11)$$

d_{icm} : La distance entre chaque point et le centre moyen

- **L'ellipse de dispersion** : permet de tenir compte de la variance d'état selon la direction considérée (anisotropie) par la définition d'une ellipse. L'orientation de l'ellipse permet d'identifier les axes de dispersion maximale et minimale.

- l'axe Y est pivoté d'un angle θ .

$$\theta = \frac{ARCTAN\{[\sum_{i=1}^n (xi - \bar{x})^2 - \sum_{i=1}^n (yi - \bar{y})^2] + [(\sum_{i=1}^n (xi - \bar{x}) - \sum_{i=1}^n (yi - \bar{y}))^2]^{1/2} + 4(\sum_{i=1}^n (xi - \bar{x})(yi - \bar{y}))^{1/2}\}}{2 \sum_{i=1}^n (xi - \bar{x})(yi - \bar{y})} \quad (12)$$

- les deux écarts-types sont calculés

$$Sx = \sqrt{\frac{\sum_{i=1}^n ((xi - \bar{x}) \cos \theta - (yi - \bar{y}) \sin \theta)^2}{n-2}} \quad (13); \quad Sy = \sqrt{\frac{\sum_{i=1}^n ((xi - \bar{x}) \sin \theta - (yi - \bar{y}) \cos \theta)^2}{n-2}} \quad (14)$$

- la longueur de chaque axe de l'ellipse correspond à deux fois l'écart type

3.4.2 Phase exploratoire

Cette phase comporte les méthodes de géostatistiques (autocorrélation spatiale globale et locale), le lissage multidimensionnel, l'analyse de la densité et la méthode de la généralisation spatiale. Une attention particulière est donnée aux méthodes d'autocorrélation spatiale et la méthode de généralisation car elles seront utilisées par la suite dans le développement de notre contribution.

3.4.2.1 L'autocorrélation spatiale

L'autocorrélation spatiale est définie comme « la corrélation d'une variable avec elle-même (autocorrélation) attribuable à l'organisation géographique des données (spatiale) » (Griffith, 1991).

Les mesures d'autocorrélation spatiale évaluent le degré de similarité entre observations en fonction de leur éloignement (voisinage). On reprend ici la première loi en géographie définie par Tobler's (Tobler, 1970) : « *Tout est lié à tout le reste, mais les zones proches sont plus liées que les zones éloignées* » (Traduction libre).

3.4.2.1.1 Autocorrélation spatiale globale

L'autocorrélation spatiale permet de mesurer la ressemblance entre voisins, en utilisant la notion de graphe de voisinage. Elle est souvent utilisée comme technique exploratoire pour indiquer si une modélisation spatiale est nécessaire (Zeitouni et al, 1999). Elle s'applique à des données quantitatives rattachées à des objets polygonaux formant un découpage. Elle se décline en deux mesures complémentaires : l'indice de Moran (Moran, 1950) et l'indice de Geary (Geary, 1954).

- **Indice de Moran (I)** : l'indice de Moran était exposé la première fois dans (Moran 1950) puis il était popularisé par Cliff et Ord (1981). L'indice de Moran mesure l'écart de la valeur d'un point et celle de ses voisins à celle de la moyenne. Il se présente comme le rapport de la covariance sur la variance :

$$I_{Global} = \frac{n}{m} \times \frac{\sum_i \sum_j wij (z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2} \quad (15)$$

- z_i = valeur de la variable au point "i" et de moyenne $\bar{z}$
- i = unité géographique de référence
- j = unités voisines du point "i", définies par la matrice

- n = nombre total des unités géographiques
- m = nombre total de paires de voisins
- W_{ij} = matrice de pondération
- **Indice de Geary (C)** : l'indice de Geary (Geary, 1954) teste si la variabilité d'une variable entre voisins (donnée par la notion de variance locale) est significativement différente de celle attendue d'un modèle aléatoire (donnée par la variance). L'indice de Geary est défini comme le rapport C de la variance locale V_{loc} et de la Variance V .

$$C = \frac{1/2m \sum_i \sum_j w_{ij} (x_i - x_j)^2}{1/2n(n-1) \sum_i \sum_j (x_i - x_j)^2} \quad (16)$$

Les constats faits à l'aide d'indices globaux sont une invitation à explorer davantage les données afin de mieux cerner les différentes configurations spatiales locales (Morency, 2006). Que ça soit l'indice I de Moran ou l'indice C de Geary, les deux se basent sur l'étude de similarité et di-similarité des valeurs voisines pour définir si le modèle spatiale est clustérisé (ressemblance des valeurs voisines), dispersé (des valeurs voisines opposées) ou alors aléatoire (absence d'autocorrélation spatiale) mais n'indiquent pas si les valeurs sont faibles ou élevées. Pour cela, une autre statistique de l'autocorrélation spatiale globale a été proposée, l'indice General G de (Getis et Ord, 1992) pour définir le rapport fort/faible.

- **Indice G de Getis-Ord** : la particularité de l'indice G de Getis-Ord par rapport aux autres indices est le fait qu'il est capable de détecter si les points sont chauds ou froids. La formule de G utilise également le produit croisé pour mesurer l'association spatiale, similaire à celle de Moran.

$$G(d) = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(d) x_i x_j}{\sum_{i=1}^n \sum_{j=1}^n x_i x_j} \quad (17)$$

- Si $G(d) < 0$ alors Les valeurs groupées sont faibles
- Si $G(d) > 0$ alors les valeurs groupées sont élevées
- Si $G(d) = 0$ alors les valeurs sont proches de la valeur nulle

La statistique $G(d)$ de Getis-Ord donne plus de renseignements sur l'autocorrélation spatiale donnée par l'indice I de Moran. Les statistiques G devraient être utilisés conjointement avec I afin d'identifier les caractéristiques des modèles non révélés par la statistique I toute seule.

3.4.2.1.2 Autocorrélation spatiale locale

Les mesures d'autocorrélation spatiale, appliquées localement, permettent de cerner les secteurs similaires et dissimilaires. Le caractère distinctif ou non d'un secteur en regard de son voisinage peut faciliter la formulation de modèles, surtout dans le choix d'un découpage spatial. L'identification, a priori, des secteurs différents peut faciliter le travail d'analyse corrélative (Morrency, 2006).

- **Indice local G_i et G_i^* de Getis-Ord** : La statistique locale Getis-Ord G_i et G_i^* est la version locale de la statistique Générale G. Elle indique comment la valeur de chaque unité est associée aux unités voisines à une distance d . A partir d'un ensemble d'entités pondérées,

identifie les points chauds et les points froids statistiquement significatifs. Il permet de renseigner si les unités spatiales sont positivement associées ou sont négativement associées. La statistique locale G_i peut être écrite comme suit (Getis et Ord, 1992) (Ord et al, 1995) :

$$G_i(d) = \left(\frac{\sum_j w_{ij}(d)x_j}{\sum_j x_j} \right)_{(18)} \text{ Avec } i \neq j ; \quad G_i^*(d) = \left(\frac{\sum_j w_{ij}(d)x_j}{\sum_j x_j} \right)_{(19)} \text{ avec } i=j$$

Avec $W_{ij}(d)$ est le poids avec la distance d

- Si $G_i(d)$ est faible (ou z-score <0) alors il indique des valeurs faibles et inférieures à la moyenne.
- Si $G_i(d)$ est élevé (ou z-score >0) alors il indique une association spatiale de valeurs élevées similaires
- Si $G_i(d) = 0$ alors les valeurs s'approchent de la valeur nulle.

Les statistiques G_i et G_i^* nous permettent de détecter les agrégats (les clusters) locales de dépendance qui peuvent ne pas apparaître lors de l'utilisation de statistiques globales (I, C ou General G).

(Getis et Ord, 1992) avaient cherché à mettre en place des indicateurs locaux d'association spatiale, mais sans lien avec les indicateurs globaux existants. Le travail d'Anselin se place dans la continuité de celui de Getis et Ord (1992) en proposant des indicateurs locaux d'Association Spatiale (Local Indice for Spatial Association) permettant le passage d'un indicateur global de la structure spatiale (indice d'autocorrélation spatiale globale) à une mesure locale des ressemblances LISA (Anselin, 1995). Les statistiques locales de Moran (LISA) donnent une mesure de l'autocorrélation spatiale pour chaque localisation.

- **Indice de LISA(Ii) d'Anselin :** Les indicateurs locaux d'association spatiale LISA (Local Indice for Spatial Association) permettent le passage d'un indicateur global de la structure spatiale (indice d'autocorrélation spatiale) à une mesure locale des ressemblances LISA (Anselin, 1995). Pour Anselin (1995), un indicateur local d'association spatiale est une statistique qui répond à deux critères :
- Le LISA de chaque observation donne une indication sur l'étendue d'un regroupement spatial significatif de valeurs similaires autour de cette observation.
- La somme des LISA pour toutes les observations est proportionnelle à un indicateur global d'association spatiale.

Les indicateurs locaux d'association spatiale (LISA), développés par Luc Anselin (Anselin, 1995) permettent de mesurer le degré de ressemblance d'une unité spatiale avec ses voisines. On peut ainsi révéler les tendances régionales tout en conservant les valeurs locales, c'est-à-dire préserver l'information relative à l'hétérogénéité interne de ces zones.

L'indice de LISA s'appuie sur la description d'un gamma qui est une statistique de produits croisés, voir (Getis, 1991) qui s'écrit de la sorte :

$$\Gamma = \sum_i \left(\sum_j w_{ij} v_{ij} \right)_{(20)}$$

On reprenant la formule du gamma, on voit que l'on peut facilement obtenir une mesure d'un gamma local par la formule (Anselin, 1995) :

$$\Gamma_i = \sum_j w_{ij} v_{ij} .$$

Gamma local est égale alors à :

$$\Gamma = \sum_i \Gamma_i$$

D'une autre part, l'indice global du I de Moran est donné par la formule suivante :

$$I_{Global} = \frac{n}{m} \times \frac{\sum_i \sum_j w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2} \quad (21)$$

Après une opération de standardisation en ligne de la matrice de voisinage on obtiendra :

$$I_{Global} = \frac{\sum_i \sum_j w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2} \quad (22)$$

L'opération de standardisation consiste à rendre les poids des voisins égaux à 1 pour chaque unité. Le résultat de la standardisation en ligne est de rendre m égal à n , et donc à simplifier l'écriture de l'indice ($n/m=1$). L'opération de standardisation en ligne ne constitue pas un artifice statistique mais une correction nécessaire afin que chaque unité ait le même poids, quel que soit son nombre de voisin. À défaut, les unités ayant le plus de voisins (unités situées au centre et/ou dans des régions denses) auront plus de poids dans le calcul de l'autocorrélation spatiale. Ceci reste d'ailleurs valable aussi bien à l'échelle globale que locale.

On obtient ainsi un I de Moran local (I_i) représenté par l'indice de LISA donnée par la formule :

$$I_i = \frac{\sum_j w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2} \quad (23)$$

De forts indices normalisés $Z(I_i)$ supérieurs et inférieurs à zéro révèlent respectivement d'importantes similarités et différences.

On peut remarquer qu'il existe un lien direct de proportionnalité entre les indices locaux et l'indice global (Oliveau, 2004) qui est de la forme :

$$\sum_i \Gamma_i = \gamma \Gamma \quad (24) \text{ avec } \gamma = n$$

Les indicateurs locaux d'association spatiale (LISA), développés par Luc Anselin (1995) permettent de mesurer le degré de ressemblance d'une unité spatiale avec ses voisines. On peut ainsi révéler les tendances régionales tout en conservant les valeurs locales, c'est-à-dire préserver l'information relative à l'hétérogénéité interne de ces zones.

$$I_i = \frac{\sum_j^m w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\sum_i^n (z_i - \bar{z})^2} \quad (28)$$

Selon (Oliveau, 2004), la moyenne des indices locaux est alors égale à l'indice global. Ce lien de proportionnalité entre l'indice global de Moran et les indices locaux des LISAs permet d'obtenir pour chaque sous région géographique une estimation de sa ressemblance avec les sous régions voisines par rapport à sa ressemblance à l'ensemble des sous régions. On distingue alors quatre cas de figures (Oliveau, 2004) :

- **Situation High –High** : Autocorrélation spatiale positive et valeur de l'indice élevé englobe les régions avec un indice fort dans un voisinage qui lui ressemble.
- **Situation High –Low** : Autocorrélation spatiale négative et valeur de l'indice élevé englobe les régions avec un indice fort dans un voisinage qui ne lui ressemble pas.
- **Situation Low –High** : Autocorrélation spatiale positive et valeur de l'indice faible englobe les régions avec un indice faible dans un voisinage qui lui ressemble.

- **Situation Low–Low** : Autocorrélation spatiale négative et valeur de l'indice faible englobe les régions avec un indice faible dans un voisinage qui ne lui ressemble pas.

Une fois qu'un niveau de signification a été défini, les valeurs pourraient être tracées sur une carte pour afficher les emplacements des points chauds: sites avec des valeurs élevées avec des voisins similaires (High-High).

Ces statistiques spatiales présente l'inconvénient de ne considérer que les relations entre objets du même thème (intra-thèmes) excluant les relations spatiales pouvant exister entre objets de thèmes différents (Zeitouni, 2006).

3.4.2.2 Analyse de la densité

Lorsque le nombre d'objets analysés est élevé et que leur distribution spatiale est hétérogène, de nombreuses localisations se superposent. Par conséquent, la visualisation cartographique de ces objets ne reflète pas l'intensité locale, ni la tendance générale de leur localisation. L'idée est de quantifier les localisations ponctuelles par des mesures de densités dans leur voisinage puis, par interpolation, d'attribuer une « intensité » au phénomène en tout lieu de l'espace. La visualisation cartographique du champ de mesures ainsi obtenu montre l'intensité du phénomène de manière plus claire. Les densités sont calculées par un balayage de l'espace par des fenêtres mobiles circulaires (Gatrell et al. 1996). Une analyse de tendance globale est obtenue par des fenêtres mobiles de rayon élevé, tandis qu'une analyse locale révélant les contrastes correspond au choix de rayon de faible taille.

L'analyse de la densité sélectionne des quantités connues de certains phénomènes et les répartit dans l'espace géographique selon la quantité mesurée à chaque localisation et la relation spatiale des localisations des quantités mesurées.

3.4.2.3 Généralisation spatiale

La méthode de la généralisation spatiale est un processus qui résume un grand nombre de données pertinentes dans une base de données à partir d'un niveau de détail conceptuel des données vers des niveaux plus élevés en remontant d'un niveau par étape de généralisation.

La généralisation spatiale est par définition (Roddik et al, 2009), le processus par lequel des hiérarchies de concepts sont utilisées pour agréger les données. La découverte de relations générales entre données spatiales et non spatiales peut être réalisée par induction orientée attribut de deux manières : (1) la généralisation à dominante spatiale (GDS) et (2) la généralisation à dominante non spatiale (NSDG) (Lu et al, 1993). La première utilise la méthode d'induction orientée attribut sur la hiérarchie spatiale pour généraliser d'abord les données spatiales ; tandis que la deuxième commence d'abord par généraliser les données non spatiales.

La sélection des deux méthodes peut être basée sur : la disponibilité des données du contexte (les concepts de l'hiérarchie), l'efficacité de la généralisation et la préférence des résultats souhaités obtenir (Han et al, 1996). D'autres types de généralisation spatiale ont également été suggérées comme de nouvelles alternatives de généralisation tel que la généralisation croisée entre les données spatiales et non spatiales et la généralisation sur de multiples cartes thématiques (Lu et al, 1993). Plusieurs algorithmes ont été proposés pour effectuer la généralisation des données

spatiales en utilisant la méthode orientée attribut comme l'algorithme GenDis proposé dans (Knor et Ng, 1996). Un exemple de généralisation de concepts est présenté dans (Lu et al, 1993), dans ce dernier, Les modèles météorologiques liés aux régions géographiques de la Colombie-Britannique peuvent être découverts par une généralisation à dominante spatiale ou non spatiale.

- La GDS fusionne de petites régions dans de grandes régions de la CB (selon l'hierarchie administrative), qui partage en conséquence les données non-spatiales associées, telles que la température et la précipitation. L'induction orientée-attribut peut être ensuite effectuée sur chaque partition pour extraire les conditions météorologiques générales associées à chaque région généralisée.
- La GDNS généralise les données climatiques, telles que la température et la précipitations, dans un petit nombre d'entrées généralisées, les zones seront ensuite groupées en les fusionnant selon l'hierarchie de concepts (climats humides et froids) ou des valeurs de la catégorie recueille les pointeurs de données spatiales associés lors de la fusion des composants non spatiaux. Ensuite, la fusion spatiale peut être effectuée sur les entités spatiales dans la même entrée généralisée.

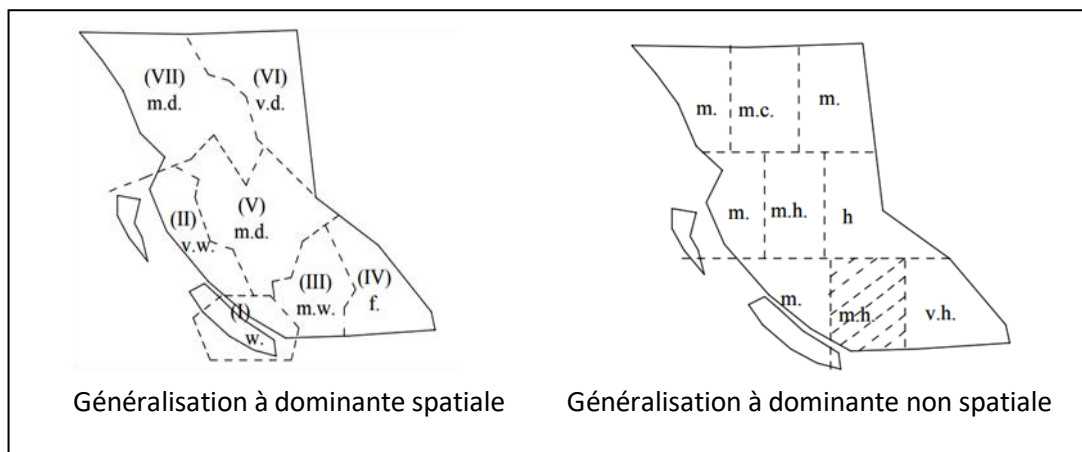


Figure 3.2 Méthode de la généralisation spatiale

Une hiérarchie de concept est une structure arborescente donnée explicitement par l'expert ou automatiquement par analyse de données et montre un ensemble des relations entre les objets. Elle décrit le passage des concepts les plus spécifiques au concept plus général de niveau supérieur. Appliquée aux données spatiales, les niveaux hiérarchiques peuvent illustrer des relations spatiales. L'auteur dans (Han, 1997) propose un algorithme pour l'extraction des relations d'hierarchie. Cette technique suppose que l'utilisateur ou l'expert doit fournir cette hiérarchie de concepts, qui n'est souvent pas possible.

En se basant sur les inconvénients de la généralisation et ses exigences, des efforts ont été faites pour élaborer des algorithmes qui n'exigent pas la proposition d'une hiérarchie de concepts préalable. Ça revient à trouver des algorithmes qui découvrent des caractéristiques directement à partir des données. C'était la motivation pour appliquer des méthodes de clustering dans l'exploration des données spatiales pour identifier les régions satisfaisant des conditions spécifiques (Wang et al, 1997).

3.4.2.4 La classification automatique (Clustering)

Ces mesures permettent d'abord d'identifier des zones de forte concentration à savoir des agrégats de points répondant à certains critères de proximité, révélant par conséquent certaines structures locales puis de raffiner cette identification par la prise en compte d'une structure spatiale de référence afin de n'identifier que les zones où l'occurrence dépasse celle qui serait normalement attendue étant donnée cette structure spatiale. Les algorithmes du clustering spatial diffèrent des algorithmes de clustering traditionnels par le fait qu'ils regroupent les objets spatiaux selon leurs localités. Pour des objets de types points, la métrique Euclidienne ou la distance de Manhattan sont suffisantes. En plus de la statistique de clustering scan statistic qui est combinée avec les outils d'autocorrelation spatiale pour générer des agrégats par l'outil SaTScan (Kulldorff, 2005), nous trouvons dans la littérature la machine GAM (Openshaw, 1996) (Geographical Analysis Machine) de Openshaw qui utilise le calcul intensif pour résoudre le clustering et qui a été utilisée pour surveiller les maladies tropicales puis elle a été optimisée en GAM/K grâce à l'utilisation de l'index spatial. Les algorithmes de clustering trouvés dans la littérature et qui peuvent être appliqués dans le domaine spatial sont catégorisés en quatre groupes:

3.4.2.4.1 Algorithmes de partitionnement

Les algorithmes de partitionnement sont des algorithmes descendants qui permettent de partitionner un ensemble de points en un nombre prédéterminé d'agrégats (k). On trouve dans cette catégorie de clustering deux groupes d'algorithmes :

- **Les algorithmes à base de K-Means** (Anil, 2010): où chaque cluster est représenté par son centre de gravité. La classification autour de centres mobiles génère une partition concernant tous les points de l'ensemble.
- **Les algorithmes à base de K-Medoids** (Kaufman, 1987): où chaque cluster est représenté par un des objets du cluster (medoid) localisé à une distance minimale des autres objets de ce même cluster. On peut citer dans cette catégorie PAM (Partitioning Around Medoids) (Kaufman, 1990) et CLARANS (Clustering Large Applications based on Randomised Search) (Ng and Han, 1994)

3.4.2.4.2 Classification hiérarchique

Les algorithmes hiérarchiques procèdent à la construction des agrégats graduellement sous forme graphique d'un arbre hiérarchique par agrégats successifs des observations qui fournissent une hiérarchie de classification (plusieurs générations de groupement). La méthode la plus connue dans cette catégorie est la méthode BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) (Zhang et al, 1996). Il existe deux méthodes de segmentation hiérarchique : ascendante et descendante.

- **Les algorithmes hiérarchiques ascendants** : considèrent, au départ, chaque objet comme un cluster. À chaque itération, deux clusters sont fusionnés, selon une certaine similarité, afin de former un nouveau cluster.

- **Les algorithmes hiérarchiques descendants** : procèdent par une méthode divisive. Ces méthodes consistent qu'à partir un seul cluster regroupant tous les objets, à diviser celui-ci en clusters plus raffinés.

La segmentation hiérarchique ascendante est plus utilisée que la descendante

3.4.2.4.3 Algorithmes basés sur la densité

Ces algorithmes sont basés dans le domaine spatial, sur l'extension de l'algorithme DBSCAN(Ester et al, 1996) en utilisant l'index spatial R*tree, ainsi que des extensions incrémentales des algorithmes DBCLASS, DBLEARN ; L'extension aux attributs non spatiaux et aux objets de forme autre que ponctuelle a été proposée dans GDBSCAN (Sander et al, 1998) en redéfinissant la fonction de similarité. Nous trouvons aussi, OPTICS (Ankerst et al, 1999) et DENCLUE (Alexander et al, 2007).

3.4.2.4.4 Algorithmes basés sur les grilles

Ces algorithmes divisent l'espace de données en cellules. Les régions jugées denses, selon la valeur d'un seuil fixé à priori, sont fusionnées pour former les clusters. Ces algorithmes sont déterministes, robustes au bruit, incrémentaux et rapides. Les algorithmes les plus connus dans cette catégorie sont : STING (STatistical INformation Grid) (Wang, 1997) et Wave Cluster (Gholamhosein, 1998).

3.4.3 Phase explicative (décisionnelle)

L'objectif des méthodes de cette phase est la description et l'explication des concentrations et des agregats reperés et detectés dans la phase exploratoire. Une interpetation des resultats d'analyses est donnée par la génération des regles (ex : regles caractéristiques, regles d'association) ou des graphes (ex : arbres de decision) permettant aussi de faire la prediction sur de nouvelles données.

3.4.3.1 Caractérisation spatiale

La méthode de caractérisation vise à ressortir les propriétés caractéristiques d'un sous-ensemble d'objets par rapport à l'ensemble des objets de la base. Son adaptation à la fouille de données spatiales (Ester et al, 1998) correspond aux propriétés caractéristiques qui s'étendent jusqu'à un degré de voisinages. La méthode de la caractérisation spatiale se base sur les paramètres suivants : i) un sous ensemble S à analyser, ii) un seuil de fréquence relative à la base, iii) une confiance définie comme la proportion d'objets de S qui satisfont le seuil de signifiante dans leur voisinage et iv) un degré de voisinage n_max. Le module Geo-Characterizer de GeoMiner proposé dans (Han et al, 1997) permet la découverte de règles caractéristiques à l'issue d'une généralisation des données spatiales et non spatiales (Lu et al, 1993).

3.4.3.2 Règles d'association spatiale

Le module Geo-associator de GeoMiner (Han et al, 1997) permet la découverte d'associations à base du concept de la généralisation et découvre les relations spatiales et non spatiales au niveau d'un concept plus général où les objets spatiaux sont comme étant des régions fusionnées depuis

un ensemble de couches thématiques donné, en utilisant certains attributs et des seuils de support et de confiance. Une amélioration de l'algorithme APRIORI a été également proposée dans (Koperski et al, 1995) pour l'évaluation du prédicat spatial en deux étapes. La première fait un test approximatif et génère des candidats pour un test précis du prédicat en seconde phase.

3.4.3.3 Classification par arbres de décision spatiale

La classification par arbre de décision spatiale se fait par extension et adaptation des algorithmes de construction des arbres de décision classiques sur les données à référence spatiale, représentées sous forme de couches thématiques. En plus de la considération des spécificités des objets spatiaux (leurs formes géométriques, leurs descriptions thématiques, les mesures métriques relatives à ces objets et leur localisation), d'autres dimensions sont en plus considérées pendant le processus de la construction de l'ADS qui est la relation de voisinage entre objets et les propriétés des objets voisins. On cherche donc à travers ce processus, à trouver les règles qui divisent l'ensemble des objets en plusieurs classes en utilisant non seulement leurs attributs descriptifs mais aussi leurs relations spatiales. Plusieurs méthodes ont été proposées pour atteindre cet objectif. Nous citons parmi d'autres :

- **La méthode de Fayyad et al**

Dans cette méthode, des arbres de décision ont été utilisés pour la classification des images stellaires afin de détecter les astres et les galaxies. Au départ, Les objets ont été classifiés par des astronomes. Sur la base de cette classification, dix bases de données d'apprentissages ont été conçues pour construire l'arbre de décision et extraire les règles. Environ 3 Terabytes d'images ont été analysées et traitées par le système FOCAS, qui sélectionne les objets et fournit des attributs de base comme : la magnitude, les régions, l'intensité, l'orientation, ... Le défaut de cette méthode est qu'elle ne s'applique pas à la base de données sous format vectoriel, souvent utilisée dans les systèmes d'informations géographiques et qu'elle n'utilise pas les données spatiales et se limite à l'utilisation des paramètres alphanumériques qui sont les attributs des images.

- **La méthode d'Ester et al**

Une méthode de classification des objets spatiaux utilisant les graphes de voisinages a été proposée par Ester (Ester et al, 1997). L'algorithme est basé sur la méthode ID3 qui s'étend à l'utilisation des données spatiales et utilise le concept de graphe de voisinage pour représenter les relations de voisinage. Les objets spatiaux sont traités comme voisins si les relations de voisinages (métriques, topologiques et de direction) sont vérifiées. Cette méthode ne garantit pas une segmentation correcte, se limite à une seule relation de voisinage et ne fait pas de distinction entre les thèmes, ce qui est essentiel en fouille de données spatiales.

- **La méthode de Koperski et al.**

Proposée par (Koperski et al, 1998), cette méthode de classification se base sur l'utilisation des prédicats spatiaux et suppose que les objets spatiaux sont caractérisés par les attributs thématiques (ex. nom d'une ville), les attributs thématiques reliés à un attribut spatial (ex. la population avoisinante), les prédicats spatiaux (ex. inside) et les fonctions spatiales (ex. la distance). Cette méthode garantit une bonne segmentation en utilisant les propriétés des objets

spatiaux, les objets des voisins et les relations spatiales exprimées par des prédicats spatiaux (Inside, Covers, Touch, Equal, Contain, Within_Distance, Nearest_Neighbor, Neighbor) mais elle a l'inconvénient d'un coût élevé de prétraitement, la perte d'information à cause de la généralisation et l'absence de choix dynamique de la distance dans les critères de discrimination.

– L'algorithme SCART

Cet algorithme a été proposé par Chelghoum et Zeitouni dans (Zeitouni et al, 2001), son objectif est de ramener le problème d'analyse de voisinages entre plusieurs couches thématiques vers une seule table à deux dimensions, exploitable par un graphe d'induction. Les auteurs ont considéré le problème de la fouille de données spatiales comme étant l'optimisation d'exécution des requêtes topologiques dans la base de données en utilisant les index de voisinage. Le tableau 5.2 (Chelghoum, 2004) compare les méthodes SCART, la méthode d'Ester et al. et la méthode de Koperski et al.

Tableau 5.2 Tableau de la comparaison des méthodes de classification spatiale

	La méthode d'Ester et al.	La méthode de Koperski et al.	La méthode proposée SCART
Bonne classification	Non	Oui	Oui
Classer les objets en fonction de leurs attributs, les attributs de leurs voisins et les relations spatiales	Oui	Oui	Oui
Le choix automatique de la bonne relation spatiale	Non	Non	Oui
Le degré de voisinage	= 1	=1	≥ 1
Distinction entre les thèmes	Non	Non	Oui

Tandis que les algorithmes de la classification par arbre de décision spatiale prédit les classes des objets spatiaux et les affectent à des catégories discrètes et non ordonnées, la régression spatiale modélise des fonctions à valeurs continues. Autrement dit, la régression est utilisée pour prédire des valeurs de données continues indisponibles plutôt que des étiquettes de classe (discrètes)

3.4.3.4 Régression spatiale

Les statisticiens sont les premiers qui ont construit des arbres de régressions (où la variable à prédire est continue), pour donner ensuite lieu à d'autres extensions au domaine spatial.

La régression spatiale découle de la *régression linéaire simple* où les données sont modélisées pour correspondre à une droite définie par l'équation : $y = x\beta + \varepsilon$ (29) ; où y est la variable dépendante ou à prédire et x la variable indépendante (aléatoire) ou prédictive, β et ε sont les coefficients de régression (β spécifie la pente de la droite et ε est l'ordonnée à l'origine). Dans cette équation, la variance de y est supposée constante. Dans le cas où une variable dépendante y , est modélisée comme une fonction linéaire de deux variables indépendantes (prédictives) ou plus, une

extension de l'équation de la régression linéaire (simple) est appliquée pour prédire y par une *régression linéaire multiple* à deux variables ou plus : $y = x\beta + z\alpha + \varepsilon$ (30).

La méthode de la régression peut être appliquée dans le domaine spatial pour modéliser les dépendances spatiales. Plusieurs modèles sont proposés ; nous citons dans ce qui suit quelques modèles :

– **La méthode (GWR Geographically Weighted Regression)**

La prise en compte de l'hétérogénéité spatiale des paramètres (des caractéristiques particulières de chaque zone géographique) signifie que les paramètres de l'équation $y = x\beta + \varepsilon$; varient selon leurs localisations géographiques. Dans le cas le plus général, le modèle à hétérogénéité spatiale des paramètres se présente comme suit (Anselin, 1990) :

$$Y_i = x_i' \beta_i + \varepsilon_i \quad (31)$$

Pour estimer les paramètres du modèle au point i (région i) on retient seulement les observations voisines qui ont un effet significatif sur i (selon la première loi de la géographie de Tobler). La méthode GWR utilise donc des pondérations W_i basées sur l'espace géographique. La forme matricielle de l'équation est donnée par :

$$Y_i = \sum_k \beta_{ki \times ki} + \varepsilon_i \quad (32)$$

– **Modèle autorégressif spatial (SAR)**

L'équation de régression spatiale est donnée par l'équation (Shekhar et al, 2003): $Y = \rho WY + X\beta + \varepsilon$ (33) ou W est la matrice de voisinage, ρ modélise d'intensité de la dépendance spatiale de Y (dépend de l'indice d'autocorrélation spatiale de y) et ε est le vecteur d'erreur. L'estimation de ρ et β se base sur le maximum de vraisemblance ou le théorème de Bayes. Si W est une matrice standardisée, la variable endogène décalée de l'équation n'est autre que la moyenne de la variable y calculée pour les régions voisines de i .

$$y_i = \rho \sum_{j=1}^N w_{ij} y_j + \sum_{q=1}^Q x_{iq} \beta_q + \varepsilon_i \quad (34)$$

– **Modèle spatial de Durbin (SDM)**

Comme le cas de la régression linéaire multiple, le modèle spatial de Durbin est appliqué quant la variable dépendante Y est prédite par deux variables indépendantes ou plusieurs. Le modèle SDM est donnée par la même équation de SAR mais augmenté à plusieurs variables prédictives. Sa forme matricielle est donnée par : $Y = \rho WY + X\beta + Z\delta + \varepsilon$ et sa forme explicite est donnée par :

$$y_i = \rho \sum_{j=1}^N w_{ij} y_j + \sum_{q=1}^Q x_{iq} \beta_q + \varepsilon_i \quad (35)$$

– Modèle avec autocorrélation spatiale des erreurs (SEM)

Ce modèle a la même équation du modèle SAR mais sans le terme correctif ρWY . L'équation du modèle SEM est donné donc par $Y = X\beta + \varepsilon$. Ce modèle se distingue par la forme du terme d'erreur qui est donnée par sa forme matricielle : $\varepsilon = \lambda W\varepsilon + u$. La forme explicite du modèle SEM est donnée par : $y_i = \sum_{q=1}^Q x_{iq} \beta_q + \varepsilon_i$ (36) et la forme explicite de l'erreur est donnée par : $\varepsilon_i = \lambda \sum_{j=1}^N W_{ij} \varepsilon_j + u_i$ (37)

3.5 Classification des méthodes de fouilles de données spatiales

Avec la multiplicité des techniques et algorithmes émergents dans le domaine spatial, le choix de la méthode de fouille de données spatiales appropriée et qui répond aux besoins de l'analyse s'avère un problème majeur. Ce choix pourrait se faire selon la nature de la relation spatiale existante entre les données de l'étude (intra thème ou inter thème). Dans (Zeitouni, 2006) on trouve le type de la fouille de données spatiale monothématique et la fouille de données spatiale multithématique.

3.5.1 Méthodes de fouille de données monothématique

Cette première approche de fouille de données spatiales considère les relations spatiales du même thème et analyse les données selon les mêmes attributs. Les méthodes existantes dans cette catégorie se distinguent selon le type d'analyse appliqué à l'étude et selon la prise en compte ou non des attributs ; on trouve :

- **Analyse de localisations sans attributs** : se base uniquement sur l'analyse de localisations pour révéler des tendances ou des concentrations comme le fait la méthode de clustering pour regrouper des clusters homogènes dans une distribution spatiale hétérogène ou l'analyse de tendances par la méthode de la densité.
- **Analyse de localisations munies de mesures numériques** : se base sur les mesures relevées sur un domaine spatial et intègre le graphe de voisinage pour calculer certains indices d'autocorrélation spatiale globale (comme celle de Moran, Geary et General G) et locale (indice de lisa et Getis-Ord local G_i et G_i^*). Les valeurs de ces indices informent de la nature du modèle spatiale (clustérisé, dispersé ou aléatoire). Ces mesures numériques peuvent être aussi analysées par régression linéaire pour découvrir des tendances spatiales pour une mesure donnée.
- **Analyse de localisations munies de catégories** : utilise uniquement les attributs catégoriels pour mesurer la corrélation spatiale inter catégories comme dans la méthode de co-localisation et découvre les règles caractéristiques à la suite d'une généralisation spatiale. Cette dernière utilise une hiérarchie de concepts définissant les données en plusieurs catégories.
- **Analyse d'individus munis de localisations** : cette catégorie de méthodes considère la localisation spatiale comme un terme correctif dans l'analyse des variables de données spatiales. illustrée par le terme ρ^*W*y dans l'équation de l'auto régression spatiale ou celui proposé dans l'algorithme de clustering multi-attribut Neighbourhood EM.

3.5.2 Méthodes de fouille de données multithématique

Ces méthodes se distinguent par leur capacité de révéler des relations inter-thèmes cachées dans des couches thématiques. Les algorithmes proposés dans cette catégorie de méthodes, cherchent à décrire ou à expliquer un phénomène par un autre ayant lieu au même endroit ou proche de cet endroit. Ils sont généralement basés sur des prédicats spatiaux interprétés comme des propriétés à prendre en compte dans le modèle à induire. Cela a été proposé dans l'algorithme de (Koperski et al. 1995) pour générer les règles d'association spatiales et dans (Koperski et al, 1998) pour faire la classification supervisée des données spatiales en proposant une liste de prédicats spatiaux. Les deux algorithmes proposés par Koperski et al se basent sur une généralisation des données selon des hiérarchies fournies par l'utilisateur et prennent en considération la description des objets spatiaux, la description d'objets liés à d'autres thèmes et les relations spatiales entre ces objets.

3.6 Travaux connexes

La particularité de l'information géo-spatiale en terme d'hétérogénéité, de complexité et d'interdépendances, a suscité la communauté scientifique spécialisée dans le domaine du GKD (Geographic Knowledge Discovery), cette branche du KDD (Knowledge Discovery in Databases), qui s'intéresse exclusivement au traitement de l'information géo-spatiale, à s'investir dans le développement de nouvelles approches afin de tenir compte de cette spécificité. De cela, est née le data mining spatial (DMS) et de nouvelles outils et méthodes ont été développés. Bien que plusieurs efforts ont été réalisés, des choses restent à faire en matière de fouille de données géo-spatiales.

Les travaux dans la FDS sont divisés en deux catégories, à savoir :

- Une première approche consistant à développer des outils, algorithmes et méthodes capables de traiter dynamiquement les corrélations entre entités géo-spatiales.
- une seconde consistant à utiliser les outils de fouille « traditionnelle » mais en prenant soin d'extraire au préalable et sous forme de données de types classiques (chaîne de caractère, numérique, booléen, etc.), les relations pouvant exister entre ces données.

3.6.1 Approche basée sur le prétraitement des données

Nous citons dans cette approche les travaux de (Chelghoum, et al., 2002) et (Zeitouni, 2006) qui visent à construire l'index de jointure spatiale puis appliquer la fouille de données géo-spatiales multi-thème c'est-à-dire l'extraction de tendances entre plusieurs couches spatiales. Ceci se résume en une fouille multi-tables ; ce qui pose problème pour les algorithmes traditionnels qui n'acceptent qu'une seule table en entrée. Pour remédier à cette situation, les auteurs proposent deux options. Soit transformer les algorithmes traditionnels afin qu'ils puissent traiter ce type de données ; ou bien ramener les données en une seule table et utiliser alors les algorithmes classiques. Nous citons également le framework géo-spatiales de (Bogorny et al) qui fournit un module de prétraitement de données géo-spatiales afin d'enrichir WEKA (Bogorny, et al., 2005). Ce Framework se subdivise en trois (3) modules : le module d'accès aux données, celui de prétraitement et enfin celui de fouille de données (Bogorny, et al., 2006). Nous citons également

dans cette catégories de travaux la plateforme d'Oracle Spatial (Appice, et al., 2000) qui rejoint l'approche de prétraitement afin d'aboutir à une structure mono-relationnelle sur laquelle ils appliquent une méthode de fouille de données classique. L'algorithme utilisé est dénommé SPADA (Spatial Pattern Discovery Algorithme) qui a été couplé avec ORACLE Spatial, grâce à un module nommé FEATEX permettant l'extraction des relations spatiales à travers le langage SQL. En fin, nous citons le travail de (Ouattara, 2010) qui a développé l'outil GeoKNIME, un nouvel outil de fouille de données géo-spatiales) qui implémente la composante spatiale au sein de l'outil de fouille de données open-source KNIME. Il s'agit de l'enrichissement de ce dernier avec la composante géo-spatiale ainsi que la mise en œuvre d'algorithmes permettant d'effectuer un Géo-clustering, une Géo-classification basée sur les K plus proches voisins et la construction d'arbres de décision spatiaux ainsi que quelques algorithmes enrichis « spatialement ».

3.6.2 Approche basée sur le traitement dynamique de l'information spatiale

L'objectif principal de cette approche est de mettre en œuvre des algorithmes ou des outils flexibles qui traitent dynamiquement la composante spatiale lors du processus de fouille sans passer par une phase de prétraitement ou de stockage préalable sur un support tiers. Le travail le plus ancien et le plus puissant que nous considérons le plus complet est le système GeoMiner (Han, et al., 1997), une extension de DBMiner qui est un outil de fouille classique. Il se compose de trois modules : le module pour la construction de cube, le module pour l'analyse en ligne OLAP et le module pour la fouille de données. Plusieurs travaux ont succédé ce travail, nous citons à titre d'exemple SubGroupMiner (Klösgen, et al., 2002) qui est un outil permettant la découverte de sous groupes spatiaux (spatial Subgroup). Nous trouvons également les travaux de (Ester et al) qui ont proposé Neighborhood Graph qui se base sur des primitives basées sur les relations spatiales que les entités entretiennent (Ester, et al., 1997) (Ester, et al., 1999). Afin d'intégrer de façon efficiente ces primitives au sein de systèmes de gestion de base de données, les auteurs proposent l'utilisation d'indices de voisinage. Dans (May, et al., 2003), les auteurs ont développé l'outil SPIN qui est une plateforme extensible N-tiers de fouille de données basée sur J2EE. L'architecture de SPIN se compose d'un serveur, des clients et d'un visualiseur spatial. Nous citons en fin de cette catégorie de travaux, le système INGENS (INductive GEographic iNformation System), un système inductif de découverte de connaissances spatiales basé sur la logique de premier ordre (Malerba, et al., 2000). Le système est construit autour de diverses couches (Map Storage, subSystem, Map convertor, Map editor, Map descriptor, Query interpreter, Learning server...). Le système INGENS a par la suite fait l'objet d'une amélioration par le développement de SDMOQL (Spatial Data Mining Object Query Language), un langage d'interrogation objet (Malerba, et al., 2002).

3.7 Conclusion

Les données à référence spatiale sont des données spécifiques par leur description géographique et leur interaction spatiale. La complexité de ce type de données provient essentiellement de leur hétérogénéité et des problèmes liés à leur représentation (mise à jour, MAUP...). Pour cela, des méthodes de fouille de données spatiales étaient développées, ces vingt dernières années, pour la gestion, l'analyse et la fouille de ce type de données.

Nous avons exploré dans ce chapitre, les différentes méthodes de FDS à travers une méthodologie qui suit une démarche systémique et décisionnelle qui va de l'analyse descriptive jusqu'à l'analyse explicative, voir prédictive par divers outils et approches. Différentes mesures et algorithmes qui supportent l'analyse des données spatiales ont été également présentés et analysés. Une attention particulière est donnée aux méthodes de FDS utilisées dans le cadre de ce travail de thèse à savoir, la méthode d'autocorrélation spatiale, la généralisation spatiale et la caractérisation spatiale. Une classification de ces méthodes est également présentée pour distinguer la FDS monothématique de la FDS multithématique. A la fin de ce chapitre, une synthèse d'état de l'art est exposée en citant quelques travaux connexes dans les deux types d'approches.

Le chapitre suivant présente notre première contribution dans ce travail de thèse à savoir la proposition de l'approche SOLAM qui intègre les méthodes de fouille de données spatiales au système SOLAP.

DEUXIEME PARTIE

CONTRIBUTIONS

Chapitre 4

Contribution 1 :

Le système SOLAP-MINING proposé

Ce chapitre présente une description détaillée du système SOLAM (Spatial On Line Analytical Mining), SOLAP Mining ou encore SOLAP sémantique qui est une extension du système SOLAP (Spatial On Line Analytical Processing) en intégrant les techniques de la fouille de données spatiales (FDS).

Plan du chapitre

1 Motivation	65
2 Introduction	65
3 Objectif visé	66
4 Approche décisionnelle proposée	66
5 Description du module SOLAM.....	68
6 Description du processus de la généralisation spatiale	69
6.1 La méthode d'Induction Orientée Attribut (AOI)	72
6.2 Principe de l'algorithme de comparaison	72
7 Expérimentation du système EPISOLAM.....	74
7.1 Délimitation de la zone d'étude	74
7.2 Formulation du problème abordé	74
7.3 Expérimentation	75
7.4 Arbre de généralisation.....	83
7.5 Discussion.....	83
8 Travaux connexes	84
9 Conclusion	85

4.1 Introduction

La prévention des épidémies est un problème majeur de la santé publique et un phénomène souvent incontrôlable. Surtout lorsqu'il s'agit de l'apparition de nouvelles maladies comme l'Ebola, virus de zika, etc. dont on ne connaît pas les comportements et des maladies à caractère endémique tel que la rougeole, le cholera, la tuberculose, etc. qui se propagent rapidement et largement. La localisation géographique des zones d'habitation sujettes à la contamination contribue à la limitation et au contrôle de la propagation de l'épidémie grâce à une stratégie de prévention et une gestion réfléchie. Dans ce contexte, le rôle du système d'information décisionnel est essentiel pour l'accès à l'information et l'aide à la décision pour les établissements de santé et les autorités locales en charge de cette prévention.

Afin d'identifier une bonne stratégie de prévention des épidémies et garantir une gestion efficace de la propagation des épidémies, il est important de concevoir un système de surveillance épidémiologique de qualité. Traditionnellement, les systèmes d'information décisionnels reposaient largement sur les entrepôts de données offrant des outils d'exploration multidimensionnelle des données et des indicateurs de la surveillance par le biais de l'analyse en ligne (traitement analytique en ligne) appelé OLAP.

En ce qui concerne les phénomènes tels que l'épidémie, les dimensions spatiale (et temporelle) sont, extrêmement, importantes pour l'analyse et la qualification de la propagation du phénomène dans le voisinage, ainsi que de son émergence ou réémergence.

Les dernières années ont vu le développement des travaux de recherche sur SOLAP ou OLAP spatial tels que ceux de (Bédard, 2004) et de (Bimonte et al., 2010) intégrant les données spatiales dans OLAP et reliant l'exploration et la visualisation cartographique. Nous avons donc expérimenté et appliqué ces approches et le système EPISOLAP (Zemri et al, 2013), a été proposé pour l'analyse épidémiologique.

Par la suite, des limites ont été détectées. L'objectif de notre première contribution est de combiner les principes du SOLAP et les méthodes de fouille de données spatiales dans le même esprit de OLAM (On Line Analytical Mining) qui veut dire "exploration multidimensionnelle", proposé dans (Han, 1997) et dans (Han et al. 1998).

4.2 Motivation

Nous avons vu dans le chapitre 1 que les systèmes d'aide à la décision spatiotemporelle SOLAP sont chargés d'effectuer l'agrégation des membres des dimensions spatiales selon les différents niveaux de granularité et selon une hiérarchie spatiale, Exemple: découpage administratif.

Les méthodes de la généralisation spatiale permettent d'explorer les données à différentes échelles sémantiques et spatiales tout en réduisant les dimensions inutiles. La question qui se pose est : *"Peut-on agréger les données spatiales dans les systèmes SOLAP en tenant en compte de l'aspect sémantique des données spatiales ?"*

L'idée est, donc, de tenir compte de la sémantique des attributs thématiques du systèmes SOLAP pour appliquer la généralisation des données spatiales. Le principe de la méthode est la

sélection et la suppression des attributs de faible importance dans le processus de caractérisation des données, produisant ainsi des zones de caractéristiques homogènes qui seront fusionnées par la suite.

4.3 Objectifs visés

L'objectif principal de l'approche que nous proposons est d'enrichir les fonctionnalités du SOLAP pour non seulement se limiter à l'exploration et la visualisation des données spatiales, mais aussi pour enrichir la fonction d'agrégation qui n'est pas nécessairement une simple moyenne ou une somme mais qui peut être basée sur la régression spatiale, la classification, l'association ou encore la caractérisation (Zemri et Hamdadou, 2019).

L'intégration entre SOLAP et la FDS conduit au développement d'un nouveau système décisionnel nommé EPISOLAP-MINING basé sur le moteur SOLAM (SOLAP Mining) que nous introduisons dans cette section. Il en découle deux autres objectifs spécifiques, à savoir :

1. Analyser les facteurs de risque épidémiologiques et les classer par ordre de pertinence, ce qui permet aux autorités d'intervenir et d'éliminer les plus influents en priorité.
2. Appliquer la fonction d'agrégation dans les systèmes SOLAP en utilisant la généralisation spatiale en prenant en considération les caractéristiques spatiales des zones administratives au lieu de faire l'agrégation selon les concepts hiérarchiques des divisions administratives seulement.

Pour atteindre ces deux objectifs, nous proposons deux solutions qui sont respectivement :

1. Calculer le gain d'information de chaque attribut dans l'algorithme AOI (Attribute Oriented Induction) (Han, 1994), classer les attributs en fonction des valeurs des gains, éliminer l'attribut correspondant au gain le plus faible et enfin procéder à la généralisation des tuples identiques correspondants aux zones voisines.
2. Utiliser la méthode de la comparaison, qui est une des méthodes de caractérisation, souvent utilisée pour distinguer une classe cible d'une ou de plusieurs classes contrastées. Par conséquent, les zones peuvent être agrégées selon leurs caractéristiques identiques.

4.4 Approche décisionnelle proposée

L'approche que nous proposons inclut une extension des systèmes SOLAP avec les techniques de la fouille de données spatiales prédictives (FDS) permettant de mettre en évidence les facteurs de risque épidémiologique. Le but est de décrire et d'expliquer les facteurs du phénomène épidémiologique observés et enregistrés dans l'entrepôt de données "EPISOLAP", afin de mieux les comprendre et par conséquent les éviter. Plus précisément, la FDS contribuera à la découverte des corrélations entre les phénomènes spatiaux et donnera une description précise des scénarios épidémiologiques ce qui devrait orienter les actions de prévention. Dans le cadre de notre étude, le couplage du SOLAP et de la FDS assure les fonctions suivantes :

- Le système SOLAP représente l'aspect spatial de l'étude épidémiologique grâce à sa capacité de détection et de localisation des épidémies.

- La FDS met en évidence certains aspects socio-environnementaux et les transformations socio-économiques qui influent sur la dynamique spatiale du phénomène épidémiologique et qui contribuent aux divers scénarios de la transmission.

Cependant, dans la littérature, les deux approches ont été proposées séparément et n'étaient pas assez intégrées. D'une part, il est regrettable que la FDS ne soit pas exécutée en ligne et ne traite que des données représentées sous forme de tableau (tableau individu/ variable) et, d'autre part, les opérateurs SOLAP classiques ne permettent que l'agrégation, la visualisation et l'exploration des données spatiales sans aucun effort d'interprétation.

Cela rend le couplage entre ces deux outils plus que nécessaire pour parvenir à cette complémentarité. Un tel couplage n'est pas simple et il reste plusieurs défis à surmonter que nous pourrions évoquer en répondant aux questions suivantes :

- Comment SOLAP et le Datawarehouse peuvent-ils intégrer des algorithmes de FDS ?
- Comment stocker dans un entrepôt de données les connaissances extraites par une méthode de FDS ?
- Comment exécuter en ligne, parfois sur de grands cubes, des algorithmes de fouille de données d'une certaine complexité et consommateurs du temps ?
- Comment modéliser de manière multidimensionnelle des données complexes (données géographiques) ?
- Comment effectuer une analyse en ligne sur les données spatiales ?

Il est évident que la mise en œuvre d'un tel couplage ne va pas être sans difficultés et devra être réalisée dans le respect de certaines contraintes. En effet, contrairement aux requêtes sur les cubes de données, la recherche basée sur SOLAP doit prendre en compte le contexte géographique. À un moment donné, l'utilisateur d'un cube affiche des faits avec différents niveaux de granularité. Cette photographie instantanée des données peut être utilisée pour faire la fouille de données spatiales. Notre approche est inspirée de la vision de Jiawei Han (Han, 1998), qui a intégré OLAP à la fouille de données dans un système cohérent qu'il a appelé OLAM (On Line Analytical Mining).

Afin de mieux positionner l'approche SOLAM par rapport aux autres approches, nous proposons une grille de mesures Analyse/Prédiction comme le montre la figure 4.1 :

1. Nous commençons par l'approche SIG qui est un système d'information géographique permettant de présenter les données spatiales sur une carte géographique et appliquer quelques analyses de base mais qui est limité en termes des méthodes de prévisions et de prédiction.
2. La FDS offre quant à elle, plus d'efforts de description et de prédiction mais seulement sur des données attributaires (sous forme de tableaux).
3. Les capacités du système SOLAP sont considérables en termes d'analyses mais insuffisantes en terme d'interprétation des résultats et de prédiction.
4. Le système SOLAM est le plus avantageux de tous ces systèmes grâce à ses capacités descriptives et prédictives issues des méthodes de la fouille de données spatiales tout en exploitant les capacités d'exploration et d'analyse du système SOLAP (un double avantage).

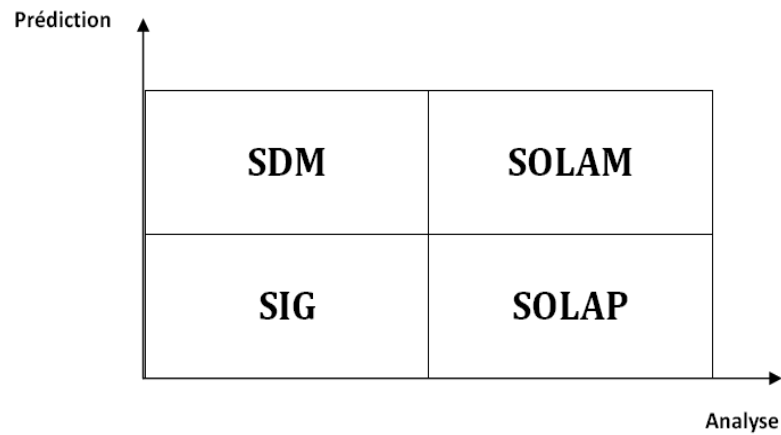


Figure 4.1 Grille de mesure Analyse / Prédiction

4.5 Description du sous système SOLAM

En se basant sur la méthode de la généralisation spatiale, nous proposons le système SOLAM qui est un sous **système** de notre système EPISOLAM MINING (Zemri et al. 2015). Un aperçu du sous système SOLAM est illustré par la figure 4.2.

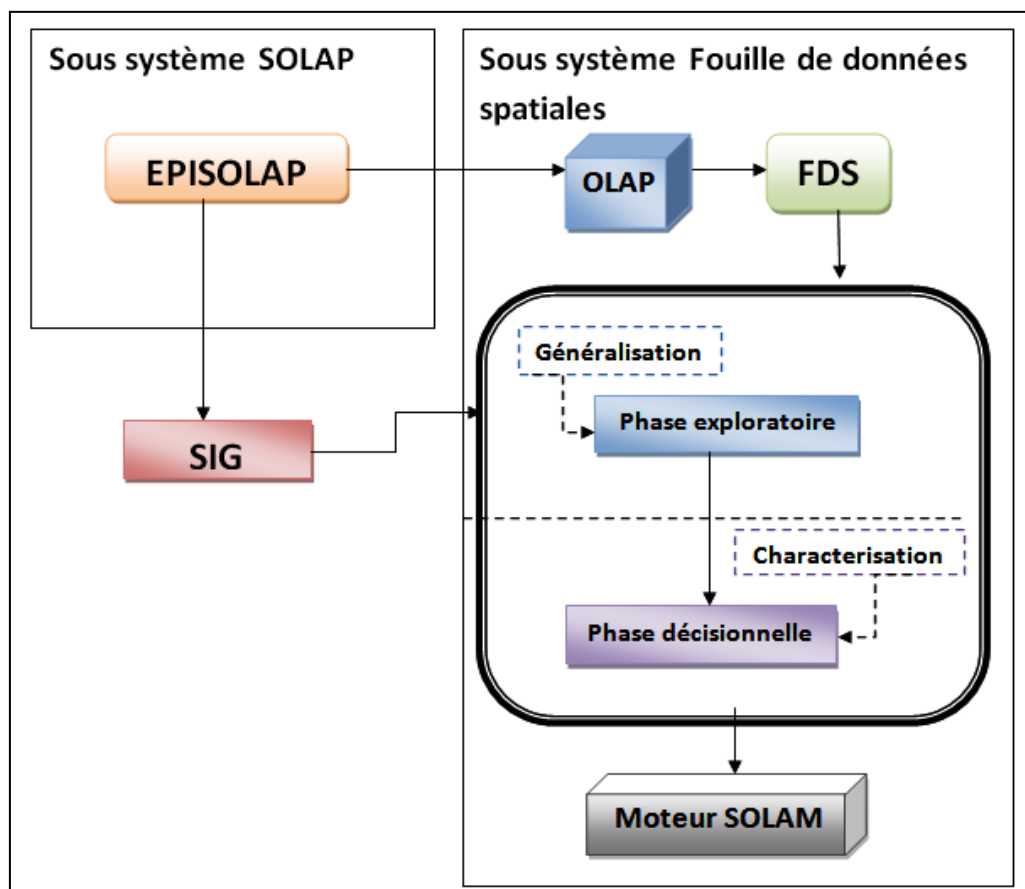


Figure 4.2 Le sous système SOLAM du système EPISOLAP_MINING

Le sous-système SOLAM proposé procède en deux phases principales :

1. **La phase exploratoire** : permet une description synthétique à l'aide des outils tels que : l'indice global d'autocorrélation, la généralisation, l'analyse de la densité, le lissage, la découverte des incohérences donnant les spécificités locales : autocorrélation locale ou analyse factorielle locale ou la recherche de clusters. Notre choix, dans cette phase exploratoire, a porté sur la méthode de **la généralisation spatiale**. Cette première phase guide la phase décisionnelle.
2. **La phase décisionnelle** : au cours de cette phase, une analyse plus détaillée est réalisée afin d'expliquer les différences ou caractériser les groupes à l'aide des techniques comme : la caractérisation, les règles de classification ou les règles d'association. Notre choix, dans cette phase décisionnelle a porté sur la méthode de **la caractérisation spatiale**.

Le choix de la méthode de la généralisation spatiale dans la phase exploratoire et celui de la méthode de la caractérisation spatiale n'était pas fait d'une manière arbitraire. Nous nous sommes repérées dans ce choix de la démarche systémique proposée dans le chapitre 3 et nous avons expérimenté ces deux méthodes vu qu'elles sont complémentaires et reposant toutes les deux sur la méthode d'induction orientée attribut qui est une méthode puissante (Han et al, 1995) et qui a fait ses preuves dans plusieurs travaux de Han et al. Nous allons explorer cette méthode avec plus de détails dans les sections suivantes.

Dans ce qui suit une description détaillée des étapes de la généralisation spatiale est donné sous forme de figures décrivant le déroulement de ce processus.

4.6 Description du processus de la généralisation spatiale

Il s'agit d'un processus d'agrégation des données spatiales en utilisant des hiérarchies conceptuelles. Il existe deux types de généralisation spatiale (Roddick et Brian, 2009) :

- **La généralisation à dominante spatiale** : regroupe, spatialement, les données sur la base d'une hiérarchie de concepts géographiques.
- **La généralisation à dominante non spatiale** : génère des unités spatiales agrégées qui partagent la même description de niveau hiérarchique supérieur.

Dans le cadre de cette thèse, nous avons choisi la généralisation à dominance non spatiale ; celle-ci n'utilise pas l'hiérarchie spatiale mais génère des localisations moins détaillées par fusion d'objets spatiaux.

Le principe de ce processus est simple, une méthode d'induction orientée attribut (AOI) (Han et al. 1995) est réalisée à l'aide des hiérarchies thématiques, tout en conservant leur description spatiale. Cette induction produit des valeurs d'attributs homogènes pour plusieurs objets. Ces objets sont ensuite fusionnés.

Le processus de comparaison se décompose en quatre sous processus, à commencer par le processus de collecte de données, l'analyse de la pertinence des dimensions, la généralisation synchrone et enfin la présentation de la comparaison. La figure 4.3 illustre les étapes 1 et 2 du processus et la figure 4.4 illustre les étapes 3 et 4.

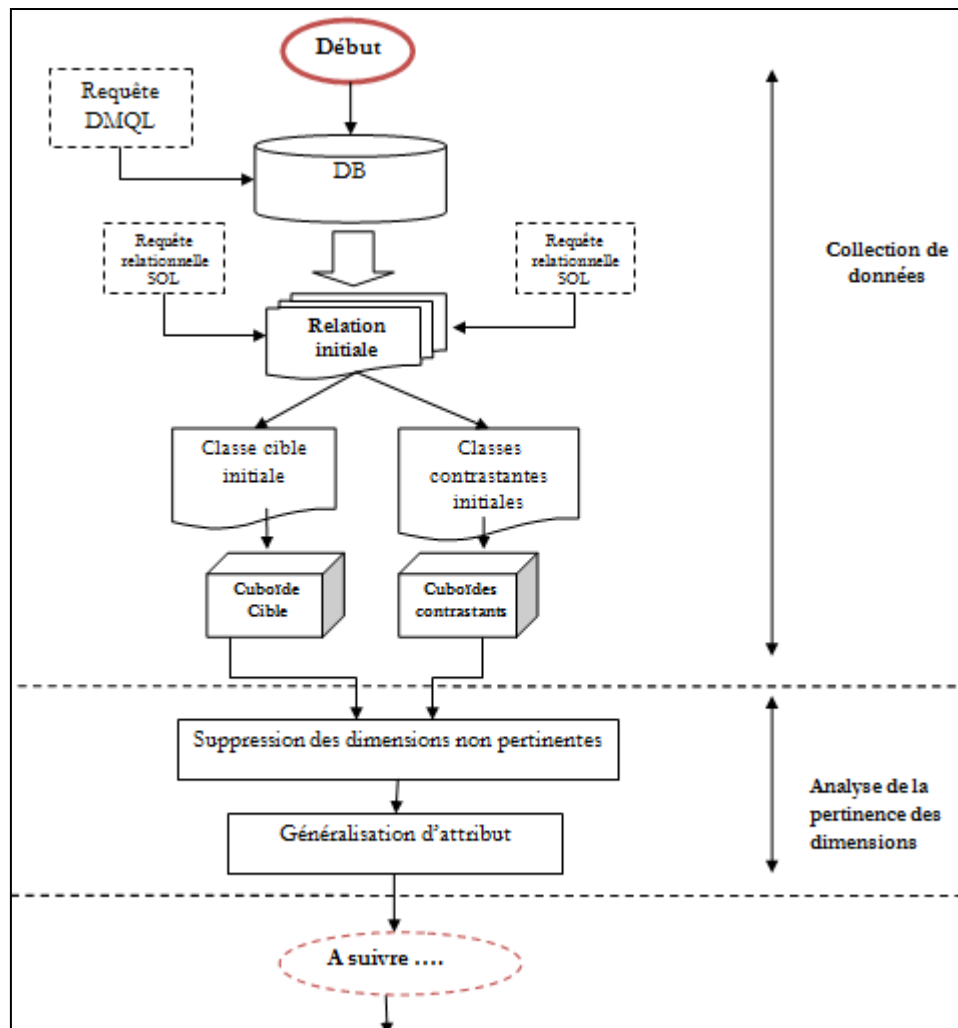


Figure 4.3 Les étapes 1 et 2 du processus de généralisation

Étape 1 : Collecte des données

Les données pertinentes de la base de données sont collectées en appliquant une requête DMQL qui partitionne le jeu de données en une classe cible et une ou plusieurs classes contrastantes (Voir la figure 4.3).

Étape 2 : Analyse de la pertinence des dimensions

Seules les dimensions pertinentes sont incluses dans la suite des analyses après l'application de l'analyse de pertinence des dimensions sur la classe cible et les classes contrastantes.

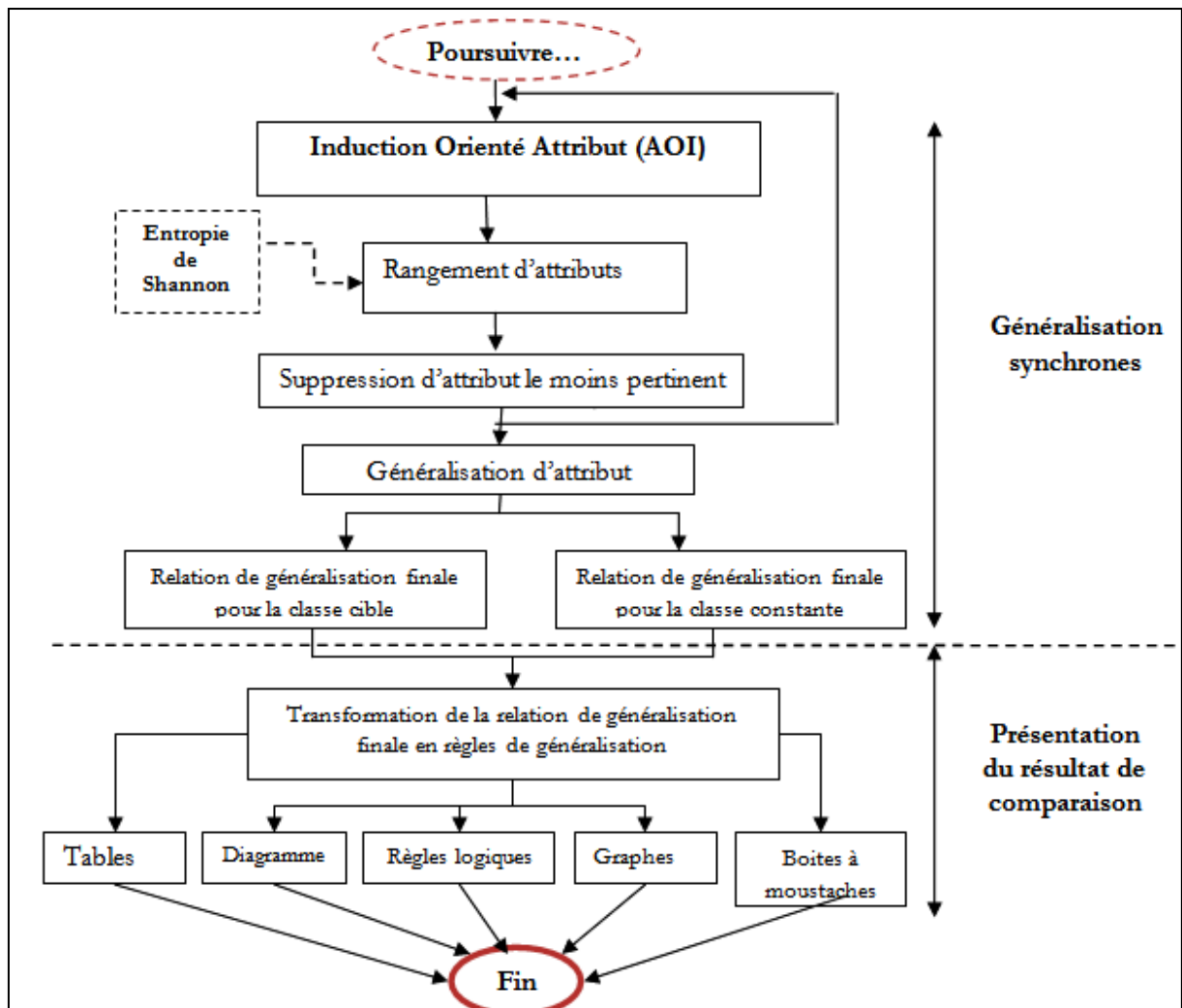


Figure 4.4 Les étapes 3 et 4 du processus de généralisation

Étape 3 : généralisation synchrone

L'algorithme AOI de base utilisé dans cette étape est celui proposé dans (Han et al. 1992). L'algorithme AOI se décompose en deux processus, à commencer par le processus de regroupement des attributs en fonction des différentes hiérarchies de concepts, puis le processus de fusion de n-uplets identiques en effectuant la généralisation des objets spatiaux correspondants. Après cela, la généralisation est effectuée sur la classe cible à un niveau contrôlé par un seuil de dimension spécifié par l'utilisateur / expert et sur la classe contrastante généralisée au même niveau. Le résultat de cette généralisation est la génération du cuboïde cible principal et du ou des cuboïdes contrastants principaux illustrés par la figure 4.4.

Étape 4 : Présentation du résultat de la comparaison

Le résultat final du processus peut être généré sous différentes formes : tableaux, diagrammes, boîtes à moustaches ou règles logiques (règles conjonctives et disjonctives). Des opérations de forages synchrones peuvent être effectuées sur la classe cible et les classes contrastées afin d'ajuster les résultats au niveau souhaité.

4.6.1 La méthode d'Induction Orientée Attribut (AOI)

La méthode d'Induction Orientée Attribut "AOI " fournit un mécanisme efficace pour générer les différents types de règles permettant la découverte de la connaissance à partir des bases de données (Warnars, 2015).

Tableau 4.1 Types de règles générées par la méthode AOI

Règles	Exemples
Règle caractéristique	Identifier les symptômes d'une maladie spécifique
Règles discriminantes (règles de comparaison)	Distinguer une maladie par rapport à une autre
Règles de classification	Classer des maladies dans des classes et fournir les symptômes de chaque catégorie
Règle d'association	Découvrir des symptômes survenant fréquemment ensemble
Règles des évolutions des données régulières	Décrire les facteurs majeurs qui influencent les fluctuations des stocks dans le temps
Règles de description des regroupements (points chauds)	Regroupement des étudiants de l'université en se basant sur différents attributs

La règle **discriminante** ou de **comparaison** est une assertion qui discrimine (ou compare) les concepts d'une classe (cible) d'une autre classe (contrastante) (Han et al. 1992). Cette règle donne un critère discriminant qui peut être utilisé sur de nouvelles données pour prédire l'appartenance à une classe.

Notre choix est donc porté sur la génération de règles de comparaison à cause des besoins de notre application dans le domaine de l'épidémiologie pour distinguer une zone à risque épidémiologique élevé (High) par rapport à une zone à risque épidémiologique faible (Low). Nous détaillerons le développement de ce choix dans le chapitre suivant.

4.6.2 Principe de l'algorithme de comparaison

SOLAM ou SOLAP sémantique est basé sur le processus de comparaison effectué par l'algorithme AOI, qui est adapté dans notre étude. Le principe de l'algorithme est le suivant:

"A chaque niveau de la hiérarchie de concepts, les unités géographiques voisines sont fusionnées si elles partagent la même description du niveau supérieur" comme le montre l'algorithme de comparaison. Cette alternative pourrait être réalisée en introduisant le gain informationnel dans le processus d'identification des attributs les plus pertinents et en prenant en compte le lien de voisinage. Cette variante d'algorithme de comparaison est adaptée à la généralisation spatiale en intégrant la matrice de voisinage comme entrée de l'algorithme.

L'algorithme de comparaison proposé procède donc comme suit :

Algorithme de comparaison des zones à risque élevé « High » avec les zones à risque faible « Low »

Entrées :

n= nombre d'attributs : entier
 m= nombre de classes : entier
 v= nombre des sous-classes : entier
 s= effectif dans les sous classes : entier
 i= itération sur le nombre d'attributs : entier
 j=itération sur le nombre de sous classes : entier
 n, m, v, s, i, j : entiers
 A1 = valeurs du premier attribut : Tableau
 Count : effectifs dans chaque région : Tableau

DEBUT

1. Le calcul de **Expected Info** I nécessaire pour classer des tuples arbitraires

$$I(s_1, s_2, \dots, s_m) = - \sum_{i=1}^m \left(\left(\frac{s_i}{s} \right) * \log \left(\frac{s_i}{s} \right) \right) // \text{Le calcul de } \mathbf{Expected Info} \text{ I nécessaire pour classer des tuples arbitraires}$$

2. Tant que le nombre des attributs est supérieur à 1

Tant que n>= 1 **faire** // Tant que le nombre des attributs est supérieur à 1

Début :

$$E(A1) = \sum_{j=1}^v \left(\left(\frac{(s_{1j}, s_{2j}, \dots, s_{mj})}{s} \right) * I(s_{1j}, s_{2j}, \dots, s_{mj}) \right) // \text{Calculer l'entropie de l'attribut A1}$$

$$\text{Gain}(A1) = I(s_1, s_2, \dots, s_m) - E(A1) // \text{Dédurre le gain informationnel de A1.}$$

Min_Gain = Gain(A₁) // Affecter cette valeur à la valeur minimale du gain Min_Gain

Pour chaque attribut A_i (2 ≤ i ≤ n) dans la relation de généralisation RG **faire** //

Pour chaque attribut A_i, le gain informationnel est calculé puis comparé à Min_Gain

Début :

$$E(A_i) = \sum_{j=1}^v \left(\left(\frac{(s_{1j}, s_{2j}, \dots, s_{mj})}{s} \right) * I(s_{1j}, s_{2j}, \dots, s_{mj}) \right)$$

$$\text{Gain}(A_i) = I(s_1, s_2, \dots, s_m) - E(A_i)$$

Si Gain(A_i) < Min_Gain **alors** Min_Gain = Gain(A_i) // Si Gain(A_i) est inférieur à Min_Gain alors Min_Gain est remplacé par Gain(A_i) et A_i est supprimé du dataset.

R=A_i

Fin

Enlever R // supprimer l'attribut qui correspond au gain minimum Min_Gain

n=n-1

Si les zones correspondantes sont voisines.

Alors fusionner les tuples identiques // Si les zones correspondantes aux tuples identiques sont voisines alors fusionner ces zones et créer un nouveau secteur

Agréger count% // Accumuler les comptes des zones fusionnées.

Créer la zone fusionnée

Fin

FIN

Nous verrons dans la section expérimentation un exemple d'usage afin de caractériser la propagation d'une épidémie.

4.7 Expérimentation du système EPISOLAP MINING

L'objectif de l'étude de cas considérée est la surveillance de la maladie de la tuberculose dans la région d'Oran, une des régions de l'ouest de l'Algérie qui souffre encore de cette maladie. Elle est classée en deuxième position selon le classement national du pays (Agadir et al, 2011).

Le challenge est de modéliser et intégrer l'impact environnemental et socio-économique dans un entrepôt de données et appliquer la fouille de données spatiale pour expliquer des scénarios prospectives et donner des interprétations précises au phénomène spatial (la propagation de la tuberculose).

Le système EPISOLAP-MINING ou EPISOLAM a été proposé comme une extension du système EPISOLAP pour répondre aux besoins grandissant des épidémiologistes en termes de prédiction et de prévention. Le processus de développement de ce système a procédé selon les étapes suivantes (Zemri et Hamdadou, 2019) :

4.7.1 Délimitation de la zone d'étude

Notre zone d'étude est composée de 26 communes de la région d'Oran. Elles représentent les 26 foyers d'épidémies de tuberculose précédemment détectées par le système "EPISOLAP". Le tableau 4.3 montre le niveau 0 de la généralisation : Gen_0 avec 26 zones.

4.7.2 Formulation du problème abordé

Nous donnons ci-dessous la liste des facteurs considérés dans les études en surveillance épidémiologique de la tuberculose. Ces facteurs seront analysés et combinés en tant qu'attributs en utilisant la méthode de **généralisation à dominante non spatiale** pour obtenir un classement des facteurs. Nous considérons dans notre étude de cas les attributs pertinents suivants : Taux d'incidence, densité de la population, nombre des constructions précaires, nombre des inactifs, Humidité

Tableau 4.2 Attributs considérés

Critères	Type	Echelle	Méthode d'évaluation	Source
Taux d'incidence	Médical	[15,115]	Nombre des cas contagieux	D.S.P
Densité de la population	Démographique	[63,9660]	Population moyenne, surface	O.N.S
Nombre des constructions précaires	Démographique	{0,...,5125}	Nombre d'habitats dans les bidonvilles	O.N.S
Nombre des inactifs	Démographique	[112,9928]	Niveau de l'éducation, âge	O.N.S
Humidité	Climatique	[1,4]	Distance par rapport à la mère	O.N.M

4.7.3 Expérimentation

L'objectif du processus est de dériver les caractéristiques générales de 26 tuples correspondants aux 26 foyers d'épidémies en utilisant une caractérisation analytique pour identifier les facteurs influents le plus sur la propagation de la tuberculose à Oran. La méthodologie a conduit à la classification des facteurs (ou attributs) en se basant sur le calcul du gain d'information de chaque attribut considéré. Gen (ai) identifie le niveau du concept hiérarchique. Gen_0 est le niveau initial de la généralisation. C'est le niveau de la généralisation avant d'appliquer aucune opération de généralisation et "ai" identifie l'attribut supprimé.

Étape 1 : Collecte de données

Au début, une requête DMQL (Data Mining Query Language) est exécutée afin de collecter les données pertinentes. L'exploration des données a pour objectif de comparer les niveaux faibles et élevés des zones épidémiologiques à l'aide des règles discriminantes.

```

Use TBC_DB
Mine comparison as "High_vs_Low outbreaks"
In relevance to name, incid_rate, density, Prec_Const,
Inact_Nbr, Humidity
For High_level_of_risck
Where class in high
Versus low_level_of_risck
Where class in low
Analyse cont%
From Outbreaks

```

Figure 4.5 Requête DMQL utilisée

La requête DMQL de la figure 4.5 décompose la relation initiale en deux classe distinctes, une concernant les zones à risque épidémiologique élevée « High_level_of_risck » et une autre concernant les zones à risque épidémiologique faible « Low_level_of_risck » en s'appuyant sur le nombre de cas dans chaque zone géographique « count% » tout en concevant les deux classes dans une même relation.

Le résultat de la requête DMQL est illustré sur le tableau 4.3

Tableau 4.3 Exemple de 26 foyers d'épidémies

N°	Name	Incidenc e rates	Density of population	Number of precarious constructions	Inactive number	Humidity
1	Oran	37	9660	5125	9928	1
2	Gdyel	44	413	101	760	2
3	Bir El Djir	36	5330	1305	2683	2
4	Hassi Bounif	49	1976	116	1356	3
5	Es Sénia	53	2144	570	1345	2
.....						
26	Ain Biya	25	89	5	395	2

Ensuite, nous divisons notre ensemble de données en deux classes, une classe cible et une classe contrastante, en exécutant deux requêtes SQL (Voir la figure 4.6). La classe cible contient le niveau de risque bas d'épidémies (Low) et la classe contrastée contient le niveau élevé de risque d'épidémies (High). Ces deux requêtes décomposent la relation initiale en deux relations séparées. Les résultats des deux requêtes SQL sont illustrés par le tableau 4.4.

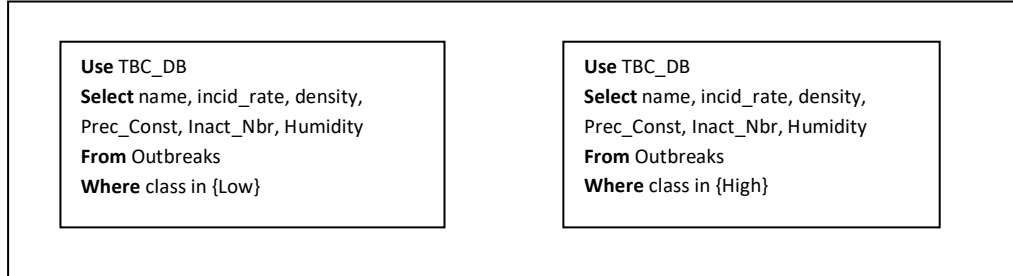


Figure 4.6 Requêtes SQL utilisées

Tableau 4.4 Relations candidates pour la classe cible et la classe contrastante

Name	Class	Inac_rate	Density	Prec_Const	Inact_Nbr	Hum	Count
A. Turck	Low	31	1005	197	671	1	12
.....							
Tafraoui		35	63	5	289	4	5
Ain kerma	High	79	71	0	122	1	6
.....							
S.Chami		55	1661	1282	2324	2	64

Étape 2 : Analyse de la pertinence de la dimension

Dans la deuxième étape, l'analyse de pertinence est appliquée à l'aide de la généralisation analytique afin d'éliminer les attributs non pertinents.

Suppression d'attributs : Avant de procéder à toute opération de généralisation, nous constatons que le nom des zones n'a aucune importance dans le processus de caractérisation analytique. Nous pouvons donc le supprimer du data set; ensuite nous généralisons les n-uplets identiques.

Généralisation des attributs : consiste à discrétiser les attributs continus en différentes catégories selon les hiérarchies de concepts de la figure 4.7 et accumuler ainsi les comptes :

- Généraliser incidence_rate à incidence_rate_C
- Généraliser la densité en densite_C

- Généraliser prec_const dans prec_const_C
- Généraliser inact_nbr dans inact_nbr_C
- Généraliser Humidity dans Hum_C

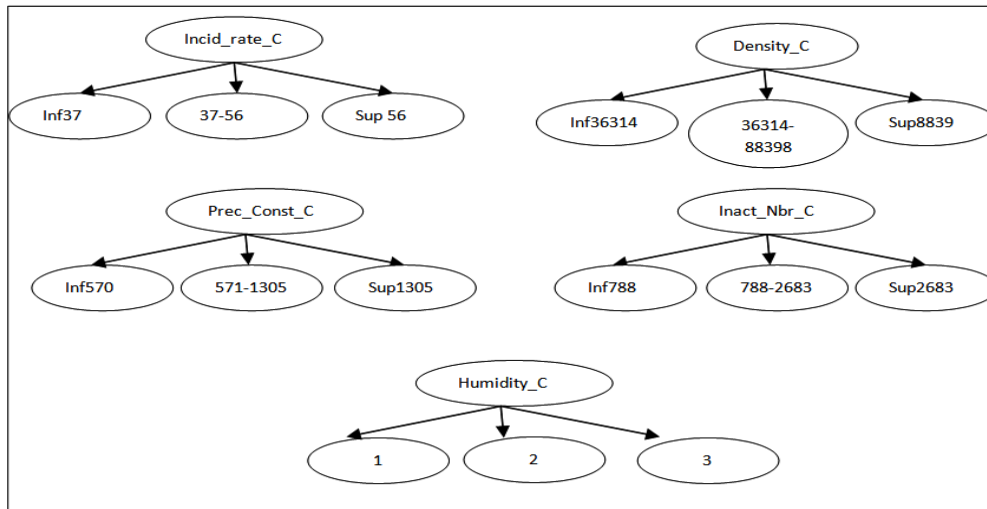


Figure 4.7 Concepts hiérarchiques

Nouvelles Relations candidates : incid_rate_C, densité_C, prec_const_C, inact_nbr_C et Hum_C.

Le jeu de données résultant des foyers d'épidémies est illustré sur le tableau 4.5 Ci-dessous, les noms des foyers d'épidémies seront donc supprimés afin de créer un nouveau tableau 4.7, le processus de l'algorithme consiste à rechercher les n-uplets identiques afin d'être fusionnés dans les nouveaux secteurs créés si et seulement si les zones correspondantes sont voisines. Par conséquent, une première opération de généralisation est effectuée sur les deux zones voisines Bir El Djir et S.Chami car leurs tuples correspondants sont identiques. Ces deux régions sont donc généralisées dans le secteur S6 qui est une nouvelle région issue de l'opération de fusion entre ces deux régions (Voir le tableau 4.5).

Tableau 4.5 Tableau du premier niveau du processus de généralisation

Name	Class	Incid_rate_Density_C	Prec_Const_C	Inact_Nbr_C	Hum_Count_C	Count	Grouped zones	
A.Turck	Low	37-56	inf36314	inf570	inf788	1	12	A. Turck
.....								
Tafraoui		37-56	inf36314	inf570	inf788	3	5	Tafraoui
A. kerma	High	sup57	inf36314	inf570	inf788	1	6	A. kerma
.....								
Bir El Djir		Sup57	36314-88398	571-1305	788-2683	2	63	S6
.....								
S.Chami		Sup57	36314 - 88398	571-1305	788-2683	2	64	S6

Étape 3 : généralisation synchrone

La généralisation synchrone est appliquée simultanément sur la classe cible et les classes contrastées à l'aide de l'algorithme de comparaison décrit dans la chapitre 2. Le déroulement de l'algorithme se fait par application du processus appelé « Relevance analysis » qui procède comme suit :

1. Calculer l'information I (expected info) dont la formule est appliquée pour target calss s1 et contrasting class s2.
2. Calculer l'entropie du premier attribut $A1=tx_incidence$.
3. Déduire le gain informationnel de A1 $Gain(A1)= Gain(tx_incidence)$

De la même manière, on calcule le gain informationnel de tous les autres attributs A_i et on supprime l'attribut qui correspond à la valeur minimale du gain. Dans notre cas, la valeur minimale est celle de la densité.

Tableau 4.6 : Tableau de gain et nombre des zones fusionnées

Removed attribute	Level	Zones	Low zones	High zones	Gain(Ai)
name	Gen_0	26	16(0)	10(1)	-
Density_C	Gen_1	25	16(0)	9(0)	0,44280177
Inact_nbr_C	Gen_2	24	16(0)	8(1)	0,503416154
Prec_Const_C	Gen_3	23	16(0)	7(1)	0,562716866
Incidence_rate_C	Gen_4	17	10(2)	7(0)	0,593837294
Hum_C	Gen_5	6	3(2)	3(2)	0,601812721
Class_C	Gen_6	2	1(1)	1(1)	-

Nous commençons par supprimer les attributs non pertinents / faiblement pertinents de la relation candidate qui est dans notre cas l'attribut densité_C. Le processus itératif est exécuté selon les différentes étapes de généralisation. Chaque étape correspond à la suppression d'un attribut par l'ordre donné dans le tableau 4.6.

Tableau 4.7 Tableau du deuxième niveau du processus de généralisation

Name	Class	Incid_rate_C	Density_C	Prec_Const_C	Inact_Nbr_C	Hum_C	Count
A.Turck	Low	37-56	inf36314	inf570	inf788	1	12
.....							
Tafraoui		37-56	inf36314	inf570	inf788	3	5
Ain kerma	High	sup57	inf36314	inf570	inf788	1	6
.....							
Gdyel		37-56	inf36314	inf570	inf788	2	17

Les foyers d'épidémies résultants sont illustrés par le tableau 4.7. L'attribut density_C sera ensuite supprimé et les n-uplets identiques seront fusionnés afin de créer le tableau 4.8. Cependant, il n'existe pas de nouveaux n-uplets identiques pour la classe contrastante dans le jeu de données des foyers d'épidémies en cours et les éléments identiques dans la classe cible ne sont pas voisins. Par conséquent, nous conservons le tableau 4.7 dans le tableau 4.8.

Tableau 4.8 Tableau du troisième niveau du processus de généralisation

name	Class	Incid_rate_C	Prec_Const_C	Inact_Nbr_C	Hum_C	Count	Grouped zones
A.El Turck	Low	37-56	inf570	inf788	1	12	A.El Turck
.....							
Tafraoui		37-56	inf570	inf788	3	5	Tafraoui
Ain kerma	High	sup57	inf570	inf788	1	6	Ain kerma
.....							
El kerma		sup57	inf570	Inf788	2	13	S7
Es sénia		sup57	inf570	788-2683	2	13	S7
Gdyel		37-56	Inf570	Inf788	2	17	Gdyel

Le jeu de données des foyers d'épidémies résultant est illustré par le tableau 4.8. Ensuite, l'attribut Inact_Nbr_C sera supprimé et les n-uplets identiques seront fusionnés afin de créer le tableau 4.9 comme suit : Il y a 2 n-uplets identiques dans la classe contrastante qui sont ensuite fusionnés dans un nouveau secteur S7.

Tableau 4.9 Tableau du quatrième niveau du processus de généralisation

Name	Class	Incid_rate_C	Prec_Const_C	Hum_C	Count	Grouped zones
A.El Turck	Low	37-56	inf570	1	12	A.El Turck
.....						
Tafraoui		37-56	inf570	3	5	Tafraoui
Ain kerma	High	sup57	inf570	1	6	Ain kerma
.....						
S6		sup57	571-1305	2	127	S8
S7		sup57	inf570	2	57	S8
Gdyel		37-56	Inf570	2	17	Gdyel

Le jeu de données des foyers d'épidémies résultant est illustré par le tableau 4.9. Ensuite, l'attribut Prec_Const_C sera supprimé et les n-uplets identiques seront fusionnés afin de créer le tableau 4.10 comme suit : Il y a 2 n-uplets identiques dans la classe contrastante qui sont ensuite fusionnés dans un nouveau secteur S8.

Tableau 4.10 Tableau du cinquième niveau du processus de généralisation

name	Class	Hum_C	Incid_rate_C	Count	Grouped zones
ATurck	Low	1	37-56	12	S1
Bethioua		1	37-56	8	S2
Arzew		1	37-56	34	Arzew
Oran		1	37-56	288	S1
El Ancor		1	inf37	2	El Ancor
M Kebir		1	inf37	6	S1
M. Hadjaj		1	inf37	4	S2
.....					
O.Telat		3	inf37	3	S3
S.B.yebka		3	inf37	2	S5
Hassi Ben	High	3	37-56	5	S4
Tafraoui		3	37-56	5	S3
H.Bounif		3	sup57	31	S4
H.Mefsouk		3	sup57	6	S5
Ain kerma		1	sup57	6	Ain kerma
.....					
Gdyel		2	37-56	17	Gdyel

Le jeu de données sur les foyers des épidémies résultant est illustré par le tableau 4.10. L'attribut Incid_rate_C sera supprimé et des nuplets identiques seront fusionnés afin de créer le tableau 4.11. Il existe 5 nouveaux n-uplets fusionnés avec deux ou trois n-uplets identiques correspondants à des zones voisines. Par conséquent, ils sont fusionnés dans les nouveaux secteurs S1, S2, S3, S4 et S5.

Tableau 4.11 Tableau du sixième niveau du processus de généralisation

name	Class	Hum_C	Count	Grouped zones
S1	Low	1	306	S1
S2		1	12	S11
Arzew		1	34	S11
El Ancor		1	2	S12
Ain El Bia		2	8	S11
Boufatis		2	3	S11
Boutlelis		3	6	S12
S3		3	8	S11
S4		3	36	S11
S5		3	8	S11
Ain kerma	High	1	6	Ain kerma
Bousfer		1	8	S10
Benfreha		3	12	S9
El Braya		3	5	S10

Misserguine	3	31	S10
S8	2	195	S10
Gdyel	2	17	S9

Le jeu de données résultant des foyers d'épidémies est illustré par le tableau 4.11. L'attribut Hum_C sera ensuite supprimé et les tuples identiques seront fusionnés afin de créer le tableau 4.12. Il y a 4 nouveaux uplets contenant plusieurs uplets identiques correspondant à des zones voisines. Par conséquent, ils sont fusionnés dans les nouveaux secteurs S9, S10, S11 et S12.

Tableau 4.12 Tableau du septième niveau du processus de généralisation

Name	Class	Count	Grouped zones
S1	Low	306	S1
S11		109	S11
S12		8	S12
Ain kerma	High	6	Ain kerma
S10		239	S10
S9		29	S9

Le jeu de données des foyers d'épidémies résultant est illustré dans le tableau 4.12. L'attribut Class_C sera supprimé et des n-uplets identiques seront fusionnés afin de créer une nouvelle table. Il existe des n-uplets avec des n-uplets identiques mais ne pouvant pas être fusionnés dans de nouveaux secteurs car ils ne correspondent pas aux zones voisines, nous conservons donc le tableau 4.12.

Étape 4 : Présentation du résultat de la comparaison

Selon le quatrième niveau de généralisation spatiale, les zones voisines S6, S7 sont fusionnées dans S8. La zone généralisée se forme au centre de la zone d'étude en raison des mêmes caractéristiques des zones fusionnées. Nous pouvons donc extraire la règle discriminante suivante décrivant le niveau élevé de risque épidémiologique :

$$\forall(x)High \leftarrow Inci_rate_C(x) \in sup57 \cap Hum_C \in 2$$

La synthèse géographique de la généralisation spatiale est illustrée selon les étapes suivantes :

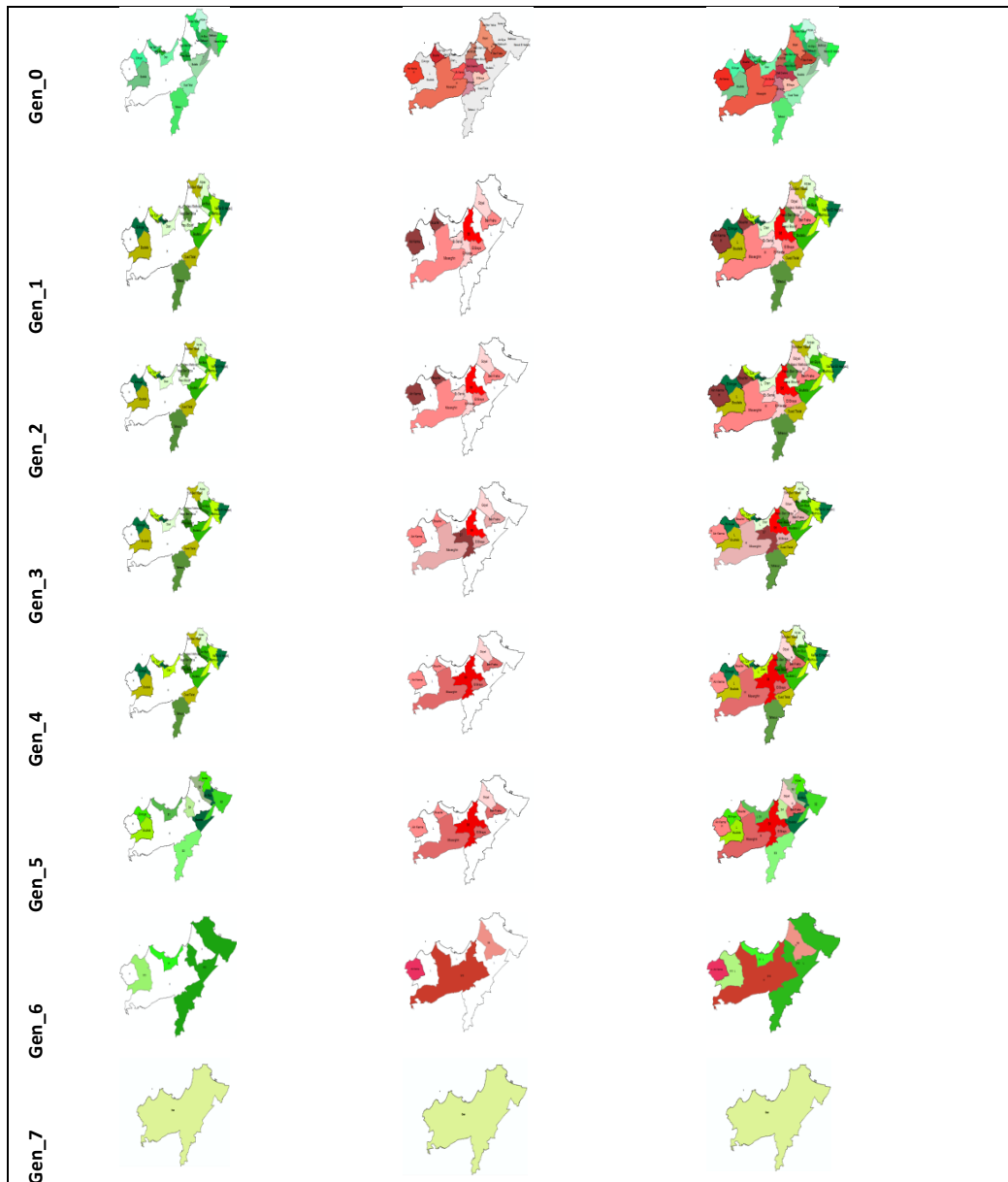


Figure 4.8 Visualisation du processus de généralisation spatiale

La figure 4.8 montre les différentes vues des régions géographiques des deux classes « Low » et « High ».

- Sur la première colonne la classe « Low » est illustrée par un dégradé de couleur verte montrant les zones à risque faible.
- La deuxième colonne, la classe « High » est illustrée par un dégradé de couleur rouge montrant les zones à risque élevée.
- La troisième colonne n'est autre que le chevauchement des deux classes High et Low sur la même carte.

Les lignes, quant à elles, montrent les différentes opérations de fusion appliquées pour chaque niveau de généralisation. Nous rappelons que chaque niveau de généralisation correspond à la suppression progressive de chaque attribut, selon son degré de pertinence (calculé par le gain informationnel de cet attribut). Nous commençons par la suppression de l'attribut le moins

pertinent ce qui correspond au niveau de généralisation Gen_1, puis le deuxième attribut le moins pertinent ce qui correspond au niveau Gen_2.. et ainsi de suite jusqu'à la dernière étape de généralisation Gen_7 qui correspond à toute la zone géographique.

4.7.4 Arbre de généralisation

L'arbre de généralisation de la figure 4.7 résume les différentes étapes de généralisation en montrant à chaque étape le niveau de généralisation correspondant à chaque suppression d'attribut ainsi que les cartes géographiques résultantes des opérations de fusion effectuées après chaque suppression d'attributs.

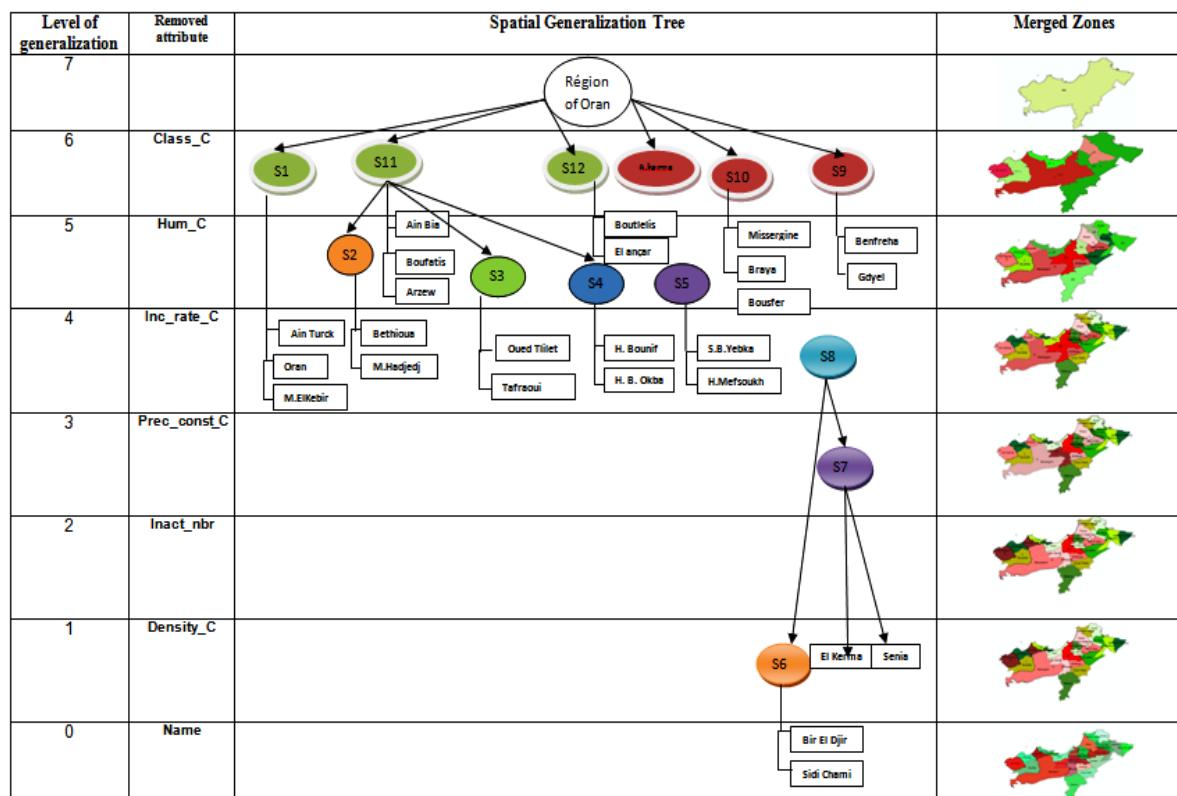


Figure 4.9 Arbre de généralisation spatiale

Nous avons repris dans la figure 4.8, les mêmes étapes de la figures 4.9 en conservant les étapes de généralisation dans la première colonne et les cartes avec chevochement de classes « High » et « Low » dans la quatrième colonne. Dans cette figure, nous avons ajouté plus de détails sur l'opération de généralisation en citant dans la deuxième colonne, l'attribut supprimé à chaque niveau ainsi que les différents secteurs issus des différentes opérations de fusions dans la troisième colonne. Les cartes de la quatrième colonne à droite représentent les régions avec un risque élevé en dégradé de couleur rouge et les zones à risque faible en dégradé de couleur verte.

4.7.5 Discussion

Nous avons initié une nouvelle approche SOLAM basée sur la généralisation spatiale appliquée en fonction des caractéristiques générales des zones géographiques à la place de l'agrégation spatiale conventionnelle appliquée dans les systèmes SOLAP.

Nous avons adopté la méthode de la **généralisation à dominante non spatiale** basée sur la méthode **d'induction orientée attribut**. Cette méthode a enrichi SOLAP pour ne plus se limiter à une simple fonction d'agrégation selon la division administrative de la zone d'étude, mais pouvant être basée sur les caractéristiques des régions géographiques. Le moteur SOLAM aide à classer les facteurs démographiques et socio-environnementaux selon leur degré de pertinence. Ainsi, les épidémiologistes peuvent concentrer leurs actions d'éradication et leurs stratégies de prévention à long terme en éliminant les facteurs par ordre de pertinence. Le système EPISOLAM présente une nouvelle fonctionnalité de sectorisation (régionalisation) permettant d'augmenter à un niveau de généralisation des zones géographiques chaque fois qu'un attribut non pertinent est supprimé de la base de données.

4.8 Travaux connexes

Le concept OLAM a été proposé pour la première fois dans (Han, 1997) où il a été suggéré comme un mécanisme qui intègre l'analyse en ligne (OLAP) avec la fouille de données. Puis d'autres systèmes en ont découlé en se basant sur la méthode d'induction orientée attribut. Le prototype DBMiner a été développé et généralisé pour de multiples fonctions d'exploration de données telles que la caractérisation (caractérisateur de DBMiner), la comparaison, la classification (classificateur de DBMiner), l'association (associeur de DBMiner), la prédiction (prédicteur de DBMiner) et le clustering. Plusieurs manières pour réaliser cette intégration ont été proposées, à savoir : cubage puis extraction, extraction puis cubage, extraction en cubes, extraction en arrière et extraction comparative.

Le processus de la généralisation est basé sur la méthode d'induction orientée attribut (AOI) proposée dans (Han et Fu, 1995) et un algorithme d'induction orienté attribut de base a été proposé dans (Han et al. 1997).

Par la suite, GeoMiner a été conçu dans (Han et al. 1997) comme l'extension de DBMiner aux données spatiales et propose trois modules : Geo-Characterizer, Geo-Comparator et Geo-Associator. Ils sont basés sur la généralisation spatiale définie dans (Roddick et Brian, 2009) en tant que processus d'agrégation de données selon les hiérarchies de concepts. La découverte des relations générales entre données spatiales et non spatiales peut être réalisée par induction orientée attribut de deux manières : (1) généralisation à dominante spatiale (GDS) et (2) généralisation à dominante non spatiale (GDNS). La première est l'induction dirigée par l'hiérarchie spatiale qui effectue la généralisation d'abord, sur les données spatiales ; tandis que la seconde est une induction orientée attribut non spatial qui effectue la généralisation sur des données non spatiales en premier (Lu et al., 1993).

D'autres types de généralisation spatiale ont également été indiqués en tant que nouvelles alternatives de généralisation telles que la généralisation alternée entre données spatiales et non spatiales et la généralisation sur plusieurs cartes thématiques. Plusieurs algorithmes ont été proposés pour réaliser la généralisation des données spatiales tel que l'algorithme GenDis dans (Knorr et Ng, 1996).

Selon les auteurs dans (Bedard et al, 2005), les outils clients actuellement utilisés pour exploiter les entrepôts de données conventionnels ne conviennent pas aux entrepôts de données

géospatiales car ils n'exploitent pas la structure géométrique des données. Dans cette optique, les technologies SOLAP ont été développées et mises en œuvre (Bedard, 2004).

Plus récemment, dans (Warnar, 2010), l'auteur a résumé la méthode de la généralisation spatiale utilisée pour générer des règles caractéristiques en utilisant la méthode d'induction orientée attribut en huit étapes séquentielles ordonnées comme suit: généralisation des données en composants décomposables plus petits; suppression d'attributs non pertinents; généralisation d'attribut en appliquant l'ascension des concepts de l'arbre; propagation des calculs; contrôle de seuil de généralisation sur chaque attribut; contrôle de seuil de généralisation sur les relations généralisées; transformation des règles et enfin, la gestion des nuplets qui se chevauchent. Cette dernière étape est spécifique à la génération des règles de classification et les règles de comparaison (discriminantes).

4.9 Conclusion

Nous avons initié dans ce chapitre une nouvelle approche SOLAM basée sur la généralisation spatiale appliquée en fonction des caractéristiques générales des zones géographiques à la place de l'agrégation spatiale conventionnelle appliquée dans les systèmes SOLAP.

Nous avons adopté la méthode de la généralisation à dominante non spatiale basée sur la méthode d'induction orientée attribut. Cette méthode a enrichi SOLAP pour ne plus se limiter à une simple fonction d'agrégation selon les divisions administratives de la zone d'étude, mais en se basant sur les caractéristiques des régions géographiques. Cette première contribution, en proposant un nouveau mécanisme SOLAM, constitue le noyau principal de cette catégorie de systèmes d'aide à la décision, ce qui mérite des études ultérieures plus complètes.

Nous suggérons d'identifier d'autres mesures et règles susceptibles de capturer des modèles de données non encore capturés par les règles de caractérisation. Un test efficace d'autocorrélation spatiale permettrait de remédier à ce manque d'informations. Les analyses d'autocorrélation globales et locales semblent très utiles pour analyser les phénomènes spatiaux ou spatiotemporels. Cela pourrait être introduit en calculant des indices géostatistiques (par exemple, les indices de Moran et de Lisa, respectivement) à chaque niveau de généralisation (voir chapitre 3).

Nous essayerons également d'introduire l'aspect temporel de l'étude de l'autocorrélation dans le chapitre 6 consacré à l'analyse spatiotemporelle en appliquant les techniques de géostatistique au système EPISOLAP-MINING afin d'analyser l'autocorrélation entre des sites à différentes périodes du pic épidémique. Il pourrait également être utile d'explorer le comportement de l'algorithme proposé lorsque le jeu de données devient très volumineux en termes de nombre de zones enregistrées et pendant de longues périodes.

Dans le chapitre 5, nous présentons notre deuxième contribution dans ce travail de thèse qui a pour principal objectif d'intégrer les méthodes d'analyse multicritères et les techniques de fouille de données dans le système SDSS_MC A_DM suggéré.

Chapitre 5

Contribution 2 :

Le système SDSS-MCA-DM proposé

Ce chapitre présente une description détaillée du système SDSS-MCA-DM proposé pour la surveillance épidémiologique de la tuberculose à Oran. Il présente une nouvelle contribution pour intégrer les des méthodes multicritères, des méthodes de fouille de données ainsi que les systèmes d'information géographique pour répondre adéquatement aux besoins des décideurs en santé publique.

Plan du chapitre

1 Introduction	87
2 Motivation.....	87
3 Objectif visé.....	88
4 Outils d'investigation	89
5 Description du modèle décisionnel proposé.....	95
6 Processus décisionnel adopté par SDSS-MCA-DM	96
6.1 Phase de structuration	97
6.2 Phase d'exploitation	98
6.3 Phase d'exécution	98
7 Modélisation UML du système décisionnel proposé.....	98
8 Expérimentation du système SDSS_MCA_DM	101
8.1 Formulation du problème décisionnel	101
8.2 Processus d'exploration de données	108
9 Travaux connexes	110
9.1 Travaux connexes sur la surveillance épidémiologique	110
9.2 Travaux connexes sur les méthodes d'analyse multicritères	111
9.3 Travaux connexes sur SOLAP	112
9.4 Travaux connexes sur le datamining	112
9.5 Synthèse.....	113
10 Conclusion	115

5.1 Introduction

Il existe de nombreux systèmes d'aide à la décision pour la surveillance épidémiologique. Ils visent tous à aider les décideurs en santé publique à contrôler et à maîtriser la situation de risque épidémiologique encourue. Leur rôle est de fournir un soutien aux épidémiologistes dans leur tâche de surveillance. Cependant, ces systèmes traditionnels d'aide à la décision se limitent à une simple analyse spatio-temporelle des données de la surveillance et à la détection visuelle des foyers d'épidémies sur une carte géographique.

Les épidémiologistes ont, souvent, besoin d'un outil automatique qui interprète les résultats de l'analyse des facteurs réels et responsables de la propagation des épidémies. Une carte de risque épidémiologique est utile en cas de pic épidémiologique quand les décideurs doivent prendre des décisions rapides et efficaces afin de minimiser les taux de mortalité et de morbidité.

5.2 Motivation

La propagation des épidémies présente encore un risque majeur pour la population dans notre société et dans le monde. Nous assistons, de nos jours, à l'apparition de nouvelles épidémies (le virus de Zika, Ebola, etc.) ou encore à la réapparition des maladies anciennes (la rougeole, le cholera, la tuberculose, etc.).

Les décideurs ont fortement besoins des systèmes puissants pour les aider dans leur tâche de surveillance épidémiologique. L'enregistrement des données périodiques de la surveillance est une étape importante voire nécessaire pour les maladies à déclaration obligatoire. Toutefois, les moyens de traitement de ce type de données souvent hétérogènes (plusieurs facteurs de risque) et provenant de plusieurs sources sont limités. S'agissant des phénomènes comme l'épidémie, les dimensions spatiales (et temporelles) sont éminemment importantes dans l'analyse et la qualification de la propagation du phénomène au voisinage, ainsi que son émergence ou réémergence.

Les dernières années ont vu le développement des travaux sur le spatial OLAP ou SOLAP permettant d'intégrer les données spatiales dans l'OLAP et de lier l'exploration et la visualisation cartographique. La démarche naturelle est donc d'appliquer ces approches et les systèmes qui ont en découlé à l'analyse épidémiologique, ce que nous avons fait lors d'un travail antérieur « le projet EPISOLAP » (Zemri, 2011). Par la suite, des limites ont été détectées : Ce système ne fournit pas une interprétation des résultats de l'analyse multidimensionnelle. Par conséquent, le système EPISOLAP a fortement besoin d'être enrichi et complété.

Afin d'atteindre ce but, nous proposons, dans ce chapitre, une combinaison de la technologie SOLAP avec les méthodes d'analyse multicritères (MAMC) ainsi que les techniques de fouille de données (FDD) offrant une analyse de données plus riche. Cette approche comprend donc deux types d'extension des systèmes SOLAP.

1. La première est d'ordre exploratoire et intègre l'identification des foyers dans SOLAP avec la priorisation des actions.
2. La seconde est d'ordre préventif et intègre l'entrepôt de données et la navigation SOLAP avec les techniques de fouille de données prédictives permettant de prévoir l'apparition ou la disparition des foyers d'épidémie.

Les problématiques majeures à traiter dans ce chapitre sont donc :

1. La géo-localisation des foyers d'épidémies dans une zone d'étude,
2. La priorisation des foyers d'épidémies pour établir un ordre d'intervention et
3. La prévention de l'apparition de nouveaux foyers d'épidémies.

La solution proposée vise à faciliter la tâche de la surveillance en adoptant les méthodes multicritères et de fouille de données automatique ce qui a donné naissance au système **SDSS-MCA-DM**. Ce système sert à trier les foyers d'épidémies dans des catégories selon le degré de risque épidémiologique des zones géographiques. Par conséquent, l'ordre d'intervention du personnel en santé sera plus efficace (surveillance curative).

Ce système intègre, également, les techniques de fouille de données, principalement les méthodes prédictives permettant de prévoir l'apparition de nouveaux foyers ou bien la disparition des foyers existants (grâce à des efforts d'éradication) selon les facteurs de risque en cours ce qui permettra de prendre des mesures préventives (surveillance préventive). Enfin, ce système permet de visualiser les résultats sur une carte de risque épidémiologique très utile aux décideurs en santé publique.

5.3 Objectifs visés

On s'intéresse dans cette étude, à la conception et au développement d'un système d'aide à la décision basé sur les méthodes d'analyse multicritères et les techniques de fouille de données. Le système proposé génère des recommandations à l'intention des décideurs sous forme d'une carte de risque épidémiologique. Cette dernière permet de classer les foyers des épidémies par catégories. L'approche adoptée repose principalement sur l'utilisation du couplage MCA-FDD (analyse multicritères et fouille de données).

Le sous-système MCA affecte les foyers d'épidémies dans des catégories distinctes qui correspondent chacune à un niveau de risque {Fort, moyen, faible} afin de faciliter l'intervention des décideurs par ordre de priorité.

Le sous-système FDD étudie la corrélation et l'association entre les critères (facteurs de risque) et génère en conséquence des règles qui prédisent et préviennent l'apparition de nouveaux foyers d'épidémies. Les objectifs spécifiques sont (Zemri et Hamdadou, 2019b) :

1. Enrichir le système suggéré **SDSS_AMC_DM** par l'intégration des méthodes d'analyse multicritères afin d'affecter les actions potentielles (les foyers) dans différentes catégories en fonction du degré de risque épidémiologique.
2. Enrichir le système proposé en intégrant des techniques de fouille de données, principalement les méthodes prédictives, et générer des règles pour prendre des mesures préventives.
3. Visualiser les résultats sur une carte de risque épidémiologique

La figure 5.1 donne une vue d'ensemble du système d'aide à la décision proposé.

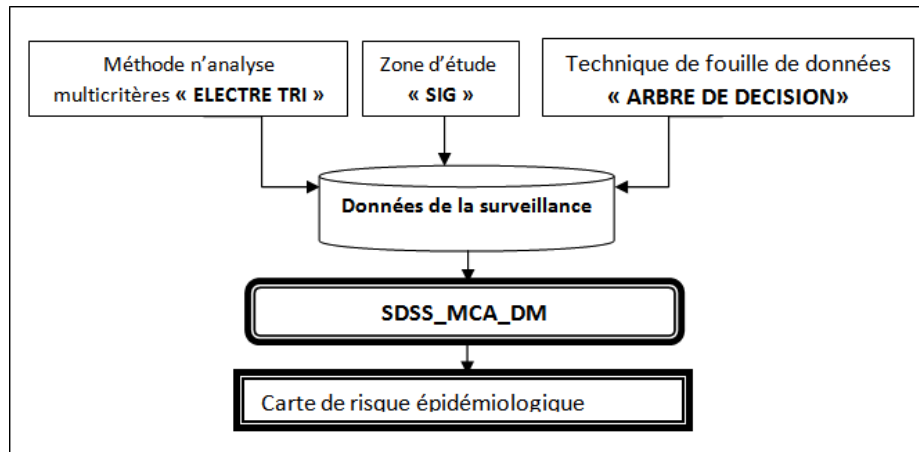


Figure 5.1.Le système d'aide à la décision proposé SDSS-MCA-DM : Vue d'ensemble

5.4 Outils d'investigation

Notre approche repose sur trois processus principaux : le processus de la surveillance épidémiologique, le processus d'analyse multicritères d'ELECTRE et le processus de fouille de données par arbre de décision. Le choix de la méthode ELECTRE et de la technique de l'arbre de décision est justifié dans ce qui suit.

5.4.1 Le processus de la surveillance épidémiologique

La surveillance épidémiologique est un processus de prise de décision. Ce processus suit les quatre principales étapes décrites dans (Zemri et al., 2013): enregistrement des données, analyse des données, interprétation des résultats d'analyses et action de la surveillance. Cette approche technique est adoptée par les épidémiologistes dans le processus de surveillance épidémiologique pour développer la prise de décision.

En général, l'aide à la décision en épidémiologie prend deux formes : l'aide à la décision opérationnelle utilisant des actions curatives (surveillance curative) et l'aide à la décision stratégique prenant des mesures proactives (surveillance préventive).

1. **L'aide à la décision opérationnelle** est la décision prise à court terme ayant un objectif curatif, dans le cas d'un pic épidémique. Dans ce cas, le décideur doit prendre des décisions rapides et efficaces pour limiter les risques de propagation.
2. **L'aide à la décision stratégique** est la décision prise à long terme ayant un objectif préventif afin de prévoir l'apparition de nouveaux foyers d'épidémies (la prédiction), en tenant compte des facteurs socio-environnementaux favorisant la contamination.

Selon les auteurs dans (Dufour et Hendrix, 2007), la surveillance peut être **active** ou **passive**, selon les méthodes utilisées. Nous pouvons, également, trouver une surveillance **exhaustive** qui est utile pour estimer les tendances des preuves sur de longues périodes et pour décrire les événements de santé rares et graves. La surveillance active fournit généralement de meilleurs résultats que la surveillance passive, mais nécessite des coûts humains et financiers élevés, dont

l'utilité doit être évaluée en fonction des objectifs. Dans la présente étude, nous nous intéressons à la surveillance préventive.

5.4.2 La méthode multicritère ELECTRE

La méthode d'analyse multicritères ELECTRE qui signifie «Élimination et Choix traduisant la réalité», traite un problème de type β de procédure d'affectation ou de classification traitant le problème de tri en attribuant chaque action, ce qui est une contribution possible dans la décision globale, à une catégorie prédéfinie. Les actions de référence permettent de segmenter l'espace des critères en catégories. Le critère est une expression quantitative ou qualitative qui examine l'action. La méthode ELECTRE est largement utilisée dans la décision multicritère pour résoudre les problèmes de classification ordinaire. Certains chercheurs tentent d'étendre la méthode ELECTRE en intégrant d'autres méthodes. Dans (Corrente et al, 2016), les auteurs étendent la méthode ELECTRE A ELECTRE-B, ELECTRE-C et ELECTRE-NC en étendant Simos – Roy – Figueira (SRF) et en adaptant le processus d'hierarchie à critères multiples (MCHP) à déterminer le poids des critères hiérarchiquement structurés.

5.4.2.1 Pourquoi ELECTRE ?

La méthode ELECTRE a quatre intérêts principaux qui permettent (Hamdadou et al, 2017) :

- Auto-juger une action potentielle indépendamment des autres actions potentielles. Cette méthode considère chaque action potentielle par sa valeur absolue selon des actions de référence prédéfinies.
- Fixer une ou plusieurs valeurs de référence telles que des normes ou des résultats minimaux nécessaires à l'acceptation ou au refus des actions.
- Considérer un grand nombre d'actions potentielles par rapport aux autres méthodes ELECTRE.
- Comparer les différentes actions potentielles, non seulement entre elles mais aussi avec des actions de référence.

5.4.2.2 Principe générale d'ELECTRE TRI

Chaque catégorie est bornée entre une limite supérieure et une limite inférieure et chaque action de référence représente la borne inférieure de la catégorie supérieure et (en même temps) la borne supérieure de la catégorie inférieure.

ELECTRE offre deux procédures principales (Ben Mena, 2001) : **Agrégation** et **Exploitation**.

Procédure d'agrégation : implique la construction de plusieurs relations de surclassement. Elle vise à comparer, de manière complète, chaque paire d'actions en utilisant deux principes : la concordance et la discordance. Dans ELECTRE, la relation de surclassement peut être interprétée comme une relation floue. L'action "a" surclasse l'action "b", notée ($a \succ b$), si "a" est au moins aussi bonne que "b" selon la majorité des critères, mais pas trop mauvaise que b selon le reste des critères. La construction de cette relation nécessite la définition d'un indice de crédibilité $\rho(aSb)$, qui caractérise la crédibilité de l'affirmation "aSb". Elle est définie en utilisant à la fois l'indice de concordance $c_j(aSb)$ et l'indice de discordance $d_j(aSb)$ pour chaque critère (g_j)

dans F. Une matrice de degrés de crédibilité est ensuite générée (matrice d'ordre n avec le degré de crédibilité $\rho(aSb)$ pour chaque paire d'actions (a, b).

Procédure d'exploitation : cette procédure est utilisée pour clarifier les recommandations (classer les actions) à partir des résultats obtenus lors de la première phase (Maystre et al., 1994) (Roy, 1993). La phase d'exploitation utilise la matrice de crédibilité (générée lors de la phase d'agrégation) pour détecter l'incompatibilité.

5.4.2.3 Les étapes de la méthode ELECTRE TRI

ELECTRE procède comme suit (Hamdadou et al, 2017) :

Étape 1 : Évaluation de l'indice de concordance

Le principe de concordance exige que la majorité des critères (en tenant compte de leur importance relative) soient en accord avec la relation de surclassement "aSb". Le remplissage de la matrice de concordance par critère est effectué en calculant les critères de concordance selon la formule (1) suivante :

$$c_j(a_i, b_k) = \begin{cases} 0 & \text{si } g_j(b_k) - g_j(a_i) \geq p_j \\ 1 & \text{si } g_j(b_k) - g_j(a_i) \leq q_j \\ g_j(a_i) - g_j(a_k) + p_j / (p_j - q_j) & \text{si } q_j < g_j(b_k) - g_j(a_i) < p_j \end{cases} \quad (1) \text{ avec}$$

- $g_j(b_k)$ est la performance de l'action de référence "k"
- $g_j(a_i)$ est la performance de l'action de référence "ai" pour le critère j
- p_j est le seuil de préférence
- q_j est le seuil d'indifférence

Étape 2: Évaluation de l'indice de concordance globale

Après avoir réalisé les matrices de concordance pour chacun des critères (m matrices de dimension n * n incluant le critère de concordance $c_j(a_i, b_k)$), une matrice de concordance globale est ensuite réalisée (matrice n * m incorporant le critère de concordance globale c_{ik}).

$$C_{ik} = C(a_i, b_k) = \frac{\sum_{j=1}^m P_j * c_j(a_i, b_k)}{\sum_{j=1}^m P_j} \quad (2)$$

Étape 3 : Évaluation de l'indice de discordance partiel

La divergence avec l'hypothèse de surclassement "aiSbk" est déterminée pour chaque critère par le calcul de l'indice de discordance $d_j(a_i, b_k)$.

$$d_j(a_i, b_k) = \begin{cases} 0 & \text{si } g_j(b_k) - g_j(a_i) \leq p_j \\ 1 & \text{si } g_j(b_k) - g_j(a_i) \geq v_j \\ (g_j(b_k) - g_j(a_i) + p_j) / (v_j - p_j) & \text{si } p_j < g_j(b_k) - g_j(a_i) < v_j \end{cases} \quad (3)$$

- $g_j(b_k)$ est la performance de l'action de référence "k"
- $g_j(a_i)$ est la performance de l'action de référence "ai" pour le critère j
- p_j est le seuil de préférence
- v_j est le seuil d'indifférence

- v_j est le seuil de véto

Les matrices de discordance sont ensuite calculées pour chaque critère (matrice $n * m$ comprenant les critères de discordance $d_j(a_i, b_k)$). Ce principe est également appelé principe de respect de la minorité. Il s'agit de vérifier que la minorité des critères refusant la relation de surclassement "aSb" ne s'opposent pas : la relation de surclassement ne devrait pas aussi être pire pour la minorité de critères.

Étape 4: Évaluation du degré de crédibilité

Le degré de crédibilité est également déterminé par:

$$\delta_j(a_i, b_k) = C(a_i, b_k) * \frac{1 - d_j(a_i, b_k)}{1 - C(a_i, b_k)} \quad (4) \text{ avec}$$

- $C(a_i, b_k)$ est l'indice global de concordance
- $\bar{F}$ est le sous-ensemble de la famille des critères F qui a les critères pour lesquels l'indice de discordance est supérieur à l'indice d'accord global.

$$\bar{F} = \{j \in F, d_j(a_i, b_k) > C_j(a_i, b_k)\}$$

Une matrice des degrés de crédibilité est ensuite calculée (matrice de dimension $n * m$ incluant les degrés de crédibilité $\delta_j(a_i, b_k)$). L'indice de coupe " λ " est défini comme le paramètre qui détermine la situation de préférence entre "a" et "bk". " λ " est toujours dans l'intervalle [0.5, 1].

Étape 5 : Établissement de la relation de surclassement

La relation de surclassement entre une action potentielle "a" et l'action de référence "b" est établie en fonction du degré de crédibilité et d'un indice de coupe constant " λ ". Il est noté $\delta(a_i, b_k) \geq \lambda$ aSbk. "P" indique la préférence ; "I" désigne l'indifférence et "R" l'incomparabilité.

"a" et "bk" sont affectés à l'une des situations suivantes :

- aIbk ssi aSbk et [bkSa]
- aPbk ssi aSbk et non [bkSa]
- aRbk ssi non[aSbk] et non [bkSa]

Étape 6 : procédure d'affectation

Avant de discuter les procédures d'affectation, sept conditions doivent être considérées : unicité, indépendance, conformité, uniformité, monotonie, stabilité et enfin aucune action potentielle ne doit être indifférente d'une action de référence.

Pour détecter l'incompatibilité, deux procédures d'affectation distinctes, appelées optimistes et pessimistes, sont généralement nécessaires. Ces derniers consistent à comparer chaque action potentielle avec des actions de référence en commençant par la plus restrictive.

- **Affectation pessimiste** : affecte l'action à une catégorie pour que l'action de référence supérieure de cette catégorie soit préférée à l'action.
- **Affectation Optimiste** : affecte l'action dans une catégorie de sorte que cette action surclasse l'action de référence inférieure de cette catégorie.

La procédure d'utilisation de la méthode ELECTRE est donnée par le diagramme illustré sur la figure 5.2 ci dessous.

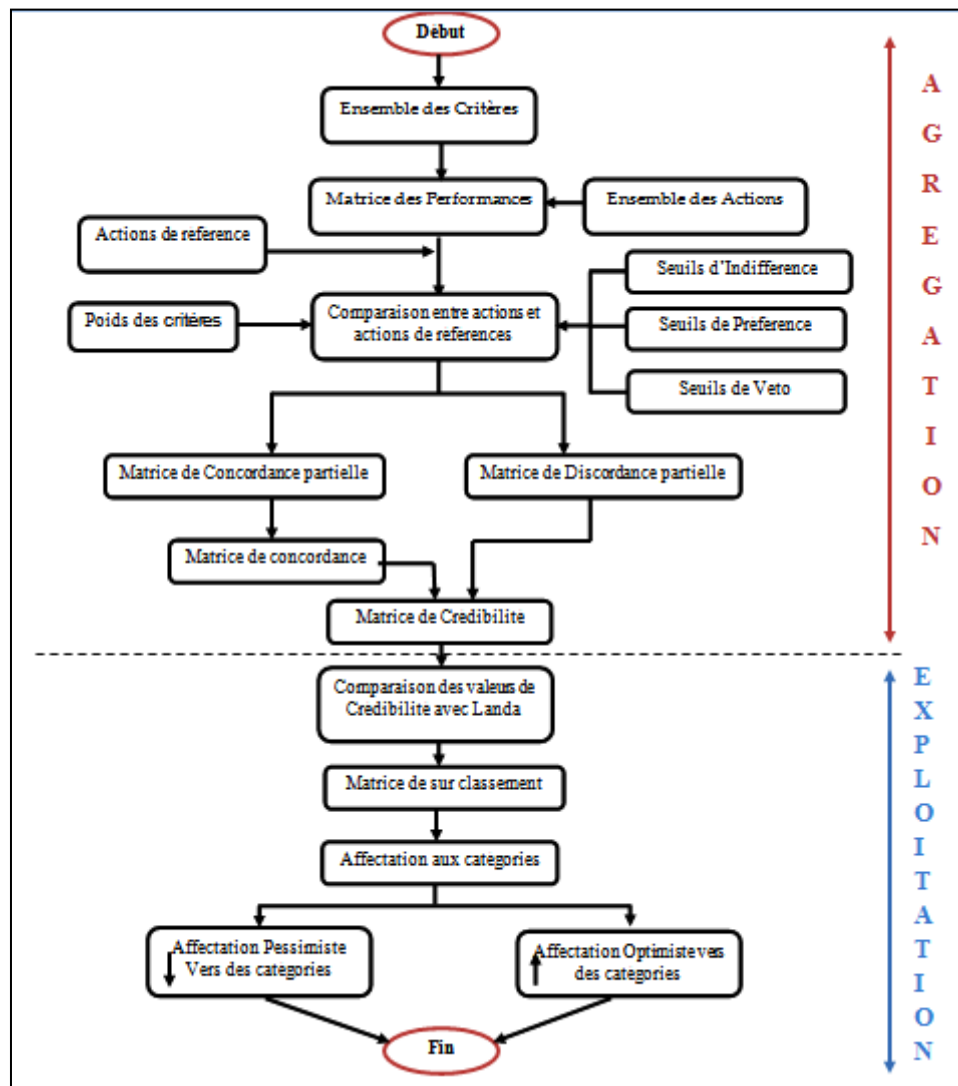


Figure 5.2 La démarche de la méthode ELECTRE

5.4.4.3 Processus de fouille de données par arbre de décision

La fouille de données (FDD) est le processus qui utilise les techniques d'apprentissage automatique et intelligent pour analyser et extraire des connaissances à partir de grands volumes de données. Le processus de FDD suit, selon (Amine, 2013) les étapes suivantes : identification du problème, préparation des données, sélection du modèle et enfin utilisation du modèle tel qu'il est présenté sur la figure 5.3.

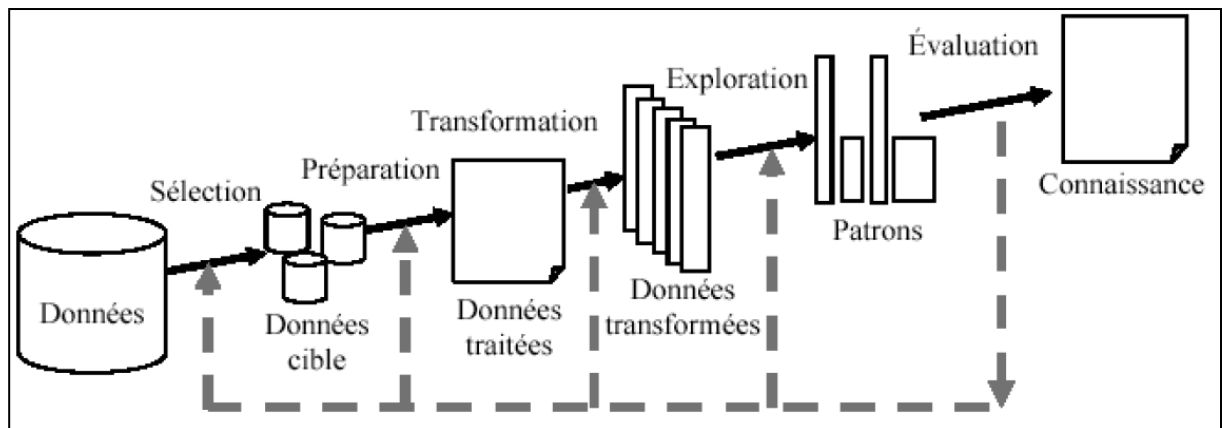


Figure 5.3 Processus de fouille de données

Le type de modèle de data mining peut être choisi selon trois points de vue (Amine, 2013) : en fonction des **objectifs** (*classification, prévision, association et segmentation*), selon le **type d'apprentissage** (*modèle prédictif et descriptif*) ou du **type de modèle** (*apprentissage supervisé et apprentissage non supervisé*). Les techniques doivent être adaptées au problème considéré (Scharff, 2004). Le tableau 5.1 présente un résumé des modèles de fouille de données existants.

Tableau 5.1 Tableau de classification des modèles de datamining

Objectif	Type d'apprentissage	Type de modèle
classification	Apprentissage supervisé	Modèles prédictifs
Prédiction		
Association	Apprentissage non supervisé	Modèles descriptifs
Segmentation		

5.4.3.1 Pourquoi l'arbre de décision ?

Notre choix a porté sur l'utilisation de la méthode de l'arbres de décision pour ses nombreux avantages. Les arbres de décision (Atmani, 2008) :

- Sont simples à comprendre et à interpréter.
- Peuvent avoir des résultats même avec peu de données précises. Des informations importantes peuvent être générées à partir d'experts décrivant une situation (ses alternatives, ses probabilités et ses coûts) et leurs préférences en matière de résultats.
- Permettent l'ajout de nouveaux scénarios possibles.
- Aident à déterminer les valeurs les plus défavorables, optimales et attendues pour différents scénarios.
- Utilisent un modèle de boîte blanche. Si un résultat donné est fourni par un modèle.
- Peuvent être combinés avec d'autres techniques de d'aide à décision.

5.4.3.2 Principe générale de l'arbre de décision

Construire un arbre de décision consiste à utiliser des attributs pour subdiviser progressivement l'ensemble d'exemples en sous-ensembles plus précis. Cela signifie que l'ensemble d'exemples constituera la racine du nœud de l'arbre et sera divisé en sous-ensembles, qui constituent des nœuds intermédiaires (Zighed & Rakotomalala, 2000). Chaque nœud représente un test sur la valeur d'un ou de plusieurs attributs, et chaque branche partant d'un nœud correspond à une ou plusieurs valeurs de ce test, puis chaque feuille de l'arbre est associée à une valeur d'attribut endogène (Zouggar, 2005).

5.4.3.3 Etapes de l'algorithme C4.5

C4.5 est un algorithme développé par Ross Quinlan. Il est utilisé pour générer un arbre de décision. C4.5 est une extension de l'algorithme ID3 antérieur de Quinlan. Les arbres de décision générés par C4.5 peuvent être utilisés pour la classification et, pour cette raison, C4.5 est souvent utilisé en tant que classificateur statistique. L'algorithme général C4.5 utilisé pour créer les arbres de décision se déroule en suivant les étapes suivantes :

1. Rechercher les cas de base.
2. Pour chaque attribut "a", calculer le gain d'informations normalisé à partir de la division en "a".
3. Soit "a_best" l'attribut avec le gain d'information normalisé le plus élevé.
4. Créer un nœud de décision qui se scinde en "a_best".
5. Revenir sur les sous-listes obtenues en scindant sur "a_best" et ajoutez ces nœuds en tant que nœuds fils.

Dans notre étude, nous avons choisi l'outil de classification J48 de weka, qui est une implémentation de l'algorithme C4.5.

5.5 Description du système décisionnel proposé

Le système d'aide à la décision proposé dans la présente étude contribue dans la prévention et la prédiction préalable de l'apparition des nouveaux foyers d'épidémie en connaissant les véritables facteurs de la propagation. Le prototype pourrait guider les efforts des épidémiologistes pour éradiquer l'épidémie. Le système SDSS-MCA-DM comprend les trois sous-systèmes suivants :

5.5.1 Le sous système d'analyse multicritères

Le sous-système d'analyse multicritères se base sur deux composants :

- **Le module de structuration** : il consiste à structurer les données de la surveillance dans un fichier ".sad" dédié au traitement multicritère.
- **Le module d'exploitation**: il consiste à exécuter la méthode multicritères ELECTRE sur le fichier ".sad".

Le processus multicritères commence par l'identification de la zone d'étude en passant par l'identification des critères et des actions, la génération du tableau de performance et des paramètres subjectifs, le traitement des structures de données par la méthode ELECTRE et se termine par la génération des classes et l'affichage des résultats d'affectation.

5.5.2 Le sous système de la fouille de données

Les classes générées par le sous-système d'analyse multicritères seront exploitées par ce sous-système pour générer un nouveau fichier de données ".arff" incluant toutes les variables permettant d'effectuer la fouille de données, complété par la catégorie du risque épidémiologique : {Elevé, moyen ou faible}. Ce sous-système exploite la méthode de l'arbre de décision en appliquant l'algorithme J48 de l'outil Weka. Ce processus se termine par la génération des règles prédictives.

5.5.3 Le sous système de visualisation

Ce sous-système comprend le système d'information géographique du système proposé SDSS-MCA-DM. Son rôle est la gestion des données spatiales et l'affichage des résultats sur une carte de risque épidémiologique.

La figure 5.4 ci-dessous présente l'architecture générale du système proposé SDSS-MCA-DM.

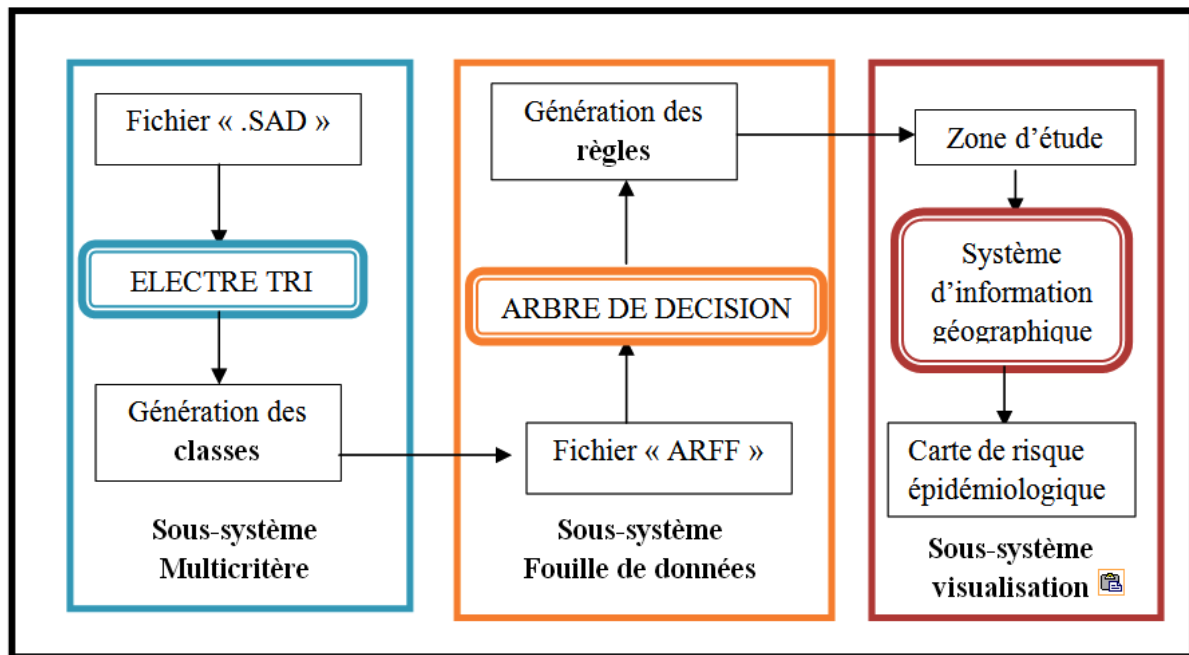


Figure 5.4 l'architecture du système décisionnel proposé SDSS-MCA-DM

5.6 Processus décisionnel adopté par SDSS_MCA-DM

L'architecture fonctionnelle du système d'aide à la décision proposé SDSS-MCA-DM, basée sur l'aide à la décision multicritères et la fouille de données, est illustrée par la figure 5.5. Ce modèle s'inspire du modèle proposé dans (Hamdadou, 2008) et repose sur les trois phases suivantes : i) phase de structuration, ii) phase d'exécution et iii) phase de d'exploitation (implémentation). Ces trois phases seront décrites dans ce qui suit.

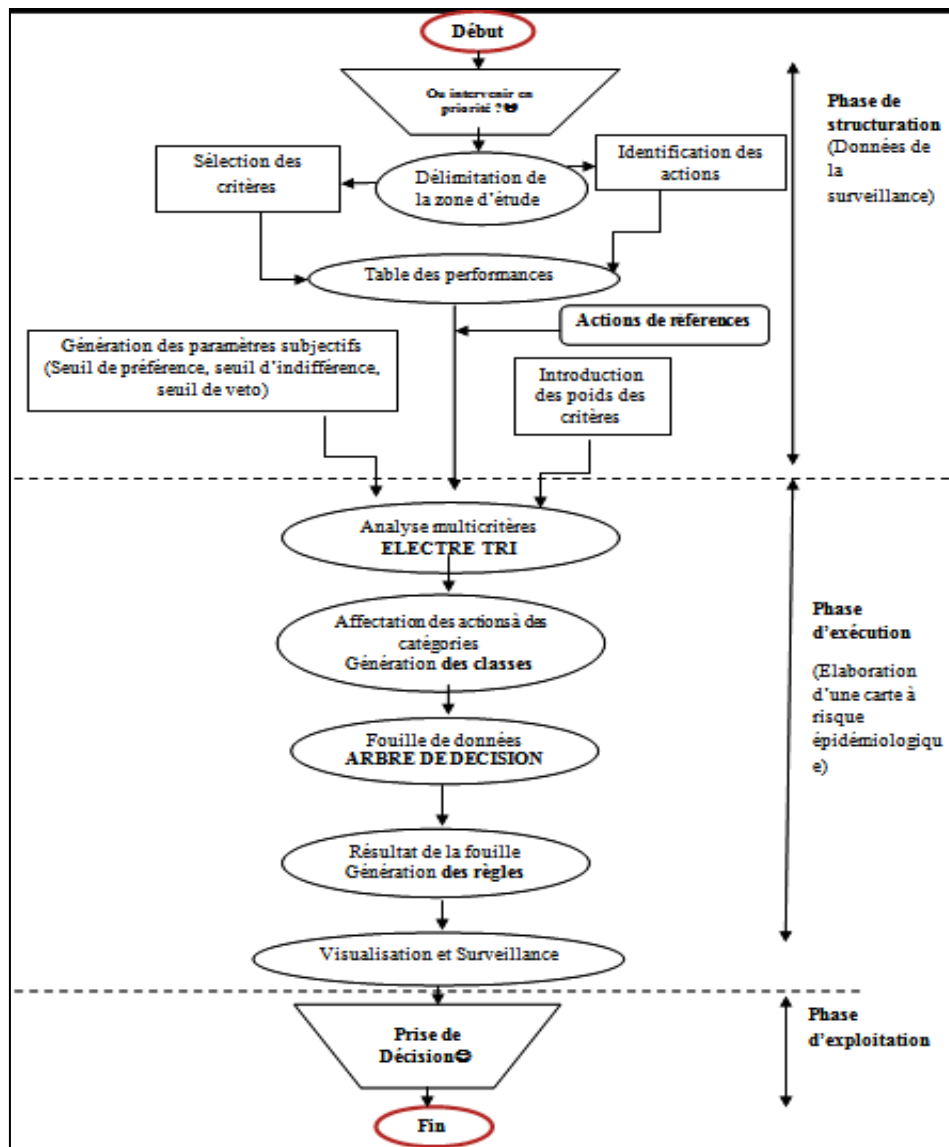


Figure 5.5 Processus décisionnel adopté par le système SDSS_AM_FD

5.6.1 Phase de structuration

Cette première phase vise à identifier le problème (localisation géographique de la zone d'étude à l'aide du SIG) et à définir les choix fondamentaux pour l'approcher. Cette phase vise également à formaliser deux éléments fondamentaux de la situation décisionnelle (Hamdadou et al, 2013) :

- **Identification des actions** : l'identification de toutes les actions potentielles est une étape très importante de toute approche d'aide à la décision, en particulier lorsque les méthodes d'analyse multicritères procèdent par agrégation partielle. L'ensemble de toutes les actions doit être complet car sa modification au cours de l'analyse peut conduire à refaire toute l'analyse multicritères.
- **Identification des critères** : la liste des critères obtenue par la collecte des facteurs correspondants doit être aussi complète que possible. Ces critères doivent être aussi liés aux contraintes et aux objectifs utilisés dans la production des règles prédictives. La

famille des critères les plus pertinents devrait vérifier les conditions d'exhaustivité, de cohérence et d'indépendance.

5.6.2 Phase d'exécution

Cette deuxième phase est le processus d'exécution des différents calculs utilisés par les méthodes adoptées par notre modèle décisionnel. Ses deux objectifs principaux sont l'évaluation des critères et l'agrégation de ces informations par une analyse multicritères.

5.6.3 Phase d'exploitation (implémentation)

Cette troisième phase comprend la mise en œuvre de la décision et le contrôle de l'exécution. La figure 5.6 résume les trois principales phases et étapes du modèle proposé.

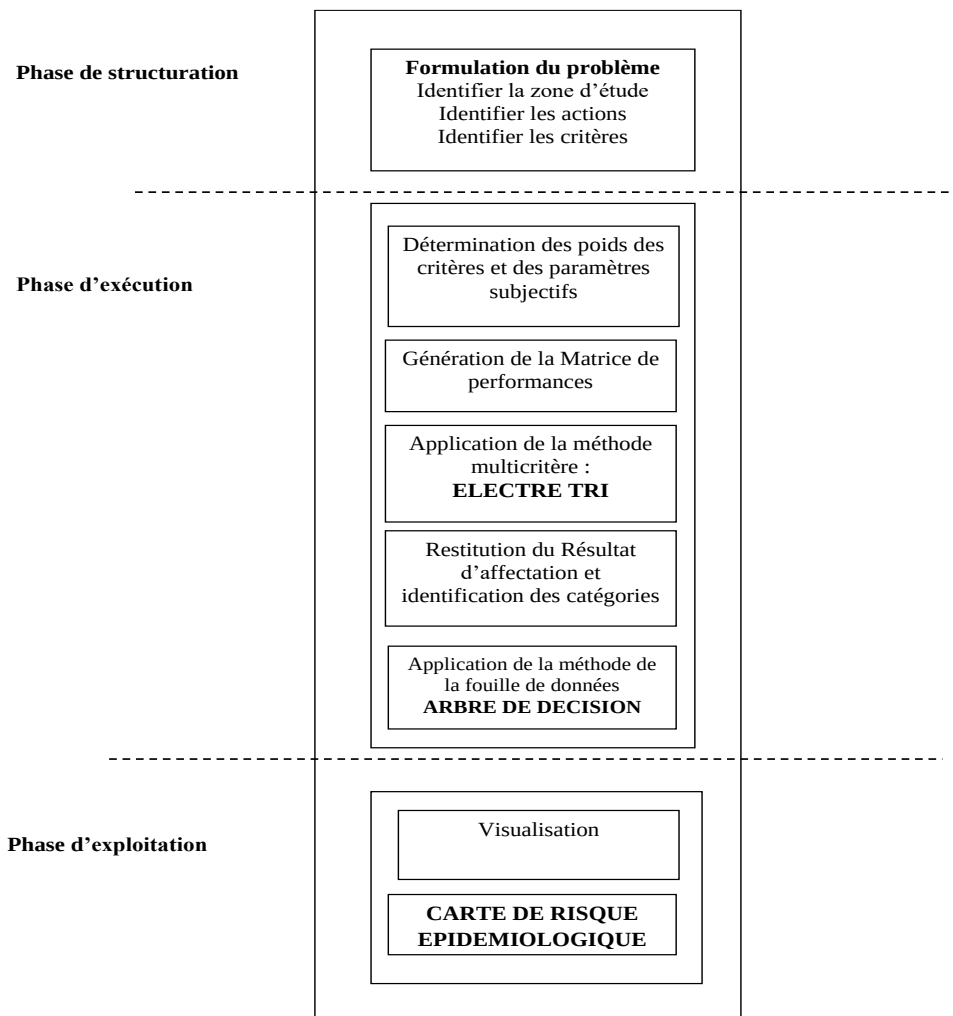


Figure 5.6 Modèle décisionnel proposé

5.7 Modelisation UML du système decisionnel proposé

Afin de représenter les interactions entre différents sous-systèmes du système d'aide à la décision SDSS-MCA-DM, nous optons pour l'utilisation du modèle UML, souvent utilisé pour décrire l'interaction entre différentes entités. La figure 5.7 présente le diagramme de séquence UML associé au système SDSS-MCA-DM.

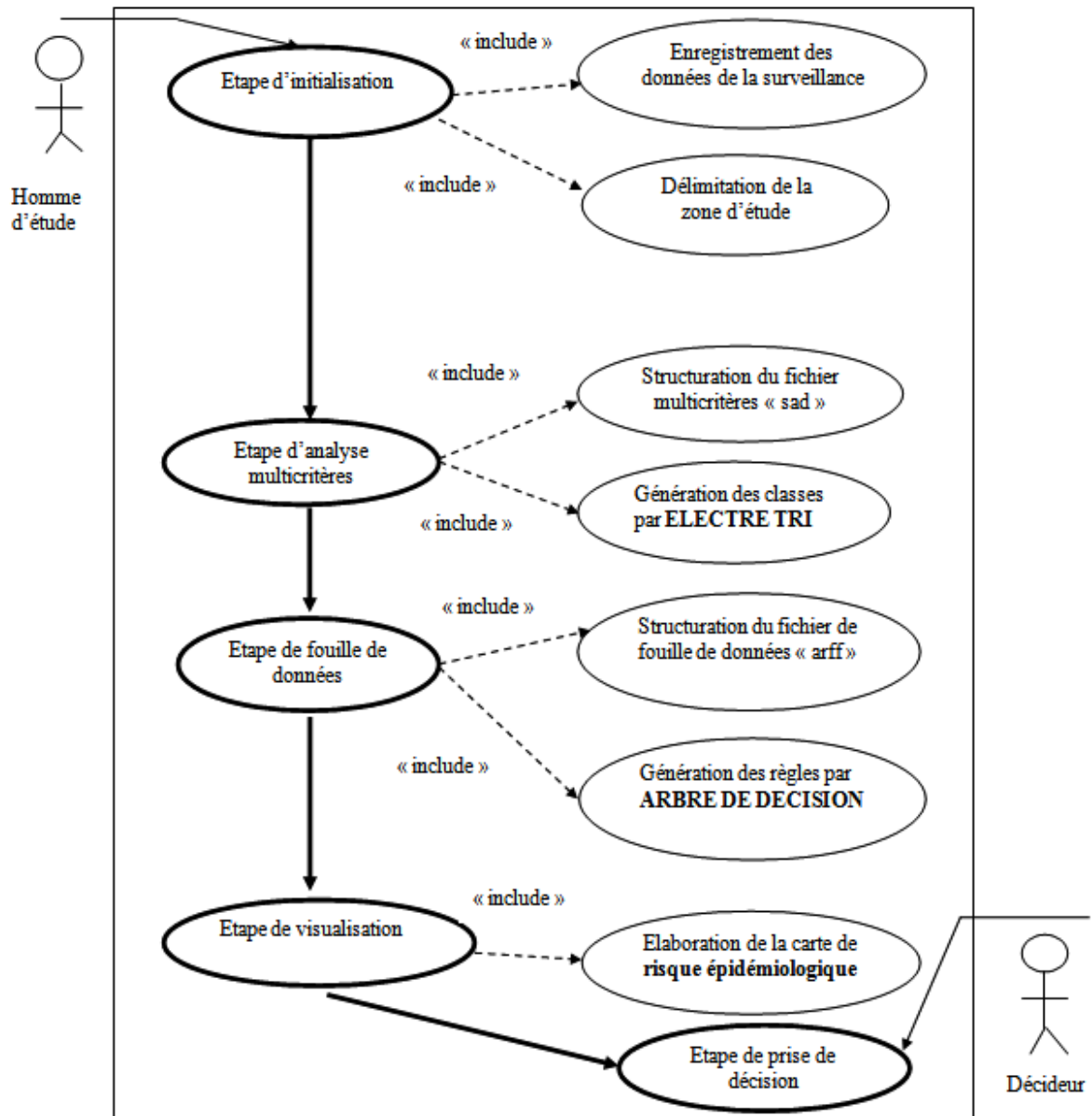


Figure 5.7.a Digramme UML du cas d'utilisation du système SDDD_MCA_DM

Le diagramme UML (de cas d'utilisation ou de séquence) illustre les différentes étapes du système SDSS-MCA-DM. Cinq étapes sont illustrées sur le tableau 5.2 :

Tableau 5.2 : Etapes des diagrammes UML (séquences et cas d'utilisation)

N°	Étapes des diagrammes UML	Explication
1	Étape d'initialisation	– enregistrer les données de la surveillance – délimiter la zone d'étude.
2	Étape multicritères	– Structurer les données de la surveillance dans un fichier multicritères "sad" – identifier les actions et les critères – déterminer les paramètres subjectifs (poids et seuils) – générer un tableau de performance – identifier les actions de référence

		– agréger et exploiter le fichier "sad" – attribuer les actions aux différentes classes, – afficher les résultats tabulaires et cartographiques.
3	Étape de fouille de données	– structurer le résultat de l'affectation dans le fichier "arff" – appliquer la technique de fouille de données de l'arbre de décision. – générer des règles de production – afficher le résultat de la prédiction de l'apparition/disparition des nouveau foyers d'épidémies/des foyers existants.
4	Étape de visualisation	– afficher les résultats d'affectation et de classification sur la zone d'étude – visualiser les classes d'épidémies sur la carte de risque épidémiologique.
5	Étape de prise de la décision	– visualiser les résultats sur la carte de risque épidémiologique. – prise des décisions par le décideur en santé publique.

Tableau 5.2 : Etapes des diagrammes UML (séquences et cas d'utilisation)

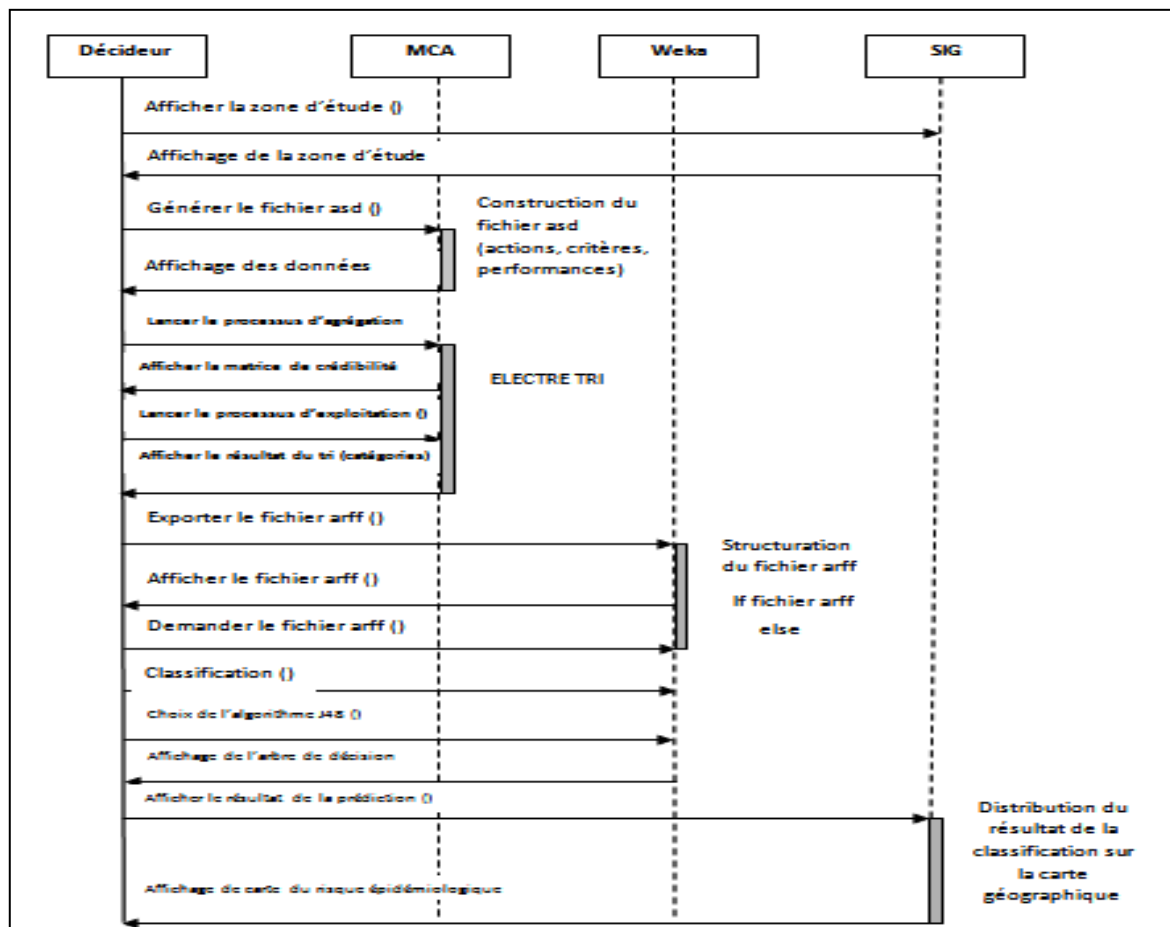


Figure 5.7.b Diagramme de séquence UML du système SDDD_MCA_DM

5.8 Expérimentation du système SDSS-MCA-DM

L'objectif du système SDSS-MCA-DM est la surveillance de la tuberculose dans la région d'Oran en Algérie.

5.8.1 Formulation du problème décisionnel

Le premier défi consiste à modéliser le problème de la surveillance épidémiologique en un problème décisionnel multicritères en prenant en compte les différents facteurs affectant la propagation de la maladie. Cette première étape est subdivisée en deux sous-phases : *la phase d'agrégation* et *la phase d'exploitation*.

5.8.1.1 Phase d'agrégation

Cette phase regroupe les différentes étapes permettant de définir les informations nécessaires pour conduire l'analyse multicritères (la zone d'étude, les actions, les critères, etc.) en terminant par la structuration de ces données dans un fichier sad.

– Délimitation de la zone d'étude

La zone d'étude est la région d'Oran située dans l'ouest de l'Algérie. Selon le manuel national antituberculeux établi en 2011, la région d'Oran occupe la deuxième place dans le classement national de la maladie de la tuberculose (Agadir et al, 2011). Certainement, des efforts énormes ont été faits depuis, pour éradiquer cette maladie mais cette région reste encore vulnérable à causes des facteurs de risques associés à cette maladie qui persistent encore aujourd'hui dans la région. La figure 5.8 montre une carte géographique d'Oran.

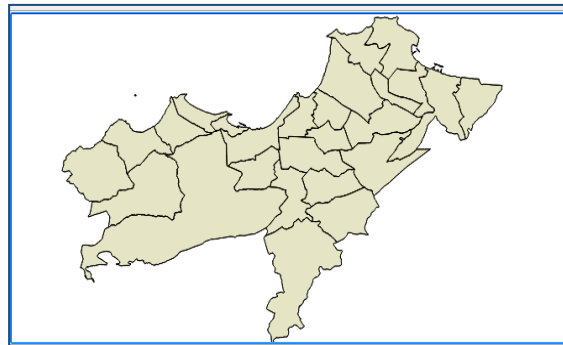


Figure 5.8 La zone d'étude

– Identification des actions

Les actions considérées dans notre analyse multicritères sont les foyers d'épidémie localisées précédemment par le système «EPISOLAP» (Zemri et al, 2013). Les foyers prévus dans notre étude, sont représentés par les 26 communes d'Oran.

– Identification des critères

En collaboration avec les décideurs et les experts en santé publique, le choix des critères était focalisé sur les cinq facteurs suivants :

- **Le risque :** Présence de cas contagieux (critère médical) : Le risque de tuberculose est défini en trois classes en fonction de la présence des cas contagieux (les cas prouvés) ou des cas non contagieux (cas non prouvés) de tuberculose mais potentiellement

contagieux. Le risque est élevé si, dans le foyer en question, il existe au moins un cas infectieux ; le risque est faible s'il n'y a pas un seul cas de tuberculose détecté comme contagieux mais qu'il y a au moins un cas de tuberculose détecté comme contagieux. Le risque est nul si aucun cas de tuberculose n'a été enregistré pendant toute l'année ("présence de cas de tuberculose" = 0). Dans notre cas, le risque est représenté par le taux d'incidence calculé à l'aide de la formule (5) et fourni par la Direction de la santé et de la population (DSP).

$$\circ \text{ Taux d'incidence} = \text{Nombre des nouveau cas} * \frac{100000}{\text{Populationmean}}^{(5)}$$

- **La densité** (critère démographique) : le surpeuplement dans une région favorise la contamination et la propagation des épidémies. Le contact direct avec le patient contaminé véhicule le virus par voie aérienne. Une population plus dense est plus vulnérable à la contamination.
- **Type d'habitat** (critère démographique) : ce critère est calculé par le nombre de constructions précaires généralement trouvées dans les bidonvilles. Dans notre étude de cas, le type d'habitat est représenté par le nombre de constructions précaires fournies par l'Office National des Statistiques (ONS).
- **Taux de chômage** (critère démographique) : la qualité de vie des individus est définie par leurs conditions personnelles et professionnelles. Il pourrait être calculé en fonction du taux de chômage de chaque région. Dans notre étude de cas, il est calculé par le nombre d'inactifs qui est fourni par l'Office National des Statistiques (ONS).
- **Humidité** (critère climatique) : L'humidité est un paramètre très important qui favorise la propagation de la bactérie (le bacille de coq) responsable de la tuberculose. Comme il est impossible de connaître le taux annuel d'humidité dans chaque municipalité d'Oran (présence de deux stations de mesure uniquement : (Senia et Arzew), nous avons proposé, en collaboration avec des spécialistes de la météorologie de l'Office National de Météorologie d'Oran (ONM), une échelle de points (1-2-3-4), qui classe les régions de la plus humide (valeur = 1) à la plus sèche (valeur = 4).

Tableau 5.3 Description des critères considérés

Critères	Type	Echelle	Méthode d'évaluation	Source
Taux d'incidence	Médical	[15,115]	Nombre des cas contagieux	D.S.P
Densité de la population	Démographique	[63,9660]	Population moyenne, surface	O.N.S
Nombre des constructions précaires	Démographique	{0,...,5125}	Nombre d'habitats dans les bidonvilles	O.N.S
Nombre des inactifs	Démographique	[112,9928]	Niveau de l'éducation, âge	O.N.S
Humidité	Climatique	[1,4]	Distance par rapport à la mère	O.N.M

– Génération de la table des performances

Le risque de la tuberculose est représenté par le risque (taux d'incidence) et la vulnérabilité de la population (densité, type d'habitat, degré d'humidité et niveau de pauvreté). Ces différents facteurs sont croisés dans le système EPISOLAP (Zemri et al, 2013). L'objectif de la modélisation multicritère du problème de la propagation de la tuberculose aboutira à la classification des zones géographiques de la plus favorables à la moins favorables pour être ou susceptible de devenir des foyers d'épidémies. Ces zones sont supposées être des zones à risque épidémiologique ou dite "épidémiologiquement dangereuses". Cette classification sera exploitée pour générer une carte de risque tuberculeux. Le résultat de l'analyse des conséquences est présenté dans une matrice des performances illustrée par le tableau 5.4.

Tableau 5.4 Table des performances

N°	Actions	inciden ce rates	The density Of population	Number of precarious constructions	Number of inactive	Humidity
1	Oran	37	9660	5125	9928	1
2	Gdyel	44	413	101	760	2
3	Bir El Djir	36	5330	1305	2683	2
4	HassiBounif	49	1976	116	1356	3
5	Es Senia	53	2144	570	1345	2
6	Arzew	47	997	120	788	1
7	Bethioua	44	169	53	342	1
8	Marsa El Hadjaj	30	253	44	304	1
9	Ain Turk	31	1005	197	671	1
10	El Ançar	17	173	82	112	1
11	HassiMefssoukh	47	499	113	346	3
....					...	...
25	Ain El Karma	79	71	0	122	1
26	Ain El Biya	25	895	5	395	2

– Paramètres subjectifs

Afin d'appliquer la méthode multicritères ELECTRE, nous introduisons quatre paramètres subjectifs : le poids (k), le seuil de préférence (p) d'indifférence (q) et de veto (v).

- **Les poids :** Chaque critère se voit attribuer un poids k qui reflète son importance par rapport aux autres critères. L'ensemble des poids des différents critères constitue l'ensemble $K = \{k_1, k_2, \dots, k_n\}$. Dans notre étude de cas, les valeurs de pondération sont générées de manière empirique en appliquant l'échelle de Saaty, basée sur un modèle mathématique développé par Thomas Saaty (Hamdadou et al, 2016). Celui-ci permet d'évaluer les différents critères et de les ordonner en fonction de leur importance relative. Après avoir comparé des paires de critères séquentiels (chaque critère est évalué par rapport à tous les autres dans une série de comparaisons), les chercheurs demandent à l'utilisateur de commander sur une échelle de (-9 à 9) l'importance relative d'un critère par

rapport à un autre. Dans la première ligne du tableau 5.5, nous avons les poids de critères générés par la méthode AHP (Analytic Hierarchy Process). Cette méthode est développée par (Saaty, 1980) et repose sur la théorie de l'utilité multi-attribut.

- **Les Seuils** : Pour chaque critère, un seuil de préférence p , un seuil d'indifférence q et un seuil de veto v sont estimés à l'aide des formules (6, 7 et 8) proposées dans (Oufella, 2009) et repris dans (Hamdadou et al, 2011) :
 - p représente la plus petite différence entre deux actions compatibles "a" et "b" avec une préférence pour "a" selon le critère g_j .
 - q représente la plus petite différence entre deux actions compatibles "a" et "b" avec une préférence pour "a" selon le critère g_j .
 - v représente pour deux actions "a" et "b" la plus petite différence $g_j(b) - g_j(a)$ incompatible avec aSb . "S" est la relation de surclassement.

Nous considérons l'estimation de l'incertitude = 20, nous avons :

- $q_j = 20/100 * \max(i,k) [g_j(a_k) - g_j(a_i)]_{(6)}$
- $p_j = 2*20/100 * \max(i,k) [g_j(a_k) - g_j(a_i)]_{(7)}$
- $v_j = 3*20/100 * \max(i,k) [g_j(a_k) - g_j(a_i)]_{(8)}, \text{ such that } q_j < p_j < v_j$

$g_j(a_i)$: est la performance de l'action a_i selon le critère j

$g_j(a_k)$: est la performance de l'action a_k selon le critère j

Le tableau 5.5 résume le calcul des trois seuils (p , q et v).

Tableau 5.5 Les paramètres subjectifs

Paramètres subjectifs	Cr1	Cr2	Cr3	Cr4	Cr5
Poids (k)	0.449	0.274	0,104	0,087	0,087
Préférence(p)	40	3864	2050	3971.2	1,2
Indifférence(q)	20	1932	1025	1985.6	0,6
Veto (v)	60	5796	3075	5956.8	1,8

– Actions de référence

Les actions de référence permettent de segmenter l'espace des actions en plusieurs catégories. Chaque catégorie est délimitée par une borne supérieure et une borne inférieure. Les actions de référence sont soit des actions réelles (appartenant à l'ensemble des actions à analyser) ou bien des actions artificielles (Maystre et al, 1994) (Roy, 1993). Ils délimitent les différentes catégories (Joerin, 1997). Les actions de référence appelées "étalons bonne" délimitent les bonnes actions alors que celles dites "étalons mauvaises" délimitent les mauvaises actions. Le décideur doit donc choisir comme actions de référence des actions qui sont, par leurs pondérations des différents critères, à la limite des deux classes.

Dans le cadre de notre étude de cas, nous avons choisi de manière aléatoire deux actions de référence et trois catégories distinctes {Elevé, Moyen ou Faible}, en suivant la méthode suivante :

Nombre d'actions de référence = Mod (Nombre d'actions / nombre de catégories) (9)

Nous avons 26 actions et 3 catégories alors, $26/3 = 8$ et le reste est 2. Cela signifie que nous avons deux actions de référence qui sont :

b1 = Marsa El Hadjedj

b2 = Ain El Biya

A priori, chacune de ces deux actions de référence délimite supérieurement et inférieurement les groupes de huit actions, ce qui fait : $26 = 3 \text{ groupes} * 8 \text{ actions potentielles} + 2 \text{ actions de référence}$.

La figure 5.9 montre le tableau de performance des actions de références considérées

Actions	criteria	Performances	reference actions		
Incidence_rate		density	Prec_Const_Nbr	Inactive_Nbr	Humidity
30.0		253.0	44.0	304.0	1.0
25.0		895.0	5.0	395.0	2.0

Figure 5.9 Actions de reference

– Structures de données

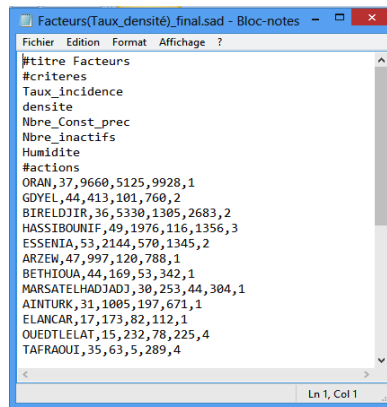
La formulation du problème de décision multicritère discutée dans le présent document est définie par le modèle (A, F, E, PS) avec :

- **A** : fait référence à toutes les actions potentielles possibles.
- **F** : désigne l'ensemble final d'attributs ou de critères, généralement contradictoires, selon lequel les actions seront évaluées.
- **E** : est l'ensemble des évaluations de performance des actions en fonction de chaque attribut ou critère, c'est-à-dire des vecteurs de performance globaux, un vecteur par action.
- **PS** : ensemble des paramètres subjectifs exprimés par le décideur.

Pour la mise en œuvre de l'outil SDSS_AM_FD, nous avons utilisé deux fichiers principaux, à savoir le fichier ".sad" pour l'analyse multicritères et le fichier ".arff" pour le processus d'exploration de données. Ces deux types de fichiers sont décrits ci-dessous.

▪ Structure du fichier " sad "

Le fichier avec l'extension ".sad" est un fichier permettant de structurer un problème de décision ; il contient les actions, les critères, la matrice des performances, les paramètres subjectifs exprimés par le décideur et les informations nécessaires au traitement décisionnel par des méthodes multicritères procédant notamment par agrégation partielle (ELECTRE). La figure 5.10 illustre la structure du fichier ".sad".



```

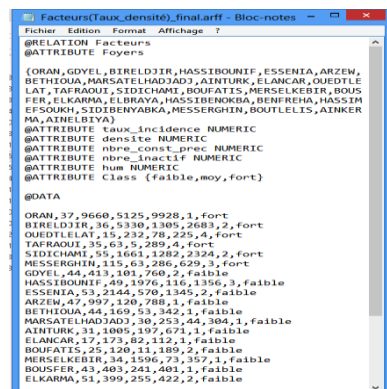
Facteurs(Taux_densité)_final.sad - Bloc-notes
Fichier Edition Format Affichage ?
#titre Facteurs
#critères
Taux_incidence
densité
Nbre_Const_prec
Nbre_inactifs
Humidité
#actions
ORAN,37,9660,5125,9928,1
GDYEL,44,413,101,760,2
BIRELDJIR,36,5330,1305,2683,2
HASSIBOUNIF,49,1976,116,1356,3
ESSENTIA,53,2144,570,1345,2
ARZEW,47,997,120,788,1
BETHIOUA,44,169,53,342,1
MARSATELHADJADJ,30,253,44,304,1
AINTURK,31,1005,197,671,1
ELANCAR,17,173,82,112,1
OUEDTLELAT,15,232,78,225,4
TAFRAOUI,35,63,5,289,4
Ln 1, Col 1

```

Figure 5.10 Format du fichier ".sad"

▪ Structure du fichier ".arff"

Ce fichier diffère légèrement du fichier ".sad" en ajoutant une nouvelle variable décrivant la classe contenant les trois valeurs {Elevé, Moyen, Faible}. Il commence par le mot clé @RELATION qui définit le début du fichier ; les autres variables sont définies par le mot clé @ATTRIBUTE et les valeurs des variables sont définies par le mot clé @DATA. Les poids, les actions de référence, les seuils de préférence, d'indifférence et de veto sont supprimés de ce fichier. La figure 5.11 montre un aperçu du fichier ".arff".



```

Facteurs(Taux_densité)_final.arff - Bloc-notes
Fichier Edition Format Affichage ?
@RELATION Facteurs
@ATTRIBUTE moyers
@ATTRIBUTE Taux_incidence NUMERIC
@ATTRIBUTE densite NUMERIC
@ATTRIBUTE nbre_const_prec NUMERIC
@ATTRIBUTE nbre_inactif NUMERIC
@ATTRIBUTE hum NUMERIC
@ATTRIBUTE Class {faible,moy,fort}
@DATA
ORAN,37,9660,5125,9928,1,fort
BIRELDJIR,36,5330,1305,2683,2,fort
OUEDTLELAT,15,232,78,225,4,fort
TAFRAOUI,35,63,5,289,4,fort
SIDICHAMI,55,1661,1282,2324,2,fort
MESSERGHIN,115,63,286,629,3,fort
GDYEL,44,413,101,760,2,faible
HASSIBOUNIF,49,1976,116,1356,3,faible
ESSENTIA,53,2144,570,1345,2,faible
ARZEW,47,997,120,788,1,faible
BETHIOUA,44,169,53,342,1,faible
MARSATELHADJADJ,30,253,44,304,1,faible
AINTURK,31,1005,197,671,1,faible
ELANCAR,17,173,82,112,1,faible
BOUFATIS,25,120,11,189,2,faible
MERSELKEBIR,34,1596,73,357,1,faible
BOUSFER,43,403,241,401,1,faible
ELKARMA,51,399,255,422,2,faible

```

Figure 5.11 Format du fichier ".arff"

5.8.1.2 Phase d'exploitation

Cette phase permet d'exploiter les données structurées dans le fichier sad pour faire les calculs nécessaires à l'analyse multicritères. Elle s'achève par le retour du résultat d'affectation des actions dans les différentes catégories.

– Evaluation de l'indice de concordance

La figure 5.12 montre la matrice de concordance concernant les actions de référence "b1" et "b2".

Mat_Concord	Mat_Discord	Credibility	optimistic assignment
		b1	b2
ORAN		0.9	1.0
GDYEL		0.61305296	1.0
BIRELDJIR		1.0	1.0
HASSIBOUNIF		0.6734928	1.0
ESSENIA		0.6692143	1.0
ARZEW		0.55	1.0
BETHIOUA		0.46672678	1.0
MARSATELHADJADJ		0.40767518	1.0
AINTURK		0.49	1.0
ELANCAR		0.19748625	0.91
QUEDTLELAT		0.27868804	0.88
TAFRAOUI		0.5466015	1.0
SIDICHAMI		1.0	1.0
BOUFATIS		0.4074236	1.0
IMERSELKEBIR		0.53499997	1.0

Figure 5.12 Matrice de concordance

- **Evaluation de l'indice de discordance**

La figure 5.13 montre la matrice de discordance concernant les actions de référence b1 et b2.

Mat_Concord		Mat_Discord		Credibility	optimistic assignment	
Incidence_rate	density	Prec_Const_Nbr	Inactive_Nbr	Humidity		
			b1		b2	
ORAN			0.0		0.0	
GDYEL			1.0		0.0	
BIRELDJIR			0.0		0.0	
HASSIBOUNIF			1.0		0.0	
ESSENIA			0.72796935		0.0	
ARZEW			1.0		0.0	
BETHIOUA			1.0		0.0	
MARSATELHADJADJ			1.0		0.0	
AINTURK			1.0		0.0	
ELANCAR			1.0		0.0	
OUEDTLELAT			1.0		0.0	
TAFRAOUI			1.0		0.0	
SIDI CHAMI			0.0		0.0	

Figure 5.13 Matrice de discordance

- Calcul de la matrice de crédibilité

La figure 5.14 montre la matrice de crédibilité concernant les actions de référence b1 et b2.

Mat_Concord	Mat_Discord	Credibility	optimistic assignment	
			b1	b2
ORAN			0.9	1.0
GDYEL			0.0	1.0
BIRELDJIR			1.0	1.0
HASSIBOUNIF			0.0	1.0
ESSENIA			0.5503467	1.0
ARZEW			0.0	1.0
BETHIOUA			0.0	1.0
MARSATELHADJADJ			0.0	1.0
AINTURK			0.0	1.0
ELANCAR			0.0	0.91
OUEDTLELAT			0.0	0.88
TAFRAOUI			0.0	1.0
SIDI CHAMI			1.0	1.0
BOUFATIS			0.0	1.0
MEDJEL KERT			0.0	1.0

Figure 5.14 Matrice de crédibilité

– Résultat de l'affectation

La figure 5.15 montre le résultat de l'affectation en adoptant, dans notre étude de cas, l'approche de l'affectation optimiste aux catégories. En effet, la catégorie cible dans notre étude est la catégorie à risque élevé et non pas la catégorie à risque faible. Nous visons à détecter les zones "épidémiologiquement" dangereuse en vue de l'éradiquer plus tard. Nous avons obtenu deux catégories {élevé et faible}, ce qui signifie qu'il n'y a aucune action dans la catégorie moyenne.

Mat_Concord	Mat_Discord	Credibility	optimistic assignment
Action	Classe		
ORAN	High		
BIRELDJIR	High		
OUEDTLELAT	High		
TAFRAOUI	High		
SIDICHAMI	High		
MESSERGHIN	High		
GDYEL	Low		
HASSIBOUNIF	Low		
ESSENTIA	Low		
ARZEW	Low		
BETHIOUA	Low		
MARSATELHADJADJ	Low		
AINTURK	Low		
ELANCAR	Low		
BOUFATIS	Low		

Figure 5.15 Résultat du tri

– Visualisation

La figure 5.16 montre le résultat du tri visualisé sur une carte de risque tuberculeux en attribuant la couleur verte aux zones à risque épidémiologique faible et la couleur rouge aux zones à risque épidémiologique élevé.

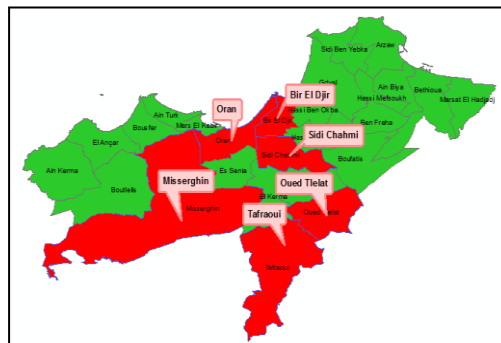


Figure 5.16 Résultat du tri

5.8.2 Processus d'exploration de données

Dans un deuxième temps, les scénarios d'impact environnemental et socio-économique seront modélisés et intégrés en appliquant la technique de l'arbre de décision pour expliquer et prédire les scénarios futurs et donner des interprétations juste aux phénomènes spatiaux.

– Arbre de décision

Le fichier généré sera exporté et enregistré avec l'extension ".arff" pour pouvoir être utilisé par l'outil weka. Nous choisissons le classificateur J48 qui exécute l'algorithme C4.5. Notre modèle épidémiologique est donné par l'arbre de décision généré et représenté sur la figure 5.17.

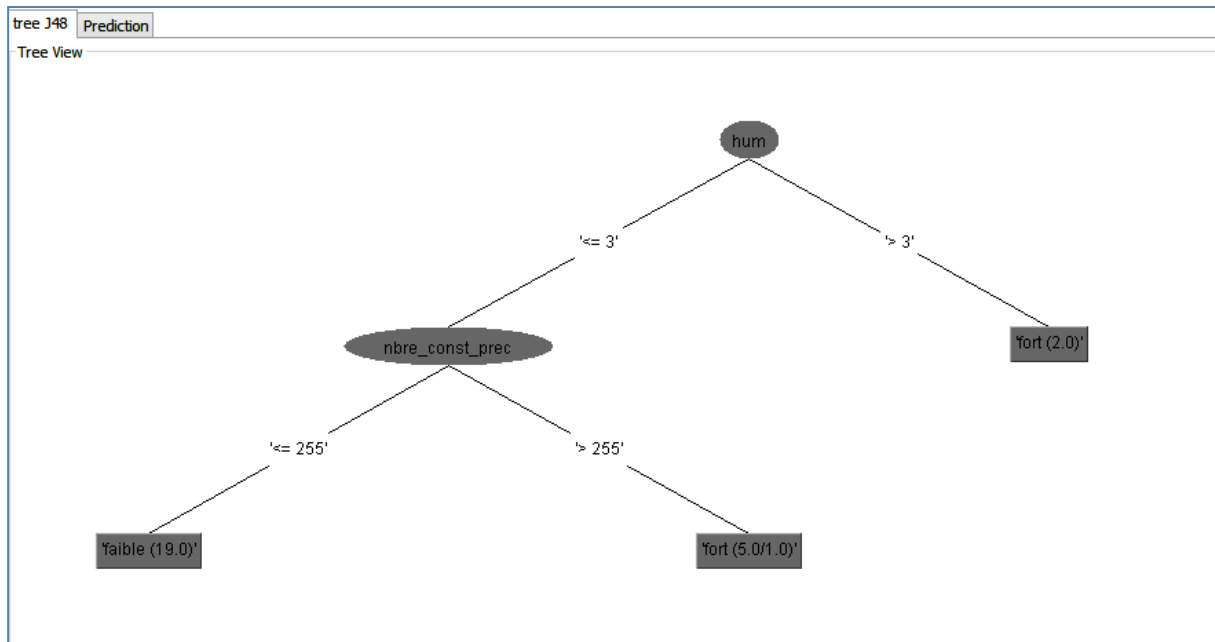


Figure 5.17 L'affichage de l'arbre de décision

Le schéma de la figure 5.17 représente un arbre de décision qui classe les foyers d'épidémies en région à risque faible ou en région à risque fort selon les règles prédictives suivante :

- **R1** : Si $hum \leq 3$ et $nbre_const_prec \leq 255$ alors degré de risque **Faible**
- **R2** : Si $hum \leq 3$ et $nbre_const_prec > 255$ alors degré de risque **Fort**
- **R3** : Si $hum > 3$ alors degré de dangerosité **Fort**

– Résultat de la prédiction

On peut visualiser le résultat de la prédiction sur la figure 5.18.

tree J48 Prediction

Communes	Classe
ORAN	High
BIRELDJIR	High
OUEDTLELAT	High
TAFRAOUI	High
SIDICHAMI	High
MESSERGHIN	High
GDYEL	Low
HASSIBOUNIF	Low
ESSENIA	High
ARZEW	Low
BETHIOUA	Low
MARSATELHADIADJ	Low
AINTURK	Low
ELANCAR	Low
BOUFATIS	Low
MERSELKEBIR	Low
BOUSFER	Low
ELKARMA	Low
ELBRAYA	Low
HASSIBENOKBA	Low
BENFREHA	Low
HASSIMEPSOUKH	Low
SIDIBENYABKA	Low
BOUTLEIS	Low
AINKERMA	Low
AINELBIYA	Low

Figure 5.18 Résultat de la prédiction

– Discussion

Si nous comparons les résultats avant et après le processus de fouille de données, nous pouvons constater que le résultat du processus multicritères était l'affectation des six actions parmi les 26 actions potentielles dans la catégorie de risque épidémiologique élevé, à savoir : {Oran, Bir El Djir, OuedTlilet, Tafraou, SidiChami et Misserguine}. Les 20 actions restantes ont été affectées dans la catégorie de risque épidémiologique faible.

Après l'application de l'apprentissage supervisé en utilisant l'approche d'arbre de décision, le résultat de la prédiction était le suivant : {Oran, Bir El Djir, **EsSenia**, Oued Tlilet, Tafraoui, Sidi Chami et Misserguine}. Une nouvelle action a été ajoutée dans la catégorie du risque épidémiologique élevé. Cela signifie que cette nouvelle action {EsSenia} est susceptible de devenir épidémiologiquement dangereuse à l'avenir en fonction des facteurs de risque favorables actuels.

– Visualisation

La figure 5.19 présente le résultat de la prédiction des zones épidémiologiques à risque. Les zones en rouge représentent les foyers épidémiologiquement dangereux présentant un niveau de risque épidémiologique élevé (catégorie élevée), tandis que les zones vertes représentent les zones à risque épidémiologique faible (catégorie inférieure).

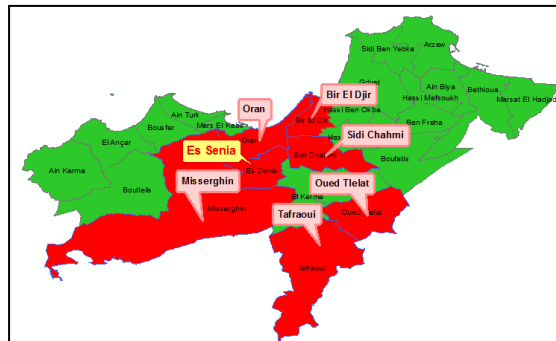


Figure 5.19 Résultat de la prédiction

5.9 Travaux connexes

Les travaux connexes que nous avons trouvé dans la littérature et qui sont reliés à notre problématique sont subdivisés en quatre sous domaines spécialisés selon les différentes thématiques abordées dans ce chapitre :

5.9.1 Travaux connexes sur la surveillance épidémiologique

Depuis 1962, plusieurs suggestions de représentations géographiques ont été proposées pour cartographier et représenter les données de la surveillance épidémiologique. Dans (Andrew et Dean, 1976), une carte ponctuelle basée sur la population a été proposée pour visualiser le comportement d'une épidémie d'infection à Salmonet-la Newport dans le comté de Pulaski. Les

auteurs dans (Noma et al, 2002) ont adopté une autre technique, appelée cartographie épidémiologique rapide de l'onchocercose (REMO) intégrée au système d'information géographique (SIG), pour délimiter les zones d'endémicité variée permettant de lutter contre l'onchocercose, l'un des défis fondamentaux à relever dans le programme africain de lutte contre l'onchocercose (APOC). Dans (Pybus et al, 2012), les auteurs démontrent que des données génétiques faciles à obtenir peuvent être utilisées pour mesurer la dynamique spatiale de populations naturelles autrement difficiles ou coûteuses à quantifier. Ils ont appliqué leur approche à la propagation du virus du Nil occidental en Amérique du Nord et ont démontré que sa propagation était déterminée de manière critique par des mouvements rares et à longue distance qu'il est peu probable de discerner lors des observations sur le terrain. D'autres chercheurs ont montré dans (Wang & Li, 2014) que presque tous les réseaux réalistes ont des propriétés de petit monde et / ou sans échelle. Dans (Lawson, 2013), une vaste gamme de logiciels disponibles pour analyser des données spatiales a été exploitée, notamment WinBUGS, SaTScan et R, pour donner un aperçu complet des principales méthodes statistiques utilisées en épidémiologie spatiale. Une autre catégorie de recherches en surveillance épidémiologique utilisant des méthodes géostatistiques a été développée dans (Wang et al, 2012) où les auteurs ont appliqué les statistiques empiriques de lissage spatial de Bayes, d'autocorrélation spatiale et d'analyse spatio-temporelle sur des cas de tuberculose signalés entre 2005 et 2010 dans la ville de Linyi en Chine. Dans cette étude, deux mesures statistiques ont été utilisées : l'indice global de Moran I calculé pour chaque année et les valeurs des indices locaux de Getis-Ord G_i^* . Dans (Munch et al, 2003), les auteurs ont étudié l'effet de divers facteurs de risque sur les lieux de transmission de la tuberculose à l'aide d'un système d'information géographique (SIG). Ils ont utilisé des méthodes statistiques pour trouver des associations significatives entre les notifications de tuberculose avec le taux de chômage, le surpeuplement et le nombre de bars dans les sous-quartiers. Leur système caractérise la transmission de la tuberculose par le surpeuplement, le taux de chômage élevé et la pauvreté. Dans (Lalvani et al, 2001), un processus épidémiologique a été suggéré par le dénombrement de cellules T spécifiques de l'antigène afin de détecter la tuberculose latente. Dans cette étude également, aucun effort de prédiction n'a été fait. Dans (Randremanana et al, 2009), des chercheurs de l'institut Pasteur ont analysé la distribution spatiale de la tuberculose à Antananarivo et ont étudié les facteurs de risque. Leur étude a démontré que l'évolution du risque d'un regroupement de tuberculose était liée à des facteurs socio-économiques et au facteur de soins des patients.

5.9.2 Travaux connexes sur les méthodes d'analyse multicritères(MAMC)

Dans (Hanafi-Bojd et al., 2012), les auteurs ont étudié les variations de la prévalence et de l'incidence de la maladie à l'aide du SIG et des méthodes géostatistiques pour déterminer la répartition géographique de la maladie. Ils ont combiné des facteurs environnementaux avec une évaluation multicritère pondérée pour planifier et cibler les interventions. Les auteurs dans (Stevens et al, 2013) ont utilisé une analyse d'aide à la décision multicritère (ADMC) avec un système d'information géographique (SIG) pour identifier les zones d'Asie susceptibles de devenir des foyers du virus de la grippe aviaire hautement pathogène (HPAIV) H5N1 chez les volailles domestiques. Cette étude a montré qu'il est possible de modéliser le risque de la présence de HPAIV H5N1 chez les volailles domestiques à une échelle continentale et a également montré que l'association du MDMC et du SIG constitue un outil pratique pour synthétiser les

connaissances sur une maladie et générer des représentations spatiales raisonnables de la variation de l'adéquation de la maladie ce qui à son tour est directement pertinent pour éclairer la conception de la surveillance de la santé animale basée sur les risques. Récemment, dans (Zemri et Hamdadou, 2014), un système d'aide à la décision basé sur une méthode d'analyse multicritères a été suggéré pour classer les foyers de la tuberculose à Oran.

5.9.3 Travaux connexes sur Spatial On Line Analytical Processing (SOLAP)

Dans (Proulx et al., 2002), les auteurs ont exploité une approche SIG dominant pour développer une application dans le domaine de la santé publique en tirant profit de la technologie SOLAP permettant d'explorer la relation entre la santé respiratoire et l'environnement. Dans les travaux de (Lebel et al, 2009), un système d'aide à la décision basé sur la technologie de JMAP Spatial OLAP de KHEOPS Technologies a été mis en œuvre par l'Institut national de santé publique du Québec pour suivre en temps réel la maladie du virus du Nil occidental. L'institut a mis en place un système entièrement capable de soutenir les efforts de la surveillance, de prévention et de lutte contre cette maladie. Dans (Zemri et al, 2013), les auteurs ont exploité des outils SIG et OLAP pour développer un système EPISOLAP de surveillance épidémiologique de la maladie de la tuberculose dans la région d'Oran en Algérie. Récemment, dans (Younsi et al, 2015), le système SOLAP a été combiné aux réseaux sociaux et aux méthodes mathématiques pour produire un système d'aide à la décision pour la surveillance de la grippe saisonnière dans la région d'Oran. Dans (Marouf et al, 2016), une approche d'intégration des données, basée sur les avantages des approches conventionnelles GAV et BAV dans les systèmes d'information décisionnels (DIS), ont été suggérées et appliquées à la surveillance épidémiologique.

5.9.4 Travaux connexes sur le data mining (DM)

Dans (Getoor, 2003), l'auteur a utilisé les réseaux bayésiens sur les données relatives à la tuberculose à San Francisco de 1991 à 1999 pour identifier la distribution des attributs et montrer la relation entre les variables. Le système a confirmé les résultats précédemment rapportés et a révélé plusieurs nouvelles associations. Dans (Mejhed et al. 2009), les auteurs ont mis au point un système d'aide à la décision basé sur les entrepôts de données et les outils de fouille de données permettant d'analyser les indicateurs épidémiologiques. Dans (Zhou et al. 2010), les auteurs ont combiné l'outil OLAP avec un processus de data mining sur des données cliniques, y compris les dossiers médicaux stockés dans un Datawarehouse, afin de créer un système d'information décisionnel permettant d'extraire et de découvrir des connaissances médicales et appuyer la décision clinique. Dans (Asha et al, 2010), les auteurs ont utilisé une technique de clustering K-means pour regrouper les données sur la tuberculose en deux groupes : la tuberculose pulmonaire (PTB) et la PTB rétrovirale (RPTB), c'est-à-dire celle liée au virus de l'immunodéficience humaine (VIH). Dans un autre travail (Rashedur et Fazle, 2011), les auteurs ont utilisé les arbres de décision et des règles d'association sur les données de surveillance hospitalière pour classer les patients tuberculeux admis en fonction de leur situation critique {faible, moyen et élevé} et les relations cachées entre différents symptômes. Dans l'étude de (Kumar et al, 2012), les auteurs ont appliqué trois types d'algorithmes de règles d'association : A Priori, A Priori MR et A Priori PT sur une base de données des personnes atteintes de tuberculose afin de détecter des associations entre les valeurs spécifiques aux variables catégorielles dans de grands ensembles de données. Dans (Santos et al, 2013), les auteurs ont utilisé un processus automatique de fouille de données

basé sur la construction d'un DW et d'une ontologie pour permettre aux décideurs en santé publique d'accéder à des informations analytiques sur les tumeurs cérébrales.

5.9.5 Synthèse des travaux d'état de l'art

Les recherches réalisées dans le domaine de la prévention et la lutte contre la tuberculose avaient pour objectif principal la détection et la recherche des facteurs réels de la tuberculose. Des épidémiologistes ont effectué des études à l'aide de méthodes statistiques pour détecter les corrélations et les associations entre les facteurs et le phénomène de la propagation de l'épidémie. Des efforts doivent être faits pour d'abord éradiquer les foyers actuels puis prévenir l'apparition de nouveaux foyers d'épidémies avec une identification préalable des causes réelles de la propagation. Dans la présente étude, des méthodes d'analyse multicritères et des techniques de data mining sont intégrées pour appuyer les efforts de prédiction et de prévention en matière de surveillance épidémiologique de la tuberculose.

Selon (Beale et al, 2008), les capacités du système en matière d'analyse épidémiologique sont plutôt limitées car elles sont davantage orientées vers la composante descriptive de la distribution de l'ampleur des risques pour la santé et de leurs déterminants. Dans cette optique, nous tentons dans ce travail de thèse de réduire certaines limitations en intégrant le contexte géographique dans la surveillance épidémiologique en plus du processus d'analyse.

Le tableau 5.6 résume les travaux connexes dans le domaine de la surveillance épidémiologique en identifiant les outils utilisés dans le développement des solutions correspondantes et compare, à la fin, notre travail par rapport à l'existant.

Table 4.5 Tableau récapitulatif des travaux liés à la surveillance épidémiologique

Travaux connexes	SOLAP	MCA	DM	DW	OLAP	GIS	Spatial Statistics	Ontology	Mathematical Methods
(Asha et al, 2010)			x						
(Getoor, 2003)			x						
(Hanafi-Bojd et al., 2012)		x				x	x		
(Kumar et al, 2012)			x						
(Lebel et al, 2009)	x			x					
Marouf et al., 2016).				x		x			
(Mejhed et al., 2009)			x	x	x				
(Proulx et al., 2002)	x			x	x	x			
(Rashedur & Fazle, 2011)			x						
(Santos et al., 2013)			x	x				x	
(Stevens and al, 2013)		x				x			
(Younsi et al, 2015)	x			x	x	x			x
(Zhou et al., 2010)			x	x	x				
(Zemri et al, 2013)	x			x	x	x			
(Zemri & Hamdadou, 2014)	x	x			x	x			
Notre approche		x	x			x			

5.10 Conclusion

Dans ce chapitre, nous avons proposé un système d'aide à la décision SDSS-MCA-DM pour la surveillance épidémiologique. Il repose essentiellement sur un couplage de trois sous-systèmes : le sous-système d'analyse multicritères, le sous-système d'exploration de données et le sous-système de visualisation. Il permet une transition facile du sous-système MCA au sous-système DM en générant un fichier adapté au processus d'exploration de données. L'avantage de notre modélisation réside dans le fait que les classes introduites dans ce fichier sont générées par un processus d'analyse multicritères. Les classes attribuées aux zones géographiques sont donc plus fiables car leur génération prend en considération les différents facteurs qui affectent la propagation de l'épidémie. Cette configuration rend le processus d'exploration de données plus crédible grâce aux classes générées automatiquement. Celles-ci sont correctement introduites par le calcul formel et sont utilisées dans l'apprentissage supervisé.

Dans le chapitre 6, nous présenterons notre troisième contribution dans ce travail de thèse à savoir l'introduction de l'indice de LISTA permettant de mesurer l'autocorrélation spatiotemporelle locale des régions géographiques voisines dans une zone d'étude.

Chapitre 6

Contribution 3 :

Proposition de l'indice de LISTA

Dans ce chapitre, nous proposons une formule pour calculer l'indice de l'autocorrélation spatiotemporelle et nous l'introduisons dans la phase exploratoire de la démarche systémique que nous avons adoptée dans le chapitre 3. Nous décrivons dans ce chapitre, les différentes étapes qui ont mené au développement de cette formule en passant par la formule de l'autocorrélation temporelle (ITA) puis l'indice croisé de corrélation temporelle (CITC) et enfin l'indice de LISTA suggéré. Nous introduisons plus en détails la matrice des indices de LISTA correspondante.

Plan du chapitre

1 Introduction	117
2 Motivation	117
3 Objectif visé	118
4 Mesure d'autocorrélation spatiotemporelle	119
4.1 Indice d'autocorrélation temporelle (ITA)	119
4.2 Indice d'autocorrélation spatiale (LISA)	120
4.3 Indice croisé de corrélation temporelle (CITC)	121
4.4 Indice d'autocorrélation spatiotemporelle (LISTA)	121
5 Matrice des indices d'autocorrélation spatiotemporelle	122
6 Expérimentation de l'indice de LISTA	123
6.1 Analyse de l'autocorrélation spatiale par indice de LISA	124
6.2 Analyse de l'autocorrélation spatiotemporelle par indice de LISTA	125
6.3 Calcul de la matrice des indices croisés de corrélation temporelle	126
6.4 Calcul de la matrice des indices d'autocorrélation spatiotemporelle	129
6.5 Discussion	130
7 Travaux connexes	132
7.1 Analyse d'autocorrélation spatiale	132
7.2 Analyse temporelle et spatiotemporelle	133
8 Conclusion	133

6.1 Introduction

Nous rappelons que nous avons adopté une méthodologie générique conçue pour analyser des données spatiales par combinaison des différentes approches d'analyse spatiale et celles de FDS dans des étapes successives et complémentaires.

La phase exploratoire de cette démarche systémique a pour objectif, la construction automatique d'une sectorisation guidée par les données de la région d'étude (voir le chapitre 3). Ce découpage spatial se base sur les valeurs des indicateurs d'autocorrélation spatiale globale (indices de Moran et de Geary) et locale (LISA d'Anselin et Local G_i^* de Getis-Ord). Ces indicateurs révèlent des structures spatiales, telles que des tendances ou des points "chauds" (un terme utilisé pour indiquer un nombre élevé de l'indice de mesure d'un phénomène spatial ex : taux d'incidence dans une étude de surveillance épidémiologique) et permettent d'assister l'analyste pour comprendre le phénomène analysé et détecter, par conséquent, des anomalies.

6.2 Motivation

Le besoin d'analyse à la fois spatial et spatio-temporel est récurrent dans certaines études spatiales comme dans l'épidémiologie, domaine d'application de nos travaux. Notre motivation dans ce travail de thèse était de se focaliser sur **la phase exploratoire** de la démarche systémique illustrée par la figure 6.1 et proposer à ce niveau une nouvelle méthode permettant de calculer l'autocorrélation spatiotemporelle locale. Nous proposons pour cela, un **indice d'autocorrélation spatio-temporelle**.

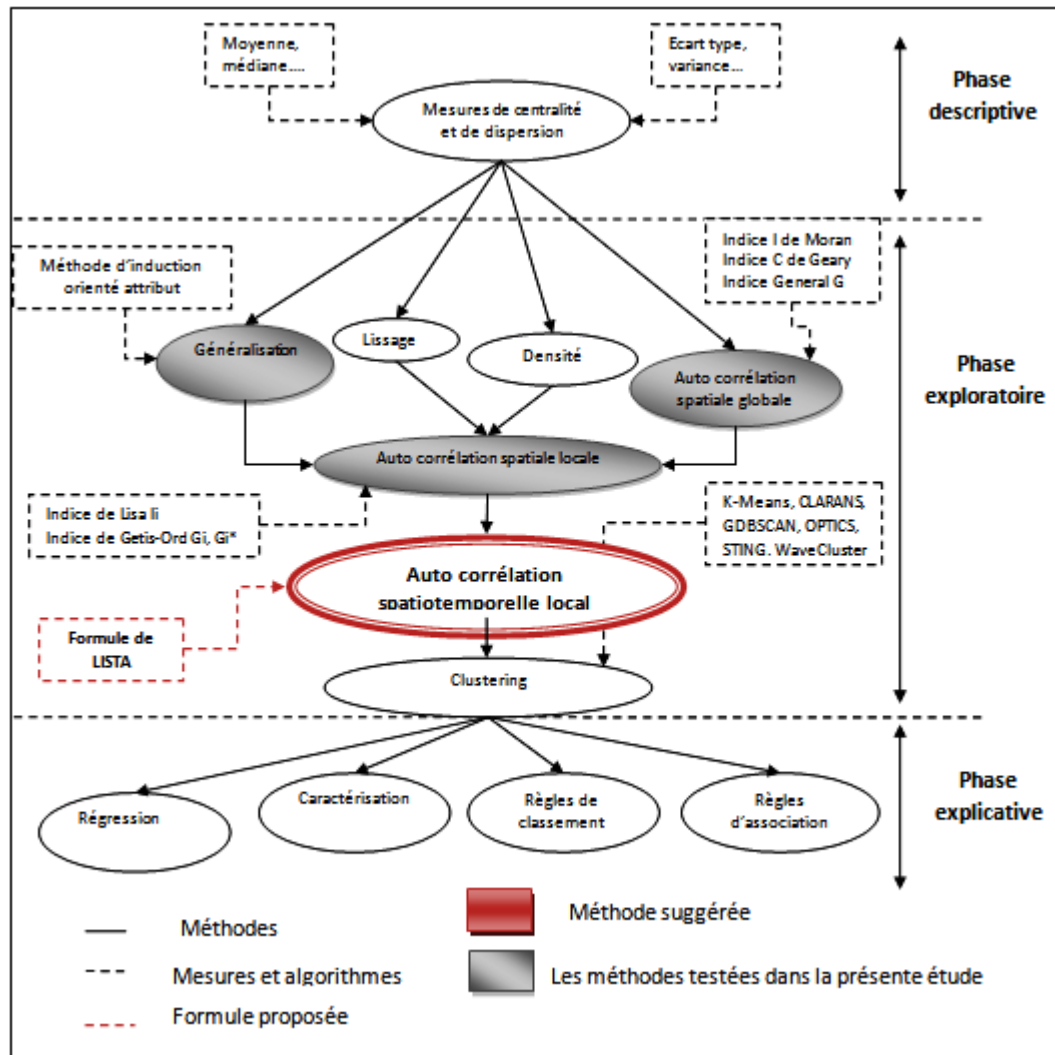


Figure 6.1 Démarche systémique de la fouille de données spatiales

La phase exploratoire se base sur le calcul des indices d'autocorrélation spatiale. Au niveau de cette phase, nous contribuons par l'introduction d'un nouvel indice qui complète l'analyse d'autocorrélation spatiale globale (calculée par l'indice de Moran) et locale (calculée par l'indice de LISA de Asselin), en vue d'enrichir l'analyse spatiale et l'étendre sur l'aspect temporel et par conséquent avoir une analyse plus complète.

Dans l'annexe C, nous menons des analyses exploratoires en appliquant ces indices sur notre étude de cas relative à la surveillance épidémiologique de la tuberculose à Oran.

6.3 Objectif visé

L'objectif principal de cette partie de l'étude est de parvenir à faire une sectorisation (régionalisation) guidée par la mesure de l'**autocorrélation spatio-temporelle** en proposant l'indice de LISTA (Local Index of Space Time Association).

Notre motivation est de vouloir généraliser l'autocorrélation **spatiale locale** à la dimension **temporelle**. L'idée est de tenir compte de la localité des mesures dans l'**espace** et dans le **temps**. On veut capturer pour un voisinage spatial et pour une fenêtre temporelle d'intervalle k ,

l'autocorrélation spatio-temporelle locale. Pour ce faire, nous introduisons trois formules qui précèdent la formulation de l'indice de LISTA suggéré :

- Indice d'autocorrélation temporelle (ITA : Indice of Temporal Autocorrelation) ;
- Indice d'autocorrélation spatiale locale (LISA : Local Indice of Spatial Temporal Autocorrelation).
- Indice croisé de corrélation temporelle (CITC : Cross Indice of Temporal Correlation) ;

La formule de LISTA permet de mesurer l'autocorrélation spatio-temporelle locale. Cette nouvelle méthode sera introduite, par la suite, dans la démarche systémique décrite dans la figure 6.1, pour compléter l'analyse exploratoire des données. L'avantage de cette méthode est qu'elle permet de mesurer l'autocorrélation à la fois spatiale et temporelle relative à une localité.

6.4 Mesures d'autocorrélation spatio-temporelle

La prise en compte de la variabilité temporelle est absente ou peu satisfaisante dans les systèmes cités dans la littérature. Les objectifs de l'analyse spatiale et l'analyse spatio-temporelle sont différents (Wang et al, 2012). La première vise à identifier les zones présentant une incidence élevée tout au long de la période d'étude, alors que la dernière analyse vise à identifier les zones avec augmentation de l'incidence dans des périodes spécifiques. Pour montrer cette différence Sharov (Sharov, 1996) définit les notions de séries temporelles, séries spatiales et séries mixtes et montrent les différences entre elles.

- **Séries temporelles** : séquence de mesures au même point de l'espace prises à différentes périodes temporelles.
- **Séries spatiales** : ensemble de mesures faites simultanément en différents points de l'espace.
- **Séries mixtes** : ensemble de séries temporelles obtenues en différents points dans l'espace.

6.4.1 Indice d'autocorrélation temporelle (ITA)

Des formules pour mesurer l'autocorrélation temporelle globale ont été utilisées dans (Cheng et al.) pour proposer un indice spatio-temporel afin de mesurer l'autocorrélation spatio-temporelle globale du flux des voitures dans un réseau routier.

Nous avons l'**autocorrélation temporelle globale** donnée par le rapport de la covariance " δk " sur la variance " σ^2 " (Cheng et al, 2012).

En appliquant la formule de l'espérance mathématique, nous aurons :

– La covariance :

$$\delta_k = E[(z_{t+k} - \bar{z})(z_t - \bar{z})] \quad (1)$$

– La variance :

$$\sigma^2 = \sum (z_t - \bar{z})^2 \quad (2)$$

$$\left. \begin{array}{l} (1) \\ (2) \end{array} \right\} \Rightarrow P_k = \frac{\delta_k}{\sigma^2} \quad (3)$$

– L'espérance :

$$E(z) = \frac{\sum_i z_i p_i}{\sum_i p_i} \quad (4)$$

Nous avons donc :

L'autocorrélation temporelle (P_k) mesure la ressemblance entre une observation relative à une unité spatiale et une antérieure par un intervalle k.

$$P_k = \sum \frac{(z_t - \bar{z})(z_{t-k} - \bar{z})}{\sum (z_t - \bar{z})^2} \quad (5)$$

Nous pouvons introduire l'indice temporel d'autocorrélation baptisé ITA (Indice of Temporal Autocorrelation) qui se calcule par la formule suivante :

$$ITA(i, k, s) = \frac{\sum_{w=k+s}^l (z_i^{w-k} - \bar{z}^{w-k})(z_i^w - \bar{z}^w)}{\sum_{w=1}^l (z_i^w - \bar{z}^w)^2} \times \frac{l}{l-k-s+1} \quad (6) \quad \text{avec } 1 \leq w-k \leq w \leq l$$

Le terme $l / (l-k-s+1)$ peut être ignoré pour l'indice local d'autocorrélation

6.4.2 Indice d'autocorrélation spatiale (LISA)

L'autocorrélation spatiale est définie comme *"La corrélation d'une variable avec elle-même (autocorrélation) attribuable à l'organisation géographique des données (spatiale)"* (Griffith, 1991). On s'intéresse dans notre étude aux deux mesures d'autocorrélation spatiale globale calculée par l'indice de Moran et locale calculée par l'indice de LISA. Nous proposons une extension de l'indice local d'autocorrélation spatiale LISA, au contexte spatio-temporel – baptisé LISTA.

Nous rappelons l'indice d'**autocorrélation spatiale locale** de LISA qui est donné par la formule suivante :

$$I_i = \frac{\sum_j w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2} \quad (7)$$

En analysant la formule de LISA de Anselin, nous constatons que nous pouvons l'étendre en intégrant l'aspect temporel.

6.4.3 Indice croisé de corrélation temporelle (CITC)

L'analyse de l'autocorrélation temporelle est analogue à celle de l'autocorrélation spatiale qui mesure la ressemblance entre des observations voisines. L'autocorrélation spatiotemporelle entre paires de voisins est donnée par l'indice croisé de corrélation temporelle CITC (Cross Index of Temporal Correlation).

Nous pouvons maintenant introduire l'indice de corrélation spatio-temporelle en intégrant la formule de **corrélation temporelle** proposée (ITA) avec celle d'**autocorrélation spatiale locale** de LISA donnée par la formule suivante :

$$CITC(i, j, k, s) = \frac{\sum_{w=k+s}^l W_{ij} (z_i^{w-k} - \bar{z}^{w-k})(z_j^w - \bar{z}^w)}{\sum_{w=1}^l (z_i^w - \bar{z}^w)^2} (8) \text{ avec } 1 \leq w-k \leq w \leq l$$

6.4.4 Proposition de l'Indice d'autocorrélation Spatio-temporelle (LISTA)

En appliquant une sommation sur j des indices croisés de corrélation temporelle, nous introduisons un nouvel indice d'autocorrélation spatiotemporelle locale « Indice de LISTA » calculé pour chaque unité spatiale i, selon un intervalle k et un temps initial s.

LISTA est défini par la formule suivante :

$$LISTA(i, k, s) = \sum_j CITC(i, j, k, s) (9)$$

On peut déduire :

$$LISTA(i, k, s) = \frac{\sum_j \sum_{w=k+s}^l W_{ij} (z_i^{w-k} - \bar{z}^{w-k})(z_j^w - \bar{z}^w)}{\sum_i \sum_{w=1}^l (z_i^w - \bar{z}^w)^2} (10) \text{ avec } 1 \leq w-k \leq w \leq l$$

- k = intervalle de temps entre périodes considérées (k=1 entre t_1 et t_2 et k=2 entre t_1 et t_3)
- w= période considérée
- i=unité spatiale
- j=unité voisine *
- l= nombre des périodes considérées (l=4 correspond à l'étude trimestrielle et l= 12 à l'étude mensuelle)
- z_i^w = mesure dans la zone considérée au temps t_w
- z_j^w = mesure dans une zone voisine à z_i au temps t_w
- W_{ij} = la matrice de voisinage pondérée

Tout comme l'indice de Moran ou de LISA, l'indice de LISTA varie entre -1 et +1.

- Si la valeur de LISTA s'approche de la valeur -1, alors les voisins ont des valeurs opposées. Les voisins sont donc dissimilaires et l'autocorrélation est négative.
- Si par contre, la valeur de LISTA s'approche de la valeur +1, alors les voisins ont des valeurs semblables. Les voisins sont donc similaires et l'autocorrélation est positive.
- En revanche, si la valeur de LISTA s'approche de 0, on considère que l'autocorrélation spatiale est nulle ou une absence d'autocorrélation.

On notera, cependant, que la valeur du LISTA peut parfois être supérieure à +1 ou inférieure à -1 (Cliff et Ord, 1981). Il n'est donc pas strictement borné entre -1 et +1.

L'autocorrélation spatio-temporelle peut être donc mesurée par l'indice de LISTA (Local Index of Spatial Temporal Association) que nous proposons comme une extension de l'indice de LISA (Local Index of Spatial Association) vu précédemment.

Nous verrons dans la section expérimentation, l'application de ces formules sur un exemple d'usage pour caractériser la propagation d'une épidémie.

Le calcul de l'indice croisé de corrélation temporelle entre toutes les paires de voisins génère une matrice étendue sur les périodes de temps considérées. Cette matrice ressemble à une matrice de voisinage mais elle contient les valeurs des indices ICTC calculés à la place des poids w_{ij} entre les zones voisines. Dans ce qui suit, nous introduisons cette matrice en montrant à travers quelques exemples le calcul des valeurs de ses cellules.

6.5 Matrice des indices de l'autocorrélation spatiotemporelle

Le calcul de l'ensemble des indices croisés de corrélation temporelle entre paires de voisins génère une matrice appelée « **matrice des indices de l'autocorrélation spatiotemporelle** ». La figure 6.2 présente une illustration de cette matrice calculée par la formule CITC entre une région i et une région voisine j selon l'intervalle k et un temps initial s .

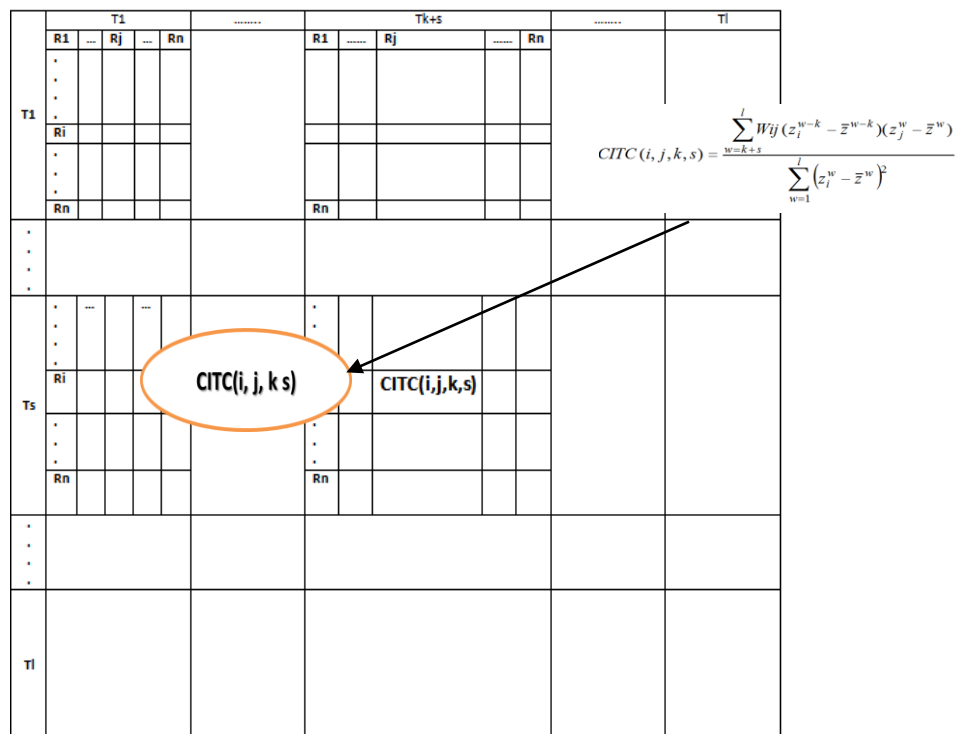


Figure 6.2 Matrice des indices de l'autocorrélation spatiotemporelle

Nous donnons dans la section expérimentation un exemple d'utilisation pour montrer comment se calcule l'indice CITC suggéré.

6.6 Expérimentation de l'indice de LISTA

Nous étudierons dans cette section, l'autocorrélation spatiale et spatiotemporelle de l'incidence de la tuberculose à Oran. Pour cela, Nous introduisons l'aspect temporel en considérant les données de la surveillance pour les quatre trimestres de l'année 2009. A ce titre, nous commençons par partitionner les données de la surveillance épidémiologique par trimestre puis nous procédons au calcul des indices d'autocorrélation spatio-temporelle entre les trimestres en appliquant la formule de LISTA proposée dans ce chapitre.

Tableau 6.1 Tableau du taux d'incidence de la tuberculose pour les quatres trimestres

N°	communes	Taux_T1	Tx_Trim2	Trim3	Trim4
0	A.El Turck	9	6	6	16
1	Ain El Bia	3	6	6	9
2	Ain El kerma	12	24	12	24
3	Arzew	6	19	10	7
4	Benfreha	11	28	28	6
5	Bethioua	11	11	6	17
6	Bir El Djir	14	25	18	15
7	Boufatis	17	8	0	0
8	Bousfer	30	7	0	22
9	Boutlelis	9	5	9	5
10	El Ancor	0	10	0	10
11	El Braya	21	43	21	21
12	El kerma	18	18	6	36
13	Es senia	23	24	9	14
14	Gdyel	14	11	11	11
15	H.Bounif	15	22	17	4
16	H.Mefsoukh	0	22	22	22
17	Hassi Ben Okba	0	18	26	0
18	M El Kebir	6	17	0	12
19	M. Hadjaj	0	8	16	8
20	Misserghin	27	64	14	37
21	O.Tlelat	6	6	6	0
22	Oran	9	14	9	6
23	S.B.yebka	0	0	0	14
24	S.Chami	25	31	13	21
25	Tafraoui	17	0	17	8

Nous commençons l'étude de l'autocorrélation par donner un aperçu global sur l'évolution de la tuberculose à Oran par trimestre dans l'année 2009.

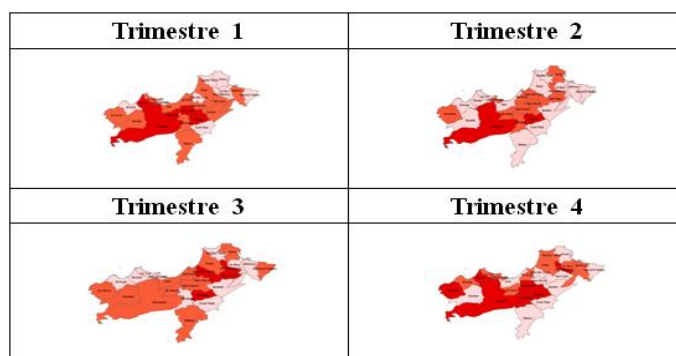


Figure 6.3 Représentation trimestrielle du TBC à Oran

Sémantique de la figure : Le dégradé de couleur rouge utilisé dans la figure 6.10 représente l'ampleur de l'autocorrélation spatiale calculée par l'indice de LISA.

- Les zones avec un rouge foncé représentent une forte autocorrélation spatiale avec des valeurs de LISA élevé (≈ 1). Ce sont des zones qui forment des agrégats (des concentrations)
- Les zones en rose claire représentent une faible autocorrélation avec des valeurs de LISA négatives. Ce sont des zones dispersées sur la carte géographique.
- Les valeurs presque nulles (≈ 0) de LISA sont ignorées car elles représentent une absence d'autocorrélation spatiale.

La représentation géographique des données de la tuberculose à Oran par trimestre permet de donner une visualisation générale des concentrations du taux d'incidence dans chacun des trimestres sans tenir compte des particularités locales des régions.

6.6.1 Analyse de l'autocorrélation spatiale par indice de LISA

Une fois qu'un niveau de signification a été défini, les valeurs de LISTA pourraient être tracées sur une carte pour afficher les emplacements des points chauds (Hotspots):

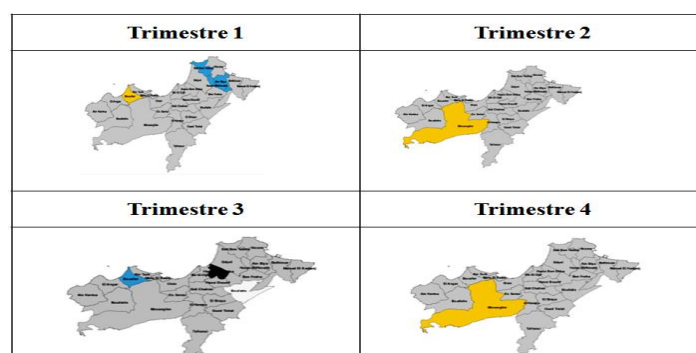


Figure 6.4 Autocorrélation spatiale locale par indice de LISA

Sémantique de la figure :

- Sites avec des valeurs élevées avec des voisins similaires (High-High) en couleur **noir**.
- Sites avec des valeurs élevées dans un voisinage avec des valeurs faibles (High-Low) en couleur **orange**.
- Sites avec des valeurs faibles dans un voisinage avec des valeurs élevées (Low-High) en couleur **bleue**.
- Sites avec des valeurs faibles dans un voisinage avec des valeurs faibles (Low-Low) en couleur **blanche**.

On s'intéresse à l'identification des zones HH qui représentent une anomalie ou un danger dans l'étude de cas proposée. Les régions identifiées en couleur noir sur la carte sont des points chauds.

La représentation géographique de l'autocorrélation **spatiale locale** donnée par l'indice de LISA permet quant à elle la détection des points chauds, par trimestre sans pouvoir capturer l'autocorrélation temporelle à travers le temps (ici les trimestres).

Nous voulons, à travers la présente étude de cas, capturer l'autocorrélation **spatio-temporelle locale**, pour un voisinage spatial et pour une fenêtre temporelle d'intervalle k . Pour cela, nous introduisons la matrice des indices de l'autocorrélation spatio-temporelle.

6.6.2 Analyse de l'autocorrélation spatiotemporelle par l'indice de LISTA

Nous définissons la matrice des indices de l'autocorrélation spatio-temporelle par une matrice des indices de l'autocorrélation spatiale multipliée par le nombre des périodes considérées par l'étude temporelle. La figure 6.5 représente une matrice des indices de l'autocorrélation spatio-temporelle locale étendue sur quatre périodes successives des trimestres (T1,T2,T3,T4).

Une cellule dans la matrice des indices de l'autocorrélation spatio-temporelle représente la mesure de l'autocorrélation spatio-temporelle entre la région i dans w et la région j dans $w - k$.

Pour calculer les valeurs de la matrice des indices d'autocorrélation spatio-temporelle, nous intégrons les considérations suivantes :

- Calculer l'indice de LISTA entre trimestres pour les mêmes régions revient à calculer l'autocorrélation temporelle entre une région z_i avec ses voisines z_j . Ces valeurs représentent les valeurs des sous matrices (T1-T2, T2-T3, T3-T4, T1-T3, T2-T4, T1-T4)
- Calculer l'indice de LISTA entre régions dans le même trimestre revient à calculer l'indice de LISA tout simplement (LISTA équivaut à LISA dans ce cas). Ces valeurs représentent les valeurs des sous matrices (T1-T1, T2-T2, T3-T3, T4-T4).
- La diagonale de toute la matrice des indices de l'autocorrélation spatio-temporelle représente les valeurs de LISA des régions z_i avec elles mêmes (ces valeurs sont ignorées et remplacées par la valeur nulle).
- Calculer l'indice de LISTA entre trimestres pour les mêmes régions revient à calculer l'autocorrélation temporelle entre une région z_i avec elles mêmes. Ces valeurs représentent les lignes qui marquent les diagonales des sous matrices (T1-T2, T2-T3, T3-

T4, T1-T3, T2-T4, T1-T4) et représentent les indices d'autocorrélation temporelle. ITA (i, k, s).

- Le reste des cellules est calculé par l'indice de LISTA proposé.

	T1	T2	T3	T4
T1	LISA	LISTA (T1-T2)	LISTA (T1-T3)	LISTA (T1-T4)
T2		LISA	LISTA (T2-T3)	LISTA (T2-T4)
T3			LISA	LISTA (T3-T4)
T4				LISA

Figure 6.5—Illustration de LISTA définie sur la matrice des indices d'autocorrélation spatio-temporelle

6.6.3 Calcul de la matrice des indices croisés de corrélation temporelle

Exemple 1 :

Dans cet exemple, On calcule l'autocorrélation spatio-temporelle entre la région R13 dans T1 et sa voisine R9 dans T2. L'intersection de la ligne 13 et de la colonne 9 de la sous-matrice T1-T2 représente la mesure donnée par l'indice CITC (13, 9, 1, 1).

$$CITC(i, j, k, s) = \frac{\sum_{w=k+s}^l W_{ij} (z_i^{w-k} - \bar{z}^{w-k})(z_j^w - \bar{z}^w)}{\sum_{w=1}^l (z_i^w - \bar{z}^w)^2} \quad (11)$$

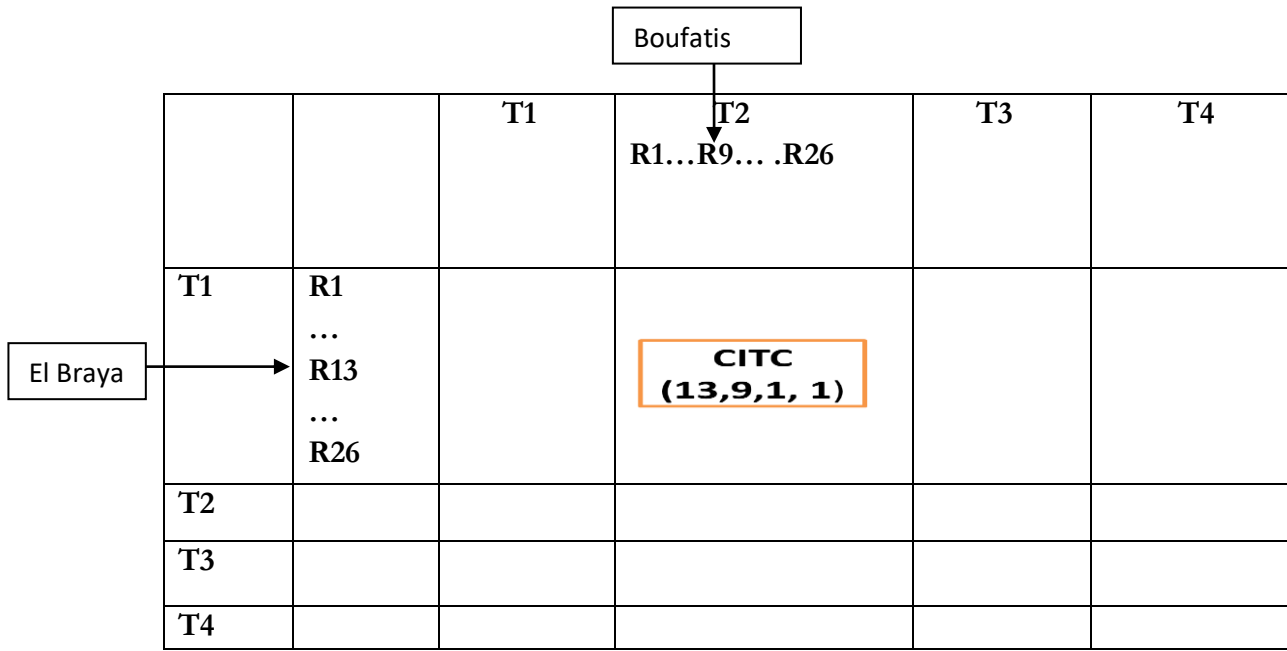


Figure 6.6 Calcul de l'indice de CITC entre T1 et T2

$$CITC(13,9,1,1) = \frac{\sum_{w=2}^4 (z_i^{w-k} - \bar{z}^{w-k})(z_j^w - \bar{z}^w)}{\sum_{w=1}^4 (z_i^w - \bar{z}^w)^2} \quad (12)$$

$$CITC(13,9,1,1) = \frac{(z_{13}^{2-1} - \bar{z}^{2-1})(z_9^2 - \bar{z}^2) + (z_{13}^{3-1} - \bar{z}^{3-1})(z_9^3 - \bar{z}^3) + (z_{13}^{4-1} - \bar{z}^{4-1})(z_9^4 - \bar{z}^4)}{(z_{13}^1 - \bar{z}^1)^2 + (z_{13}^2 - \bar{z}^2)^2 + (z_{13}^3 - \bar{z}^3)^2 + (z_{13}^4 - \bar{z}^4)^2}$$

$$CITC(13,9,1,1) = \frac{(z_{13}^1 - \bar{z}^1)(z_9^2 - \bar{z}^2) + (z_{13}^2 - \bar{z}^2)(z_9^3 - \bar{z}^3) + (z_{13}^3 - \bar{z}^3)(z_9^4 - \bar{z}^4)}{(z_{13}^1 - \bar{z}^1)^2 + (z_{13}^2 - \bar{z}^2)^2 + (z_{13}^3 - \bar{z}^3)^2 + (z_{13}^4 - \bar{z}^4)^2}$$

$$CITC(13,9,1,1) = \frac{(-10)(-36) + (-32)(-45) + (-32)(-53)}{1019 + 99 + 1019 + 1001} = 1,1043427$$

Exemple 2 :

Dans cet exemple, On calcule l'autocorrélation spatio-temporelle entre la région R13 dans T2 et sa voisine R9 dans T3. L'intersection de la ligne et de la colonne représente la mesure donnée par l'indice CITC (13, 9, 1, 2).

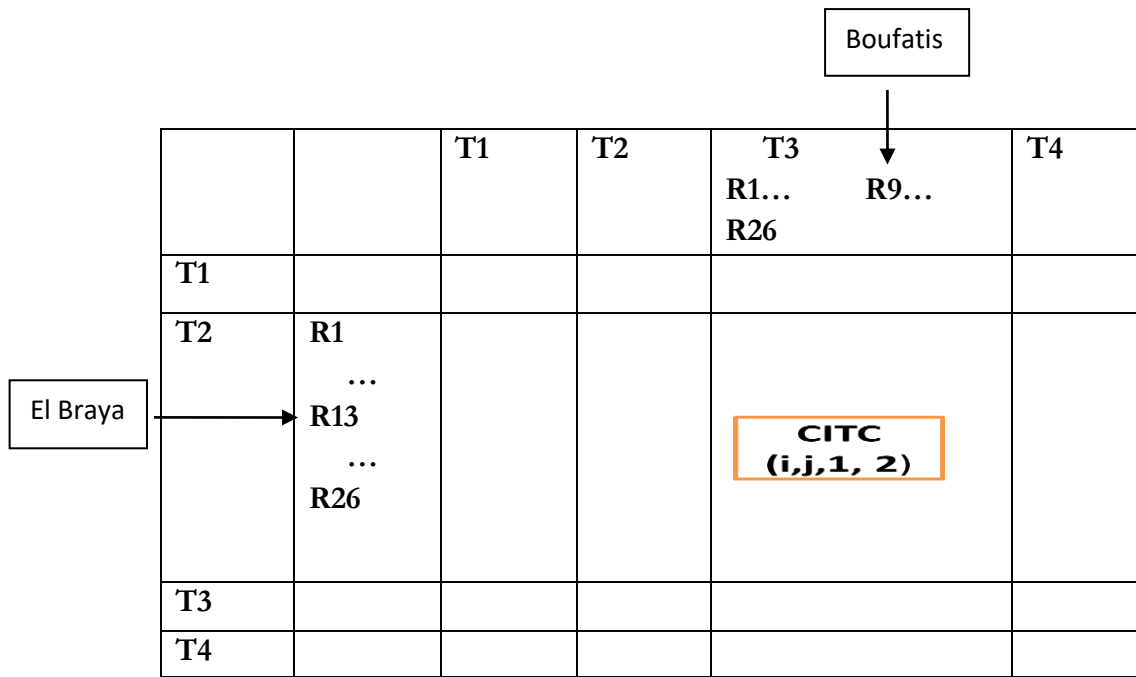


Figure 6.7 Calcul de l'indice de CITC entre T2 et T3

$$CITC(13, 9, 1, 2) = \frac{\sum_{w=3}^4 (z_i^{w-k} - \bar{z}^{w-k})(z_j^w - \bar{z}^w)}{\sum_{w=1}^4 (z_i^w - \bar{z}^w)^2}$$

$$CITC(13, 9, 1, 2) = \frac{(z_{13}^{3-1} - \bar{z}^{3-1})(z_9^3 - \bar{z}^3) + (z_{13}^{4-1} - \bar{z}^{4-1})(z_9^4 - \bar{z}^4)}{(z_{13}^1 - \bar{z}^1)^2 + (z_{13}^2 - \bar{z}^2)^2 + (z_{13}^3 - \bar{z}^3)^2 + (z_{13}^4 - \bar{z}^4)^2}$$

$$CITC(13, 9, 1, 2) = \frac{(-32)(-45) + (-32)(-53)}{1019 + 99 + 1019 + 1001} = 0,99936265$$

On définit une matrice des **poids de voisinage modifiée avec le degré d'influence entre zi et zj k unités de temps plus tard**. Le calcul des indices CITC pour toutes les régions i j donnera la matrice des indices d'autocorrélation spatio-temporelle illustrée sur la figure 6.8.

Matrice des indices croisés d'autocorrélation spatio-temporelle 26*4 × 26*4

	T1	T2	T3	T4
T1	CITC (i,j,0, 1)	CITC (i,j,1, 1)	CITC (i,j,2, 1)	CITC (i,j,3, 1)

T2		CITC (i,j,0,2)	CITC (i,j,1, 2)	CITC (i,j,2, 2)
T3			CITC (i,j,0, 3)	CITC (i,j,1, 3)
T4				CITC (i,j,0, 4)

Figure 6.8 Illustration de la matrice des indices CITC

6.6.4 Calcul de la matrice des indices d'autocorrélation spatiotemporelle

La somme sur j des indices CITC, nous donnera les valeurs de LISTA pour chaque région i selon un temps initial s et selon un intervalle k.

Matrice des indices de LISTA $26*4 \times 4$

	T1	T2	T3	T4
T1	LISTA (i, 0, 1)	LISTA (i, 1, 1)	LISTA (i, 2, 1)	LISTA (i, 3, 1)
T2		LISTA (i, 0, 2)	LISTA (i, 1, 2)	LISTA (i, 2, 2)
T3			LISTA (i, 0, 3)	LISTA (i, 1, 3)
T4				LISTA (i, 0, 4)

Figure6.9 Illustration de la matrice des indices de LISTA

Les représentations cartographiques des indices d'autocorrélation spatio-temporelles calculés par indices de LISTA sont représentées sur la figure 6.10

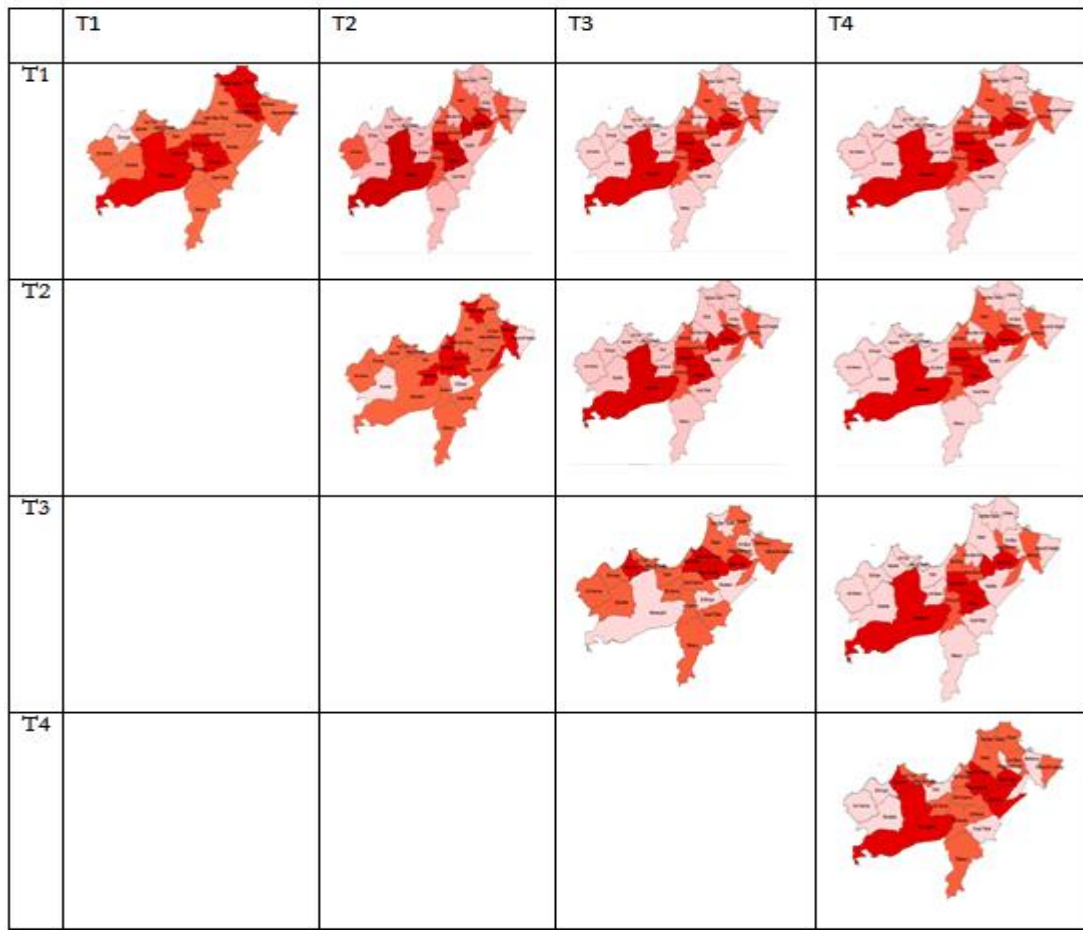


Figure 6.10 Représentation cartographique de l'autocorrélation spatio-temporelle de la tuberculose à Oran par trimestre calculée par l'indice de LISTA

Sémantique de la figure : Le dégradé de couleur rouge utilisé dans la figure 6.10 représente l'ampleur de l'autocorrélation spatiotemporelle calculée par l'indice de LISTA.

- Les zones avec un rouge foncé représentent une forte autocorrélation spatiotemporelle avec des valeurs de LISTA élevé (≈ 1)
- Les zones en rose claire représentent une faible autocorrélation avec des valeurs de LISTA négatives (≈ -1)
- Les zones qui ont des valeurs presque nulles (≈ 0) marquent une absence d'autocorrélation spatiale et sont donc ignorées.

6.6.5 Discussion

Nous pouvons distinguer clairement qu'il existe une forte autocorrélation spatio-temporelle entre les trimestres voisins (T1-T2, T2-T3, T3-T4) et une faible autocorrélation spatio-temporelle entre les trimestres non consécutifs (T1-T3 et T2-T4). Ce résultat confirme d'une part que le phénomène de propagation de l'épidémie dépend bien du temps et d'autre part, indique que sa propagation est bien visible d'un trimestre au trimestre suivant, mais pas lorsque ces trimestres sont séparés d'un ou plusieurs trimestres (valeur de k). Plus l'intervalle k est petit plus l'autocorrélation spatiotemporelle est significative.

Afin de montrer l'utilité de l'indice de LISTA par rapport à l'indice de LISA, nous prenons les exemples de quelques valeurs représentées sur le tableau 6.2. Ces valeurs sont extraites des sous matrices de la matrice des indices de l'autocorrélation spatio-temporelle de la figure 6.9. Ces valeurs sont calculées par la formule de l'indice CITC (i,j,k,s) décrite ci-dessus.

Tableau 6.2 Exemples de valeurs de CITC (i, j, k, s)

Region z_i	Region z_j	CITC (i,j,k,s)
Elbraya(T1)	Boufatis(T2)	1,1043427
El ancor(T1)	Misserguine(T2)	1,0717802
Misserguine(T1)	Bousfer(T3)	1,0296758
Misserguine(T2)	Ain Turck(T3)	1,0066348
Misserguine(T2)	Boutlelis(T3)	1,00266175
Misserguine(T2)	ElKerma(T3)	0,8254962
Misserguine(T2)	EsSenia(T3)	0,71585897
Boufatis(T3)	ElBraya(T4)	0,5335526
Misserguine(T2)	AinTurk(T4)	0,9613501
Misserguine(T2)	Boutlelis(T4)	0,9677061
Misserguine(T2)	ElKerma(T4)	0,7492245
Misserguine(T2)	EsSenia(T4)	0,6356142
Ain Turck(T3)	Bousfer (T4)	0,22570931
Ain Turck(T3)	M El Kebir (T4)	0,24214585
Ain Turck(T3)	Misserguine(T4)	0,29826996

Nous constatons, à travers les résultats de l'étude de l'autocorrélation spatio-temporelle entre les régions voisines de la zone d'étude, qu'il y avait de fortes autocorrélations entre les régions z_i dans les trimestres (T1,T2,T3) et les régions z_j dans les trimestres (T2,T3,T4). Par exemple la région Elbraya dans T1 est fortement autocorrélée avec la région Boufatis dans T2. La même chose pour Elancor dans T1 avec Missserguine dans T2 avec les indices de CITC respectifs **1,1** et **1,07**. De plus, Missserguine dans T2 est fortement autocorrélée avec sa zone voisine Ain Turck dans T3. Cela signifie que l'épidémie s'est propagée de Missserguine vers Ain Turk mais cette dernière n'est pas autocorrélée avec ses zones voisines (Bousfer, M El Kebir et Missserguine) dans T4 avec les indices d'autocorrélation de CITC respectifs **0,23 ; 0,24 ; 0,30**. Cela signifie que la propagation de l'épidémie a été limitée dans le temps au trimestre T3 et ne s'est pas étendue aux zones voisines pendant le trimestre suivant T4. Par contre, la région de Missserguine en trimestre T2 était fortement autocorrélée avec ses voisines (Ain Turck, Boutlelis, El Kerma, et Es Senia) avec des valeurs significatives de CITC, respectives : **0,9613501 ; 0,9677061 ; 0,7492245 ; et 0,6356142**. Ceci signifie que l'épidémie s'est propagée rapidement entre les trimestres T3 et T4, de la région de Missserguine à ses régions voisines. On observe aussi que la région de Missserguine a été contaminée par la région d'El Ancor au cours du trimestre T2, puis elle a été la cause de la contamination de toutes ses zones voisines plus tard lors du trimestre T4.

Les scénarios qui ont été révélés grâce au calcul de l'indice de LISTA n'ont pas été détectés auparavant par une analyse d'autocorrélation spatiale seule. L'indice de LISTA est capable de détecter des tendances à travers le temps car il opère sur plusieurs couches temporelles. Une région dans T1 peut être autocorrélée avec une région voisine dans T2 même si ces deux régions ne sont pas forcément autocorrélées spatialement dans le même trimestre T1 ou T2.

Au final, nous nous sommes limités dans ce travail au niveau trimestre pour tester l'applicabilité de l'indice de LISTA proposé. Une analyse plus fine au niveau mensuel mérite d'être explorée afin de discerner l'autocorrélation spatio-temporelle entre les mois et détecter d'autres scénarios pour savoir à quel moment la propagation a été déclenchée et à quel endroit. Dans nos futurs travaux, nous proposerons une analyse plus fine en descendant dans la granularité temporelle aux niveaux mois, semaine, jour pour révéler les agrégats spatio-temporels non détectés dans l'étude trimestrielle. L'échelle d'observation aura certainement un impact sur la compréhension du phénomène étudié.

6.7 Travaux connexes

Nous avons rassemblé dans cette section, les travaux connexes les plus pertinents qui sont en relation avec notre étude en épidémiologie. Il existe la catégorie des travaux qui utilisent des analyses spatiales pour identifier les zones présentant une incidence élevée tout au long de la période d'étude et d'autres études qui touchent plus des analyses spatio-temporelles pour identifier les zones avec augmentation de l'incidence dans des périodes spécifiques.

Nous distinguons dans ce qui suit ces deux catégories de travaux connexes : l'analyse d'autocorrélation spatiale et l'analyse temporelle et spatio-temporelle.

6.7.1 Analyse d'autocorrélation spatiale

Les méthodes d'autocorrélation spatiale et spatio-temporelle sont largement utilisées pour trouver des relations entre les phénomènes spatiaux et découvrir des modèles cachés à l'œil de l'analyste à première vue de la représentation géographique des données. Dans (Tsai et al, 2009) les auteurs ont exploité l'indice d'autocorrélation globale I de Moran et l'indice d'autocorrélation locale de Getis-Ord pour décrire et cartographier les agrégats spatiaux et les régions où sont situés, les 20 principales causes de décès à Taiwan en 2006. Ils ont pu découvrir grâce à cette analyse que la prévalence de la tuberculose correspond à la localisation des cantons autochtones. Les chercheurs dans (Chaikaew et al, 2009) ont pu définir des points chauds de l'épidémie de la diarrhée dans la province de Chiang Mai située dans le nord de la Thaïlande entre 2001 et 2006. Les chercheurs ont exploité les indicateurs locaux de l'association spatiale (LISA) puis ils ont appliqué l'interpolation de distribution en utilisant la méthode de la densité pour créer une surface continue représentant la densité de l'indice de morbidité de la diarrhée DMBR (*Diarrhea MorBidity Rate*). Ils ont pu enfin découvrir que les villages dans les régions du milieu et du nord ont révélé des incidences élevées. Dans (Wang et al, 2012), les chercheurs ont utilisé l'indice I de Moran pour discerner l'autocorrélation globale et détecter le modèle de la répartition spatiale de la tuberculose dans la ville de Linyi, en Chine entre 2005 et 2010 puis ils ont fait recours à l'indice de Getis-Ord G_i^* pour identifier la localisation des points chauds. Récemment, dans (Runzi et al, 2017), des analyses d'autocorrélation spatiale et spatio-temporelle ont été utilisées pour détecter des agrégats spatio-temporels des oreillons dans la province de Shandong en Chine entre 2005 et 2014. Le test de I de Moran était significatif et l'analyse des indicateurs locaux de l'autocorrélation spatiale (LISA) a révélé des agrégats spatiaux importants avec une incidence élevée. Une autre étude comparative des trois méthodes, scan statistique, l'indicateur local d'autocorrélation spatiale (LISA) et la statistique locale G_i a été faite sur les données du cancer des poumons dans le nord-est de la région d'Ohio entre janvier 1994 et Mai 2006 (Sikder et al, 2007). Afin de détecter le rôle de la population migrante dans la transmission de la tuberculose, les auteurs dans (Jia et al, 2008),

ont utilisé l'autocorrélation spatiale sur la population permanente et la population migrante entre 2000 et 2006 à Pékin en Chine.

6.7.2 Analyse temporelle et spatio-temporelle

Dans la littérature, nous retrouvons de nombreuses études intégrant l'aspect temporel de l'étude dans l'analyse spatiale. Les séries temporelles sont couramment utilisées pour modéliser la variabilité temporelle, surtout quand il s'agit du suivi sur de longues périodes. C'est le cas dans (Reis et Mandl, 2003) pour réaliser des prévisions sur l'utilisation des services d'urgence. Les auteurs dans (Andrienko et al, 2004) ont proposé CommonGis qui a été utilisé dans le domaine de la foresterie pour suivre le développement de 2600 arbres dans une période de 100 ans avec un intervalle de 5 ans. Certains travaux visent à supporter les séries temporelles longues ou multi-variées. D'autres études ont intégré l'aspect spatial et temporel pour avoir des études plus complètes. Le *Clustering scan statistic* proposé par (Kulldorff, 2005) et implémenté dans l'outil SaTScan, forme des clusters en explorant des fenêtres cylindriques avec une base géographique circulaire et dont la hauteur correspond au temps. Cette technique a été exploitée par plusieurs travaux tels que (Runzi et al, 2017), (Wang, 2012) et (Sikder et al, 2017). Une extension de l'autocorrélation spatiale aux données temporelles a également été proposée dans (Hardisty et Klippel, 2010) et implémentée dans l'outil LISTA-Vis. Le prototype a été utilisé pour surveiller la pandémie du virus H1N1 en 2009 aux Etats-Unis. Dans (Cheng et al, 2011), une proposition de l'extension de l'indice de Pearson pour intégrer l'aspect temporel a été proposée et appliquée à la surveillance du flux routier dans la ville de Londres. Récemment, dans (Jay et Shengwen, 2017), une tentative pour étendre l'indice de Moran tenant compte du temps a été proposée après plusieurs autres propositions similaires. Bien que ce dernier reflète une mesure globale et locale de corrélation tenant compte du temps, aucune étendue temporelle n'a été définie pour mesurer l'intervalle entre périodes considérées. Dans ce chapitre, nous introduisons l'indice de LISTA, une extension de l'indice de LISA telle que la matrice de voisinage tient compte d'un intervalle de voisinage temporel plus ou moins large, donné en paramètre.

6.8 Conclusion

Dans ce chapitre, une extension de l'indice de LISA pour intégrer l'aspect temporel dans un processus de surveillance épidémiologique a été proposée et appliquée à la surveillance de la tuberculose à Oran. Cette proposition a été faite en proposant la matrice des indices de l'autocorrélation spatio-temporelle permettant de considérer le temps et l'espace en même temps.

Le calcul de l'indice de LISTA dans les différentes zones géographiques de la région d'étude a montré l'utilité de cette mesure dans la détection des tendances spatiales des régions locales dans différentes échelles temporelles. Nous nous sommes limités, dans cette thèse, au niveau trimestriel mais théoriquement, nous pouvons aller en profondeurs dans la dimension temporelle jusqu'au niveau jour (reste à considérer la disponibilité des données de la surveillance). Ceci permettra de détecter les modèles journaliers des régions considérées. Le réajustement des niveaux de détails de la dimension temporelle reste relatif au niveau de la dangerosité de l'épidémie étudiée ainsi qu'à la vitesse de la propagation.

Conclusion générale et perspectives

Les systèmes d'aide à la décision spatiale et spatiotemporelle sont des systèmes efficaces dans la gestion et le contrôle des problèmes à caractère spatial et spatiotemporel tel que le phénomène de la propagation des épidémies. Les entrepôts de données, les systèmes d'information géographiques et les systèmes SOLAP constituent les principaux éléments de tels systèmes décisionnels. La fouille de données spatiales est une nouvelle démarche qui vise à exploiter les données à référence spatiale dans le processus d'exploration des données et la découverte de la connaissance. Dans cette thèse, nous nous sommes intéressées à la problématique de l'intégration des techniques de datamining spatial dans l'analyse multidimensionnelle spatiale afin de mettre en place un système décisionnel spatial (SDSS) plus complet et étendu sur l'aspect temporel au même titre que l'aspect spatial de l'étude. Différentes méthodes et outils ont été intégrés pour supporter, gérer, traiter, et analyser ce type de données complexes dans les entrepôts de données. Nos recherches sont guidées par une étude de cas sur la surveillance épidémiologique et nos principales contributions sont résumées ci-après.

1. Bilans et contributions

- **Le système SDSS_MCA_DM.** L'une des solutions proposées dans cette thèse pour améliorer la détection des cas consiste à proposer le système SDSS_MCA_DM qui intègre d'une part, les techniques de data mining afin d'expliquer la propagation de l'épidémie et prédire l'apparition de nouveaux foyers. D'autre part, il fait appel aux méthodes d'analyse multicritères pour prendre en compte les différents critères agissant sur la propagation, permettant ainsi une évaluation prédictive du risque pour la santé. Le système d'aide à la décision proposé comprend un sous-système de fouille de données, un sous-système d'analyse multicritères et un sous-système de visualisation. L'objectif de cette approche est de comprendre la dynamique de la tuberculose à Oran sur la base des facteurs socio-environnementaux (démographiques, climatiques, etc.) afin de créer un indice de vulnérabilité de la population vivant dans la région et établir une **carte de risque de la tuberculose**. Ce risque est défini par la présence simultanée de l'aléa et de la vulnérabilité de la population. L'évaluation des risques est basée sur l'apparition fréquente d'un ou plusieurs paramètres autour des cas. Ainsi, il est possible de créer un indice de risque basé sur les paramètres identifiant la présence du danger (Exemple: la présence des cas contagieux) et les paramètres identifiant la vulnérabilité de la population (Exemple: surpeuplement, localisation, type d'habitat ou taux de chômage).
- **Le système SOLAM (Spatial On Line Analytical Mining), SOLAP Mining ou SOLAP sémantique.** Une extension du système SOLAP Spatial On Line Analytical

Processing) en intégrant les techniques de la fouille de données spatiales (FDS). L'approche consiste à intégrer le système EPISOLAP, adapté à la surveillance épidémiologique avec la méthode de généralisation des données spatiales en tenant en compte non seulement de la hiérarchie spatiale mais aussi de l'aspect sémantique des autres attributs permettant ainsi l'évaluation prédictive des risques pour la santé en présence simultanée du danger de contamination et la vulnérabilité de la population exposée. L'architecture proposée est une plateforme décisionnelle intégrée pour la découverte de connaissances à partir des bases de données spatiales et les données de la surveillance. Les méthodes de généralisation spatiale permettent d'explorer les données à différentes échelles sémantiques et spatiales tout en réduisant les dimensions inutiles. Le principe de la méthode consiste à sélectionner et à supprimer des attributs de faible importance dans la caractérisation des données, afin de produire des zones de caractéristiques homogènes qui seront ensuite fusionnées.

- **Proposition d'une démarche systémique** qui va de l'analyse descriptive en passant par l'analyse exploratoire pour finir par l'analyse explicative (prédictive) appelée aussi décisionnelle. Nous avons structuré, à travers cette démarche, les différentes méthodes et techniques existantes pour analyser les données à référence spatiale en suivant une logique d'exploration de ce type de données.
- **Proposition de l'indice de LISTA (Local Indice of Space Time Autocorrélation)** pour mesurer et analyser l'autocorrélation spatio-temporelle locale à une zone et l'ajouter comme une nouvelle méthode dans la phase exploratoire de la démarche systémique proposée. Cet indice fonctionne comme un instrument de mesure permettant d'assister l'analyste en lui fournissant une représentation interactive des indicateurs d'autocorrélation spatiotemporelle pour assister sa réflexion à trouver des relations et à détecter des anomalies. Nous avons définie également la matrice des indices d'autocorrélation spatiotemporelle qui aboutit à une représentation géographique des indices de LISTA obtenus.

2. Perspectives

Nous avons exploré, à travers cette étude, différentes techniques de fouille de données (arbre de décision) et de fouille de données spatiales (généralisation et autocorrélation spatiales et spatiotemporelle) ainsi que des méthodes d'aide à la décision multicritères (ELECTRE TRI et PROMETHEE II). Beaucoup reste à faire pour compléter l'étude abordée dans ce modeste travail. Nos futurs travaux concernent chacune des contributions proposées dans cette thèse que nous avons séparées en perspectives à court terme et perspectives à long terme :

2.1 Perspectives à court terme

- **Le système SDSS_MCA_DM** : dans la première contribution de cette thèse, nous avons traité le problème de **la surveillance prédictive** par l'intégration de la méthode multicritères ELECTRE TRI pour affecter les foyers d'épidémie dans trois catégories distinctes {Fort, Moyen, Faible} en vue de préparer l'étude prédictive utilisant la méthode de l'arbre de décision. Cette approche permet de planifier les prochains efforts d'éradication et de prise en charge des malades. Dans un travail antérieur (Zemri et al,

2014), décrit dans l'annexe B, nous avons traité **la surveillance curative** en classant les foyers d'épidémie par la méthode d'analyse multicritères PROMETHEE II. Ce classement aide à établir un ordre d'intervention et à limiter le risque de la propagation. Le défi consiste à intégrer les systèmes curatif et préventif dans un seul système global et cohérent pour la surveillance épidémiologique.

- **Le système SOLAM (Spatial On Line Analytical Mining), SOLAP Mining ou SOLAP sémantique :** dans notre seconde contribution, nous avons montré l'efficacité du système EPISOLAM en termes de généralisation spatiale basée sur la méthode d'induction orientée attribut. Ce système fournit des informations pertinentes aux décideurs en santé publique en supprimant les attributs inutiles dans chaque niveau d'hierarchie et en ne laissant que les attributs ayant une grande importance. Seules les méthodes de caractérisation et de comparaison ont été testées et exploitées dans ce travail de thèse. Néanmoins, la méthode d'induction orientée attribut pourrait être appliquée sur d'autres méthodes de fouille de données spatiales telle que la classification spatiale, l'association spatiale et le clustering spatial. Dans nos futurs travaux, nous essayerons d'expérimenter ces méthodes et d'autres afin de compléter notre étude actuelle.
- **Autocorrélation spatiotemporelle par indice de LISTA :** des efforts pour intégrer l'aspect temporel ont été proposés pour supporter l'auto corrélation des données spatiotemporelles. Nous pouvons, également, proposer une analyse des séries spatiotemporelles en se basant sur l'indice de LISTA. Des techniques d'analyse des séries spatiotemporelles pourraient être appliquées en intégrant l'aspect spatial et temporel des données pour analyser l'évolution spatiotemporelle des données de la surveillance.

2.2 Perspectives à long terme

- **Démarche systémique pour la fouille de données spatiales :** Cette démarche utilise divers méthodes, mesures ou algorithmes. Il reste à intégrer véritablement de tels outils et les appliquer afin de suivre d'autres études de cas réels en vue de valider notre démarche générique.
- **Traiter le problème de l'interactivité du système décisionnel proposé :** D'autres techniques de fouille de données spatiales méritent d'être améliorées, notamment en ce qui concerne l'aspect visuel de l'étude permettant d'intégrer l'aspect spatial et temporel d'une manière plus interactive afin de concevoir un système de surveillance spatio-temporel en temps réel.
- **Intégration des systèmes proposées :** nous devons penser à comment peut on regrouper toutes les contributions autour d'un seul système en suivant une logique d'intégration pour assurer une complémentarité entre les sous-systèmes.

TROISIEME PARTIE

ANNEXES

Epidémiologie

Dans cette annexe, nous présentons le domaine de l'épidémiologie et ses pratiques. Notamment, la surveillance et ses indicateurs

A.1 Introduction

On date les débuts de l'épidémiologie moderne aux premières enquêtes type cas-témoins ou cohorte. La principale cause de la baisse de la mortalité par maladies infectieuses a été les progrès de l'hygiène et de l'alimentation dans les pays industrialisés au 20ème siècle. S'ajoute à cela l'avènement des antibiotiques et des vaccins.

Initialement (et dans l'esprit de beaucoup) l'épidémiologie est associée à l'étude des maladies infectieuses. Aujourd'hui l'épidémiologie couvre des domaines beaucoup plus variés comme les maladies chroniques : cardiovasculaires, cancers, neuro dégénératives, la santé au travail, santé de la reproduction, effets de l'environnement sur la santé...etc.

A.2 Définition de l'épidémiologie

L'épidémiologie est une discipline de la santé publique qui s'intéresse à l'état de santé des populations plutôt que celle des individus (médecine clinique) et fournit des méthodes pour mesurer et identifier les causes des maladies (Hannekens et al, 1987).

Le dictionnaire de l'épidémiologie, Last, 1988 la définit comme « *l'étude de la répartition et des déterminants des événements de santé dans les populations, et l'application des résultats au contrôle des problèmes de santé* ». Ces déterminants sont : les facteurs de risque qui implique une causalité directe et les facteurs associés qui peuvent être causals ou non.

L'épidémiologie est la science de l'estimation du risque en santé et permet de répondre à deux questions très simples (Fontanet, 2014) :

- Quel est le risque d'être (de devenir) malade dans une population donnée ?
- Quelle est l'augmentation du risque associée à certaines caractéristiques : gène, comportement, environnement, ...?

A.3 Etapes d'une étude épidémiologique

Pour mener une étude en épidémiologie, les épidémiologistes suivent une démarche statistique qui se compose des étapes suivantes :

- Formulation d'une hypothèse
- Méthode d'échantillonnage et de suivi
- Anticipation des biais possibles
- Calcul de taille d'échantillon
- Recueil des données
- Analyse statistique
- Communication des résultats

A.4 Pratique de l'épidémiologie

Les études en épidémiologie se distinguent en deux catégories : épidémiologie **descriptive** et épidémiologie **analytique**.

A.4.1 Epidémiologie descriptive

Cette catégorie concerne principalement la surveillance épidémiologique.

A.4.1.1 Définition de la surveillance

La surveillance est le processus continu et systématique de recueil, analyse, interprétation et dissémination de données descriptives permettant le **suivi temporel** des événements de santé dans une population. (Adapté de SB Thacker, 1992).

La surveillance repose sur des mesures de la population : les maladies à déclaration obligatoire, les registres, les réseaux de cliniciens (ex : grippe), les réseaux de laboratoire, les enquêtes transversales par échantillonnage représentatif et sondage d'opinion (ex: Enquêtes Démographiques de Santé).

A.4.1.2 Types de la surveillance

On distingue deux types de surveillance reference:

1. **Surveillance active** : ou l'enquêteur va chercher l'information. Si possible, en temps réel, données disponibles rapidement : **réactivité** (système de surveillance informatisé).
2. **Surveillance passive** : repose sur la remontée des données (exemple : les maladies à déclaration obligatoire)

A.4.1.3 Objectifs de la surveillance

La surveillance a pour objectifs principaux (Fontanet, 2014) :

- La détection d'une épidémie : Franchissement d'un seuil épidémique et la surveillance syndromique.
- L'estimation fiable des grandes tendances (augmentation, stabilité, baisse) du nombre de nouveaux cas d'une maladie pour une période donnée.
- L'évaluation de l'histoire naturelle d'une épidémie ;
- L'évaluation de l'impact d'une intervention (exemple : dépistage, contrôle).

A.4.1.4 La modélisation mathématique

La modélisation mathématique intervient en complément de la surveillance quand la mesure n'est plus possible (ex : émergence) et permet de détecter des tendances (analyse des séries

temporelles), définir des seuils épidémiques, prédire le devenir de l'épidémie ou estimer l'impact d'une intervention.

A.4.1.5 Propriétés d'un système de surveillance

Un système de surveillance doit respecter les propriétés suivantes (Fontanet, 2014) :

- **Sensible** : détecte les épidémies débutantes ou les tendances
- **Réactif** : résultats disponibles rapidement pour les autorités de santé
- **Intégré** : peu de surcharge de travail
- **Acceptable** : pour les patients comme pour les personnels soignants
- **Représentatif (exhaustif)** : les tendances repérées sont celles de la population générale

A.4.2 épidémiologie analytique

L'investigateur utilise une démarche scientifique pour étudier la **relation causale** entre facteur d'exposition et maladie. Il existe quatre groupes d'études analytiques (Bouyer et al, 2009):

A.4.2.1 Etude d'observation

Celles où l'investigateur observe la relation telle qu'elle existe dans les populations d'étude. Il existe quatre types dans cette catégorie : études de corrélation écologiques, enquêtes transversales, étude cas-témoin et étude de cohorte.

– Etudes de corrélation écologiques

Une étude de corrélation écologique a pour unité d'analyse des groupes d'individus, intéressante pour l'élaboration d'hypothèses mais elle est la plus fragile pour mettre en évidence une association exposition maladie car une association au niveau des populations n'existe pas nécessairement au niveau individuel. Il est également difficile de contrôler les facteurs de confusion.

Exemple illustratif : les mangeurs de la viande rouge ont plus tendance d'avoir un cancer du colon non pas à cause de la consommation de la viande rouge mais parce qu'ils consomment moins de fibres, et les fibres ont un effet protecteur contre le cancer du colon.

– Enquêtes transversales

Les enquêtes transversales sont réalisées pour estimer la prévalence d'une maladie en population. Le recueil simultané de données sur la maladie et des facteurs d'exposition peut conduire à regarder les associations entre exposition et maladie. Mais c'est une approche considérée aussi comme peu fiable à cause de la possibilité de changement des comportements avec le développement de la maladie. Exemple : étudier l'association entre la présence de la nicotine dans les ongles et infarctus du myocarde.

– Études cas-témoins

Conceptuellement, l'étude cas-témoin est un échantillonnage d'une cohorte virtuelle ou un groupe de malades est constitué, et la proportion d'exposés est comparée entre ce groupe et un

groupe témoin. Les sujets sont inclus dans l'étude selon qu'ils ont la maladie (les cas) ou non (les témoins). Exemple : études sur le cancer du poumon et le tabac.

Les témoins doivent être représentatifs de la population source dont sont originaires les cas. L'investigateur compare la proportion d'exposés (« **l'odds** » d'exposition) entre les cas et les témoins.

– Études de cohorte

Sont les études considérées comme les plus fiables en épidémiologie analytique. Un groupe d'individu est suivi, et la survenue de la maladie est comparée entre les « exposés » et les « non-exposés ». Tous sont indemnes de la maladie étudiée en début de cohorte. Le suivi peut s'étaler sur de longues périodes (souvent plusieurs années) afin de déterminer s'ils vont ou non développer la maladie d'intérêt. Ils sont particulièrement appropriés pour la mise en évidence de l'association temporelle entre exposition et maladie pour avoir une mesure précise de l'exposition. Les indicateurs utilisés sont : **incidence cumulative** (IC) dans une cohorte fermée et **taux d'incidence** (IT) dans une cohorte ouverte.

La puissance d'une étude de cohorte est plus élevée quand (Fontanet, 2014) :

- Le nombre d'événements (plutôt que le personne-temps de suivi) est élevé.
- Le personne-temps de suivi est reparti de façon équilibrée entre exposés et non-exposés.
- L'étendue des valeurs de l'exposition est élevée.

Les études de cohorte peuvent être **prospectives** ou bien **rétrospectives** définies selon que la maladie étudiée a déjà été observée ou non chez les sujets de l'étude au moment où l'investigation commence :

- **Prospective** : la maladie n'a pas encore été observée.
- **Rétrospective** : la maladie a déjà été observée. Exemple : Association entre « Benzidine » amine aromatique utilisée comme colorant et cancer de la vessie.

Exemple : «Nurses Health Study» ref. <http://www.channing.harvard.edu/nhs/>

- Recrutement :
- 1976 : 1ère cohorte de 122 000 infirmières âgées de 30 à 55 ans
- 1989 : 2ème cohorte : 117 000 infirmières âgées de 25 à 42 ans
- 2010 : 3ème cohorte
- Exposition : questionnaires sur contraception orale et nutrition
- Maladies : questionnaire sur les événements de santé tous les deux ans

– Etude cas-témoin « nichée » (nested case-control study)

Etude de cohorte où l'exposition n'est déterminée que pour un nombre restreint de sujets qui représentent les cas dans l'étude cas-témoins et un groupe de témoins non malades au moment du diagnostic des cas. Cette méthode a le grand avantage pour économiser le temps et l'argent si l'exposition est difficile à mesurer Exemple : dans l'étude de cohorte la Nurses Health Study:

I.4.2.2 Etudes d'intervention (expérimentale)

Celles où l'investigateur intervient en assignant le facteur d'exposition à un groupe d'individus, et en comparant la survenue de la maladie entre ce groupe et un groupe témoin. Nous citons dans cette catégorie les essais cliniques ou essais randomisés.

- **Essais randomisés** : connues aussi sous le terme essais cliniques car elles se rapprochent des expériences de laboratoire des sciences fondamentales et correspondent à des études de cohorte où l'exposition est assignée de façon aléatoire entre deux groupes de participants. Tout est fait pour que les deux groupes soient parfaitement comparables en dehors de l'exposition.

Chacune de ces trois dernières méthodes présentent des puissances et des faiblesses que nous citons dans le tableau suivant (Fontanet, 2014).

Tableau A.1 Avantages et inconvénients des études en épidémiologie

Types d'études	Avantages	Inconvénients
Etude de cohorte	– Séquence temporelle entre exposition et maladie clairement établie. – Qualité de mesure de l'exposition – Permet d'étudier les facteurs d'exposition peu répandus en les «sur-représentant» dans l'effectif de la cohorte. – Permet l'étude de la relation entre un facteur d'exposition et plusieurs maladies. – Adaptée aux expositions rares (ex: amiante)	– Longue : suivi prolongé (plusieurs années) – Nombre élevé de sujets – lourdes logistiquement et coûteuses. – Un taux élevé de perdus de vue dans une étude de cohorte entraîne systématiquement une baisse de puissance de l'étude
Études cas-témoins	– Appropriées pour les maladies rares. – Appropriées pour les maladies à incubation longue. – Plusieurs facteurs d'exposition peuvent être étudiés pour une même maladie. – Nécessitent moins de sujets et sont moins coûteuses comparées aux études de cohorte	– Choix du groupe témoin – Qualité de mesure de l'exposition
Essai cliniques (randomisés)	– La méthodologie la plus rigoureuse disponible pour l'évaluation d'une relation entre exposition et «guérison».	– Difficulté pour la « généralisation » des résultats au delà de la population d'étude. – des considérations éthiques et de faisabilité.

Le schéma de la figure A.1 résume les différents types des études en épidémiologie analytique.

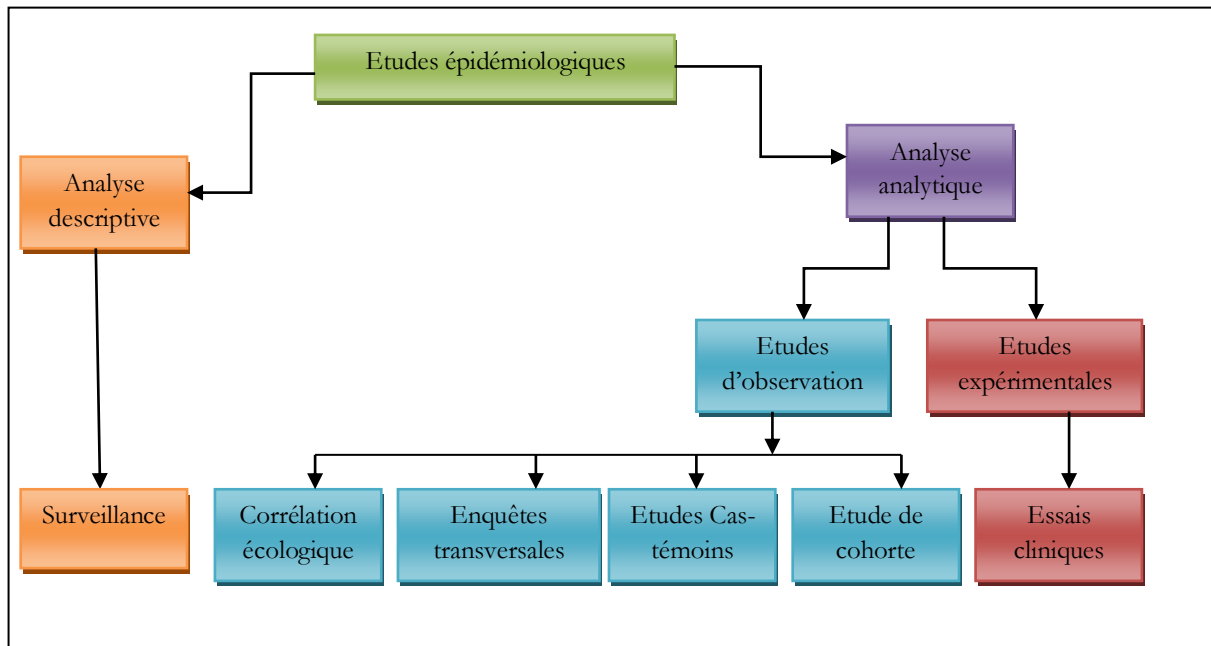


Figure A.1 Types d'études en épidémiologie

A.5 Niveau de preuve et causalité

Il est important de prouver la puissance d'une association retrouvée dans une étude épidémiologique. Ces évaluations sont complexes et font encore l'objet de développement méthodologique. Les études en épidémiologie sont classées selon leurs niveaux de preuves, du niveau le plus bas au niveau le plus élevé selon le classement suivant :

1. Etudes de corrélation écologique
2. Enquêtes transversales
3. Etude cas-témoins
4. Etudes de cohortes
5. Essais cliniques (randomisés)

Les essais randomisés est l'étude dont les résultats sont les plus probants (Bouyer et al, 2009).

A.6 Lien de causalité

Après avoir assuré que l'exposition précède la maladie et afin de démontrer le lien de causalité entre une exposition et une maladie il faut respecter cinq critères décrits dans ce qui suit (Fontanet, 2014) :

1. Mesure de l'intensité (la force) de la relation entre exposition et maladie par le calcul du risque relatif.
2. Absence d'effet de confusion en faisant la distinction entre facteur de risque (relation causale) et facteur associé (causalité non prouvée)

3. la relation doit être une relation dose-effet : l'effet augmente avec l'augmentation de la dose de l'exposition.
4. la confirmation des résultats par d'autres études.
5. trouver une explication biologique plausible au phénomène observé.

A.7 Indicateurs

Cette section a pour objectif de présenter les principaux indicateurs qui mesurent la présence et l'intensité d'une maladie en population (**prévalence et incidence**), et ceux qui mesurent l'association entre exposition et maladie (**risque relatif et risque attribuable**). Nous définissons dans ce qui suit chacun de ces indicateurs :

Le risque est défini sur (Wikipedia, 2008) comme une perte potentielle, identifiée et quantifiable, inhérente à une situation ou une activité, associée à la probabilité de l'occurrence d'un événement ou d'une série d'événements. Le risque d'être malade en épidémiologie est estimé principalement selon deux mesures **la prévalence et l'incidence** :

- **Prévalence (P)** : proportion de sujets atteints à un instant donné dans une population.
- **Incidence (I)** : nombre de nouveaux cas sur une période de temps donnée dans une population et s'exprime en personnes-temps. (Exemple : une incidence de 2,3 cas pour 100 personnes-années correspond à 2. 0,19 cas pour 100 personnes-mois).
- **Incidence cumulative (IC)** : proportion de personnes qui deviennent malades sur une période donnée (IC = total de la population à risque).
- **Le taux d'attaque** : une proportion qui a pour dénominateur le nombre de sujets à risque de la maladie (Nombre de cas d'une maladie / population donnée sur une période courte et définie). S'utilise en cas d'épidémie à développement rapide (exemple : taux d'attaque est approprié à Une épidémie de grippe dans une école).
- **Le taux de morbidité** : Nombre de nouveaux cas d'une maladie / population à risque pendant une période donnée et s'exprime en cas par personnes-temps.
- **Le taux mortalité** : Nombre de décès / population à risque pendant une période donnée (exprimé en personnes années).
- **Le taux de létalité** : Nombre de décès / nombre de malades. n'est pas un taux, c'est une proportion.
- **Ratio Morbidité/ Mortalité** : SMR (standardized morbidity/mortality ratio) s'exprime par Taux de morbidité/ Taux de mortalité. s'exprime sous forme d'un ratio

Les mesures d'association

Les mesures d'association permettent de quantifier la relation entre un facteur d'exposition et une maladie mais ne garantissent pas pour autant que l'association soit causale. On peut choisir entre deux mesures (Bouyer et al, 2009) :

- **Le risque relatif** : mesure l'augmentation (diminution) du risque d'être malade chez les exposés comparés aux non-exposés. Il est calculé par le rapport des risques entre exposés et non exposés ($R1/R0$)
- **Le risque attribuable** : proportion des cas dans la population attribuable à une exposition donnée. Il est calculé par la différence des risques entre exposés et non exposés ($R1-R0$)

L'expression du risque relatif prend une forme différente selon le type d'étude utilisé.

– **Dans les études de cohorte :**

- Ratio des incidences cumulatives si la cohorte est fermée.
- Ratio des taux d'incidence si la cohorte est ouverte.

En absence de perdus de vue, le ratio des incidences cumulatives pour une exposition délétère est inférieur au ratio des taux d'incidence

– **Dans les études cas-témoins :** La mesure de choix du risque relatif dans une étude cas-témoins est l'« odds ratio » qui est une estimation non biaisée du ratio des taux d'incidence de la cohorte virtuelle à l'origine des cas et des témoins de l'étude. On compare les « odds » d'être exposés, et non les proportions d'exposés, entre cas et témoins. L'odds-ratio se calcule comme suit :

- Odds d'être exposés pour les cas = probabilité d'être exposés divisée par la probabilité de ne pas être exposés pour les cas.
- Odds d'être exposés pour les témoins = probabilité d'être exposés divisée par la probabilité de ne pas être exposés pour les témoins.
- Odds-ratio = ratio de ces deux odds

Selon les valeurs que peut prendre l'odds ratio on peut savoir si l'exposition a un effet protecteur, délétère ou absence d'association

- L'effet protecteur d'une exposition se caractérise par un risque relatif d'être malade entre exposés et non exposés < 1
- L'effet délétère d'une exposition se caractérise par un risque relatif d'être malade entre exposés et non exposés > 1
- absence de risque associée à l'exposition se caractérise par un risque relatif $= 1$

L'odds d'être exposé chez les cas divisé par l'odds d'être exposé chez les témoins est égal à l'odds d'être malade chez les exposés divisé par l'odds d'être malade chez les non-exposés.

Quand l'exposition a plus de deux catégories, le choix de la catégorie d'exposition de référence pour le calcul des odds-ratios est basé sur la catégorie pour laquelle l'interprétation des odds-ratio sera la plus parlante.

A.8 Conclusion

Nous avons survolé à travers cette annexe le domaine de l'épidémiologie axée sur les comportements des individus malades (ex : tabac, nutrition...). Aujourd'hui l'épidémiologie s'ouvre à la génétique, à l'environnement et au « big data ». S'ajoute à cela le besoin d'intégrer la dimension sociale à l'épidémiologie en plus de l'approche déterministe reposant sur l'étude moléculaire, possible grâce aux progrès de la biologie « marqueurs » ; ce qui crée des modèles à plusieurs niveaux

Analyse multicritères dans EPISOLAP

Dans cette annexe, nous présentons une application de la méthode multicritères Prométhée pour ranger les foyers d'épidémie dans la région d'Oran.

B.1 Introduction

En raison de l'évolution démographique et de la situation socio-économique de l'Algérie, de nouveaux progrès restent à faire pour réaliser les objectifs du plan de relance de la lutte contre la tuberculose (2006-2015) qui font partie des Objectifs du Millénaire pour le Développement (OMD), dans la nouvelle stratégie « Halte à la tuberculose » recommandée par l'OMS depuis 2006 : « Stopper l'augmentation de l'incidence de la tuberculose et commencer à la réduire sur tout le territoire national » (Agadir et al. 2011). Ce travail s'intègre dans cette optique pour aider dans la réalisation de tels objectifs en élaborant un système d'aide à la décision spatiale (SDSS) capable de porter une aide dans la surveillance épidémiologique.

B.2 Conception multicritères

Le modèle décisionnel proposé basé sur l'aide à la décision multicritères est largement inspirée de celui proposée par (Hamdadou, 2008) et repris dans (Zemri et al, 2012). Ce modèle se base sur les trois phases suivantes : la phase de structuration, la phase d'exécution et la phase d'exploitation.

Les principales phases et les étapes du modèle proposé sont illustrées sur la figure B.1.

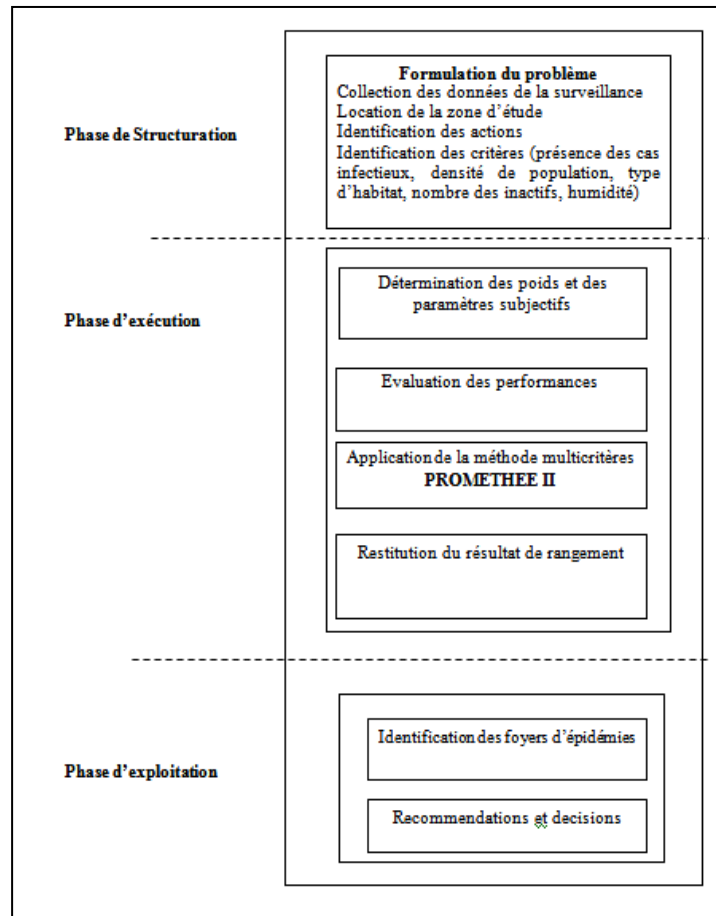


Figure B.1 Modèle décisionnel multicritères proposé

1. La phase de structuration

Cette première phase vise à identifier les paramètres à prendre en compte telles que la situation géographique de la zone d'étude à l'aide d'un SIG et les différents critères et actions. Cette phase vise, principalement, à formaliser deux éléments fondamentaux de la situation décisionnelle :

- **Identification des actions** : l'identification de toutes les actions potentielles est une étape très importante dans toute approche d'aide à la décision, en particulier lorsque les méthodes d'analyse multicritères procèdent par agrégation partielle. Il est très important que l'ensemble de toutes les actions soit complet car sa modification au cours de l'analyse peut entraîner une récurrence de l'analyse multicritères.
- **Identification des critères** : la liste des critères obtenue en rassemblant les facteurs correspondants (sous-critères) devrait être aussi complète que possible. Ces critères doivent être liés à des contraintes et des objectifs utilisés dans les activités de production. La famille de critères les plus pertinents doit vérifier les conditions de l'exhaustivité, la cohérence et l'indépendance.

2. La phase d'exécution

Cette deuxième phase est le processus d'analyse de l'étude. Ses deux principaux objectifs sont l'évaluation des critères, puis l'agrégation de ces informations par une analyse multicritères en

exploitant les méthodes de la classification multicritères à savoir (les méthodes de la famille **PROMETHEE**)

3. La phase d'exploitation

Cette troisième phase est principalement le résultat de l'acceptation sociale. Toutefois, elle comprend également la mise en œuvre de la décision et le contrôle de cette exécution.

L'architecture fonctionnelle du système décisionnel proposé basé sur l'aide à la décision multicritères est illustrée sur le diagramme de la figure 3.7.

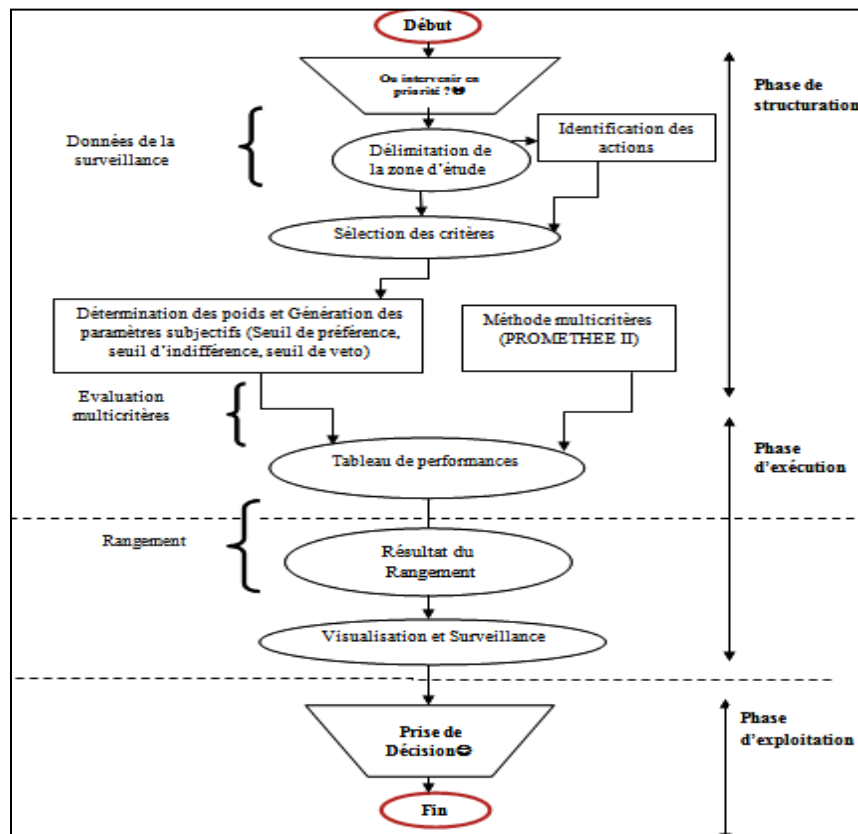


Figure B.2 organigramme associé au modèle décisionnel proposé

B.3 La méthode multicritères PROMETHEE

PROMETHEE est l'acronyme de Preference Ranking Organisation METHod for Enrichment Evaluation. La méthode PROMETHEE relève de la problématique de rangement^y procédant par agrégation partielle. Tout comme les méthodes de la famille Electre, PROMETHEE construit une relation de surclassement, en s'appuyant sur la comparaison des actions deux à deux, son but est donc de ranger les actions de la meilleure à la moins bonne. Il en existe quatre versions assez proches dont les deux premières rencontrent assez bien de succès grâce, notamment, à la disponibilité d'un logiciel commercial qui comporte une interface réussie. PROMETHEE se distingue d'Electre par le fait qu'elle construit une relation de surclassement vaguée traduisant une intensité de préférences : on peut considérer que les méthodes PROMETHEE sont à mi chemin entre l'approche de surclassement dont elles conservent la relation S et les méthodes de MAUT dont elles utilisent les méthodes de construction des

fonctions d'utilité partielles g_i . Le principe général de fonctionnement de PROMETHEE (I et II) est donné par l'organigramme illustré par la figure B.3.

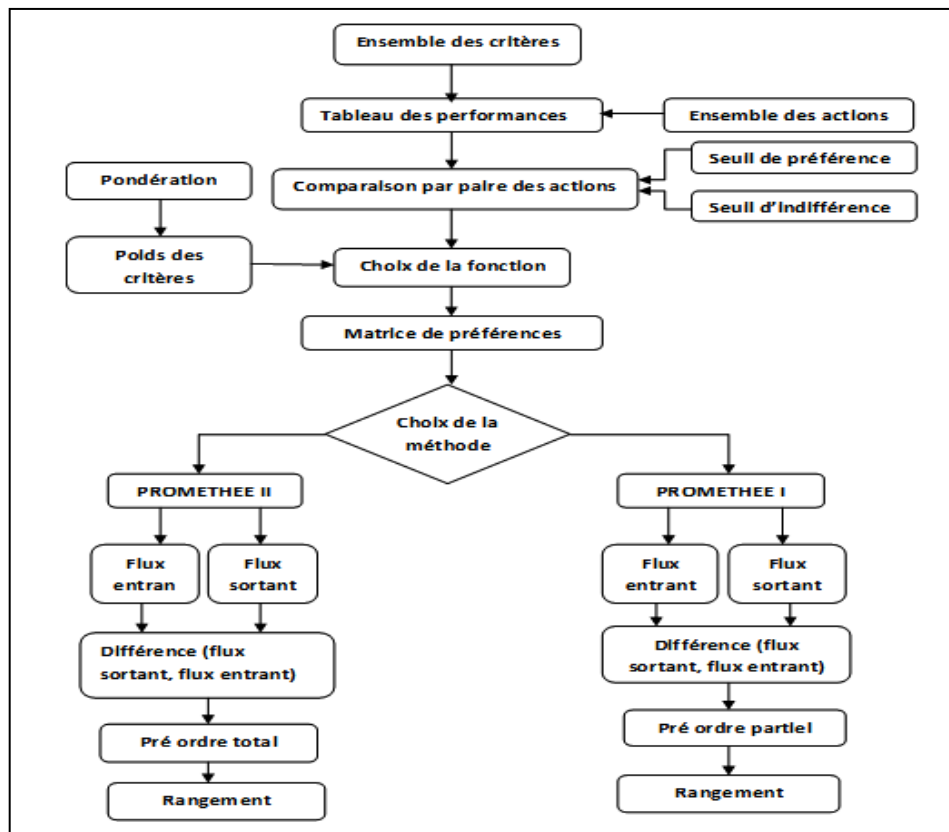


Figure B.3 Le principe général de fonctionnement par PROMETHEE (I et II)

Le choix de PROMETHEE (Préférence Ranking Organisation Method for Enrichment Evaluations) est justifié par les avantages suivants (Hamdadou et al, 2012) :

- La simplicité et l'aspect intuitif de la méthode
- La puissance de sa fonction de préférence
- La simplicité de la phase d'exécution de la méthode

La méthode PROMETHEE I fournit à l'utilisateur un classement des différentes actions (foyers). Le problème est que cette méthode est qu'elle ne classe pas toutes les actions. Certaines actions peuvent rester incomparables. La méthode PROMETHEE II permet d'enlever cette incomparabilité. Le principe de cette méthode est de créer un procédé numérique de comparaison de chaque action par rapport à toutes les autres. Ainsi, il est possible de calculer la plus (mérite) ou moins (démérite) de chaque action par rapport à tous les autres. Le résultat de cette comparaison permet la classification ordonnée des actions. La mise en œuvre du procédé revient à exécuter les trois étapes suivantes :

1. **Choix des critères généraux** : Chacun des critères $C_1, C_2 \dots C_m$ est associé à un critère généralisé basé sur une fonction de préférence et les effets d'échelle sont éliminés.
2. **Détermination d'une relation de surclassement** : Dans une deuxième phase, il est nécessaire de déterminer une relation de surclassement à travers un indice de préférence

qui quantifie les préférences du décideur. L'intensité de la préférence est calculée comme suit :

$$\begin{cases} p(d) = 0 \text{ si } d \leq q_j, \\ p(d) = (d - q_j) / (p_j - q_j) \text{ si } q_j < d \leq p_j \text{ et} \\ p(d) = 1 \text{ sinon} \end{cases}$$

- « a » et « b » étant deux actions potentielles,
- « d » est la différence entre la performance de « a » et la performance de « b » ($g_j(a) - g_j(b)$).
- q_j est le seuil d'indifférence,
- p_j est le seuil de préférence.

3. Evaluation des préférences : L'évaluation de la préférence du décideur est assurée par l'inclusion de flux entrants et sortants.

- Calcul du seuil de préférence : $\pi(a, b) = \sum W_j * P_j(a, b) / \sum W_j$; W_j est le poids du critère j.
- Calcul des flux entrants : $\phi^+(a) = \sum \pi(a, x)$
- Calcul des flux sortants : $\phi^-(a) = \sum \pi(x, a)$
- Calcul des flux globaux : $\phi(a) = \phi^+(a) - \phi^-(a)$

L'organigramme suivant illustre, en détails, la démarche d'utilisation de Prométhée II.

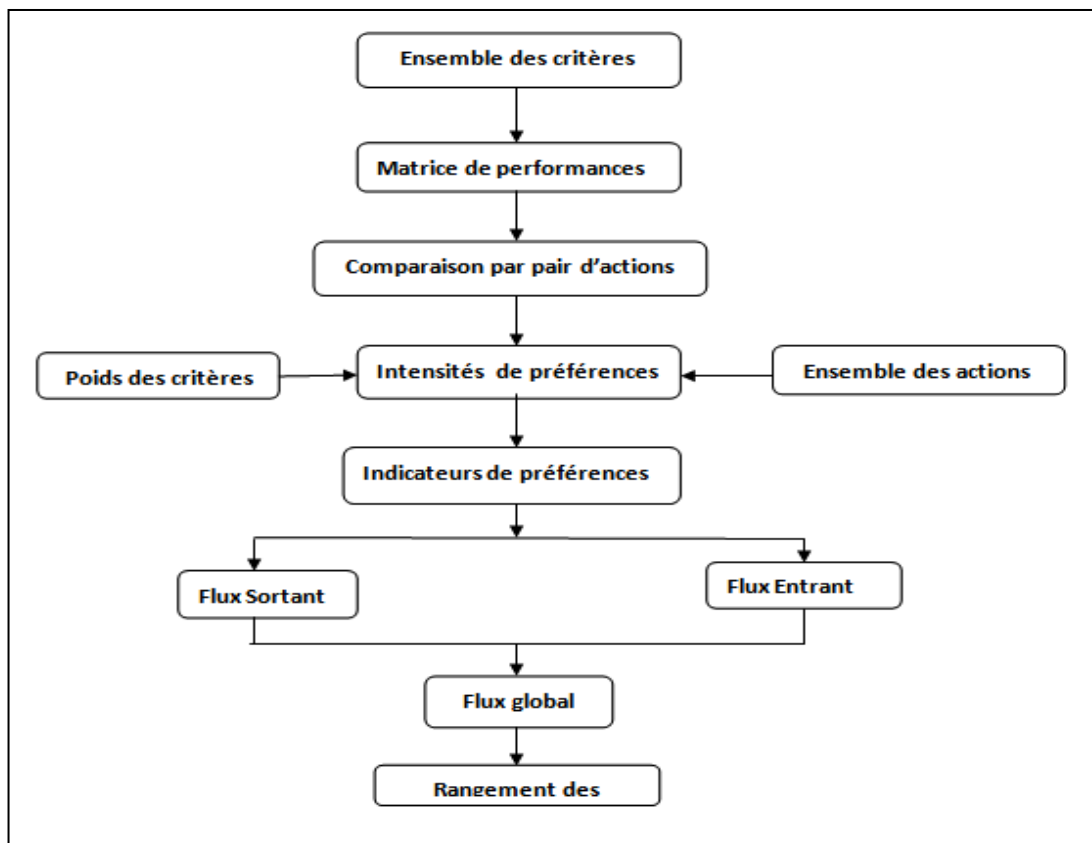


Figure B.4 Le principe général de fonctionnement de PROMETHEE II

B.4 Expérimentation

L'étude de cas est la surveillance de la maladie de la tuberculose dans la région d'Oran en Algérie.

B.4.1. Identification de la zone d'étude : Le terrain d'étude est la région d'Oran en Algérie qui est une des régions de l'ouest de l'Algérie touchée par la maladie de la tuberculose (Agadir et al 2011).

B.4.2. Identification des actions : Les actions considérées dans notre étude multicritères sont les foyers d'épidémies représentées par les 26 communes de la wilaya d'Oran.

B.4.3. Identification des critères : Notre choix dans cette conception multicritères était porté sur les cinq critères suivants :

- **Aléas : Présence de cas contagieux (Critère Médicale) :** L'aléa de la tuberculose est défini en trois classes en fonction de la présence des cas contagieux (les cas prouvés) et de la présence de cas de tuberculose non contagieux (non prouvés) mais potentiellement contagieux. L'aléa est fort si, dans le foyer considéré, il y a au moins un cas contagieux ; l'aléa est faible si il n'y a aucun cas de tuberculose détecté comme contagieux mais qu'il y a au moins un cas de tuberculose recensé dans le foyer ; l'aléa est nul si aucun cas de tuberculose n'a été recensé pendant toute l'année (probabilité de l'évènement "présence d'un cas de tuberculose" = 0).
- **La densité (Critère démographique) :** Il existe dans la littérature peu d'informations concernant la « vulnérabilité sociale » et les données socio-économiques qui peuvent y être associées. Parmi les facteurs socio-économiques liés à la présence de la tuberculose, et selon la disponibilité de ces paramètres vulnérables, Notre choix était porté sur le paramètre de « la densité de la population » qui nous paraisse le plus significatif. En effet, une surpopulation dans une région donnée favorise la contamination et la propagation des épidémies y compris la tuberculose qui suffit un contact direct avec le sujet atteint de la maladie pour transporter le virus par voix aérienne. Une population plus dense est donc plus vulnérable d'être contaminé en présence bien sur de l'aléa qui est ici la présence du cas contagieux.
- **Typologie d'habitat (Critère démographique) :** Ce critère est calculé par le nombre de constructions précaires occupant généralement les bidonvilles.
- **Nombre des inactifs (Critère démographique) :** Le nombre des inactifs est calculé dans notre étude par le nombre des personnes sans emploi dans les régions.
- **Humidité (Critère climatique) :** L'humidité du climat est un paramètre très important qui favorise la propagation des bactéries (par exemple, le bacille de coq responsable de la maladie de la tuberculose). Etant dans l'impossibilité d'avoir l'humidité annuelle moyenne de toutes les régions d'Oran (présence de deux stations de mesure seulement : Sénia et Arzew), nous avons identifié, à l'aide d'un spécialiste dans la météo au bureau national de la météo à Oran, une échelle de quatre points (1-2-3-4) qui classe les régions de la plus humide (mesure = 1) à la plus sèche (mesure = 4).

B.4.4. Génération du tableau de performances : Les couches d'informations intervenant dans notre modèle du risque tuberculeux, c'est-à-dire l'aléa et la vulnérabilité de la population : la densité, type d'habitation, le taux de l'humidité et le niveau de la pauvreté, croisés dans le système SOLAP, faisant apparaître les zones au niveau desquelles les interactions entre une population

vulnérable face à la tuberculose et des patients atteints de tuberculose qui sont susceptibles de transmettre la maladie (Prouvé/Non prouvé) sont les plus intenses.

B.4.5. Poids des critères

Dans le cas fréquent, l'analyse des conséquences des actions potentielles conduit à la construction de plusieurs critères multiples est l'analyse des critères permet d'apporter des réponses au problème. Pour une action donnée, et pour chaque critère un seuil de préférence p , d'indifférence q et le seuil de veto v sont estimés, en sachant que les méthodes PROMETHEE n'exploitent pas le seuil de veto qui est un paramètre subjectif exprimé par le décideur. Chaque critère est affecté d'un seuil de préférence p , d'un seuil d'indifférence q et d'un poids k reflétant sa contribution à la décision finale. Le résultat de l'analyse des conséquences est présenté dans un tableau de performances. Pour des raisons de simplification, nous avons choisi que: p (PREF) = q (INDI) = 1.

B.4.6. Résultat du rangement

L'objectif de cette modélisation multicritères du problème de la propagation de la maladie de la tuberculose va aboutir à une classification des foyers d'épidémies de la plus favorable à la moins favorable. Ces zones sont supposées être les plus risquées ou encore les plus « dangereuses épidémiologiquement » pour la tuberculose humaine ce qui permettra d'élaborer une carte de risque tuberculeux. La figure 3.10 montre une carte de risque qui donne une classification des foyers d'épidémies de la région la plus à risque épidémiologique à la région la moins à risque épidémiologique en utilisant une légende. Le résultat de classement a été visualisé sur une carte à l'aide du SIG Map Info Professional 11.0.

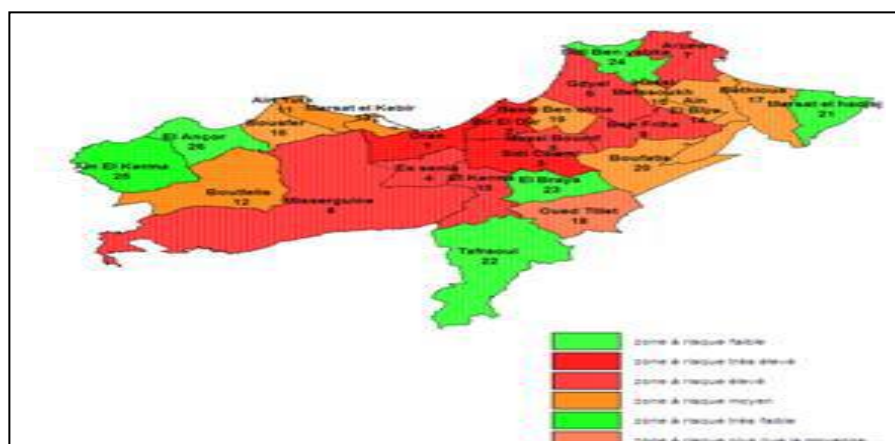


Figure B.5 Carte de risque tuberculeux avec Map Info Professional 11.0

B.4.7. Discussion

Nous pouvons remarquer que le résultat du rangement sur une carte de risque épidémiologique permet de présenter des informations pertinentes aux décideurs en santé publique leur permettant de voir les relations entre les phénomènes, tout en les incitant à découvrir les connaissances et générer, en conséquence, la bonne décision pour agir efficacement dans le temps et dans l'espace. Nous avons initié, à travers cet exemple d'application, un premier modèle prédictif utilisant l'analyse multicritères pour établir une carte de risque épidémiologique aidant les décideurs en santé publique à prendre les dispositifs nécessaires pour éviter un risque pour la santé. L'approche peut se généraliser à divers autres cas d'analyse spatiale impliquant une problématique de décision multicritères.

B.5 Conclusion

Grâce à cette étude de cas, nous avons pu discerner l'efficacité du modèle décisionnel proposé permettant de fournir des informations utiles pour la surveillance, destinées aux décideurs en santé publique, sous la forme d'une carte de risque épidémiologique. Le système décisionnel proposé permet aux épidémiologistes de voir et détecter les relations entre les phénomènes, tout en les encourageant à découvrir et à prendre une décision efficace.

Ce processus de prévision généré par l'intégration du SIG et la méthode Prométhée, constitue un modèle prédictif utilisant une analyse multicritères pour développer une carte de risque. Cette dernière aide les décideurs en santé publique à prendre les dispositifs nécessaires pour éviter un risque pour la santé.

Annexe C

Autocorrélation spatiale

Dans cette annexe, nous présentons les expérimentations que nous avons effectuées pour analyser les différents phénomènes spatiaux étudiés au cours de cette thèse.

C.1 Introduction

La méthodologie générique proposée dans le chapitre 3 présente une panoplie de méthodes et mesures permettant d'explorer les données à référence spatiale en suivant une démarche logique et systémique. Dans ce qui suit, nous reprenons cette démarche structurée en trois phases : d'abord **descriptive** (en appliquant des méthodes de statistiques descriptives) puis **exploratoire** (en analysant les données monothématiques) et en fin **explicative ou prédictive** (sur des données multithématiques) dans un but d'aide à la décision.

C.2 Mesures des statistiques descriptives

Ces mesures représentent principalement les mesures de centralité et de dispersion. Ces derniers permettent de comparer rapidement des nuages de points propres à des segments clés de la population. L'effet discriminant ou non de l'espace peut être observé.

C.2.1 Mesures de centralité

Le centre moyen est le descripteur le plus simple d'une distribution ; c'est l'équivalent de la moyenne, en deux dimensions. Les points colorés sur la carte de la figure C.1 représentent les centres moyens pour chaque attribut considéré dans notre étude de la surveillance épidémiologique.

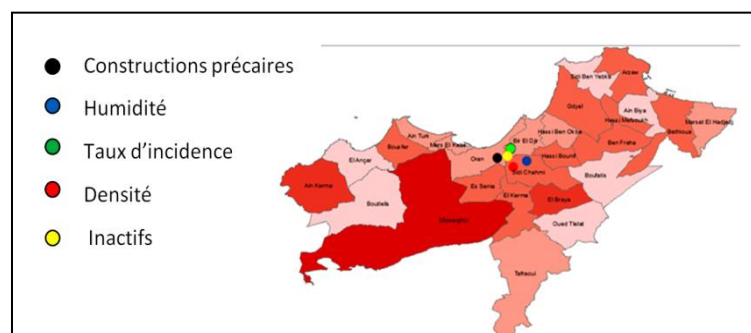


Figure C.1 Centres moyens des facteurs de la tuberculose à Oran en 2009

C.2.2 Mesures de dispersion

C.2.2.1 Ecart type

La première mesure de dispersion consiste à mesurer l'écart type (*Standard Deviation*) dans chacun des axes (x,y), de façon distincte pour les quatre attributs. Les cercles colorés sur la carte de la figure C.2 représentent les écarts type de la distance pour chaque attribut considéré dans notre étude de la surveillance épidémiologique.

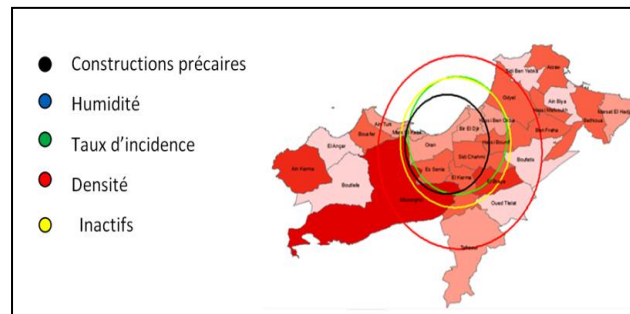


Figure C.2 Ecart type de la distance des facteurs de la tuberculose à Oran 2009

C.2.2.2 Ellipse de dispersion

La dispersion d'un nuage de points peut toutefois présenter certaines tendances spatiales. L'ellipse de dispersion (EDS : *Standard Deviational Ellipse*) permet de tenir compte de cette anisotropie. Selon cette mesure, les écarts types dans les deux directions (x,y) sont orthogonaux et définissent une ellipse.

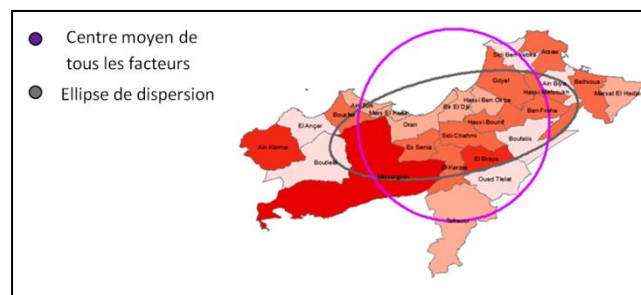


Figure C.3 Ellipse de dispersion des facteurs de la tuberculose à Oran en 2009

C.3 Les Mesures de l'autocorrélation spatiale

La notion d'autocorrélation spatiale mesure essentiellement la ressemblance entre observations voisines. Pour expliquer ce concept, (Asselin, 1999) parle de « *coïncidence de valeur avec la coïncidence de localisation* ».

Avant de commencer notre analyse, nous donnons sur la figure C.4 une représentation géographique de la tuberculose à Oran durant toute l'année 2009.

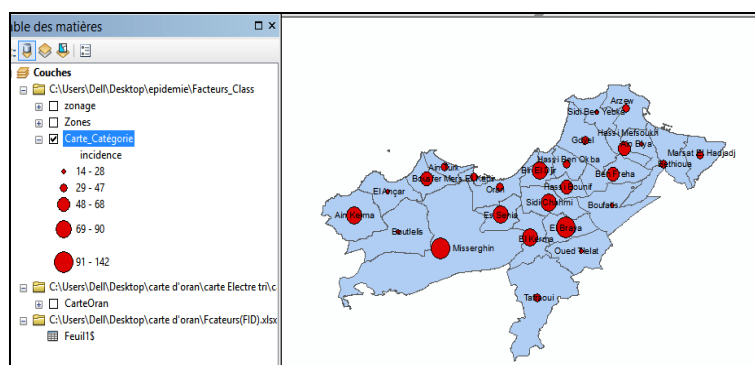


Figure C.4. Représentation géographique de l'incidence de la tuberculose à Oran en 2009

C.3.1 Mesures d'autocorrélation spatiale globale

De multiples statistiques permettent de mesurer l'autocorrélation spatiale. Nous utilisons les deux plus simples étant l'indice de Moran (I) et l'indice de Getis-Ord (G). Ces indices permettent de valider l'existence d'autocorrélation spatiale pour une variable donnée, et ce pour l'ensemble de la zone d'étude.

C.3.1.1 Indice de Moran (I)

L'indice I de Moran se base sur l'étude de similarité et di-similarité des valeurs voisines pour définir si le modèle spatial est clustérisé (ressemblance des valeurs voisines), dispersé (des valeurs voisines opposées) ou alors aléatoire (absence d'autocorrélation spatiale)

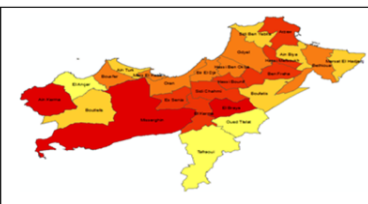
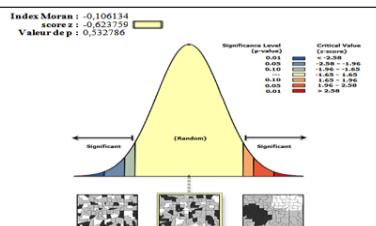
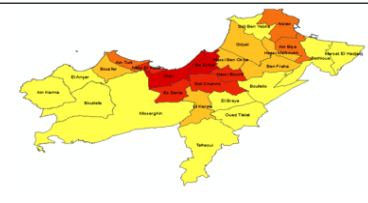
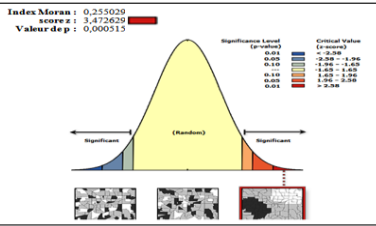
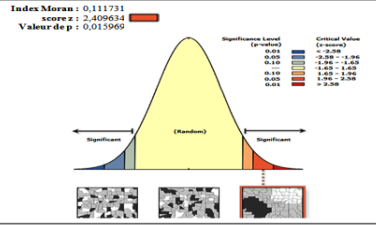
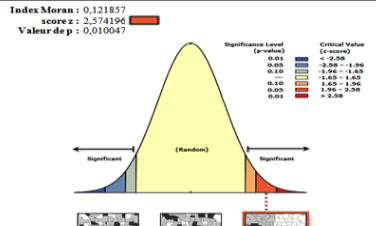
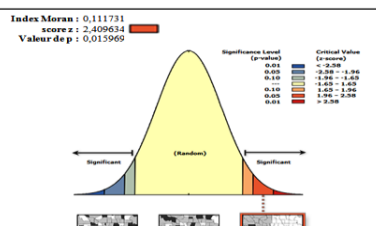
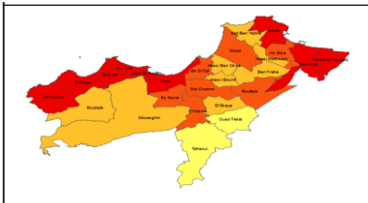
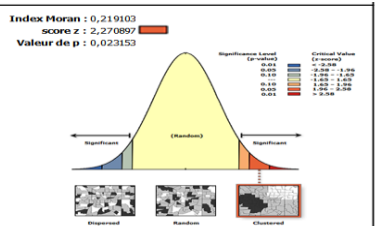
Nous présentons dans le tableau C.1 l'autocorrélation spatiale globale calculée par l'indice de Moran pour l'incidence de la tuberculose à Oran dans l'année 2009.

Tableau C.1 Indice d'autocorrélation spatiale globale (I de Moran) du Taux d'incidence de la tuberculose à Oran en 2009

Année	Moran's I	Z score	P value	Pattern
2009	-0,111722	-0,687827	0,491562	Random

Dans ce qui suit, nous considérons les cinq facteurs de l'étude et nous calculons pour chacun d'entre eux l'indice de Moran permettant de dire si le modèle est clustérisé ou aléatoire. Le tableau C.2 montre la comparaison des mesures de l'autocorrélation spatiale globale calculée par l'indice de Moran (I) pour les cinq facteurs de la tuberculose à Oran.

Tableau C.2 Tableau de comparaison des indices d'autocorrélation spatiale globale calculée par I de Moran, pour les 5 facteurs considérés

Facteurs considérés	Modèle Obtenu après calcul de l'indice de Moran (I)		Type du modèle
Taux d'incidence	 Incidence rate	Index Moran : -0,106134 score z : -0,623759 Valeur de p : 0,532786 	Aléatoire
Densité	 Density	Index Moran : 0,255029 score z : 3,472629 Valeur de p : 0,000515 	Clusterisé
Nombre des inactifs	 Nombre des inactifs	Index Moran : 0,111731 score z : 2,409634 Valeur de p : 0,015969 	Clusterisé
Nombre de constructions précaires	 Nombre de constructions précaires	Index Moran : 0,121857 score z : 2,574196 Valeur de p : 0,010047 	Clusterisé
Nombre d'inactifs	 Nombre des inactifs	Index Moran : 0,111731 score z : 2,409634 Valeur de p : 0,015969 	Clusterisé
Humidité	 Humidity	Index Moran : 0,219103 score z : 2,270897 Valeur de p : 0,023153 	Clusterisé

Nous présentons, dans ce qui suit, l'analyse de l'autocorrélation spatiale globale (Les indices I de Moran) trimestriels et mensuels.

Tableau C. 3 Autocorrélation spatiale globale par trimestre

Trimestre	Moran's I	Z score	P value	Pattern
01	0,116640	1,382678	0,166764	Random
02	0,079151	1,165877	0,243664	Random
03	0,131816	1,520746	0,128324	Random
04	0,031426	0,043347	0,516338	Random

Tableau C.4 Autocorrélation spatiale globale par mois

Mois	Moran's I	Z score	P value	Pattern
01	0,192556	2,052067	0,040163	Clustered
02 ^a	-	-	-	-
03	-0,158386	-1,056879	0,290567	Random
04	0,162858	1,919038	0,054979	Clustered
05	-0,102656	-0,555965	0,578235	Random
06	-0,168144	-1,179566	0,238173	Random
07	0,199760	2,107700	0,035057	Clustered
08	0,282290	2,299866	0,0251456	Clustered
09	-0,024562	0,149468	0,881184	Random
10	0,049953	0,914838	0,360277	Clustered
11	-0,150812	-1,012623	0,311240	Random
12	-0,122284	-0,746621	0,455292	Random

a. Les services de santé ont enregistré 0 cas pour le mois de février

C.3.1.2 Indice G de Getis-Ord

Selon (Levine, 2004), L'indice de Moran est généralement considéré plus robuste et moins sensible aux variations locales. En effet, L'indice I de Moran n'indique pas si les valeurs sont faibles ou élevées.

Pour cela, une autre statistique de l'autocorrélation spatiale globale a été proposée, l'indice General G de (Getis et Ord, 1992) pour définir le rapport fort/faible.

- Si $G(d) < 0$ alors les valeurs groupées sont faibles
- Si $G(d) > 0$ alors les valeurs groupées sont élevées
- Si $G(d) = 0$ alors les valeurs sont proches de la valeur nulle

Tableau C.5 Indice d'autocorrélation globale (G de Getis-Ord) du Taux d'incidence de la tuberculose à Oran en 2009

Année	General G	Z score	P value	Pattern
2009	0,000021	-0,907713	0,364030	Random

C.3.2 Mesures d'autocorrélation spatiale locale

Les mesures d'autocorrélation spatiale, appliquées localement, permettent de cerner les secteurs similaires et dissimilaires. Le caractère distinctif ou non d'un secteur en regard de son voisinage peut faciliter la formulation de modèles, surtout dans le choix d'un découpage spatial. L'identification, a priori, des secteurs différents peut faciliter le travail d'analyse corrélative

C.3.2.1 Indice de LISA(Ii) de Anselin

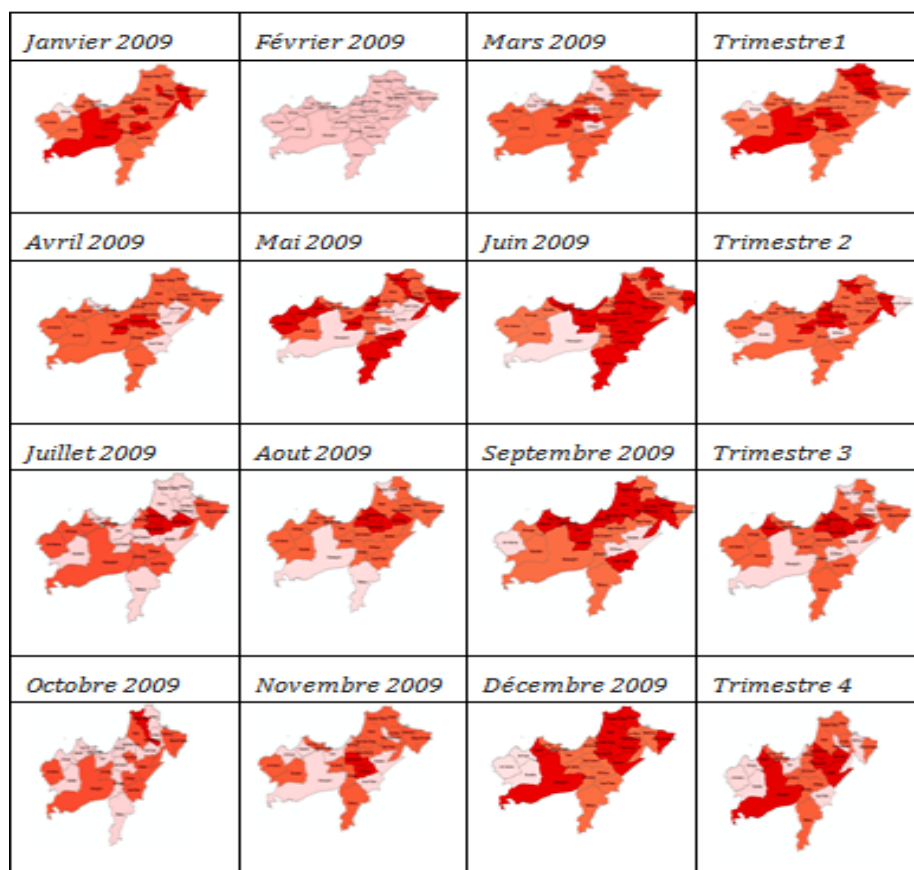


Figure C.5 Autocorrélation spatiale locale de la tuberculose à Oran par mois synthétisé par trimestre calculée par indice de lisa (Ii)

La figure C.6 donne la représentation cartographique de l'indice d'autocorrélation spatiale locale pour différentes études de cas.

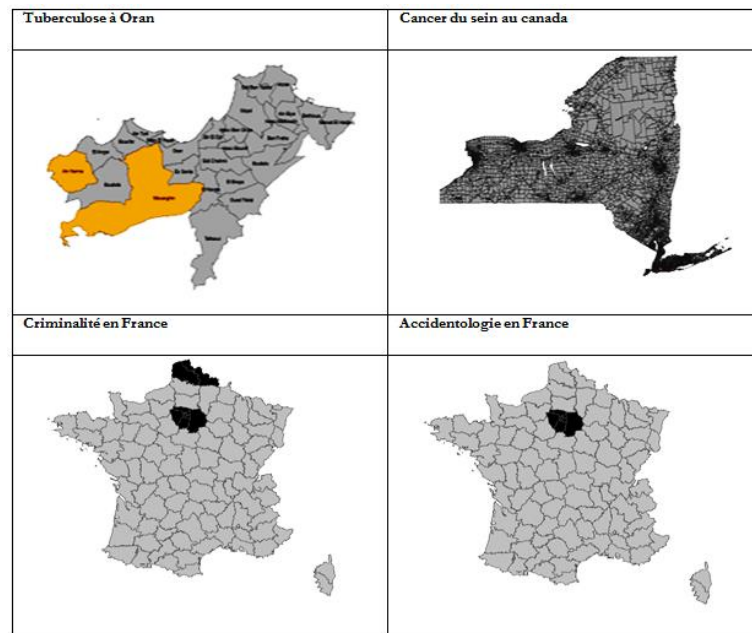


Figure C.6 Analyse des clusters par auto corrélation locale par indice de lisa

La moyenne des indices locaux est égale à l'indice global (Voir le chapitre 3). Ce lien de proportionnalité entre l'indice global de Moran et les indices locaux des LISAs permet d'obtenir pour chaque sous région géographique une estimation de sa ressemblance avec les sous régions voisines par rapport à sa ressemblance à l'ensemble des sous régions. On distingue alors quatre cas de figures :

- **High-High** : Sites avec des valeurs élevées avec des voisins similaires en couleur **noir**
- **High-Low** : Sites avec des valeurs élevées dans un voisinage avec des valeurs faibles (High-Low) en couleur **orange**.
- **Low-High** : Sites avec des valeurs faibles dans un voisinage avec des valeurs élevées (Low-High) en couleur **bleue**.
- **Low-Low** : Sites avec des valeurs faibles dans un voisinage avec des valeurs faibles (Low-Low) en couleur **blanche**.

On s'intéresse à l'identification des zones HH qui représentent une anomalie ou un danger dans les études de cas proposées.

– Surveillance de la tuberculose à Oran(Algérie)

Une analyse plus fine au niveau mois et trimestre pour la surveillance de la tuberculose à Oran utilisant l'indice de LISA est représentée sur la figure C.7.

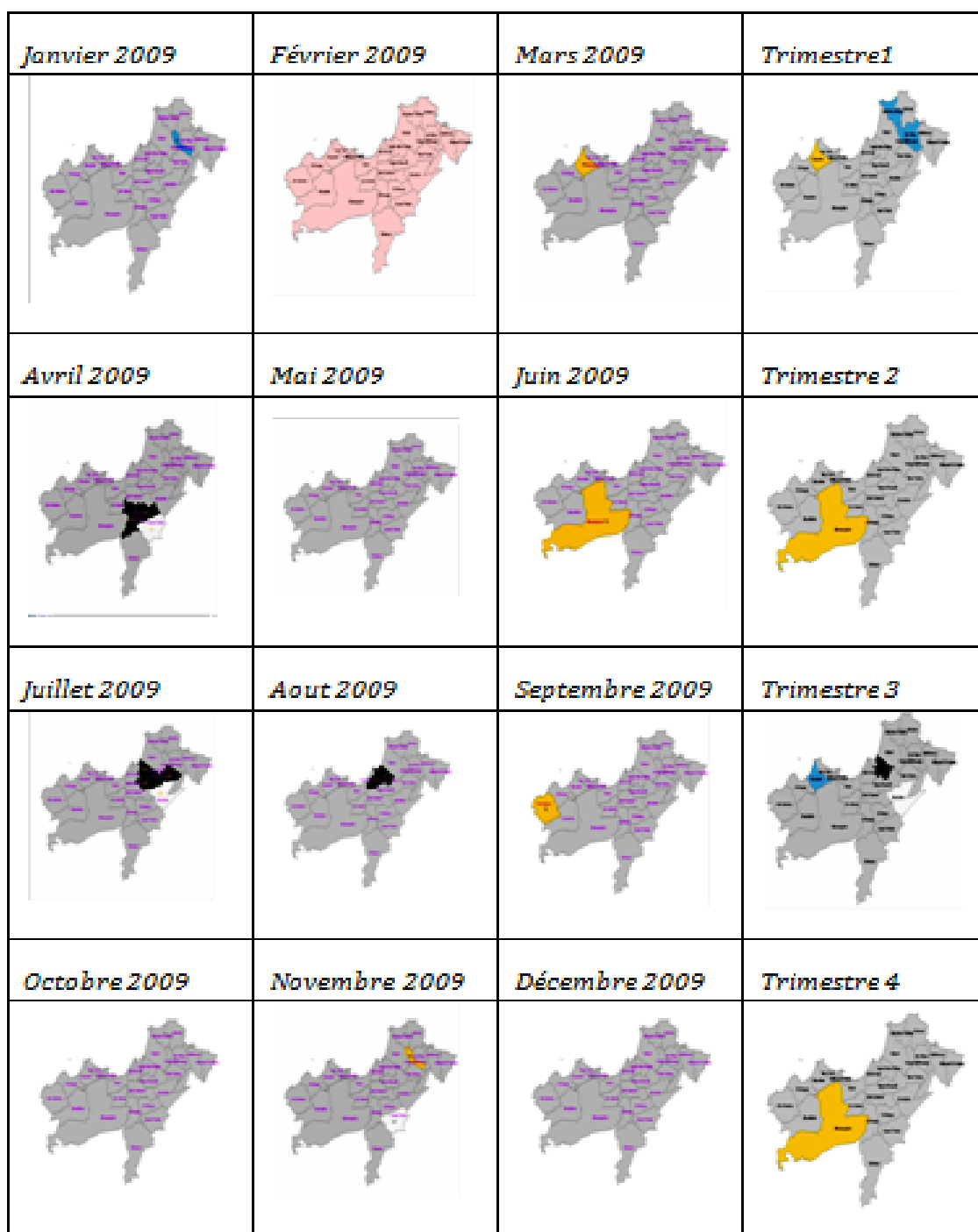


Figure C.7 Autocorrélation spatiale locale de la tuberculose en 2009 à Oran par mois synthétisée par trimestre

– Analyse des accidents de la route en France par années

Une analyse plus fine au niveau année puis mois pour la surveillance des accidents de la route en France utilisant l'indice de LISA est représentée par années et par mois sur les figures C.8 et C.9 respectives.

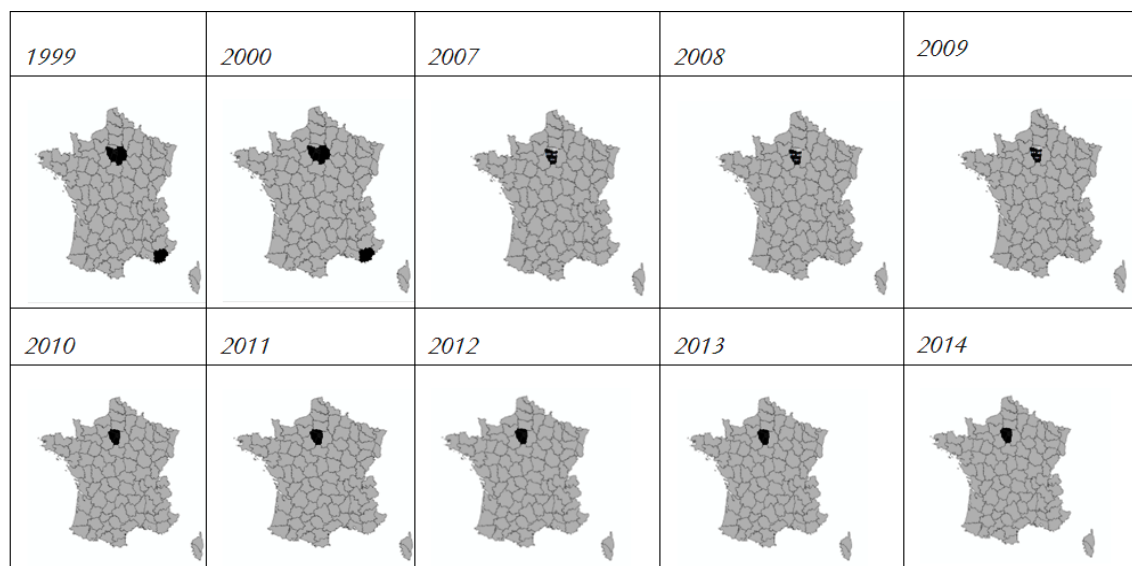


Figure C.8 Autocorrélation locale des accidents de la route en France par année

– Analyse des accidents de la route en France par mois

Une analyse mensuelle plus fine dans l'année 2009 donne les clusters illustrés sur la figure C.8.

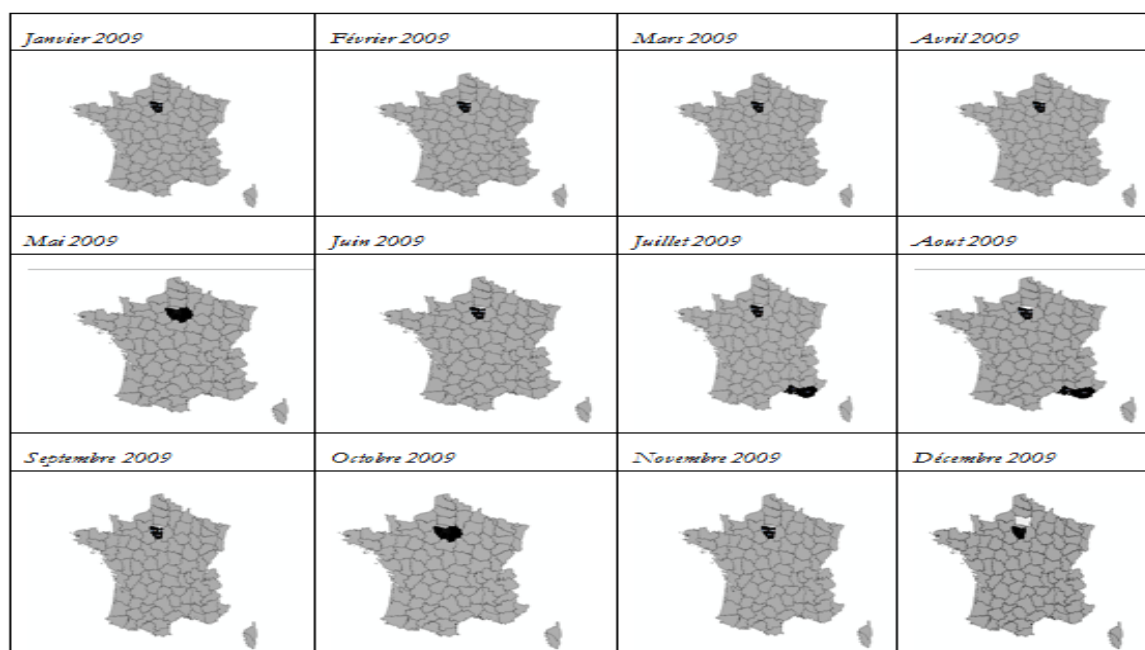


Figure C.9 Autocorrélation locale des accidents de la route en France en 2009 par mois

C.3.2.2 Indice local Gi et Gi* de Getis-Ord

Les statistiques Gi et Gi* nous permettent de détecter les agrégats (les clusters) locaux de dépendance qui peuvent ne pas apparaître lors de l'utilisation des statistiques globales (I, C ou General G). La figure C.8 représente la représentation cartographique de l'indice d'autocorrélation spatiale local calculé par l'indice de Gi de Getis-Ord pour différentes études de cas.

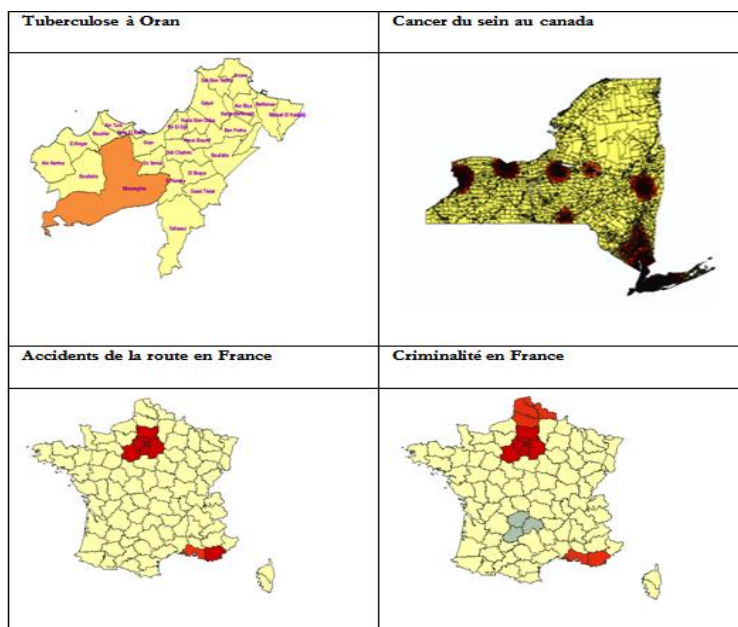


Figure C.10 Analyse de points chauds avec la statistique Gi* de Getis-Ord

– Surveillance de la tuberculose à Oran(Algérie)

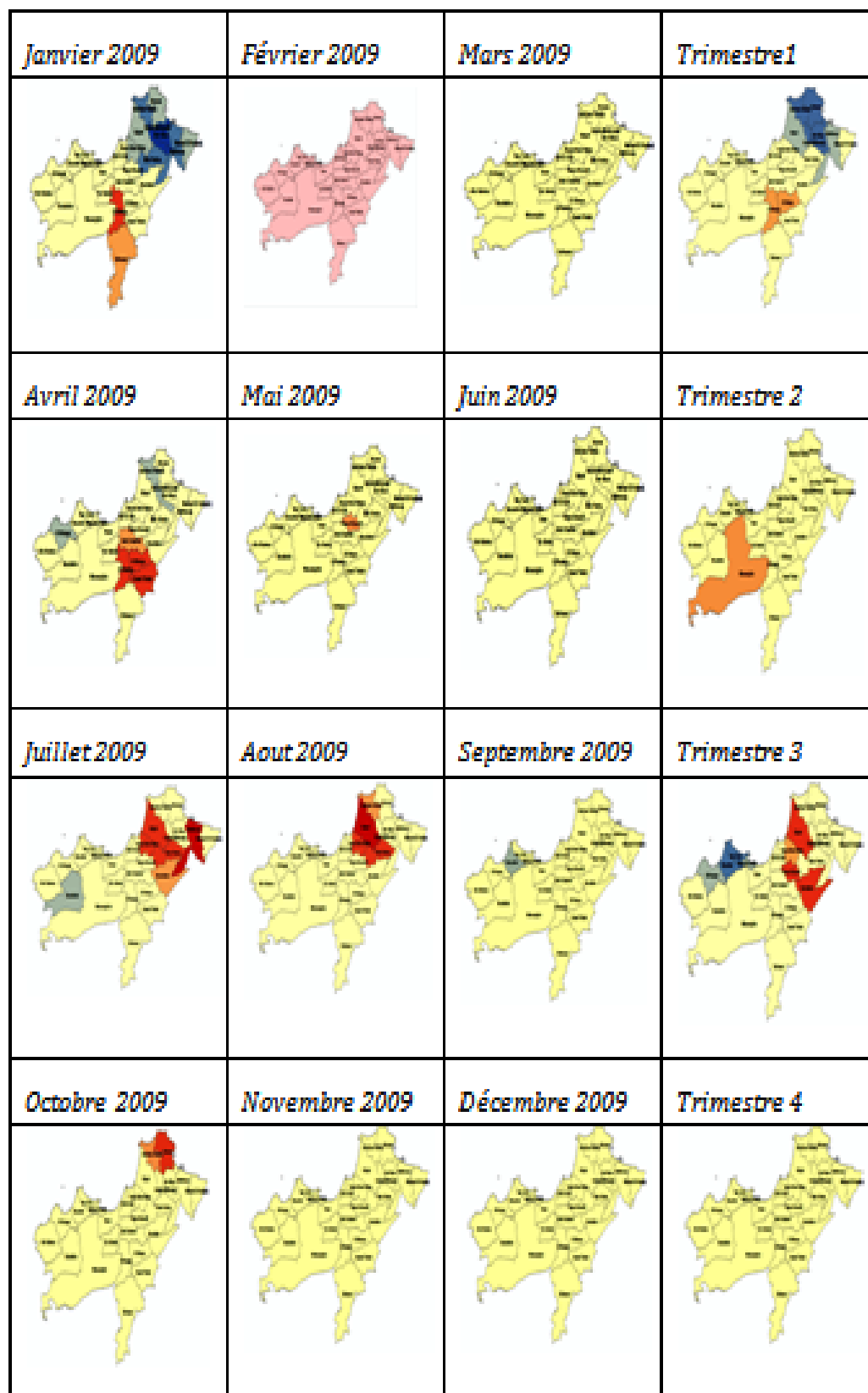


Figure C.11 Analyse de points chauds de la tuberculose à Oran par mois synthétisés par trimestre

– Analyse des accidents de la route en France par années

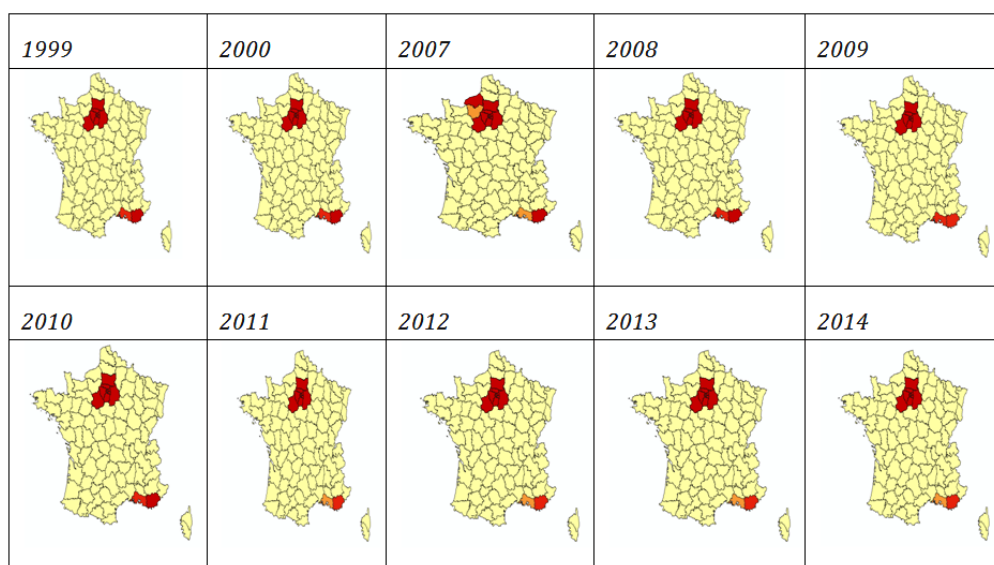


Figure C.12 Analyse de points chauds des accidents de la route en France par année

– Analyse des accidents de la route en France par mois

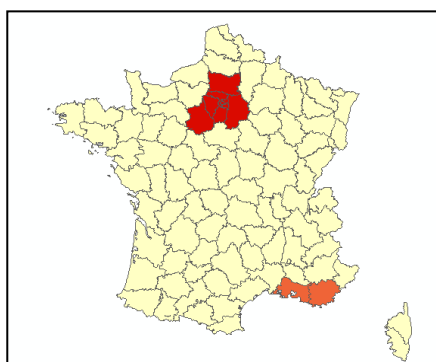


Figure C.13 Analyse de points chauds des accidents de la route en France pendant l'année 2009

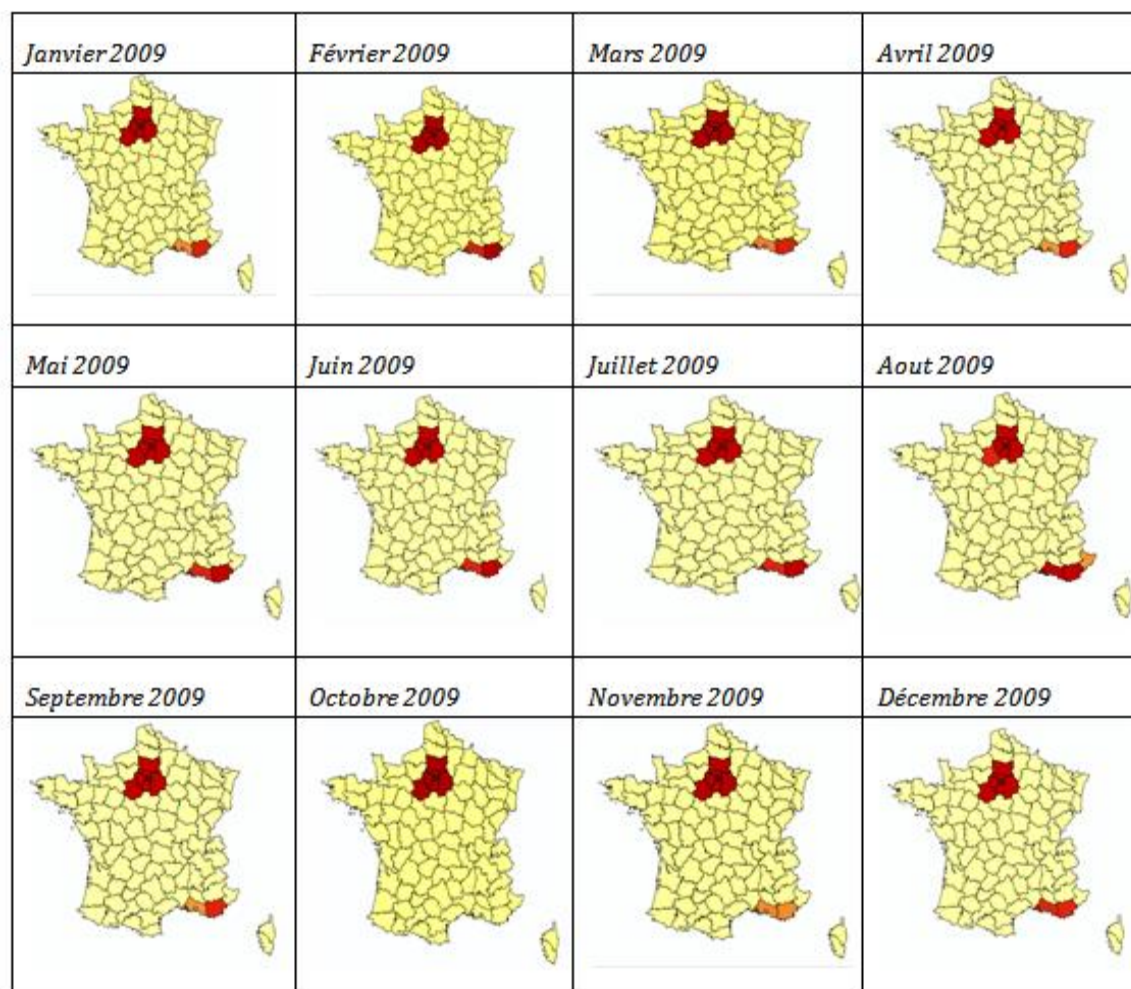


Figure C.14 Analyse de points chauds des accidents de la route en France par mois pendant l'année 2009

– Analyse de la criminalité en France par années

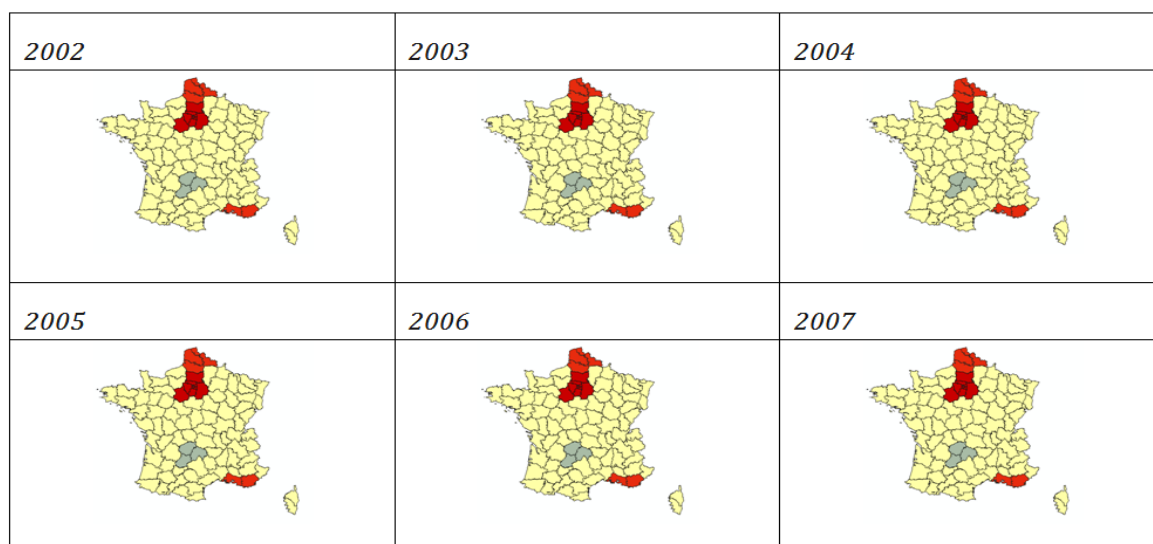


Figure C.15 Analyse de points chauds de la criminalité en France par année

C.3.2 Mesures d'autocorrélation spatiotemporelle locale

C.3.2.1 Autocorrélation temporelle entre trimestres

Par application de la formule de l'indice d'ITA (voir chapitre 3), nous obtenons le tableau C.6

– Surveillance de la tuberculose à Oran(Algérie)

Tableau C.6 Autocorrélation temporelle entre trimestres

Régions par trimestre	R1 (T1-T2)	R2 (T1, T3)	R3 (T1, T4)	R1 (T2-T3)	R2 (T2-T4)	R1 (T3, T4)
01	0,78	0,49	0,21	0,51	0,49	0,23
02	0,75	0,50	0,25	0,48	0,50	0,23
03	0,71	0,50	0,24	0,47	0,50	0,24
04	0,69	0,49	0,29	0,47	0,49	0,27
05	0,54	0,42	0,38	0,34	0,42	0,23
06	0,77	0,50	0,22	0,52	0,50	0,24
07	0,68	0,49	0,30	0,46	0,49	0,27
08	0,76	0,48	0,21	0,58	0,48	0,31
09	0,80	0,41	0,11	0,64	0,41	0,25
10	0,75	0,50	0,25	0,50	0,50	0,25
11	0,73	0,50	0,24	0,49	0,50	0,24
12	0,52	0,42	0,32	0,42	0,42	0,32
13	0,74	0,45	0,12	0,49	0,45	0,16
14	0,74	0,47	0,22	0,58	0,47	0,33
15	0,76	0,50	0,24	0,52	0,50	0,26
16	0,66	0,47	0,31	0,47	0,47	0,29
17	0,63	0,46	0,29	0,34	0,46	0,17
18	0,56	0,43	0,37	0,31	0,43	0,19
19	0,72	0,50	0,24	0,51	0,50	0,27
20	0,69	0,48	0,29	0,40	0,48	0,20
21	- 0,03	0,32	0,16	0,08	0,32	0,25
22	0,73	0,50	0,26	0,50	0,50	0,26
23	0,72	0,50	0,27	0,50	0,50	0,27
24	0,77	0,49	0,21	0,49	0,49	0,21
25	0,71	0,47	0,23	0,55	0,47	0,33
26	0,73	0,50	0,22	0,48	0,50	0,22

Les valeurs de lista calculées entre trimestres par la formule proposée dans le chapitre 3, sont présentées sur le tableau C.7.

Tableau C.7 Autocorrélation spatiotemporelle entre trimestres

	LISTA (T1-T2)	LISTA (T1-T3)	LISTA (T1-T4)	LISTA (T2-T3)	LISTA (T2-T4)	LISTA (T3-T4)
A.El Turck	1,40	0,96	0,40	1,08	0,96	0,58
Ain El Bia	2,10	1,44	0,83	1,32	1,44	0,67
Ain El	2,60	1,76	0,87	1,73	1,76	0,87
Arzew	2,45	1,63	0,81	1,52	1,63	0,70

Benfreha	6,84	4,64	2,59	4,54	4,64	2,38
Bethioua	3,12	2,12	1,20	2,07	2,12	1,09
Bir El Djir	3,32	2,36	1,53	2,25	2,36	1,31
Boufatis	2,04	1,45	0,92	1,44	1,45	0,87
Bousfer	2,00	1,45	0,68	1,36	1,45	0,73
Boutlelis	1,21	0,94	0,46	0,84	0,94	0,48
El Ancor	1,61	1,10	0,48	1,17	1,10	0,60
El Braya	6,94	4,86	2,33	5,01	4,86	2,80
El kerma	4,16	3,00	1,54	3,02	3,00	1,77
Es senia	2,28	1,67	0,77	1,65	1,67	0,92
Gdyel	2,69	1,91	1,25	1,61	1,91	0,83
H.Bounif	3,28	2,29	1,48	2,26	2,29	1,31
H.Mefsoukh	4,25	2,92	1,66	2,72	2,92	1,39
Hassi Ben	2,05	1,45	0,92	1,39	1,45	0,80
M El Kebir	1,43	1,05	0,52	0,99	1,05	0,56
M. Hadjaj	0,66	0,42	0,18	0,45	0,42	0,21
Misserghin	11,80	7,40	3,21	8,30	7,40	3,90
O.Tlelat	1,50	1,02	0,50	1,02	1,02	0,50
Oran	2,06	1,51	0,78	1,54	1,51	0,94
S.B.yebka	1,38	0,96	0,55	0,89	0,96	0,47
S.Chami	7,46	5,00	2,60	5,39	5,00	2,96
Tafraoui	1,43	0,94	0,42	0,96	0,94	0,44

C.4 Mesures de classification automatique (Clustering)

Les mesures de classification automatique permettent d'abord d'identifier des zones de forte concentration d'observations puis de raffiner cette identification par la prise en compte d'une structure spatiale de référence afin de n'identifier que les zones où l'occurrence dépasse celle qui serait normalement attendue étant donnée cette structure spatiale. La figure C.16 représente un exemple de regroupement par K means, des zones de la région d'Oran par rapport à l'incidence de la tuberculose dans ces zones et selon le nombre de groupes k.

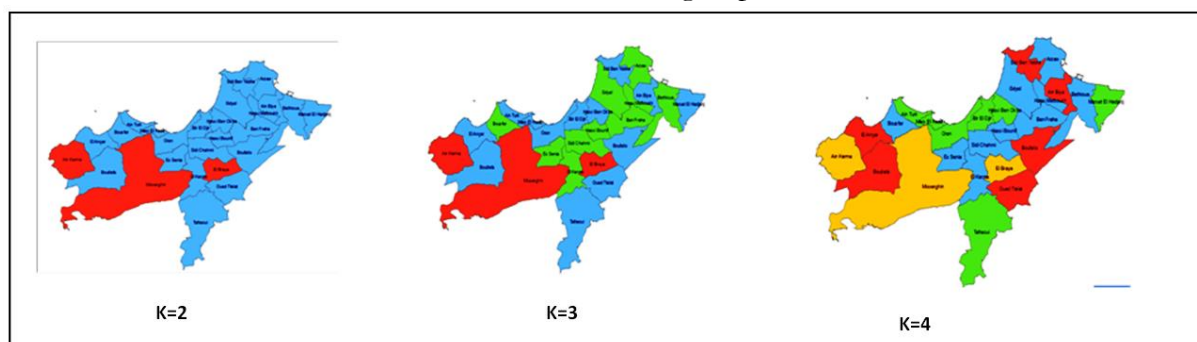


Figure C.16 Analyse de regroupement en 2, 3 et 4 groupes distincts

C.5 Conclusion

Les méthodes expérimentées dans cette annexe suivent la démarche systémique proposée dans le chapitre 3. L'application de ces méthodes permet de concrétiser la logique exploratoire de la méthodologie générique proposée.

Ces méthodes spatiales, intégrées dans une démarche exploratoire, ont la capacité d'aller progressivement dans l'analyse des données spatiales et découvrir au fur et à mesure les connaissances utiles pour répondre aux objectifs des utilisateurs.

Liste des références

- (Agadir et al, 2011) Agadir, F., Ali Pacha, S., Anane, T., Baough, L., Benkara, A., Boulahbal, F., Chaulet, P., Halassa, S.A., Hannoun, D., Haouichat, H., L'Hadj, M., Larbaoui, D., Nafti, S., Ouagani, D., Yala, D., Zidouni, N. *Manuel de la lutte antituberculeuse à l'usage des personnels médicaux*. Rapport du ministère de la santé de la population et de la réforme hospitalière. 2011.
- (Amine, 2013) Amine, A. *Data Mining*. Professeur de l'université Tahar Moulay, Saida. Présentation à l'école d'hiver sur les applications de l'informatique industrielle, réseaux et génie logiciel, 09-12 Décembre 2013.
- (Atmani, 2008) Atmani. B. *Extraction et Gestion de Connaissances par Apprentissage Automatique à partir des données empiriques*. Support de Cours « Apprentissage Automatique » destiné aux étudiants de Magister Diagnostic, Aide à la décision et interaction homme machine, 2009.
- (AnandR, 1995) AnandR. P. *Multi-Criteria Methods in River Basin Planning: a Case study*, Water Science and Technology Journal, Volume 31, pp.261-272, 1995.
- (Alatrasta, 2013) Alatrasta H.S. *Extraction de relations spatio-temporelles à partir des données environnementales et de la santé*. Thèse de doctorat. UMR TETIS - Territoires, Environnement, Télédétection et Information Spatiale. Université de Montpellier II, 2013.
- (Alexander et al ,2007) Alexander. H et Hans-Henning. G, *Fast clustering based on kernel density estimation*. In proceedings of the International symposium on intelligent data analysis, pp. 70-80. Springer, 2007.
- (Amanzougarene, 2014) Amenzougarene. F. *Extension du modèle multidimensionnel aux faits qualitatifs. Application à l'analyse en ligne des gènes des chantiers urbains*. Thèse de doctorat de l'université de Versailles Saint Quentin en Yvelines, 2014.
- (Andrew et Dean, 1976) Andrew G et, Dean A. G. *Population-based spot maps: an epidemiologic technique*. In the American journal of public health, Volume 66(10), pp. 988-989, 1976.
- (Andrienko et al, 2004) Andrienko. G, Andreinko. N and Gatalaky. P. *Visual Mining of Spatial Time Series Data*. In Proceedings of the conference of Knowledge Discovery in Databases, pp. 524-52. 2004.
- (Anil, 2010) Anil K.J : *Data clustering: 50 years beyond k-means*. Journal of Pattern Recognition Letters, Elsevier. Volume 31, Issue 8, pp. 651-666. 1 June 2010.
- (Ankerst et al, 1999) Ankerst. M, Marku. M.B, Kriegel H.P et Sander.J. *OPTICS : ordering points to identify the clustering structure*. In Proceedings of ACM Sigmod Record, pp. 49-60, 1999.

Liste de références

- (Anselin, 1990)** Anselin, L. *Spatial dependence and spatial structural instability in applied regression analysis*. Journal of Regional Science. Volume 30, pp. 185-207, 1990.
- (Anselin, 1995)** Anselin, L. *Local indicators of spatial association - LISA*. The Geographical Analysis journal, Volume 27, Issue 2, pp. 93-115, 1995.
- (Anselin, 1999)** Anselin, L. *Spatial Econometrics*. Book chapter 14 of the Companion to Theoretical Econometrics, pp. 310-330, 1999.
- (Ansoff et Anthonio, 2004)** Ansoff H. I et Anthonio. P.H. *Strategic Management of Technology*. In journal of Technology Analysis and Strategic Management Journal. Volume 16, Issue 2, pp. X-X. 2004.
- (Argyris et Schön, 1978)** Argyris. C and Schön. D. A., *Organizational learning: a theory of action perspective*. In Addison-Wesley Publishing Company, 1978.
- (Asha et al, 2010)** Asha,T., Natarajan, S., et Murthy, K.N.B. *Data Mining Techniques in the Diagnosis of Tuberculosis*. In proceedings of Diagnosis and knowledge extraction, Artificial, ACM Bangalore Conference, pp.1-6, Bangalore, India, 2010.
- (Attonaty et al, 1991)** Attonaty. J. M., Chatelin. M. H., Poussin. J. C and Soler. L. G., *Advice and decision support systems in agriculture : new issues*. In Proceedings of Decision Support Systems conference, Bruges, Belgium, 1991.
- (Beale et al, 2008)** Beale, L., Abellan, J. J., Hodgson, S., et Jarup, L. *Methodologic issues and approaches to spatial epidemiology*. Environmental health perspectives journal, Volume 116(8), pp. 1105-1110, 2008.
- (Bédard et al, 2005)** Bédard. Y, Proulx M.J, Rivest. S, *Enrichissement de l'OLAP pour l'analyse géographique : exemples de réalisations et différentes possibilités technologiques*. In Proceedings of the conference « les entrepôts de données et analyse en ligne » , Lyon 2005.
- (Bedard, 1997)** Bédard. Y. *Spatial OLAP*. 2eme Forum annuel sur la R-D, GéomatiqueVI : Un monde accessible, Montréal, 1997.
- (Bédard, 2004)** Bédard. Y. *Amélioration des capacités décisionnelles des SIG par l'ajout d'un module SOLAP*. Keynotes à l'Université de Provence, Centre de Mathématiques et Informatique, LSIS, Marseille, 8 avril 2004.
- (Ben et Mandl, 2003)** Reis B.Y and Mandl.K.D *Time Series modeling for syndromic surveillance*. Medical Informatics and Decision Making Journal , Volume 3 , Issue 2, pp. 1472-6947, 2003.
- (Ben Mena, 2000)** Ben Mena. S. *Introduction aux méthodes multicritères d'aide à la décision*. Journal of Biotechnol Agron Society Environment, Volume 4, Issue 2, pp 83-93, 2000.
- (Ben Mena, 2001)** Ben Mena. S. *Une solution informatisée à l'analyse de sensibilité d'Electre III*, journal of Biotechnol Agron. Society.Environment, Volume. 5 (1), pp. 31-35, 2001.

- (Bernier et al, 2007)** Bernier.E, Bédard. Y, Thierry Badard, Frédéric Hubert. *UMapIT (Unrestricted Mapping Interactive Tool) : Merging the datacube paradigm with an occurrence-based approach to support on-demand web mapping*. From the book of International Perspectives on Maps and the Internet. Chapter 13. pp.187-204, 2007.
- (Besalloua et Hamdadou, 2018)** Bensalloua A.C, Hamdadou. D. *Users Integrity Constraints in SOLAP Systems. Application in Agroforestry*. International Journal of Interactive Multimedia and Artificial Intelligence, Volume. 5, N° 1, pp 47-56, 2018.
- (Bimonte et al. 2007)** Bimonte. S, Tchounikine A. and Miquel. M. *Spatial OLAP : Open Issues and a Web Based Prototype*. In Proceedings of 10th AGILE International Conference on Geographic Information Science, Aalborg University, Denmark, 2007.
- (Bimonte, 2007)** Bimonte, S. *Intégration de l'information géographique dans les entrepôts de données et l'analyse en ligne : de la modélisation à la visualisation*. Thèse de doctorat à l'institut national des sciences appliquées de Lyon, 2007.
- (Bimonte, 2010)** Bimonte. S, Tchounikine A., M. Miquel, Pinet. F. *When Spatial Analysis Meets OLAP : Multidimensional Model and Operators*. International Journal of DataWarehousing and Mining, Volume 6(4), pp. 33-60, 2010.
- (Bouamrane et al, 2012)** Bouamrane. K., Hamdadou, D., Yachba. Kh. Guenachi. Kh. *Towards a decision support system, application: itinerary road modification for road transport of hazardous materials*. International Journal of Decision Sciences, Risk and Management, Volume. 4, No.3/4, pp. 175 – 196, 2012.
- (Boulema, 2006)** Boulema. C. *Eléments de proposition à la mise en place d'une base de données urbaine dans une collectivité locale*. In Proceedings of the conference AUGC, pp. 679–696, 2006
- (Bourgine et Lemoigne, 1990)** Bourguine. P and Lemoigne. J. L. *Les bonnes décisions sont-elles optimales adéquates ?* In the Proceedings of XII congrès européen de recherche opérationnelle, Athènes, Grèce, 1990.
- (Brans et Mareschal, 2001)** Brans. J.P. et Mareschal. B. *Prométhée-Gaia : Une méthodologie d'aide à la décision en présence de critères multiples*. In the journal of Statistiques et mathématiques appliquées, Edition de l'université de Bruxelles, pp 192, Belgique, 2001.
- (Chaikaew et al, 2009)** Chaikaew. N, Tripathi. N.K, Souris M. *Exploring spatial patterns and hotspots of diarrhea in Chiang Mai. Thailand*. International journal of Health Geographics, volume 8, pp. 8-36, 2009.
- (Chakhar et al, 2005)** Chakhar. S., et Pusceddu. C., *Un processus pour la prise de décision spatiale*. In Proceedings of ROADEF'05 conference, pp. X-X. Tours, France, 2005
- (Chaudhuri, 1998)** Chaudhuri S. *Data Mining and Database Systems : Where is the Intersection ?*. In Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Volume 21(1), pp. 4-8, 1998.

Liste de références

- (Chelghoum, 2004)** Zeitouni, K., Chelghoum, N., *Un problème de fouille de données Multi tables*. Thèse de doctorat à l'université de Versailles Saint-Quentin-en-Yvelines, 2004.
- (Cheng, 2011)** Cheng T, Haworth J, Wang J. *Spatio-temporal autocorrelation of road network data*. In the journal of Geographical Systems, Volume 14, Issue 4, pp. 389-413, 2011.
- (Clif et Ord, 1981)** Cliff, A. D., and J. K. Ord. *Spatial Promsm: Models and Applications*. London: Pion Press, 1981.
- (Corrente et al, 2016)** Corrente, S., Greco, S., et Slowinski, R. *Multiple criteria hierarchy process for ELECTRE Tri methods*. The European Journal of Operational Research, Volume 252(1), pp. 191-203. 2016.
- (Courbon, 1993)** Courbon. J. C. *Systèmes Interactifs d'Aide à la Décision*. In Sfez L. (Ed.), «Dictionnaire Critique de la Communication , PUF, Paris. France, 1993.
- (Cyert et March, 1963)** Cyert. R. M., and March. J. G.,. *A Behavioural Theory of the Firm*. New Jersey: Prentice-Hall, 1963.
- (Daminani et Spaccapietra, 2006)** Damiani. M.L et Spaccapietra. S. *Spatial Data Warehouse Modelling*. Book chapter 1 of Processing and Managing Complex Data for Decision Support edited by Idea Group Publishing, 2006.
- (Dufour et Hendriks, 2007)** Dufour, B., et Hendriks, P. *Surveillance épidémiologique en santé animale*. Book, 2nd edition Edited by AEEMA et Quae, editions, Paris, France, 2007.
- (Duvallet, 2001)** Duvallet. C. *Des systèmes temps réel et distribuées : modélisation par agents*. Thèse de Doctorat de l'université du Havre, 2001.
- (Ester et al, 1996)** Ester. M, Kriegel. H.P, Sander.J, et Hu. X. *A density-based algorithm for discovering clusters in large spatial databases with noise*. In Proceedings of KDD conference, pp. 226-231, 1996.
- (Ester et al, 1997)** Ester M., Kriegel H.P., Sander J., *Spatial Data Mining: A Database Approach*, in proceedings of the 5th Symposium on Spatial Databases, Berlin, Germany, 1997.
- (Ester et al, 1998)** Ester M., Frommelt A., Kriegel H.P., Sander J., *Algorithms for Characterization and Trend Detection in Spatial Databases*. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, pp. 44-50. New York, 1998.
- (Evans, 1984)** Evans, G., *An overview of techniques for solving multiobjective mathematical programs*. Journal of Management Sciences, Volume. 30, 11, pp. 1268-1282, 1984.
- (Fontanet, 2014)** Fantanet. A. *Concepts et méthodes en épidémiologie*. MOOC d'épidémiologie du Pr. A Fontanet @lecnam @institutpasteur. Hébergé sur @funmooc depuis 2014.

Liste de références

- (Ferrahi et al, 2016) Ferrahi.I, Bimonte. S, et Boukhalifa.K. *Conception logique et physique des hiérarchies spatiales non strictes dans les entrepôts de données spatiales*. In proceedings of SAGEO Nice, pp ????, France. 6-9 décembre 2016.
- (Franklin 1992) Franklin, C. *An Introduction to Geographic Information Systems: Linking Maps to Databases*, Journal of Database, Volume 15, Issue 2, pp. 13-21, April 1992.
- (Geary,1954) Geary, R.C. *The contiguity ratio and statistical mapping*. Journal The incorporated Statistician, volume.5, n°3, pp.115–145. 1954.
- (Getis et Ord, 1992) Getis, A. Ord, K.J., *The analysis of spatial association by use of distance statistics*, Geographical analysis Journal, Volume 24, No. 3, pp. 189-206, 1992.
- (Getis, 1991) Getis, A. *Spatial Interaction and Spatial Autocorrelation: A Cross-Product Approach*. Journal of Environment and Planning , Volume. 23, pp. 1269-77, 1991.
- (Getoor, 2003) Getoor, L. *Understanding tuberculosis epidemiology using structured statistical models*. Journal of Artificial Intelligence in Medicine , Volume 30. pp, 233–256, 2003.
- (Gholamhosein et al , 1998) Gholamhosein.S, Surojit C.J and Aidong. Z *Wavecluster : A multi-resolution clustering approach for very large spatial databases*. VLDB journal, Volume 98, pp. 428-439, 1998.
- (Ginting, 2000) Ginting R. *Intégration du système d'aide à la décision multicritères et du système d'intelligence économique dans l'ère concurrentielle*, Thèse de doctorat de l'université de droit et des sciences d'Aix Marseille, 2000.
- (Glossaire, 2004) *Glossaire du Guide Canadien des incidences sur la santé* – Santé Canada, 2004.
- (Gatrell et al. 1996) Gatrell A., Bailey T., Diggle P., Rowlingson B. *Spatial point pattern analysis and its application in geographical epidemiology*, Journal of Transactions of the Institute of British Geographers, Volume 21, pp. 256-274, 1996.
- (Griffith et Arnrhein, 1991) Griffith D. A et Arnrhein C. G. *Statistical Analysis for Geographers*. A book published by Pearson; Édition : Facsimile, 1 novembre 1990.
- (Guernier, 2006) Guernier V. *Combiner analyse spatiale et épidémiologie pour l'aide à la décision dans la lutte contre la tuberculose en Guyane française*. Thèse de doctorat l'université d'Orléans, 2006.
- (Guran et al, 1998) Guran M., Mehanna A., Hussein B. *On-Line Analytical Mining for Advanced Business Intelligence Solution*. In Scientific Bulletin, Series C, Volume 71, Issue 2, pp. 75-90, 2009.
- (Hall, 1984) Hall. R. I., *The natural logic management policy making its implications for the survival of an organisation*. Journal of Management Science, Volume. 30, pp. 905-927, 1984.

- (Hamdadou et Bouamrane, 2011) Hamdadou. D et Bouamrane. K. *A Multi Criteria Group Decision Support System for Industrial Diagnosis*. Journal of Computer Science, INFOCOMP, Volume 10, Issue 3, pp. 12-24, 2011.
- (Hamdadou Bouamrane, 2014) Hamdadou. D, Bouamrane. K, Merouane. A. *Un système interactif d'aide multicritères à la décision pour le diagnostic industriel*. In proceedings of the International Conference on engineering of Industrial Safety and Environment, ICISE'14,. Published in the International journal of Decision Sciences, Risk and Management, 2014.
- (Hamdadou Bouamrane, 2016) Hamdadou.D, Bouamrane.K. *A Spatial Group Decision Support System: Coupling Negotiation and Multicriteria Approaches*. Intelligent Decision Technologies Journal, Volume. 10, Issue 2, pp. 129-147, 2016.
- (Hamdadou al., 2007) Hamdadou, D. Labed.K et Beldjilali.B. *Proposal for a Decision-making process in Regional planning: GIS, Multicriterion Approach, Choquet's Integral and Genetic Algorithms*. In proceedings of the conference ERIMA07, European Research on Innovation Management Alliance, pp. 51-59. Biarritz, France, 2007.
- (Hamdadou, 2008) Hamdadou. D. *Un système d'aide à la décision en Aménagement du Territoire : une approche Multicritère et une approche de négociation*. Thèse de doctorat, Université d'Oran, 2008.
- (Hamdadou et al, 2017) Hamdadou D, Bouamrane K et Naoui O. *Méthodologie Multicritères d'Aide à la Décision Cours et Exercices*. Polycopié de Cours destiné aux étudiants de Master Informatique, Université Oran 1, 2017.
- (Han et Fu, 1995) Han, J. and Yongjian F, Y. *Exploration of the power of attribute-oriented induction in data mining*. Journal of Advances in Knowledge Discovery and Data Mining, Volume 16, pp. 399-421, 1995.
- (Han et al, 1992) Han.J, Cai.Y et Cercone. N. *Knowledge Discovery in Databases: An Attribute-Oriented Approach*. In proceedings of the 18th VLDB Conference, pp. 547-559, Vancouver, British Columbia, Canada, 1992
- (Han et al, 1997) Han J., Koperski K., and Stefanovic N. : *GeoMiner : A system Prototype for Spatial Data Mining*. In proceedings of the international conference on Management of Data (SIGMOD'97) Record (ACM Special Interest Group on Management of Data), Volume 26, Issue 2, pp. 553-556 Tucson, Arizona, System prototype demonstration, 1997.
- (Han et al, 1998) Han J., Chee S., Chiang J. *Issues for On-Line Analytical Mining of Data Warehouses*. Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'1998), Seattle, Washington, USA, 1998.
- (Han, 1994) Han.J. *Towards efficient induction mechanisms in database systems*. Journal of Theoretical Computer Science, Volume 133, Issue 2, pp. 361-385, 1994.
- (Han, 1997) Han J. *OLAP Mining: An Integration of OLAP with Data Mining*. Book chapter of Data Mining and Reverse Engineering, Springer Book Archive, pp. 3-20, 1997.

- (Han, 1998) Han, J. *Towards on-line analytical mining in large databases*. In ACM Sigmod Record journal, Volume 27, Issue1, pp. 97-107, 1998.
- (Han et al, 2012) Han,J, Kamber. M, et Pei. J. *Data Mining. Concepts and Techniques*. Third Edition of the Book Morgan Kaufmann series in Data Management Systems, 2012.
- (Hanafi-Bojd et al., 2012) Hanafi-Bojd, A., Vatandoost, H., Oshaghi, M., Charrahy, Z., Haghdoost, A., Zamani, G., Abedi, F., Sedaghat, M., Soltani, M. et Shahi, M. *Spatial analysis and mapping of malaria risk in an endemic area, south of iran : a GOSg based decision making for planning of control*. Journal of Acta tropica, Volume 122, Issue1, pp. 132-137, 2012.
- (Hardisty et Klippel, 2010) Hardisty F, Klippel A. *Analysing spatio-temporal autocorrelation with LISTA-Viz*. In the international journal of Geogr Inf. Sci. Volume 24, Issue 10, pp. 1515-1526, 2010.
- (Horvitz, 1989) Horvitz. E., *Reasoning about beliefs and actions under computational resource constraints*. Journal of Uncertainty in Artificial Intelligence. Volume. 3, pp. 301-324, Elsevier, 1989.
- (Jayet, 2001) Jayet H. *Économétrie et données spatiales : une introduction à la pratique*. Journal of Cahiers d'Economie et Sociologie Rurales, Volume 58, pp. 105-129, 2001.
- (Jia et al, 2008) Jia ZW, Jia XW, Liu YX, Dye C, Chen F, Chen CS, Zhang WY, Li XW, Cao WC, Liu HL. *Spatial analysis of tuberculosis cases in migrants and permanent residents, Beijing, 2000-2006*. Journal of Emerg Infect Dis, Volume 14, Issue 9, pp. 1413-1419, 2008.
- (Joerin, 1998). Joerin. F. *Décider sur le territoire : Proposition d'une approche par l'utilisation de SIG et de AMC*. Thèse de doctorat, Ecole Polytechnique Federale de Lausanne. Suisse, 1997.
- (Kaufman et Rousseeuw, 1987) Kaufman.L et Rousseeuw. P. *Clustering by Means of Medoids*. In Reports of the Faculty of Mathematics and Informatics. Faculty of Mathematics and Informatics, 1987.
- (Kaufman et Rousseeuw, 1990) Kaufman. L and Rousseeuw. P.J. *Finding Groups in Data : an Introduction to Cluster Analysis*. Book published by John Wiley et Sons, 1990.
- (Keen et al, 1978) Keen P, Scott-Morton M. *Decision Support Systems : an organizational perspective*. Book published by Addison- Wesley Publishing, 1978.
- (Klen, 1997) Klein M. *Capturing Geometry Rationale for Collaborative Design*. In Proceedings of the Sixth IEEE, Workshop on Enabling Technologies : Infrastructure for Collaborative Enterprises (WET ICE'97), pp. 24-28, Cambridge, MA, USA, USA, 18-20 june 1997.

Liste de références

- (Knorr et Ng, 1996). Knorr. E. M. and Ng. R. *Extraction of spatial proximity patterns by concept generalization*. In proceedings of the 2nd international conference on Knowledge Discovery and Data Mining (KDD'96), pp. 347-350, Portland, OR, USA, August 1996.
- (Koperski et Han, 1995) Koperski K. and Han J. *Discovery of Spatial Association Rules in Geographic Information Databases*, In proceedings of Advances in Spatial Databases, SSD'95, pp. 47-66, Portland, ME, August, 1995.
- (Koperski et al, 1998) Koperski K., Han J., Stefanovic N., *An Efficient Two-Step Method for Classification of Spatial Data*, In proceedings of International Symposium on Spatial Data Handling, SDH'98, pp. 45-54, Vancouver, Canada, July 1998.
- (Kulldorff, 2005) Kulldorff. M. *SaTScanTMv6.0 : Software for the spatial and space-time scan statistics*. Journal of Information Management Services (computer program). Version 2.1 Bethesda, MD : National Cancer Institute, 2005.
- (Kumar et al, 2012) Kumar, P., Angwal, D. et N. Puri. *Diagnosis of Tuberculosis using association rule method*. Journal of Information and Operations Management. Volume 3, Issue 1, pp. 133-135. 2012.
- (Laaribi, 1999) Laaribi. A. *SIG et analyse multicritère*. Edition Hermes, Science Publications, Paris, 2000.
- (Lalvani et al, 2001) Lalvani A1, Pathan AA, McShane H, Wilkinson RJ, Latif M, Conlon CP, Pasvol G, Hill AV. *Rapid detection of Mycobacterium tuberculosis infection by enumeration of antigen-specific T cells*. Journal RespirCrit Care Med. Volume 163, Issue 4, pp. 824-8, 2001.
- (Lawson, 2013) Lawson, A. B. *Statistical methods in spatial epidemiology*. John Wiley et Sons. 2013.
- (Lebel, 2009) Lebel. G. *surveillance du virus du Nil occidental*. Institut national de santé publique du Québec, 2009.
- (Levine et Pomerol, 1989) Levine P., Pomerol J. C., *Systèmes interactifs d'Aide à la décision et systèmes experts-coll*, Traité des nouvelles technologies, série décision assistée par ordinateur, Book, Edition Hermès, Paris, 1989.
- (Levine, 2004) Levine N. *CrimeStat II: A Spatial Statistics Program for the Analysis of Crime Incident Locations*. Published in Houston/Washington DC, Ned. Levine and Associates. The National Institute of Justice, 2004.
- (Lu et al, 1993) Lu. W, Han. J and Ooi. B. C. *Discovery of general knowledge in large spatial databases*. In proceedings of Far East Workshop on Geographic Information Systems, pp. 275-289, Singapore, June 1993.
- (Malinowski et Zimanyi, 2004) Malinowski. E. et Zimanyi. E. *OLAP Hierarchies : A conceptual Perspective*. In proceedings of 16th International Conference, CAiSE'04. Conference of Advanced Information Systems Engineering, pp. 477-491, Riga, Latvia, June 7-11, 2004.

Liste de références

- (Marouf et al, 2016) Marouf, D., Hamdadoud, D., et Bouamrane, K. *The MAV-ES Data Integration Approach for Decisional Information Systems (DIS): Case on Epidemiologic Monitoring*. International Journal of Healthcare Information Systems and Informatics (IJHISI), Volume 11, Issue 4, pp. 33-55, 2016.
- (Mayster et al, 1994) Maystre, L.Y., Pictet, J. et Simos, J. *Méthodes multicritères Electre, Description, conseils pratiques et cas d'application à la gestion environnementale*. Published in Gérer l'environnement. Préfacé par Bernard Roy. Presses Polytechniques et Universitaires Romandes, Lausanne, Suisse. Paris, Janvier 1994.
- (Mejhed et al, 2009) Mejhed, H., Boussaa, S., et Mejhed, N.El. *Data warehouse and the deployment of data mining process to make decision for leishmaniasis and development of mathematical models for phlebotomine sandflies seasonality in marrakech city*. In proceedings of WSEAS Transactions on Biology and Biomedicine, Volume 6, Issue 1, pp.12-25, January 2009.
- (Moran, 1950) Moran, P.A.P., *Notes on continuous stochastic phenomena*, Biometrika journal, Volume. 37, Issue 12, pp. 17-23. 1950.
- (Morency, 2006) Morency, C. *Etude de méthodes d'analyse spatiale et illustration à l'aide de micro données urbaines de la grande région de Montreal*. Les Cahiers Scientifiques du Transport Journal, Volume 49, pp. 77-102. 2006.
- (Mottier, 1999) Mottier, V. (1999). *Un composant logiciel pour les systèmes informatisés de gestion des réseaux d'assainissements*. Thèse de doctorat, Ecole polytechnique fédérale de Lausanne, Suisse, 1999.
- (Mousseau, 1993) Mousseau, V. *Problèmes liés à l'évaluation de l'importance relative des critères en aide multicritère à la décision*. Thèse de doctorat, UFR Sciences de l'organisation, Université Paris-Dauphine, 1993.
- (Munch et al, 2003) Munch Z. L, Van Lill S. W, Booysen C. N, Zietsman H. L, Enarson D. A, Beyers N. *Tuberculosis transmission patterns in a high-incidence area: a spatial analysis*. International Journal of Tuberc Lung Dis. Volume 7; Issue 3, pp. 271-7. , 2003.
- (Ng et Han, 1994) Ng R., Han J., *Efficient and effective clustering method for spatial data mining*. In proceedings of the International conference on very large database, pp 144-155. Santiago, Chile, September 1994.
- (Noma et al, 2002) Noma, M., Nwoke, B. E. B., Nutall, I., Tambala, P. A., Enyong, P., Namsenmo, A. et Seketeli, A. *Rapid epidemiological mapping of onchocerciasis (REMO): its application by the African Programme for Onchocerciasis Control (APOC)*. Published in Annals of Tropical Medicine et Parasitology, Volume 96, pp. 29-39, 2002.
- (Oliveau, 2004) Oliveau, S. *Modernisation villageoise et distance à la ville en Inde du Sud*, Thèse de doctorat, Université Paris 1 Panthéon Sorbonne, 2004.
- (Openshaw, 1996) Openshaw, S. *Geographical Information Systems and tropical diseases*. Journal of Transactions of the royal society of tropical medicine and hygiene. Volume 90, pp. 337-339. 1996.

Liste de références

- (Ord et Getis, 1995) Ord JK, Getis A. *Local spatial autocorrelation statistics: distributional issues and an application*. Journal of Geographical Analysis, Volume 27, pp. 286-306. 1995.
- (Ouattara, 2010) Ouattara. M. Fouille de données : *Vers une nouvelle approche intégrant de façon cohérente et transparente la composante spatiale*. Mémoire de Maîtrise en sciences géomatiques. La Faculté des études supérieures de l'université Laval, Québec. 2010.
- (Paul et al, 2016) Paul C.M, Flavie L. Goutard, F. R, Davun H, Weerapong T, François L. R, Annelise T. *Quantitative assessment of a spatial multicriteria model for highly pathogenic avian influenza H5N1 in Thailand, and application in Cambodia*. Science Reports Journal. Volume 6, Article Number : 31096, pp. X-X, 2016.
- (Pedersen et al. 2009) Pedersen. T.B, Junmin G, Arie Shoshani, C, Jensen. S. *Object-extended OLAP querying*. Journal of Data et Knowledge Engineering. Volume 68, pp. 453-480, 2009.
- (Pictet, 1996) J.Pictet. *Dépasser l'évaluation environnementale, procédure d'étude et insertion dans la décision globale*. Published in Collection Meta, Presses Polytechnique et universitaires Romandes, Lausanne, Suisse, 1996.
- (Proulx et al, 2002) Proulx, M-J., Bédard,Y., Nadeau.M., Gosselin, P., et Lebel, G. *Géomatique et santé environnementale résultant du projet*. Published in ICEM/SE. Montréal, Quebec, Canada, 2002.
- (Pybus, 2012) Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., ... et Delwart, E. L. *Unifying the spatial epidemiology and molecular evolution of emerging epidemics*. In proceedings of the National Academy of Sciences. Volume 109, Issue37, pp. 15066-15071, 2012.
- (Randremanana et al, 2009) Randremanana RV, Sabatier P, Rakotomanana F, Randriamanantena A, Richard V. *Spatial clustering of pulmonary tuberculosis and impact of the care factors in Antananarivo City*. Journal of Tropical Medecine International Health. Volume 14, Issue 4, pp. 429-37, 2009.
- (Rashedur et Fazle, 2011) Rashedur M. R et Fazle R.M.H. *Using and comparing different decision tree classification techniques for mining*, published in Hospital Surveillance data, pp. 11421–11436, 2011.
- (Rivest et al, 2001) Rivest. S, Bedard. Y et Marchand. P. *Toward better support for spatial decision making : defining the characteristics of spatial on line analytical processing (SOLAP)*. Geomatica Journal. Volume 55, Issue 4, pp. 539-555, 2001.
- (Roddick et Brian, 2009) Roddick J.F and Brian G. L. *Spatio-Temporal Data Mining Paradigms and Methodologies*. The book of Geographic Data Mining and Knowledge Discovery. Second edition, 2009.
- (Roy et Bouyssou, 1993) Roy. B. et Bouyssou. D. *Aide multicritère à la décision : Méthodes et cas*. Published in Production et techniques quantitative appliqués à la gestion. Economica, Paris. France, 1993.
- (Roy, 1985) Roy.B. *Méthodologies multicritère d'aide à la décision*. Journal of Politiques et management publique. Volume 4, Issue 3, pp. 138-140, 1985.

- (Roy, 1997) Roy.B. *Un chaînon manquant en RO-AD, les conclusions robustes*. Published in Cahiers du Laboratoire d'Analyse Multicritères et Modélisation de Système pour l'Aide à la Décision, LAMSADEE, Volume 144. Université de Paris Dauphine, Paris, 1997.
- (Runzi et al, 2017) Runzi L, Shenghui C, Cheng L, Shannon R, Jin C, Qinqin X, Xiaodong L, Yanxun L, Fuzhong X, Qing X et Xiujun L. *Epidemiological Characteristics and Spatial-Temporal Clusters of Mumps in Shandong Province, China, 2005–2014*, Journal of Scientific Reports, Volume 7, Article number: 46328,pp. X-X, 2017.
- (Saaty et al, 1984) Saaty, L. Dahan, L. Hirsch G. *Décider face à la complexité : une approche analytique multicritère d'aide à la décision*, published in Business and Economics, 1984.
- (Russel, 1999) Russel. S., *Rationality and Intelligence*. In Foundations of rational agency journal. Applied logic series., M. Wooldridge and A. Rao, Eds.: Kluwer Academics Publishers Journal, Volume.14, pp. 11-33, 1999.
- (Sander et al, 1998) Sander J., Ester M., Kriegel H.-P., Xu X., *Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications*. International journal of Data Mining and Knowledge Discovery, Kluwer Academic Publishers. Volume. 2, Issue 2, pp 169-194, 1998.
- (Santos et al, 2013) Santos, R., Malheiros, S., Cavaleiro, S. etDe Oliveira, J. P. *A data mining system for providing analytical information on brain tumors to public health decision makers*. Journal of Computer methods and programs in biomedicine. Volume 109, Issue 3, pp. 269-282, 2013.
- (Scharff, 2004) Scharff, C. *Fouilles de Données, Introduction à la fouille de données*, published in IFI, Juin 2004.
- (Scharling, 1985) Scharling. A. *Décider sur plusieurs critères, panorama de l'aide à la décision multicritère*. Published in Collection Diriger l'entreprise, Presse Polytechnique et Universitaire Romandes, Lausanne, Suisse, 1985.
- (Sharov, 1996) Sharov A. *Quantitative Population Ecology*. Cours of Departement of Entimology. Virginia Technical University, 1996.
- (Shekhar et Chawla, 2003) Shekhar Sh, Chawla S, *Spatial Databases: A Tour*, published in Prentice Hall. 2003.
- (Sikder, et Woodside, 2007) Sikder I et Woodside J. *Detection of Space-Time Cluster*. In Proceedings of International Conference on Information and Communication Technology. ICICT'07, pp. 139 - 143, Dhaka, Bangladesh, 7-9 March 2007.
- (Simon, 1982) Simon. H. A. *From substantive to Procedural Rationality*. In *Models of Bounded Rationality* book. Volume. 2, pp. 424-443, 1982.
- (Stevens et al, 2013) Stevens. K. B., Gilbert. M. et Pfeiffer, D. U. *Modeling habitat suitability for occurrence of highly pathogenic avian influenza virus H5N1 in domestic poultry in Asia: A spatial multicriteria decision analysis approach*. Journal of Spatial and Spatio-temporal Epidemiology, Volume 4, pp. 1-14, 2013.

- (Taher, 2008) Taher. O.A. *Spatial On-Line Analytical Processing (SOLAP) : Overview and Current Trends*. In proceedings of the International Conference on Advanced Computer Theory and Engineering, ICACTE'08, pp. 1095-1099, 2008
- (Taleb Zouggar, 2005) Taleb Zouggar. T S. *Graphes d'induction et réseaux de neurones : une nouvelle technique hybride pour la sélection et la segmentation des données*. Mémoire de Magister en Informatique. Université Es-Senia, Oran, 2005.
- (Tchounikine et al, 2005) Tchounikine. A, Miquel. M, Laurini. R, Taher. A, Bimonte. S, Baillot. V. *Panorama de travaux autour de l'intégration de données spatio-temporelles dans les hypercubes*. In proceedings of 1 ère journée francophone sur les entrepôts de données et l'analyse en ligne EDA, pp. 21-33, 2005.
- (Tsai, 2009) Tsai PJ, Lin ML, Chu CM, Perng CH. *Spatial autocorrelation analysis of health care hotspots in Taiwan in 2006*. Journal of BMC Publ Health, Volume 9, pp. 464. 2009.
- (Tukey, 1977) Tukey. J. *Exploratory Data Analysis*. Addison-Wesley, Book, Volume 691, 1977.
- (Vansnick,1990) Vansnick, J.C., *Measurement theory and decision aid* in the journal of Bana e Costa (ed.), Readings in Multiple Criteria Decision Aid, Springer-Verlog, Berlin, pp. 81-100, 1990.
- (Vincke, 1989) Vincke. P. *L'aide multicritère à la décision*. Published in Editions de l'Université de Bruxelles, Bruxelles, 1989.
- (Viswanathan et Schneider, 2011) Viswanathan. G et Schneider. M. *OLAP Formulation for Supporting Complex Spatial Objects in Data Warhouses*. In proceedings of the 13th International Conference, DaWaK'11, pp.39-50, Toulouse, France, August 29-September 2, 2011.
- (Viswanathan et al, 2010) Viswanathan, G., Schneider, M. : *BigCube : A MetaModel for Managing Multidimensional Data*. In proceedings of the 19th International Conference On Software Engineering and Data Engineering (SEDE), pp. 237-242, 2010.
- (Wang et Li, 2014) Wang, L., et Li, X. *Spatial epidemiology of networked metapopulation: An overview*. Chinese Science Bulletin, Volume 59, Issue, 28, pp. 3511-3522, 2014
- (Wang et al, 1997) Wang. W, Yang. J, Muntz. R. *Sting : A statistical information grid approach to spatial data mining*. Published in Technical report CSD- 97006, computer science department, university of California, Los angeles, 1997.
- (Wang et al, 2012) Wang, T., Xue, F., Chen, Y., Ma, Y., et Liu, Y. *The spatial epidemiology of tuberculosis in Linyi City, China, 2005–2010*. Journal of BMC public health, Volume 12, Issue1, Article number 885, pp. X-X, 2012.
- (Warnars, 2010) Warnars. S. *Measuring Interesting rules in characteristic rule*. In proceedings of the 2nd International Conference on Soft Computing, Intelligent System and Information Technology (ICSIT), pp. 152-156. Bali, Indonesia, 2010.
- (Warnars, 2015) Warnars. S. *Mining Patterns with Attribute Oriented Induction*. In Proceedings of the International Conference on Database, Data Warehouse, Data Mining and Big Data (DDDDBD), pp. 11-21, Jakarta, Indonesia 2015.

- (Xu et Sheu, 2012) Xu. C. et Sheu. P.C-Y. *SOLAP Based on Novel Spatial Dimentions*. Journal of Software Engineering and Knowledge Engineering AISC, Y. Wu(Ed) : Volume 114, pp.383-391, 2012.
- (Younsi et al, 2009) Younsi F, Hamdadou.D. *Un système interactif d'aide à la décision en AT, Télédétection et approche multicritère*. In proceedings of CHIA'09 Conférence Internationale sur l'Informatique et ses Applications, pp. X-X, SAIDA 2009.
- (Younsi et al, 2015) Younsi, F. Z., Hamdadou, D. et Boussaid, O. *A surveillance and spatiotemporal visualization model for infectious diseases using social network*. International Journal of Decision Support System Technology, Volume 7, Issue 4, pp. 1-19, 2015
- (Zeitouni et al, 2001) Zeitouni K. et Chelghoum N., *Spatial Decision Tree - Application to Traffic Risk Analysis*. In proceedings of International Conference on Computer Systems and Applications, pp. 203-207. Beirut, Lebanon, June 26-29, 2001.
- (Zeitouni, 2006) Zeitouni, K., *Analyse et extraction de connaissances des bases de données spatiotemporelles*. Habilitation à Diriger des Recherches Spécialité. Informatique, Université de Versailles Saint-Quentin-en-Yvelines, 2006.
- (Zeitouni, 2006) Zeitouni. K. *Fouille de données (Data Mining)*. Partie V – Fouille de données complexes. Notes de cours de Master COSY à l'Université de Versailles Saint-Quentin., 2006.
- (Zeitouni, et al, 1999) Zeitouni. K et Yeh. L. *Le Data Mining Spatial et les bases de données spatiales*. Journal of Revue internationale de géomatique. Volume 9, Issue 4, pp. 389 - 423, 1999.
- (Zemri, 2011) Zemri. F.A. *Vers un Système Interactif d'Aide à la Décision pour la Surveillance Epidémiologique : Couplage SOLAP et Datawarehouse*. Mémoire de magister en informatique, 2011.
- (Zemri et al. 2013) Zemri, F. A., Hamdadou, D et Bouamrane, K. *Towards a Spatio-Temporal Decision Support System for Epidemiological Monitoring: coupling SOLAP and Datawarehouse*. In proceedings of the International Conference on Software Engineering, Databases and Expert Systems (SEDEXS'12), November 13-14. Settat- Morocco. 2013.
- (Zemri et Hamdadou, 2014) Zemri F.A et Hamdadou, D. *Integration of Multi Criteria Analysis Methods to a Spatio Temporal Decision Support System for Epidemiological Monitoring*. In Proceedings of the International Conference on Advanced Aspects of Software Engineering. ICAASE'14, Article Number 52, pp. X-X, Constantine, Algeria, 2014.
- (Zemri et al, 2015). Zemri F. A, Hamdadou D et Zeitouni K. *Vers un système d'Aide à la Décision Multicritères et Spatiotemporel pour la Surveillance Epidémiologique*. In proceedings of Extraction et Gestion des Connaissances, EGC'2015. Actes de l'atelier GAST – Gestion et Analyse de données Spatiales et Temporelles, Article Number 6, pp. 57-72, Luxembourg. 2015.
- (Zemri et Hamdadou, 2019a) Zemri F.A et Hamdadou D. *SOLAM : A Novel approach of Spatial Aggregation in SOLAP*. In International Journal of Interactive Multimedia and Artificial Intelligence. IJMAI. ISSN : 1989-1660 EISSN : 1989-1660.
https://www.ijimai.org/journal/sites/default/files/files/2019/10/ijimai20195_7_10_pdf_10434.pdf

Liste de références

- (Zemri Hamdadou, 2019b)** et Zemri F.A, Hamdadou. *Integration of Data Mining techniques in Multi Criteria Spatial Decision Support System for Epidemiological Monitoring*. In proceedings of INTIS'2019, Article number 89, pp. X-X, Tanger (Maroc), 20, 21 décembre 2019.
- (Zhang et al, 1996)** Zhang. T, Ramakrishnan. R, and Livny. M. *BIRCH : An Efficient Data Clustering Method for Very Large Databases*. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 103-114, Montreal, Canada, 1996.
- (Zhoo et al, 2010)** Zhou, X., Chen, S., Liu, B., Zhang, R., Wang, Y., Li, P., Guo, Y.,Zhang, H., Gao, Z. et Yan, X. *Development of traditional chinese medicine clinical data warehouse for medical knowledge discovery and decision support*. Journal of Artificial Intelligence in Medicine, Volume 48, Issue 2, pp. 139-152, 2010.
- (Zighed Rakotomalala, 2000)** et Zighed, D. A. et Rakotomalala, R. *Graphs of induction, Training and Data Mining*, published in Hermes Science Publication, . 2000.

Résumé

L'objectif visé dans cette thèse est de proposer une nouvelle formulation SOLAP (Spatial On Line Analysis Processing) intégrant les techniques de la FDS (Fouille de Données Spatiales) et les MADMC (Méthodes d'Aide à la Décision Multicritères). A travers la présente étude, nous évaluons d'une part, les avantages du SOLAP dans notre système pour la surveillance épidémiologique « EPISOLAP », et d'autre part nous exploitons les avantages des techniques de la FDS et des MADMC dans l'évaluation du risque sanitaire. Ce risque encouru est mesuré par la présence simultanée de l'aléa et de la vulnérabilité de la population menacée. L'outil EPISOLAP quant à lui, servira pour la détection et la localisation des foyers d'épidémies. Cette proposition d'intégration de ces outils d'aide à la décision puissants sera réalisée dans le cadre d'une seule plate forme décisionnelle intégrée d'une manière cohérente et transparente. Notre préoccupation majeure est de fournir plus d'explications sur les facteurs réels et responsables de la propagation des épidémies ainsi que de son émergence ou réémergence. Enfin, nous pensons que l'exploitation des capacités de la FDS qu'offrent les méthodes et techniques d'extraction de connaissances à partir des données nous permettra d'avoir dans une perspective à court terme une meilleure précision de la prédiction et ainsi des résultats plus convaincants.

Mots clé :

SOLAP (Spatial On Line Analysis Processing); Fouille de données spatiales(FDS); Système d'aide à la décision spatiale(SADS); Aide à la décision multicritères (ADMC); Surveillance épidémiologique (ED); Entrepôt de données; Généralisation spatiale; Caractérisation spatiale; Méthode d'Induction orientée attribut (IOA); Autocorrélation spatiotemporelle.