Discovery of chorems by spatial data minig

Ibtissem Cherni

UNIVERSITE DE TUNIS
INSTITUT SUPERIEUR DE GESTION

UNIVERSITE DE LYON
INSTITUT NATIONAL DES SCIENCES APPLIQUEES

Découverte de chorèmes par fouille de données spatiales

THESE EN COTUTELLE
En vue de l'obtention du Doctorat en

Informatique de Gestion (Tunisie)
Informatique et mathématique (Lyon)

Présentée et soutenue publiquement par
Ibtissem CHERNI
Le 11-09-2015

Membres du Jury:

M.Rached BOUSSEMA	Professeur à l'ENIT de Tunis, Tunisie	Président
M.Mourad ABED	Professeur à l'Université de Valenciennes, France	Rapporteur
M. Jalel AKAICHI	Professeur à l'ISG de Tunis, Tunisie	Rapporteur
M. Thomas DEVOGELE	Professeur à Université de Tours, France	Examinateur
Mme.Hajer BAAZAOUI	MDC Habilité, Université de La Manouba, Tunisie	Examinatrice
M. Sami FAIZ	Professeur à l'ISAMM de La Manouba, Tunisie	Directeur de Thèse
M. Robert LAURINI	Professeur à l'INSA de Lyon, France	Directeur de Thèse
Mme. Sylvie SERVIGNE	MdC à l'INSA de Lyon, France	Co directrice de Thèse

---

Thèse préparée aux

- Laboratoire d'Informatique en Image et Systèmes d'Information d'Information UMR CNRS 5205, Lyon, France
- Laboratoire de Télédétection et des Systèmes d'Informations à Références Spatiales, Tunis, Tunisie

A mes parents
A mes sœurs et mon frère
A ma fille YASMINE
A tous ceux qui me sont chers

Remerciements

Découverte de chorèmes par fouille de données spatiales

Résumé

La motivation de notre thèse est basée essentiellement sur le concept de « Chorème ». Ce néologisme inventé par Roger Brunet [11] désigne à la fois un élément et une structure d'un système. Les chorèmes partent d'éléments géométriques simples comme le point, la ligne, le vecteur, le réseau pour former des sémantiques plus complexes. Ces combinaisons créent des représentations de tout type : de la représentation simple d'un lieu important à travers un point, jusqu'aux flux d'échange qui existent entre des zones à l'aide de cercles, lignes, couleurs, textures, flèches, etc.

Notre travail vise à définir des solutions cartographiques afin de mieux représenter les informations géographiques extraites à partir du contenu de bases de données géographiques, qui se réfèrent à la fois aux objets statiques et aux phénomènes dynamiques. La représentation visuelle dans une carte simplifiée des informations extraites de cette analyse devient une solution pour résoudre le problème d'une complexité encore plus grande, surtout lorsqu'il s'agit de domaines comme la politique, l'économie et la démographie. Nous proposons une solution basée sur le concept de chorème et sur sa capacité à résumer les scénarios impliquant des objets statiques et des phénomènes dynamiques en les associant avec des notations schématiques visuelles.

Notre méthodologie a pour objectif d'extraire les motifs qui servent à construire les résumés visuels de base des données géographiques. Ces motifs sont les suivants : Les clusters (regroupements géographiques), les faits, les flux, les co-localisations, les contraintes topologiques et les informations extérieures.

Il se trouve, cependant, que le nombre des motifs extraits de la première phase est souvent important, nous procédons alors à le réduire en se basant sur l'élimination des connaissances inutiles d'un point de vue de l'expert et en même temps, exclure ceux qui sont redondants.

Pour la phase de visualisation, nous proposons deux choix : la visualisation des résultats basée sur la technique des treillis de concepts et la visualisation sous forme de chorèmes.

D'une manière générale, notre approche comprend trois phases :

1. La première phase concerne l'extraction de patterns à partir de la fouille de données et notamment la fouille de données spatiales,
2. La seconde est dédiée à l'identification des patterns les plus importants,
3. La dernière phase est allouée à la visualisation de résumés visuels.

Un prototype a été implèmenté permettant de valider cette approche.

Mots-Clés: Extraction, résumés visuels, fouille des données géographiques, chorèmes, visualisation, motifs.

Chorems discovery from spatial datamining

Abstract

Our dissertation motivation is basically based on the "chorem" concept. This neologism, invented by Roger Brunet [11], points out to both an element and a structure of the system. A "chorem" starts from a simple geometrical element, such as a dot, a line, a vector, a network.. to form much more complex semantics. These combinations create all types of interpretation, starting from a simple representation of an important place through a given point, to flows of exchanges which exist between zones, with the help of circles, lines, colours, textures, arrows, etc.

The aim of our work is to define a cartographic solutions in order to better represent geographic information extracted from the content of geographic database which refers to statistic objects and dynamic phenomena at the same time. The visual representation, in a simplified map, of the information extracted from this analysis, becomes a solution to an even more complex problem, especially when the domains at stakes are political, economic, and demographic.

We propose a solution based upon the concept of the "chorem" and its capacity to epitomize the scenarios involving statistic objects and dynamic phenomena, relating them to visual schematic rating. The objective of our methodology is to extract the patterns which may serve as a basis to construct visual display of a geographic database. These patterns are the following: clusters, facts, flows, co-localizations, topological constraints, and exterior information.

Nevertheless, the number of patterns extracted during the first phase is seldom important, so, we procede to reduce it being based on the elimination of the unnecessary knowledge from the view point of an expert, at the same time, we exclude those which are redundant. As far as the phase of visualization is concerned, we propose two alternatives: results based on the technique of the concepts lattices and the direct visualization as chorems.

Generally, our approach comprises 3 phases:

1- The first as being the extraction of patterns through spatial data mining
2- The second is dedicated the identification of the most important patterns
3- And the last phase concerns visualization of the results.

To validate this approach, a prototype was implemented.

KEY WORDS: Extraction, visual summary, spatial datamining, chorem, visualization, patterns.

Table des matières

Chapitre 1

Introduction générale

1 Introduction générale

Ce chapitre présente le domaine de recherche, le but et l'organisation de la thèse. En particulier, il identifie les motivations et la problématique qui la sous-tendent et expose les objectifs de ce travail tout en décrivant son organisation.

1.1 Motivations

Un système d'information géographique (SIG) est un outil technologique qui se donne comme objectif de comprendre l'espace géographique afin de prendre des décisions intelligentes.

Les SIG organisent les données géographiques afin qu'une personne qui lise une carte puisse choisir les données nécessaires pour un projet ou une tâche spécifique. Une carte thématique possède une sorte de « table des matières » qui permet au lecteur d'ajouter des couches d'information à une carte de base du monde réel. Avec la capacité de combiner une variété d'ensembles de données d'un nombre illimité de manières, le SIG est un outil indispensable dans la plupart des domaines de la connaissance allant de l'archéologie à la zoologie.

Un bon SIG [16] est en mesure de traiter des données spatiales à partir de sources différentes et de les intégrer dans un projet de carte ou de raisonnement spatial. Certaines données sont recueillies sur le terrain par des appareils de positionnement qui associent un couple de coordonnées (latitude et longitude) à un objet géographique. Les cartes SIG ont la caractéristique d'être interactives. Sur l'écran de l'ordinateur, les utilisateurs peuvent afficher une carte SIG dans n'importe quelle orientation, changer la nature des informations contenues dans la carte afin de choisir de cartographier, par exemple, les routes, leur nature et la façon de les représenter. Ils peuvent également choisir quels autres éléments ils souhaitent voir figurer à côté de ces routes tels que les plantes rares, ou les hôpitaux Certains systèmes SIG sont conçus pour effectuer des calculs sophistiqués pour faire le suivi des orages ou prédire les modèles d'érosion. Les applications SIG peuvent être intégrées dans des activités communes telles que la vérification d'une adresse.

Afin d'accomplir régulièrement des tâches liées au travail d'exploration scientifique de la complexité de notre monde, les SIG donnent aux utilisateurs la possibilité de devenir plus productifs, plus conscients et également donner des citoyens plus sensibles à la planète Terre.

Donc, dans notre recherche pour construire une ″ terre numérique ″ – un accès global à toutes les données possibles sur les lieux à la surface du globe terrestre et du sous-sol – des chercheurs et des praticiens sont confrontés à de nombreux défis concernant, notamment le développement des systèmes de visualisation avec des interfaces conviviales qui permettent l'analyse, la modélisation et la simulation des données, au delà de la simple vision de ces dernières.

Les bases de données géographiques contiennent les informations nécessaires à la compréhension de notre environnement ; ainsi elles nous aident à prendre des décisions par rapport à notre entourage. Bien sûr, ces informations ont besoin d'une représentation claire et facile à comprendre pour pouvoir aider à la prise des décisions. Dès lors, nous avons besoin de cartes qui donnent une vision synthétique d'ensemble et qui intègrent des résumés visuels décrivant facilement l'information importante à expliquer.

La motivation de cette thèse est basée essentiellement sur le concept de ″ Chorème ″. Un chorème est la représentation schématique d'un espace géographique, avec des caractéristiques importantes que nous souhaitons représenter sur une carte afin de les mettre en évidence pour une étude ou obtenir une meilleure compréhension. Ce néologisme inventé par Roger Brunet (1980) [11], ″ chorème ″, vient du mot grec *chôra*, qui signifie étendue, lieu, contrée et du suffixe *ème* qui désigne à la fois un élément et une structure d'un système. Les chorèmes partent d'éléments géométriques simples comme le point, la ligne, le vecteur, le réseau pour former des sémantiques plus complexes. Celles-ci créent des représentations de tout type : de la représentation simple d'un lieu important à travers un point, jusqu'aux flux d'échange qui existent entre des zones à l'aide de cercles, lignes, couleurs, textures, flèches, etc.

Chaque chorème est un dessin qui a sa propre forme et sa propre signification. La signification peut correspondre à un processus ou bien à la représentation de la dynamique d'un certain lieu. Par conséquent, un chorème est un outil puissant qui permet de représenter la connaissance que l'on possède sur un certain lieu, ceci grâce à sa capacité à symboliser et à encapsuler une méthodologie et son interprétation correspondante. Nous pouvons ainsi montrer des situations climatiques, géographiques, économiques, sociologiques, géologiques, agronomiques, etc. basées sur leur contexte spatial, statistique et temporel, grâce à la combinaison de plusieurs chorèmes pour constituer une carte chorématique.

C'est à partir de l'étude des chorèmes que nous avons été motivée à contribuer à l'élaboration d'un système pour représenter des situations comme celles citées précédemment à partir des résultats de la fouille de données sur une base de données géographiques. Les situations les plus communes à représenter sont celles relatives à l'étude de la structure et la dynamique de la population, les concentrations urbaines ou l'interaction entre les systèmes naturels et sociaux.

Pour obtenir les représentations souhaitées, il est nécessaire de faire une analyse en profondeur de la structure et des aspects les plus importants du lieu à représenter

La figure 1.1 montre une comparaison entre une carte traditionnelle de la France et une carte de France contenant des chorèmes, dans laquelle des aspects différents ont mis en évidence :

- La forme simplifiée des limites,
- Les villes les plus importantes,
- Les zones avec différents niveaux de développement,
- Les flux qui représentent les principales routes nationales.

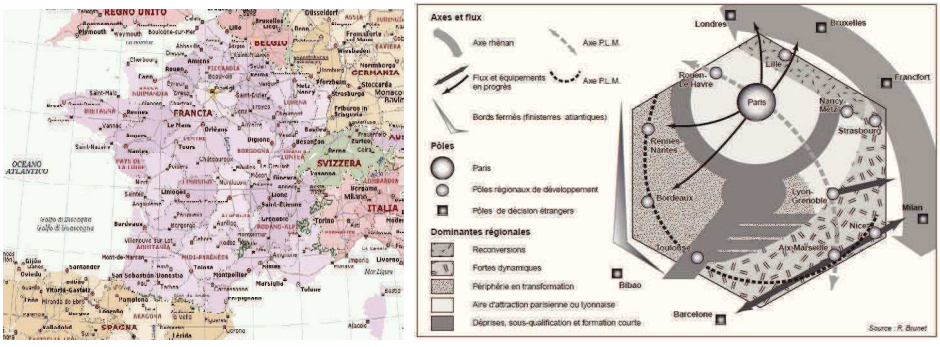


Figure A.1: Carte traditionnelle de France Figure B.1: Carte de France avec des chorèmes.

Figure 1.1 Comparaison entre une carte traditionnelle de la France et une carte chorématique de la France [27]

1.2 Objectifs de la thèse

L'objectif de ce travail de recherche est de mettre au point une méthode interactive de génération de résumés visuels, inspirés par les chorèmes, à partir d'une base de données géographiques. Après une étape de fouille de données spatiales, les faits les plus saillants seront extraits, puis visualisés. Dès lors, un décideur pourra avoir une vision globale simplifiée du contenu d'une base de données géographiques ou bien d'un nouveau lot de données qu'il vient de recevoir.

D'une manière générale, notre approche comprend trois composantes : la première est l'extraction de patterns à partir de la fouille de données et notamment la fouille de données spatiales, alors que la seconde, c'est l'identification des patterns les plus importants et enfin la visualisation de résumés visuels. L'approche proposée peut se présenter comme une amélioration du système conçu et réalisé par Karla Lopez [65], mais en allant plus loin. Alors que son travail s'appuie sur la fouille de données spatiales avec des fonctions d'Oracle et le système Subdue, pour l'extraction des patterns, l'utilisateur intervient et choisit les plus importants de ces derniers ; ensuite le résultat est codé en ChorML1 et visualisé avec le système de visualisation de Vincenzo Del Fatto [27]. Notre contribution s'appuie aussi sur la fouille des données spatiale, Nous avons choisi de travailler avec l'algorithme de *k*-means et de s'inspirer de l'algorithme Apriori pour développer notre propre système d'extraction des patterns, qui seront visualisés dans une carte chorématique grâce au système de visualisation ou sous forme de trellis de concepts.

1.3 Contributions de la thèse

Les chorèmes sont des représentations schématisées des territoires, de sorte qu'ils puissent fournir un bon point de départ pour des analyses poussées depuis des bases de données spatiales. Notre projet vise ainsi à définir des solutions cartographiques afin de mieux représenter les informations spatiales extraites à partir du contenu de bases de données, qui se réfèrent à la fois aux objets statiques et phénomènes dynamiques. Après une analyse géographique d'une base de données spatiales, le volume des résultats obtenus peut être énorme et difficile à exploiter. La représentation visuelle dans une carte simplifiée des informations extraites de cette analyse est une solution pour résoudre le problème d'une complexité encore plus grande, surtout lorsqu'il s'agit de domaines comme la politique, l'économie et la démographie. Comme déjà dit, nous proposons une solution basée sur le concept de chorème et sur sa capacité à résumer les scénarios impliquant des objets statiques et des phénomènes dynamiques, en les associant avec des notations schématiques visuelles. Cette solution offre aux utilisateurs experts, une nouvelle classe de modèles décrivant des positions, des évolutions et des faits faciles à comprendre et à expliquer aux utilisateurs non experts intéressés par les mêmes préoccupations. De manière plus précise, nos contributions sont les suivantes :

- La proposition d'une méthodologie d'extraction de chorèmes
- La mise en œuvre d'un sous-système d'extraction de patterns grâce à la technique de la fouille de données spatiales appliquée aux données géographiques,
- La réalisation d'un générateur du langage ChorML : Chorem Markup Language,
- L'établissement d'un sous-système d'identification des motifs importants,
- La visualisation des résultats obtenus par le sous-système d'extraction.

La visualisation des résultats sous forme des chorèmes a été faite en collaboration avec l'étudiante Sarra Hasni [51] dans le cadre de son travail pour obtenir le Diplôme du Master de l'Université de Jendouba (Tunisie).

1.4 Organisation de la thèse

Outre l'introduction générale (chapitre 1), ce mémoire s'articule autour de cinq chapitres organisés comme suit :

- Le chapitre 2 est dédié à l'état de l'art qui présente les concepts et définitions touchant les deux grands domaines de notre recherche : d'abord, les systèmes d'information géographiques (SIG) et la fouille des données (FD), la FD spatiale avec une étude comparative de quelques algorithmes et finalement nous terminons par un état de l'art sur les résumés visuels.
- Le chapitre 3 décrit les cartes chorématiques, présente une étude cognitive des chorèmes, le langage ChorML ainsi que le module de génération automatique de ce langage ChorML
- Le chapitre 4 décrit notre approche d'extraction et de visualisation des résumés visuels de BDG. Notre méthodologie permettant, tout d'abord, l'extraction des motifs à partir d'une base de données géographiques, ensuite la visualisation des patterns extraits sous forme de graphes de treillis ou de cartes chorématiques.
- Le chapitre 5 décrit la structure du système proposé, d'abord avec la description à travers son architecture, puis les applications et les résultats de la base de données. avec plus de précision, nous décrivons dans ce chapitre notre approche qui comprend trois phases. La première concerne l'extraction de patterns à partir de la fouille de données et notamment la fouille de données spatiales, alors que la seconde est dédiée à l'identification des patterns les plus importants. La dernière phase est allouée à la visualisation de résumés visuels sous forme de treillis de concept, de graphe ou de carte chorématique.
- Le chapitre 6 conclut cette thèse en résumant nos contributions et en déterminant l'orientation des travaux futurs en ouvrant sur quelques perspectives de recherche.

Chapitre 2

Etat de l'art

2 Etat de l'art

L'objectif de ce chapitre est de présenter le contexte du travail basé sur les systèmes d'information géographique (SIG), la fouille de données, la fouille de données spatiales, les chorèmes et les cartogrammes.

2.1 Les systèmes d'information géographiques

2.1.1 Introduction

Un système d'information géographique (SIG) [16] est un système d'information permettant de créer, d'organiser et de présenter des données alphanumériques spatialement référencées, autrement dit géo-référencées, ainsi que de produire des plans et des cartes.

Ses usages couvrent les activités géomatiques de traitement, de partage et de diffusion de l'information géographique. La représentation est généralement en deux dimensions, mais un rendu 3D ou une animation présentant des variations temporelles sur un territoire sont possibles.

Par sa capacité à stocker des données localisées nombreuses et variées, à en conserver l'historique, à les superposer et les croiser, le SIG permet de réaliser des états des lieux et des bilans réguliers sur un territoire et sur un ou plusieurs thèmes.

2.1.2 Les principaux composants d'un SIG

Les concepts définis, nous nous attacherons à décrire les principaux composants d'un SIG, au travers de leurs aspects logiques et organisationnels.

2.1.2.1 Les référentiels cartographiques

Tout support d'information géographique [48] pour être exploitable, il doit préciser le référentiel géographique auquel seront rattachées les informations. Plusieurs types existent, l'un est direct, il a un caractère géométrique (latitude, longitude) et l'autre est indirect, il se réfère à un ou plusieurs autres référentiels, il s'agit en particulier de référentiel administratif.

2.1.2.2 La représentation graphique

La géographie fait appel, à travers de la cartographie, aux vertus représentatives de l'image. Qu'il s'agisse du fond topographique ou des objets que le SIG représente, cette dimension graphique est l'axe essentiel pour l'utilisateur. Deux modes techniques permettent de mettre en œuvre cette représentation : le mode raster et le mode vecteur comme présentés dans la Figure 2.1. Il est possible de passer d'un mode à l'autre : on parle alors de vectorisation ou de rastérisation. Ce pont possible entre les deux modes facilitent entre autre l'acquisition des données : une carte numérisée peut être ensuite vectorisée avec identification des formes qu'elle contient.

- **Le mode raster ou mode tramé (ou matriciel)**

La surface de l'objet est composée par des points jointifs ou pixels. De leur résolution dépend la finesse de la représentation. La position est définie par rapport à la maille de la matrice nécessitant de repasser par un deuxième système de référence pour localiser en absolu le pixel. Ce mode de représentation est le plus proche de l'informatique. Chaque pixel porte une information identifiant sa couleur et l'entité à laquelle il est rattaché. Ainsi une ligne ou une surface sont elles-mêmes définies par l'ensemble des pixels contigus dont la caractéristique de rattachement est identique. Plusieurs couches d'information composées de pixels peuvent être superposées représentant chacune un thème particulier. A ce stade, la description des objets est implicite. Un lien peut être établi entre le fichier raster et une table de données, il permet la description explicite des pixels, mais aussi le traitement des informations du fichier graphique. La relation spatiale entre les objets est implicite.

- **Le mode vecteur**

L'image est décrite par un ensemble d'objets : les SIG retiennent trois primitives de base qui permettent de recomposer la géométrie des objets, il s'agit des objets ponctuels, des objets linéaires et des objets surfaciques.

Un objet ponctuel sera localisé par un seul triplet de coordonnées *(x, y, z)*. Un objet linéaire est une suite ordonnée de points. Chaque point est relié au suivant par un segment de ligne définie mathématiquement. Un objet surfacique est défini comme étant l'intérieur de son contour. Il est donc délimité par un objet linéaire fermé sur lui-même. On peut par extension définir des spécialisations d'objet surfacique. Par exemple un objet surfacique à trou est défini comme un objet surfacique dont une partie intérieure est délimitée par un objet linéaire fermé. Un objet volumique est un objet entouré d'objets surfaciques.

La description des objets est explicite. Une couche d'informations regroupe un ensemble d'objets qu'on souhaite représenter simultanément. La position des objets est exprimée par des coordonnées attachées à un système de positionnement.

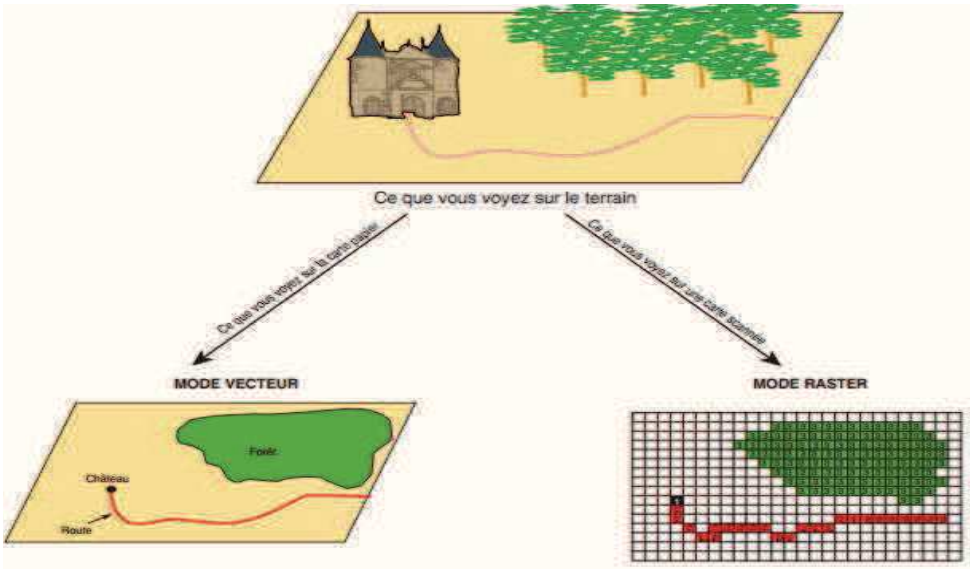


Figure 2.1 Mode de représentation de l'information géographique dans un SIG [16].

- **Relations spatiales entre objets**

Les relations spatiales entre objets sont, soit de type booléen (intersection, inclusion, adjacence par exemple), soit de type « flou » lorsque les critères de la relation doivent être précisés (par exemple, la notion de proximité). Dans ces deux cas les relations peuvent être soit implicites (recalculées à chaque usage) ou explicites (calculées une fois et stockées). Les deux modes de relations spatiales sont utilisées entre les objets. Bien que, par nature, la description des formes issues du mode raster soit implicite, il est possible de rendre leurs relations explicites. C'est la théorie des graphes qui fournit les outils nécessaires à cette transformation.

2.1.3 Les fonctionnalités d'un SIG

On peut rapidement décrire les fonctions attendues d'un SIG [14], la littérature dans le domaine évoque les 5 A d'un SIG :

- *Abstraire* : Le module d'abstraction regroupe les outils de définition des données. A ce titre des fonctionnalités de conception du schéma conceptuel des données peuvent y être intégrées. D'autres fonctions permettent de construire les dictionnaires de données et de contraintes à partir du SCD (Schéma conceptuel de données).
- *Acquérir* : Ce module intègre deux types d'outils, les fonctions d'importation de données, et les fonctions de numérisation. Ces fonctions sont complétées par des outils de géo-référencement, et de contrôle sémantique.
- *Archiver* : Ce module s'appuie sur le support de stockage d'informations évoqué au paragraphe précédent, pour les données sémantiques voir pour les données graphiques, l'utilisation d'un logiciel de CAO/DAO est une alternative possible pour gérer ces dernières données. Les fonctions d'interrogation sont traitées par un langage déclaratif qui transforme les termes de la requête de l'utilisateur en élément d'algèbre relationnel.
- *Analyser* : Ce module contient les fonctions qui différencient les SIG entre eux. A ce titre on peut remarquer les fonctions de manipulation de données qui ne génèrent pas de nouvelles connaissances, les fonctions d'analyse, ce sont celles qui évoluent le plus vers des outils d'aide à la décision, et dernier domaine de développement qui sera traité dans un des paragraphes suivants, la notion de généralisation.
- *Afficher* : ce module intègre tous les outils de restitution des traitements, leur finalité tient dans la matérialisation physique des phénomènes spatiaux, et de leur interaction avec les données sémantiques, un mode hypertexte peut être retenu. L'affichage peut être interactif.

2.1.4 Les domaines d'application des S.I.G

Un système d'information géographique s'intéresse aux relations possibles entre entités et territoires. Une entité est assimilable à un objet. Elle sera donc décrite et localisée par des attributs, et des relations avec d'autres entités. On pourra s'intéresser :

- La gestion d'un objet en particulier : Ce premier usage permet une forme d'accès à l'objet soit par le système d'information classique, soit par l'interface graphique. C'est un premier usage possible du SIG, il offre une alternative au système de requête classique. Il offre de plus la représentation cartographique des résultats des requêtes.
- L'évaluation : Ce mode d'utilisation regroupera les fonctions de recensement, qu'il s'agisse d'un objet spécifique, d'un ensemble d'objets appartenant à une classe d'objets localisés sur un territoire donné.

- La recherche : Dans ce mode d'utilisation, ce sont les fonctions, d'exploration qui permettront de découvrir des relations entre données seules ou entre données et localisation qui seront mises en œuvre, mais aussi des fonctions de simulation et de raisonnement (avec prise en compte d'interaction entre les objets).

Ces trois modes d'utilisation ne sont pas exclusifs l'un de l'autre. On peut facilement imaginer qu'il y ait une progressivité qui conduise au passage d'un mode à l'autre. Les utilisations de SIG pourront être liées à la gestion d'objet, au simple recensement, à la recherche de relations entre objets, à la recherche de tendances et aussi d'évolutions dans le temps. Ces usages peuvent être directement liés à l'activité et aux finalités de l'organisation qui l'utilisera, mais d'une certaine manière à la maturité du système d'information.

2.1.5 L'organisation d'un SIG

Comme tout système d'information, la mise en place, la gestion, l'exploitation d'un SIG nécessite une démarche de projet et des moyens à mettre en œuvre. A chaque étape, une organisation spécifique sera mise en œuvre : pilotage, étude, groupes de travail mais aussi formation et évaluation composeront les clés de cette mise en œuvre. Cette démarche prend ici une dimension particulière au regard de la multiplicité des types de données qu'un SIG peut accueillir.

2.1.6 Conclusion

Dans cette partie nous avons mis en évidence l'importance des systèmes d'information géographique. Leur puissance à analyser les données géographiques est la plus forte raison pour développer notre système, puisque l'utilisateur pourra obtenir des informations précises pour le décideur et notre objectif est de les montrer d'une façon facile à comprendre pour tous les deux.

2.2 La fouille des données

2.2.1 Introduction

Les outils de collecte automatique des données et les bases de données conduisent à d'énormes masses de données stockées dans des entrepôts ; cette croissance explosive des données et des bases de données a généré un besoin urgent de nouvelles techniques et d'outils qui peuvent intelligemment et automatiquement transformer les données en informations utiles et donc en connaissance. Par conséquent, l'extraction de données est devenue un domaine de recherche en forte croissance.

La fouille de données, qui est aussi appelée découverte de connaissances dans les bases de données, se définit comme un processus d'extraction non trivial de données implicites, jusque-là inconnues pour produire des informations utiles (telles que les règles de connaissance, les contraintes, les régularités) à partir de données dans les bases de données.

2.2.2 Définition de la fouille de données

On trouve dans la littérature plusieurs définitions formelles de la fouille de données. Certaines sont admises par la communauté des bases de données et d'autres par les statisticiens. Nous citons celles données par [42], [43] et [52]. D'après [42], la fouille de données est l'extraction automatique de connaissances intéressantes et intelligibles cachées dans les bases de données. Pour [43], la fouille de données est un ensemble de techniques permettant d'extraire, depuis une base de données historiques, des modèles par raisonnement statistique afin de décrire le comportement actuel et/ou de prédire le comportement futur. Les modèles peuvent être soit fonctionnels soit logiques. Pour [52], la fouille de données est un processus non élémentaire de mise à jour de relations, corrélations, dépendances, associations, modèles, structures, tendances, classes, facteurs obtenus en navigant à travers de grands ensembles de données.

La fouille de données est donc un ensemble de méthodes permettant d'extraire des connaissances dans des bases de données. Le mot connaissance est compris comme étant un ensemble de relations entre les données (règles, tendances, similitude, …). Cette connaissance doit être nouvelle, utile, non triviale, non redondante, intéressante, relativement certaine pour que l'utilisateur puisse lui accorder une certaine confiance, et exprimée par un modèle simple et compréhensible [40]. Un panorama de méthodes de fouille de données est donné dans la section suivante.

2.2.3 Les techniques de fouille de données

L'émergence des méthodes de fouille de données provient de l'évolution conjuguée des techniques de la statistique, de l'analyse de données, des bases de données et des algorithmes d'apprentissage automatique. En théorie, il n'existe pas de frontière nette entre ces domaines et la fouille de données. Toutes les méthodes de fouille de données trouvent leur fondement dans ces domaines. En revanche, en pratique, il existe une différence certaine. En effet, la fouille de données est de nature exploratoire et traite des gros volumes de données tout azimut alors que les autres domaines sont confirmatoires et exploitent des données structurées et souvent de taille plus faible.

Les méthodes de fouille de données peuvent être classées, selon leurs tâches, en deux catégories : les méthodes réalisant des tâches descriptives et les méthodes réalisant des tâches prédictives (Figure 2.2), la première catégorie est orientée vers la découverte et permet une analyse exploratoire. Elle a pour but d'explorer, de décrire et de définir les données. La deuxième est à

caractère décisionnel. Elle cherche à prédire une donnée particulière, dans le but de prendre une décision, en se basant sur des connaissances antérieures. Il n'existe pas une frontière nette entre ces deux familles. Certaines méthodes prédictives peuvent être descriptives et vice versa. Les méthodes relevant de l'analyse descriptive réalisent trois fonctions [91] : (i) résumer les données, (ii) regrouper les données, et (iii) chercher les dépendances entre les données. Dans l'analyse prédictive, plus orientée par l'utilisateur que la phase descriptive, on distingue deux principales tâches : (iv) la régression et (v) la recherche des règles de classement. Toutes ces tâches sont décrites ci-dessous. Une présentation plus détaillée est donnée dans [6], [49] et [64].

Les méthodes de fouille de données peuvent être aussi selon le type d'apprentissage utilisé [37] : (i) Apprentissage supervisé (Fouille supervisée) qui est un processus dans lequel l'apprenant reçoit des exemples d'apprentissage comprenant à la fois des données d'entrée et de sortie, les exemples d'apprentissage sont fournis avec leur classe (valeur de sortie prédite). Le but de cette méthode est de classer correctement un nouvel exemple (généralisation utilisée principalement en classification et prédiction et (ii) l'apprentissage non supervisé (Fouille non supervisée) : processus dans lequel l'apprenant reçoit des exemples d'apprentissage ne comprenant que des données d'entrée sans notion de classe. Le but de cette méthode est de regrouper les exemples en « paquets » (clusters) d'exemples similaires (on peut ensuite donner un nom à chaque paquet) utilisés principalement en association et segmentation.

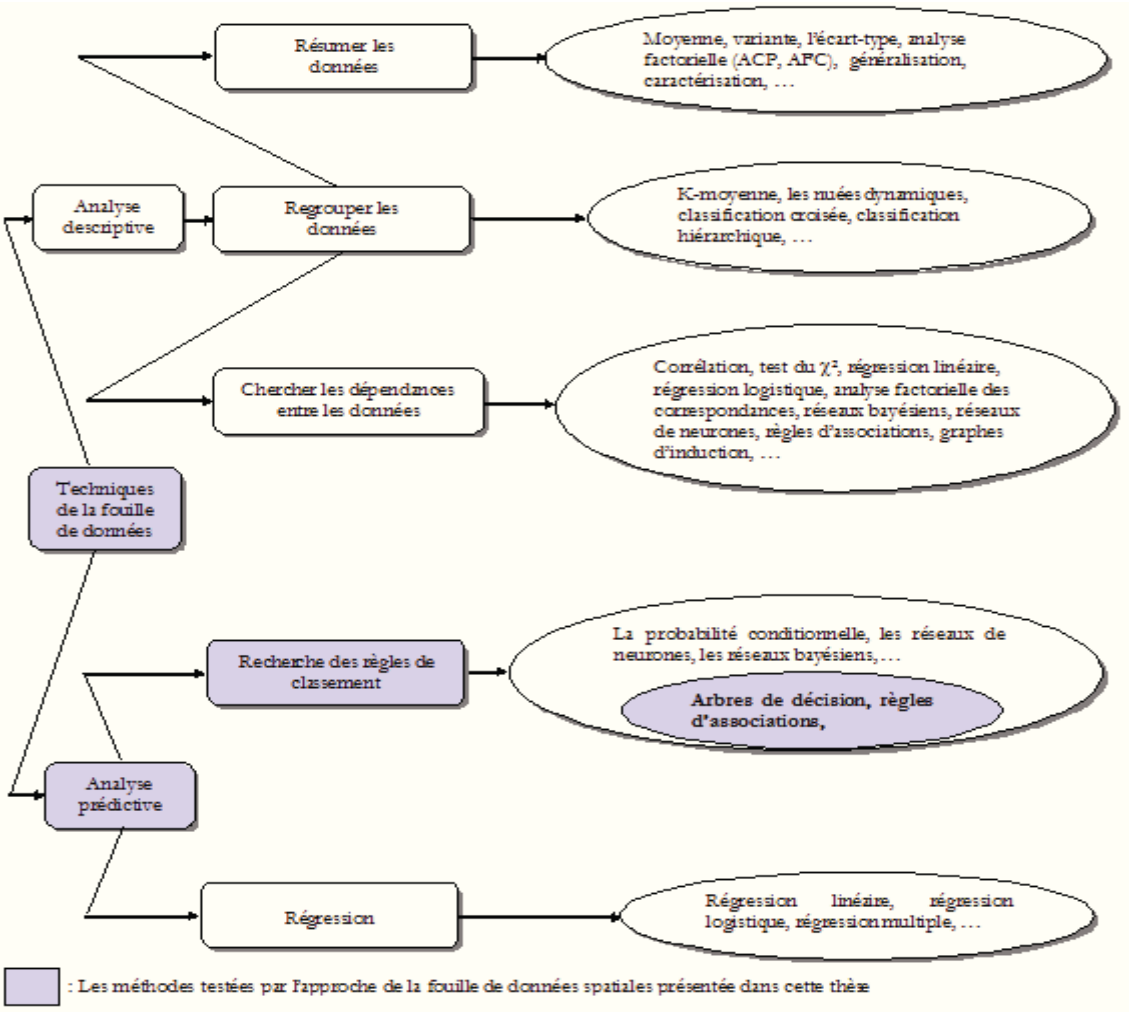


Figure 2.2 Techniques de fouille de données [70].

2.2.4 Quelques méthodes de fouille de données

	Supervisée	Non supervisée
Segmentation		• *k* moyennes (*k*-means) • *k* plus proches voisins (PPV raisonnement à partir de cas) • Réseaux de neurones avec cartes de Kohonen
Classification / Prédiction	• Arbres de décision Réseaux de neurones avec Perceptron • Modèles / réseaux bayésiens • Machines à vecteur supports (SVM) • Programmation logique inductive	• *k* plus proches voisins (PPV raisonnement à partir de cas) • Règles temporelles • Recherche de séquences • Reconnaissance de formes
Prédiction	• Arbres de décision • Réseaux de neurones avec Perceptron	• *k* plus proches voisins (PPV raisonnement à partir de cas)
Association		• Règles d'association

Tableau 2.1 Techniques de fouille de données

2.2.4.1 Classification supervisée par arbre de décision

La classification est une tâche de fouille de données. Elle est la plus utilisée car elle intervient dans plusieurs domaines d'activité comme banque, médecine, etc. Son but est d'assurer la prédiction d'un attribut cible nominal de type chaine de caractère sur la base d'une connaissance préalable des données qui leur sont fournies en entrée [78]. Parmi les algorithmes accomplissant la classification, on cite :

Les arbres de décision : est une technique de classification supervisée et automatisée. Le principe des arbres de décision consiste à classifier des données à partir d'un jeu d'exemples déjà classées. Elle permet d'anticiper la classe d'un attribut non classée selon une base de connaissances [70]. Plusieurs algorithmes ont été proposés notamment CART, ID3, etc. [78].

La génération des arbres de décision se fait à partir d'un échantillon d'apprentissage en deux phases : construction de l'arbre par divisions récursives puis élagage de l'arbre depuis les feuilles afin de réduire sa taille. La construction consiste à déterminer une séquence de nœuds. La définition d'un critère de sélection de la meilleure division, d'une règle de terminaison, et d'affectation de classe sont donc nécessaires. Pour le choix d'une division, différentes mesures ont été proposées dans la littérature, nous nous limitons à deux: la fonction Gini et la fonction Entropie.

Soit *p* un ensemble de données à partitionner, appelons *pi* la fréquence relative de la classe *i* dans *p*.

L'entropie sert à détecter l'attribut le plus discriminant. Elle est donnée par la formule ci-après:

$$E(p) = -\sum_i p_i \log_2 p_i$$

L'indice de Gini est le suivant :

$$G(p) = 1 - \sum_i p_i^2$$

L'algorithme d'apprentissage pour les arbres de décision est le suivant [78] :

Entrée : *échantillon* ***S***
Début
Initialiser l'arbre courant à l'arbre vide ; la racine est le nœud courant
Répéter
Décider si le nœud courant est terminal
Si *le nœud est terminal* ***alors***
Lui affecter une classe
Sinon
Sélectionner un test et créer autant de nouveaux fils qu'il y a de réponses possibles au test
Finsi
Passer au nœud suivant non exploré s'il existe
Jusqu'à *obtenir un arbre de décision*
Fin

Un exemple de la tâche de classification est la prévision météo. En effet, pour prévoir le temps qu'il fera dans une zone géographique donnée, on prend en considération un ensemble de variables explicatives comme l'humidité, la vitesse du vent, la température de la veille et la position de la zone géographique étudiée par rapport aux pôles. Chaque fois qu'on aura alors besoin de prédire la température, il suffira de passer à cette fonction ou modèle les paramètres appropriés pour obtenir les résultats désirés.

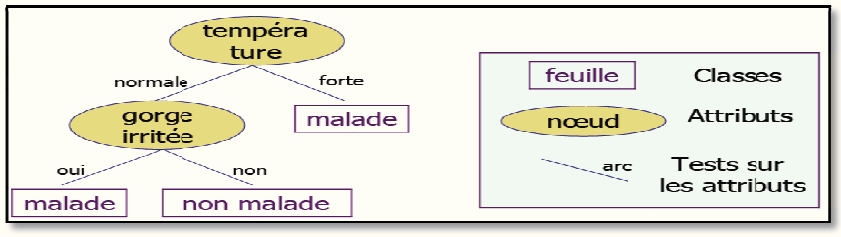


Figure 2.3 Exemple d'arbre de décision.

2.2.4.2 La méthode ID3 (Induction Decision Tree)

ID3 [64] ou arbre d'induction est la désignation d'un algorithme ancien de construction de graphes d'induction arborescents et dans lequel le critère de sélection des variables est le gain informationnel que nous introduisons ci-dessous. De façon analogue à toutes les autres méthodes à base d'arbre, ID3 construit une succession de partitions sur l'échantillon d'apprentissage de plus en plus fines. Il part de la partition grossière située à la racine de l'arbre et cherche, parmi les *p* variables, celles qui donnent la " meilleure partition " nouvelle au sens du critère " gain informationnel ". Sur chaque élément de cette partition, on répète le processus de segmentation comme s'il s'agissait de la racine, sans se préoccuper de ce qui se passe sur les autres sommets de l'arbre. On déclare qu'un sommet est saturé lorsqu'il n'existe aucune variable qui permet d'engendrer une sous-partition qui puisse améliorer la valeur du critère utilisé. Le processus s'arrête si tous les sommets sont saturés. L'algorithme peut être enrichi par d'autres paramètres de saturation comme l'introduction d'une contrainte d'admissibilité sur les sommets, qui permet d'éviter l'apparition de sommets dont les effectifs seraient trop faibles pour être significatifs sur le plan statistique.

Le gain informationnel (critère de segmentation), appelé également le pouvoir informatif, s'interprète comme étant la quantité d'information apportée par la variable utilisée lors du passage d'une partition à une autre. Il est exprimé par une variation d'entropie et plus spécifiquement l'entropie de Shannon.

2.2.4.3 La méthode C4.5

C4.5 [79] est un descendant de l'ID3. En utilisant toujours le même principe de calcul du gain informationnel, C4.5 propose d'introduire un autre facteur visant à pénaliser la prolifération des sommets. Il désavantage les variables qui ont beaucoup de modalités et évite ainsi un émiettement trop rapide de la population. Le critère élaboré, baptisé " le ratio du gain ", exprime le rapport entre le gain informationnel et la distribution des individus sur la partition produite par une variable. Il est calculé selon la formule suivante :

$$G_{ratio} = [H_s(S_j) - \{\sum_{k=1}^{\alpha} (n^j_{.k}/n)\,[-\sum_{i=1}^{m} ((n^j_{ik}/n^j_{.k})\log_2(n^j_{ik}/n^j_{.k}))]\}] \,/\, [-\sum_{k=1}^{\alpha}((n^j_{.k}/n_{.j})\log_2(n^j_{.k}/n_{.j}))]$$

avec:

- $H_s(S_j)$ est l'entropie de Shannon au sommet S_j qu'on vient d'éclater

  $$H_s(S_j) = [-\sum_{i=1}^{m} ((n_{ij}/n_{.j})\log_2(n_{ij}/n_{.j}))]$$

  m : le nombre de classes,

  n_{ij} : l'effectif de la classe C_i dans le sommet S_j.

  $n_{.j}$: l'effectif du sommet S_j.
- α est le nombre de modalité de la variable de division = nombre de nœuds issus du nœud S_j.
- $n^j_{.k}$ est l'effectif du sommet S_j^k $(k = 1,\ldots \alpha)$.
- n^j_{ik} est l'effectif de la classe C_i dans le sommet S_j^k.

Figure 2.5 : Formule mathématique de calcul du gain ratio [79].

C4.5 permet aussi d'élaguer les arbres, de simplifier les règles et de gérer les données manquantes. L'idée d'élagage consiste à développer l'arbre au maximum et ensuite d'appliquer une procédure d'élagage qui vise à supprimer les sous-arbres ne vérifiant pas une certaine condition qui repose sur le taux d'erreur. Le principe revient à se demander si les sommets terminaux issus d'un même sommet père conduisent à un taux d'erreur moyen meilleur. La règle consiste à dire que si la

moyenne des taux d'erreurs sur les sommets issus d'un sommet *S* est plus grande que le taux d'erreur en S, il faut alors supprimer les descendants du sommet *S.*

Pour les données manquantes, C4.5 élabore plusieurs stratégies. La plus simple et la plus utilisée consiste à créer une modalité supplémentaire correspondant aux valeurs manquantes et de ne rien changer dans la formule de calcul du critère de segmentation. Ainsi, le processus de construction d'arbre reste inchangé.

2.2.4.4 CART: Classification And Regression Tree

CART [9] est une méthode de construction d'arbre de décision binaire. A la différence de ID3 et de C4.5, elle choisit non seulement une variable de division mais aussi une modalité particulière de cette variable. Pour la construction de l'arbre, elle suit la même démarche que ID3 et C4.5, sauf que pour le choix de la meilleure segmentation sur un sommet, elle s'appuie sur l'indice de Gini pour un problème de deux classes et sur le critère de Twoing pour un problème de plusieurs classes.

Pour éviter des arbres trop profonds, CART utilise deux stratégies. La première, à l'instar des autres méthodes, fixe des paramètres pour arrêter le développement du graphe. Ces paramètres peuvent être une contrainte sur le critère de segmentation, une contrainte sur le nombre d'éléments dans un nœud ou une limitation de la profondeur du graphe. La deuxième développe l'arbre jusqu'à sa taille maximale et ensuite utilise une procédure d'élagage qui le ramène le plus près possible de la " bonne taille ". On entend par un arbre de " bonne taille ", un arbre qui contient le moins possible de nœuds feuilles et qui classe le mieux possible les objets d'apprentissage.

2.2.4.5 Les règles d'association

L'association est connue sous le nom d'analyse du panier de la ménagère. La tâche d'association en fouille vise à dire quelles sont les variables qui se regroupent ensemble. Selon [26], les règles d'association présentent une approche automatique pour la découverte des relations entre des objets. Il s'agit de trouver des règles du type $X ==> Y$ [support, confidence] où X et Y sont des ensembles d'items disjoints. Où X est l'antécédent et Y est le conséquent. Avec :

- *Support* $P(X ==> Y) = P(X \text{ et } Y)$: décrit la probabilité d'existence de X et Y au sein de même jeu de données. Il présente le pourcentage des transactions qui contiennent tous les antécédents et conséquents.
- Confiance $P(X ==> Y) = P(Y \mid X) = P(X \text{ et } Y)/P(X)$: décrit quant à elle la probabilité d'existence de Y dans l'ensemble de données contenant X pour indiquer le pourcentage des transactions de X qui contiennent Y.

Le nombre maximal d'associations est $K*2(k-1)$, avec k attributs prenant une valeur binaire. À titre d'exemple pour les algorithmes effectuant une tâche d'association, on note l'algorithme Apriori [24].

Dans cet algorithme, un ensemble d'objets (produits) ayant un support supérieur à *MinSup* est qualifié de fréquent. Le problème est de trouver les ensembles fréquents de taille K (K-ensemble fréquent).

L'algorithme Apriori est le suivant [24] :

Nécessite: un support seuil **s**
  $L_1 \leftarrow$ liste des items dont le support est **> s**
  $i \leftarrow 1$
  répéter
    i + +
    à partir de L_{i_1}, déterminer l'ensemble C_i des EIF candidats comprenant i items.
    $L_i \leftarrow \emptyset$
    pour tout élément e $\in C_i$ **faire**
      si support (e) > seuil **alor**s
        ajouter e à L_i
      fin si
    fin pour
  Jusque $L_i \neq \emptyset$;

Un exemple d'utilisation des techniques d'association est la découverte des liens conditionnels entre les produits vendus dans un supermarché. On peut aussi écrire que chaque fois que la viande hachée est achetée, à 80% les pâtes sont aussi achetées. On note donc une association entre la produit viande hachée et les pâtes achetées avec un taux de confiance égal à 80%. Alors, on dispose le rayon " pâtes " au voisinage de celui concernant " la viande " pour amener le client à ne pas fournir d'effort pour aller dans le rayon " pâtes ". En revanche, d'autres pensent qu'il faut mettre " la viande " et les " pâtes " le plus loin possible de manière à ce que les clients passent devant d'autres produits.

2.2.4.6 Réseaux de neurones

Les réseaux neurones [70] (ou réseaux connexionnistes) utilisent l'analogie avec l'architecture physiologique du cerveau humain : les neurones sont des entités élémentaires qui reçoivent des signaux en entrée et transmettent à d'autres neurones des signaux de sortie qui résultent d'une combinaison des signaux d'entrée. Les premiers neurones d'entrée sont reliés à certaines valeurs des attributs d'un objet. Par exemple pour la reconnaissance d'images, ce sont les pixels allumés ou éteints. Les neurones de sortie indiquent la valeur finale de la décision, c'est-à-dire la classe de l'objet. Des neurones intermédiaires sont organisés en couches et l'ensemble constitue un réseau.

Pendant la phase d'apprentissage, des objets sont présentés au réseau et lorsque la réponse diffère de la classe supervisée, un algorithme de rétro-propagation modifie les comportements des neurones intermédiaires. Techniquement, un tel réseau calcule des équations d'hyper-plans séparateurs des classes, selon un algorithme de descente de gradient.

Les réseaux de neurones ont connu un succès rapide, particulièrement pour le traitement des images. Cependant, il est très difficile d'expliquer comment la décision est rendue par ces réseaux, du fait de la grande complexité de leur architecture.

2.2.4.7 Réseaux bayésiens

Fondés sur la notion de probabilité conditionnelle, les réseaux bayésiens [84] permettent de calculer la distribution conjointe sur un ensemble de données à l'aide de procédés stochastiques.

Leur architecture est obtenue par apprentissage à partir des données, mais cette étape reste la partie difficile de la mise en œuvre de cette technique. En revanche, les décisions rendues par ces

réseaux sont plus compréhensibles que celles fournies par les neurones.

La structure de ce type de réseau est simple : un graphe dans lequel les nœuds représentent des variables aléatoires, et les arcs (le graphe est donc orienté) reliant ces dernières sont rattachées à des probabilités conditionnelles. Notons que le graphe est acyclique : il ne contient pas de boucle.

Les arcs représentent des relations entre variables qui sont soit déterministes, soit probabilistes. Ainsi, l'observation d'une ou plusieurs causes n'entraîne pas systématiquement l'effet ou les effets qui en dépendent, mais modifie seulement la probabilité de les observer. L'intérêt particulier des réseaux bayésiens est de tenir compte simultanément de connaissances a priori d'experts (dans le graphe) et de l'expérience contenue dans les données.

Les domaines d'utilisation principaux sont : le diagnostic (médical et industriel), l'analyse de risques, la détection de spams, le datamining, la détection de fraudes, l'exploitation du retour d'expérience, la modélisation et la simulation de systèmes complexes, la détection d'intrusions, TextMining, l'analyse de BioPuces [http 31] et l'analyse de trajectoires de santé. Pour résumer, un réseau bayésien est un modèle probabiliste graphique permettant d'acquérir, de capitaliser et d'exploiter des connaissances, né du besoin de créer des systèmes experts à base de probabilités.

Construire un réseau bayésien revient donc à : (i) Définir le graphe du modèle, (ii) Définir les tables de probabilités de chaque variable, conditionnellement à ses causes (iii) regrouper les probabilités suivant le graphe.

Le graphe est aussi appelé la " structure " du modèle, et les tables de probabilités ses " paramètres ". Généralement, la structure est définie par des experts et les tables de probabilités calculées à partir de données expérimentales. Il est possible d'utiliser des algorithmes tels que K2, le recuit simulé ou encore certains algorithmes génétiques pour construire le réseau.

2.2.5 Conclusion

La fouille de données est un domaine en pleine expansion avec de nombreux résultats de recherche nouveaux et des systèmes ou prototypes mis au point récemment. Les chercheurs et développeurs dans de nombreux domaines ont contribué à faire progresser la théorie et les techniques de fouille de données.

La fouille de données est confrontée à deux points critiques au niveau des données à cause :

- du volume de données étudiées, il faut faire un passage à l'échelle pour passer :
  - Des algorithmes d'apprentissage standards qui ne sont applicables, en pratique, que sur des volumes de données relativement faibles.
  - Aux algorithmes qui prennent en compte de très grands volumes de données généralement stockées sur disque : le temps d'accès à ces données est nettement plus long que si elles étaient en mémoire centrale.

- de la qualité des données, il faut résoudre les problèmes dus à l'absence de données (données manquantes) et aux bruits (données erronées).

2.3 La fouille de données spatiales (FDS)

2.3.1 Introduction

La fouille de données spatiales, ou la découverte de connaissances en base de données spatiales, se réfère à l'extraction des connaissances implicites, les relations spatiales, ou d'autres motifs non explicitement stockés dans des bases de données spatiales.

La difficulté principale de la FDS est la prise en compte de la géométrie (plane ou sphérique) et de la topologie.

Un défi crucial pour l'exploration de données spatiales est l'efficacité des algorithmes d'extraction de données spatiales en raison de l'énorme quantité de données spatiales, de complexité des types de données spatiales et des méthodes d'accès.

Les méthodes de la fouille de données spatiales peuvent être appliquées à extraire des connaissances intéressantes et régulières dans de grandes bases de données spatiales, et ce dans de nombreuses applications dans les systèmes d'information géographique. La découverte de connaissances à partir des données spatiales peut être de formes diverses, telles que les règles caractéristiques et discriminantes, l'extraction et la description de structures importantes ou des groupes, des associations du territoire, etc.

2.3.2 Définition de la fouille de données spatiales

La fouille de données spatiales (FDS) est une branche de la fouille de données. Elle est née du besoin d'exploiter dans un but décisionnel des données à caractère spatial, produites, importées ou accumulées au fil du temps, susceptibles de délivrer des informations ou des connaissances par le moyen d'outils exploratoires. Sa spécificité par rapport à la fouille de données traditionnelle est qu'elle considère les propriétés de voisinage et les relations spatiales [84], [85] *Koperski et al* [56] définit la fouille de données spatiales comme la découverte automatique, depuis des grandes bases de données spatiales, de connaissances nouvelles, utiles, implicites et précédemment inconnues.

Tout comme la fouille de données traditionnelle, la fouille de données spatiales est un composant d'un processus plus global de transformation des données en connaissances. Elle doit être précédée par une étape de préparation des données et suivie par une étape d'évaluation des modèles découverts. En théorie, cette démarche en trois étapes est linéaire, mais en pratique, la fouille de données spatiales effectue quelques allers-retours entre ces étapes pour améliorer et enrichir la connaissance. La fouille de données spatiales n'est pas totalement automatique. Elle interagit avec l'utilisateur qui la contrôle et la guide. Elle intègre également des informations supplémentaires récupérées de la base de connaissances. En plus des connaissances classiques extraites de la FDS engendrera de nouveaux types de connaissances dans laquelles interviendront la géométrie et la topologie (par exemple la co-localisation).

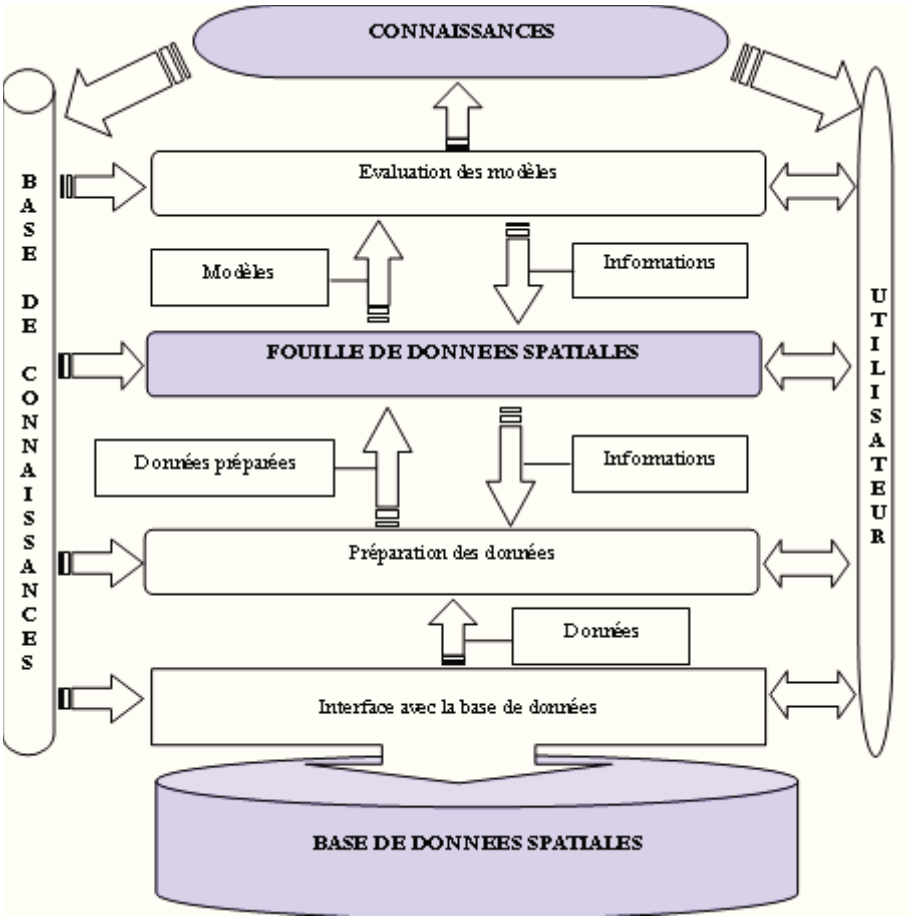


Figure 2.7 : Place de la FDS dans le processus de découverte des connaissances [70]

Comme dit précédemment, un objet spatial est une entité du monde réel (par exemple, une ville, un pays, une route, un lac, un arbre,…) représenté dans une base de données spatiales par une structure. Cette structure contient à la fois des données sémantiques et des données spatiales. Les données sémantiques, appelées aussi les attributs *non-spatiaux* ou *alphanumériques,* décrivent qualitativement ou quantitativement les propriétés de l'objet (Ex : le nom d'une commune, la hauteur d'un bâtiment,…).

Rappelons que les données spatiales décrivent la géométrie de l'objet spatial, sa localisation dans l'espace et les relations spatiales qui le relient aux autres objets. La localisation est la position géographique de l'objet. Elle est repérée, selon un système de projection donné, par la latitude, la longitude et éventuellement l'altitude de l'objet. Les relations spatiales désignent les liens de voisinage entres les objets.

2.3.3 Spécificités de bases de données spatiales

Une base de données spatiale est un ensemble d'objets spatiaux organisés et regroupés dans des couches thématiques. Chaque couche représente un thème et regroupe un ensemble d'objets spatiaux partageant les mêmes propriétés. Chaque objet est défini par des données sémantiques et des données spatiales. Une donnée spatiale regroupe la forme ou la morphologie de l'objet, sa localisation géographique et parfois les relations spatiales qui le relient aux autres objets.

2.3.3.1 Caractéristiques des données spatiales

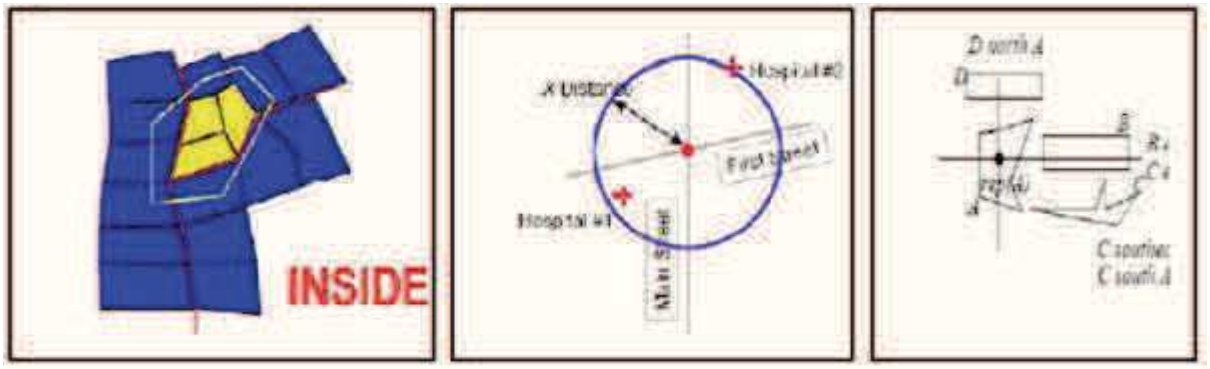
Les relations spatiales traduisent une caractéristique essentielle du monde réel. Les indices d'auto-corrélation spatiale définis en analyse spatiale considèrent uniquement les interactions sur une couche thématique. Or, en règle générale, les couches thématiques sont fortement corrélées. Une carte des précipitations et de la densité de population seraient, par exemple, non corrélées si l'on reste au niveau de couches séparées, mais corrélées si on travaille sur plusieurs couches : en effet, la densité de population dépend de la production agricole qui, elle même, est liée aux précipitations. Par conséquent, nous distinguons deux types de relations spatiales : (i) celles qui lient les objets d'une même classe (appelées intra-thème) et (ii) celles qui sont associées à plusieurs classes (appelées inter-thèmes).

2.3.3.2 Définition d'un objet spatial

Un objet spatial est une entité du monde réel (par exemple, une ville, un pays, une route, un lac, un arbre, etc.) représenté dans une base de données spatiales par une structure. Cette structure contient à la fois des données sémantiques et des données spatiales [46]. Les données sémantiques, appelées aussi les données *non-spatiales* ou *alphanumériques,* décrivent qualitativement ou quantitativement les propriétés de l'objet (Ex : le nom d'une commune, la hauteur d'un bâtiment. etc.). Les données spatiales décrivent la géométrie de l'objet spatial, sa localisation dans l'espace et les relations spatiales qui le relient aux autres objets. Le terme géométrie désigne la forme ou la morphologie d'un objet (par exemple : point, ligne, polygone, cercle, etc.). La localisation est la position géographique de l'objet. Elle est repérée, selon un système de projection donné, par la latitude, la longitude et éventuellement l'altitude de l'objet...

2.3.3.3 Les Relations spatiales

La notion de dépendance spatiale est très importante pour les données géographiques et est connue comme la 1ère loi en géographie**.**
Les relations spatiales constituent la base des descriptions linguistiques de configurations spatiales. Ces relations sont généralement classées en différentes catégories : relations topologiques, distances, direction. C'est-à-dire, relations qui désignent les liens de voisinage entre les objets. La figure 2.8 donne des exemples des relations spatiales.



Relation topologique Relation de distance Relation de direction

Figure 2.8 Les relations topologiques, de distance et direction

• Les relations topologiques

Les relations topologiques sont les relations qui restent invariables sous des transformations topologiques, elles sont préservées si les deux objets changent d'échelle sont translatés ou pivotés simultanément. Les définitions formelles sont basées sur les frontières, les intérieurs et les compléments des deux objets connexes voir SS 4.4.2.2.2.

• Les relations de distance ou de proximité (co-localisation)

Les relations de distance sont ces relations comparant la distance de deux objets à une constante donnée utilisant un des opérateurs arithmétiques. La distance entre deux objets, ensemble de points, peut alors simplement être définie par la distance minimum entre leurs points.

• Les relations de direction

Pour définir la relation de direction *objet-2* R *objet-1*, nous distinguons l'objet source *objet-1* de l'objet cible *objet-2* de la relation R de direction. Il y a plusieurs possibilités pour définir des relations de direction selon le nombre de points qu'ils considèrent dans la source et l'objet de destination.

2.3.3.4 La jointure spatiale

Une jointure attributaire [http 30] consiste à lier à une couche des données provenant d'une table ou d'une autre couche. On se base pour cela sur les données attributaires.

Pour faire une jointure, il est possible de se baser sur la position des éléments et non plus sur leurs données attributaires : il s'agit alors d'une jointure spatiale, mais il faut tenir compte des erreurs de mesure. Dès lors le critère de jointure sera :

$$Dist\ (x_1, x_2) < \mathcal{E}\ x \wedge Dist\ (y_1, y_2) < \mathcal{E}\ y$$

Par contre, la jointure spatiale ne peut se faire qu'entre deux couches SIG, de type point, ligne ou polygone. Il est possible par exemple de partir d'une couche de polygones et d'une couche point, et de lier à chaque polygone les données attributaires du point contenu par ce polygone.

On note que comme pour une jointure attributaire, les données qui seront jointes sont toujours les données attributaires.

2.3.4 Les techniques de fouille de données spatiales

La fouille de données spatiales représente un catalogue de méthodes. Ces méthodes sont pour la plupart des extensions de celles de la fouille de données traditionnelle intégrant les critères spatiaux.

Elles utilisent d'une manière intensive les relations spatiales car ces dernières mettent en évidence l'influence de voisinage entre les entités spatiales. Toutes ces méthodes trouvent leurs fondements dans plusieurs domaines : la statistique spatiale, la fouille de données et les bases de données spatiales.

Tout comme les techniques de fouille de données, les techniques de fouille de données spatiales peuvent être classées, selon leurs tâches, en deux catégories : les méthodes réalisant des tâches descriptives et des méthodes réalisant des tâches explicatives. La première tâche est un complément en amont de la deuxième. Elle est orientée découverte et permet de décrire les données. La deuxième affine les résultats obtenus dans la première phase en cherchant à expliquer les écarts ou les caractéristiques des groupes. Cette explication est ensuite utilisée pour la prise de décision en cherchant à prédire la valeur d'une donnée particulière. Il n'existe pas, comme en fouille de données traditionnelles, une frontière nette entre ces deux familles. Certaines méthodes réalisant des tâches prédictives peuvent réaliser aussi des tâches descriptives et réciproquement.

Les méthodes de la fouille de données spatiales présentent une extension de celle traditionnelle. Elles sont décrites d'une manière claire dans ce qui suit.

2.3.4.1 Clustering spatial

Comme est définie dans la section fouille de données, le regroupement (clustering spatial) présente une méthode de classification non supervisée, qui regroupe des objets dans des classes. Son but est de maximiser la similarité intra-classes et de minimiser la similarité interclasses.

Pour leur transposition au domaine spatial, les algorithmes de clustering s'appuient sur une mesure de similarité d'objets localisés suivant leur distance métrique. Toutefois, le résultat du regroupement en spatial n'est pas tant de former des classes que de détecter des concentrations anormales.

Exemple :

En utilisant le clustering pour détecter un point chaud dans l'étude de criminalité, ou des zones à risques en accidentologie [90].

2.3.4.2 Règles associatives spatiales

Une règle associative est une implication de la forme " si A alors B " [44] ou plus formellement notée : *AB, [s%; c%]* où *A* et *B* sont des ensembles de prédicats spatiaux et non spatiaux, s% est le support de la règle, et c% est sa confiance. Les règles associatives servent à trouver des associations entre des propriétés des objets et celles de leur voisinage. L'extension de la découverte de règles d'association aux données spatiales pour générer des règles.

Exemple :

La règle suivante est une règle associative spatiale [44]:
Est-un (*X*, " école ") proche-de (*X*, " station de bus ") proche-de (*X*, " marché ") [20%; 80%].
Cette règle exprime que 80% des écoles qui sont proches des stations de bus sont également à proximité des marchés, et que 20% des données appartenant à un tel cas.

2.3.4.3 Classification spatiale

La recherche de règles de classement vise à structurer un ensemble d'objets en classes d'objets ayant des propriétés communes [40]. Cette tâche est réalisée par apprentissage supervisé. L'extension au domaine spatial a été définie par l'extension aux propriétés de leurs voisins jusqu'à un ordre *N* de voisinage.

Exemple :

Avec les algorithmes de classification spatiale, il est possible de trouver une règle de type :
Si population élevée et type de voisin = route et voisin de voisin = aéroport Alors puissance économique élevée (à 95%). Exemple : classer les accidents selon 3 classes impliqués (piéton, 2 roues, véhicules) selon les propriétés des accidents et des objets voisins.

2.3.5 Avantages et inconvénients d'algorithmes de fouille de données spatiales

Dans le tableau 2.2 nous présentons les avantages et les inconvénients de quelques algorithmes de fouille des données spatiales.

Technique de FDS	Méthodes	Avantages	Inconvénients
Classification spatiale	Algorithme basé sur la méthode **ID3** [64]	Ces méthodes considèrent non seulement les propriétés des objets à classer, mais aussi les attributs des objets voisins et modifient l'algorithme en conséquence. Le degré de voisinage dépasse le niveau 1.	Le défaut majeur de cette méthode est qu'elle ne garantit pas une segmentation correcte menant à des sous populations non disjointes. Cette méthode est limitée également à une seule relation de voisinage. Enfin, elle ne fait pas de distinction entre les thèmes.
	Index de jointure spatiale [70]	Grâce à l'index de jointure spatiale, la classification peut se baser désormais sur une représentation directement en relationnel. Nous permet de prendre en compte l'organisation en couches thématiques et les relations spatiales propres aux données spatiales.	L'inconvénient majeur de cette méthode est son temps d'exécution qui est considérablement élevé.
	SCART [9]	SCART classe les objets spatiaux selon à la fois leurs attributs, les attributs de leurs voisins et les relations spatiales. contrairement aux méthodes existantes, elle effectue un choix automatique de la "bonne" relation de voisinage.	L'inconvénient de cet algorithme est qu'il constitue une méthode naïve trop coûteuse en temps d'exécution qu'il faut évidemment éviter.

Techniques de FDS	Méthodes	Avantages	Inconvénients
Règles d'association spatiale (RAS)	L'algorithme **SAR** [56]	Permet la distinction entre les thèmes, découverte d'associations entre non seulement les propriétés des objets à analyser, mais aussi les objets voisins et les relations spatiales sous forme de prédicats spatiaux	L'inconvénient majeur est génère qu'il un très grand nombre des règles d'association même pour des contextes d'extraction raisonnable. L'utilisation des prédicats spatiaux mais pas les fonctions spatiales (du moins, elles nécessitent leur transformation en prédicats). La généralisation des données spatiales (ce qui engendre une perte d'informations détaillées).
	L'algorithme **ARGIS** (Association Rules in **GIS)** [http28]	Cet algorithme est basé sur la notion de table de liens.	
	RASMA [http29]	RASMA permet de générer des règles qui englobent en même temps des prédicats spatiaux et des prédicats non spatiaux ce qui n'est pas le cas pour RAS. RASMA offre la possibilité de créer des agents qui coopèrent ensemble pour diminuer le temps d'exécution de RAS.	Son utilisation nécessite la transformation coûteuse des données relationnelle en un ensemble de faits exprimés en logique du premier ordre. Difficulté de distribuer les tâches sur les agents.

Techniques de FDS	Méthodes	Avantages	Inconvénients
Clustering	Clustering par grille : **STING** [71] Clustering par partitionnement : ***k*-means, PAM,** **CLARA, CLARANS** [71]	Méthode hiérarchique qui consiste de deviser le territoire en plusieurs zones de taille plus réduite afin de minimiser la complexité de la recherche. ***k*-means** Rapidité **:** il ne compare pas toutes les observations entre elles mais par rapport aux centres de classes. Permet de détecter les valeurs extrêmes et de les isoler Est pratique quand il y a un très grand nombre d'observations (des milliers). ***k*-medoids(PAM)** Plus robuste que *k*-means Plus insensible aux " outliers **"**	Elle ne fournit aucune description sur la relation des classes et leurs relations avec les caractéristiques des objets. ***k*-means** N'est pas applicable en présence d'attributs qui ne sont pas du type intervalle (moyenne=?) On doit spécifier *k* (nombre de clusters) Les clusters sont construits par rapports à des objets inexistants (les milieux). Ne peut pas découvrir les groupes non-convexes ***k*-medoids (PAM)** -Beaucoup plus coûteuse que K-means -Plus de calculs -Efficace uniquement dans le cas de données de petite taille.

Techniques de FDS	Méthodes	Avantages	Inconvénients
	Clustering hiérarchique: **BIRCH, CURE** [55]	**CLARNS** Données numériques de petite dimension Classes de meilleures qualités que celles obtenues avec PAM et CLARA. Obtention de classes sphériques de forte densité et bien séparées. **CLARA** Peut traiter des échantillons de taille beaucoup plus grande que ceux traités par PAM. **Birch** Trouve les clusters en une seule passe sur la BD. **Cure** S'adapte bien à la géométrie des clusters Représentant multiples. Déplacement du bord vers le centre des représentants	**CLARNS** On peut traiter des échantillons de taille beaucoup plus grande que ceux traités par PAM **CLARA** Efficacité de CLARA dépend la taille des échantillons tirés au hasard considérés **Birch** Ne considère que les données numériques et est sensible à l'ordre des enregistrements **Cure** Eliminer les exceptions (points aberrants) Regrouper les clusters partiels

Technique de FDS	Méthodes	Avantages	Inconvénients
	Clustering par densité, parmi les algorithmes on a **DBSCAN** [39]	**DBSCAN** Un cluster est l'ensemble maximal de points connectés Il découvre des clusters non nécessairement convexes. Permet de détecter les anomalies et les objets qui ne forment avec les autres aucun cluster	Nécessite un index spatial choisir

Tableau 2.2 Comparaison des algorithmes de FDS.

En conclusion de cette analyse rapide, nous constatons donc que le clustering spatial et plus précisément l'algorithme *k*-means est la méthode la plus appropriée pour extraire des connaissances qui ont correspondu aux structures de motifs que nous voulons extraire et qui sont présentés dans le chapitre 4. Avec la distance euclidienne, *k*-means nous permet de sélectionner des zones avec un ensemble maximal de point connectés qui partagent les mêmes caractéristiques c'est le contraire aux autres algorithmes qu'ils ne permettent pas de détecter les valeurs extrêmes et de les isoler et qui sont beaucoup plus coûteuses.

2.3.6 Conclusion

La fouille de données spatiales est un champ prometteur de la recherche avec de nombreuses applications dans le domaine de la géomatique. Bien que ce champ disciplinaire soit assez jeune, un certain nombre d'algorithmes et de techniques a été proposé pour découvrir différents types de connaissances à partir de données spatiales. La variété des sujets encore inexplorés et des problématiques encore ouvertes rend attrayante et stimulante la découverte de connaissances dans les bases de données spatiales.

2.4 Etat de l'art sur les résumés visuels

De nombreux travaux [11], [35], [25], qui tournent autour de la chorématique, ont été effectués depuis la publication de l'article séminal de Roger Brunet en 1986.

Une autre direction de recherches est celle des résumés de base de données. S'il est classique de fournir un résumé d'un document sous divers noms (abstract pour une publication, synopsis, résumé exécutif d'un document pour un décideur, etc.), il devient urgent de générer des résumés de base de données. Des travaux ont porté sur la création de résumés de structure (par exemple, extraction d'un schéma simplifié de structure de données) ; mais dans notre cas, il s'agira non pas de résumer la structure, mais le contenu. Pour aller plus loin, un des objectifs souhaités est non pas de résumer textuellement une base de données géographique, mais d'offrir des résumés visuels pour donner une idée des problématiques à certaines échelles. Nous citons comme exemple de ces résumés, chorèmes et les cartogrammes.

2.4.1 Les chorèmes : Résumés visuels de base de données géographiques

Un chorème [11] est une représentation schématique d'un espace géographique, avec des caractéristiques importantes que nous souhaitons représenter sur une carte et mettre en évidence pour une étude ou une meilleure compréhension.

2.4.1.1 Evolution des chorèmes

Les chorèmes et la chorématique ont participé à des progrès observés dans la cartographie, en introduisant une certaine rigueur, notamment dans les figurés.

Les chorèmes sont pédagogiques en ce sens qu'ils fournissent une lecture plus aisée de l'espace, et avec le temps, ils sont connus une grande évolution comme ils ont appliqués dans la plupart des domaines. Dans les paragraphes ci-dessous nous avons détaillé cette évolution.

- **Les chorèmes de Roger Brunet**

Roger Brunet a suggéré une approche fondée sur les modèles graphiques. Il propose un tableau de 28 chorèmes [11] donné par la figure 2.9 et il affirme que construire des représentations du monde permet de mieux l'analyser. Pour pouvoir comprendre les différentes situations d'un certain endroit, il est nécessaire de faire des représentations de celui-ci et ensuite les comparer avec celles d'autres lieux.

Les chorèmes ont pour but de présenter des territoires avec tous les aspects qui les caractérisent. Parmi ces aspects nous pouvons trouver les individus, les entreprises, les états, les organisations et les acteurs qui font partie du territoire, mais également les changements et les évolutions. Nous pouvons ainsi décrire entre autres la séparation, l'intégration, le mouvement, la rupture et l'asymétrie.

Brunet a fait connaître une série de lois qui démontrent que les espaces à représenter ont des formes géographiques particulières, ce qui permet de répondre à toutes les critiques qui voient en eux seulement une manière d'imposer une modélisation spatiale déterministe.

A travers les chorèmes, nous pouvons identifier les structures et les processus les plus importants d'une situation spécifique lesquels aideraient à prendre des décisions aux parties chargées d'administrer un territoire. Chaque partie peut être intéressée par des informations différentes selon la situation à administrer. Par conséquent, les chorèmes à produire doivent être différents selon les nécessités de l'utilisateur.

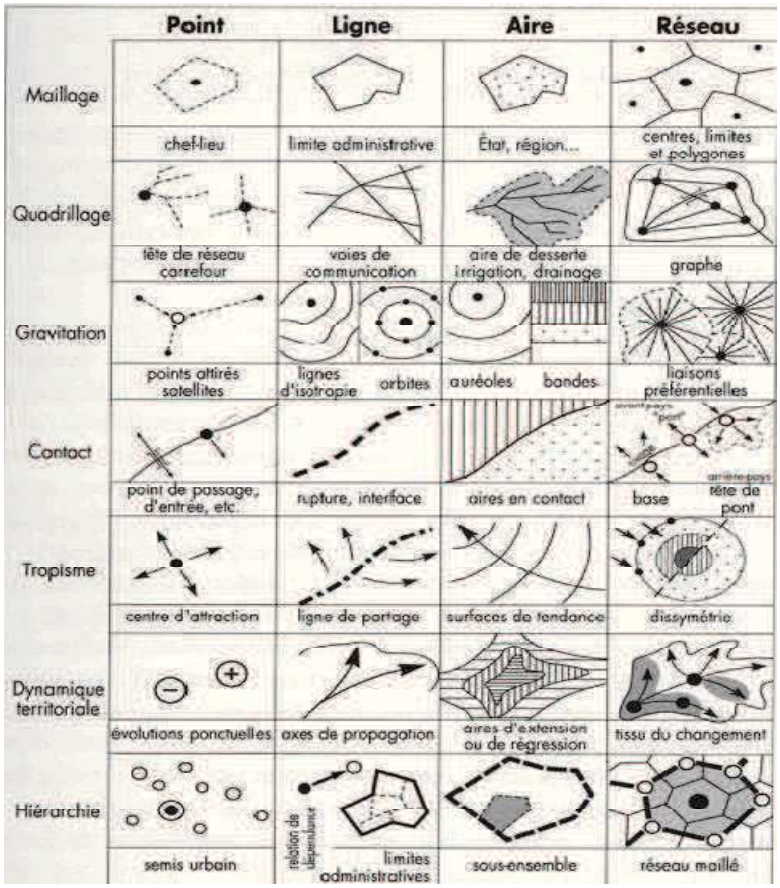


Figure 2.9 Les 28 chorèmes de Roger Brunet [11].

- **Les chorèmes de César Ducruet**

César Ducruet [35] a présenté une table des chorèmes (cf. Figure 2.10), dans les colonnes, trois modes d'organisation spatiale (le point, la ligne et la surface) et deux niveaux de complexité spatiale (le système et le modèle) sont distingués. Dans les lignes, plusieurs entrées couvrent la majorité des chorèmes déjà existants, plus un certain nombre de nouveaux choèmes, mais il n'est pas nécessaire de les présenter avant leur application aux stratégies réseaux urbaines.

Les modèles spatiaux classiques sont la région, le modèle de gravitation, le modèle d'endroit central et le modèle de périphérie de base. Chacun d'entre eux étant une combinaison de plusieurs chorèmes.

D'autres modèles spatiaux sont construits sur quelques variantes, telles que les secteurs (quartiers irréguliers), polynucléaires (fusion de plusieurs centres), l'extra version, les échanges (les effets de la spécialisation et l'emplacement d'interface) et de la décentralisation (planification de la politique de déconcentration accompagnée).

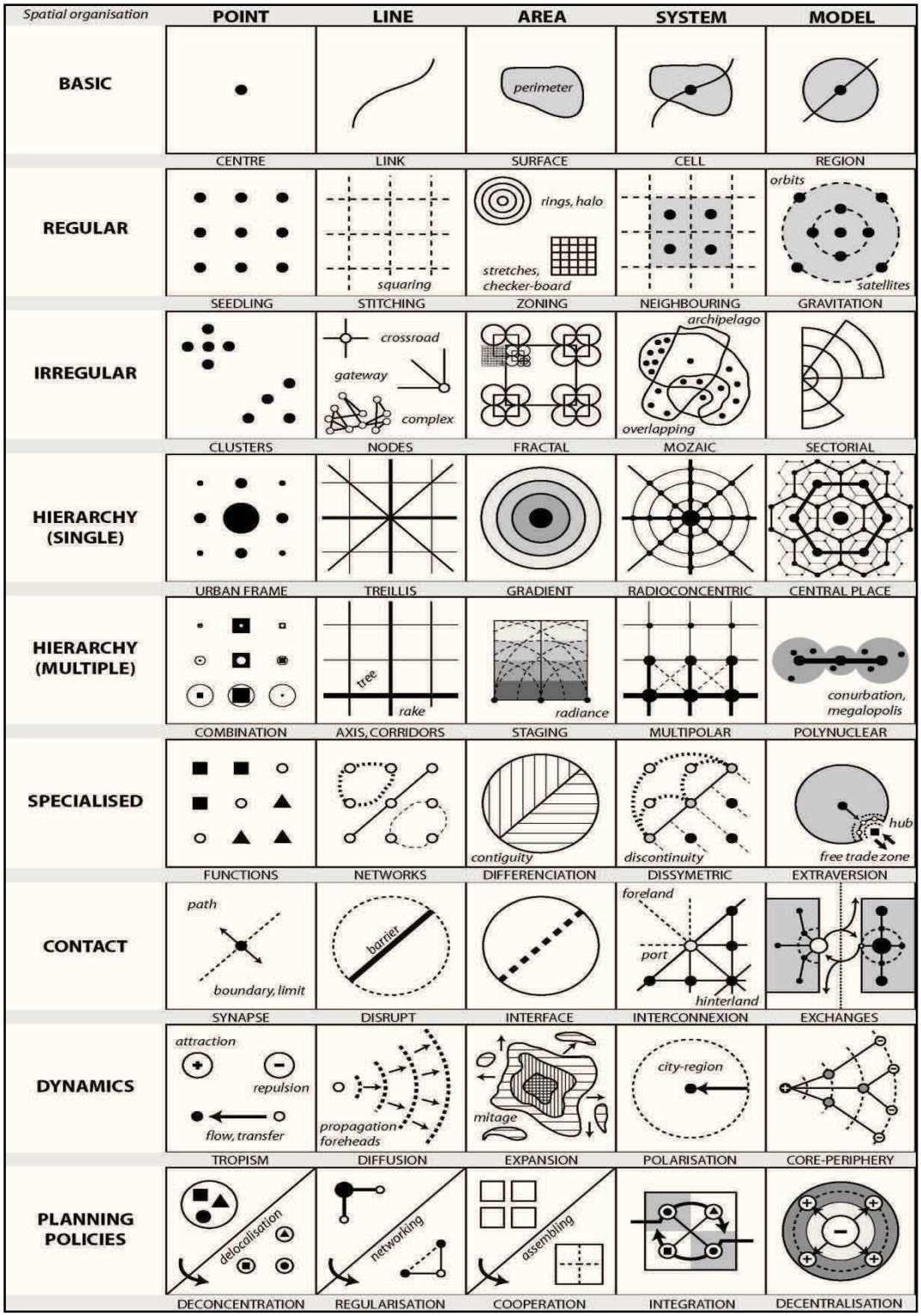


Figure 2.10 Les chorèmes de César Ducruet [34].

Le principe de chorèmes est la représentation simplifiée de la réalité afin de montrer des phénomènes qui ne peuvent être révélées par habitude des outils graphiques comme des cartes, des diagrammes et l'analyse statistique. Chaque chorème est une étape de démonstration.

A la fin, la combinaison harmonieuse des chorèmes conduit à une carte de synthèse avec les principales structures et la dynamique.

- **Les chorèmes de Cheylan**

Comme illustré par la figure 2.11, Jean-Paul Cheylan [25] applique les chorèmes à l'activité agricole et la gestion de l'espace rural.

Pour lui, les chorèmes fournissent un langage commun aux chercheurs et aux acteurs concernés et facilitent le dialogue. Tout d'abord, il s'agit d'identifier quelles sont les entités pertinentes que l'on retient et quelles sont les principales relations, il y a aussi une phase de négociation pour obtenir un consensus collectif autour des chorèmes retenus pour rendre compte du phénomène étudié, ensuite il faut donner voir les connaissances que l'on a sur une situation et faire comprendre les mécanismes, interpréter les processus encours. Le choix des chorèmes lève les ambiguïtés du discours.

Une fois le dialogue est établi entre les partenaires, on peut « jouer » avec ces chorèmes, les combiner pour tester des hypothèses et élaborer des scénarios d'évolution.

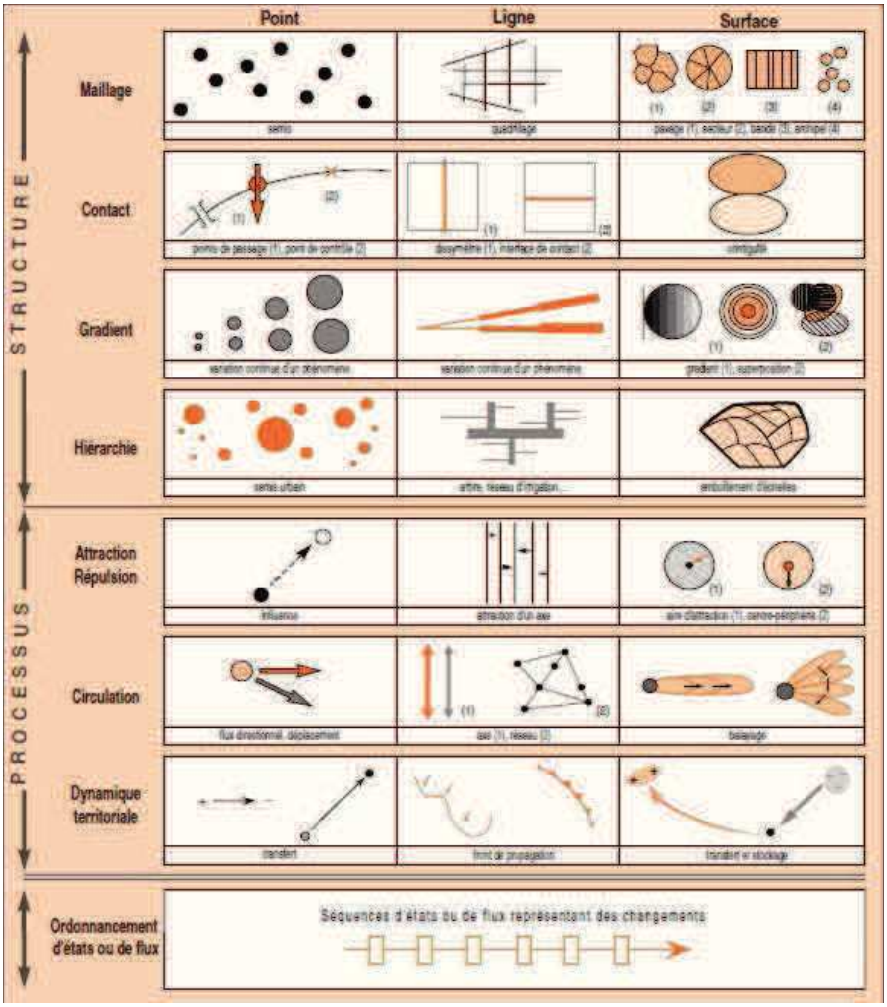


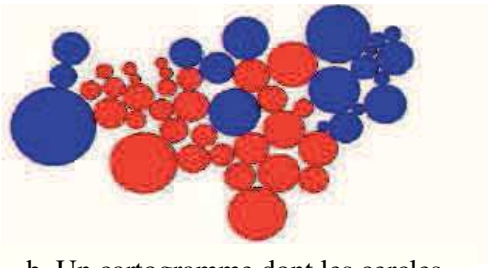
Figure 2.11 Les chorèmes de Cheylan [25].

2.4.2 Les cartogrammes : résumés visuels d'une table de base de données géographiques

Les premiers cartogrammes connus datent du XIXe siècle, ils étaient conçus à la main. A partir des années 1960, le cartographe Waldo Tobler [http 32] a développé et appliqué des méthodes informatiques de génération de ces résumés visuels.

Les *cartogrammes*, appelés communément *anamorphoses cartographiques*, offrent une représentation cartographique très intéressante permettant de figurer les diverses valeurs prises par une variable et faire facilement des comparaisons visuelles entre les territoires. La géométrie de l'espace de la carte est déformée afin de se conformer aux informations relatives à la variable représentée. Ce type de carte ne représente alors plus la réalité géographique, mais la réalité du phénomène. Ces anamorphoses sont donc utilisées en cartographie statistique pour montrer l'importance d'un phénomène donné.

a. Carte conventionnelle des USA montrant le résultat des sélections présidentielles



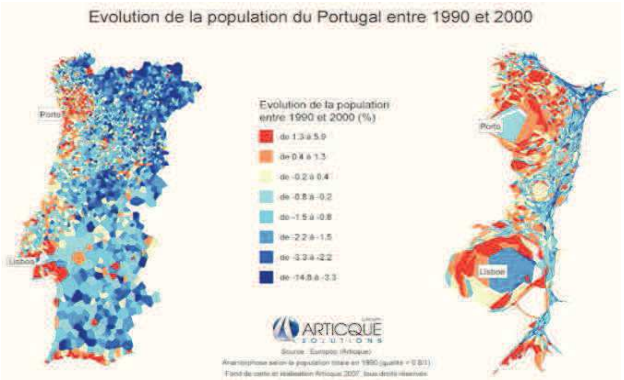
b. Un cartogramme dont les cercles sont proportionnels à la population plutôt qu'à la surface du territoire

Figure 2.12 Différence entre carte conventionnelle et cartogramme [31].

La figure 2.12b présente un exemple d'utilisation d'un cartogramme pour schématiser le résultat des sélections présidentielles des USA proportionnellement à la population plutôt que du territoire. La simplicité d'analyse de cet événement à travers le cartogramme est bien démontrée par rapport à la carte conventionnelle correspondante dans la figure 2.12 a. Nous pouvons réaliser de plus des cartogrammes pour percevoir les effets des réseaux de transport car les moyens de transports déforment l'espace.

2.4.2.1 Classification des cartogrammes

Nous distinguons principalement deux catégories de cartogrammes : les cartogrammes de surface et les cartogrammes de distance. La figure 2.13 présente deux exemples de ces cartes.



a. Cartogramme de surface du Portugal

b. Cartogramme de distance de la France

Figure 2.13 Cartogramme de surface et cartogramme de distance [31].

- *Cartogramme de surface* (Figure 2.13 a) : Nous appelons parfois un cartogramme de surface une carte à valeur-par-surface ou carte « isodémographique » dans le cas de cartogramme de population. Celui-ci représente la taille des différents pays du monde (ou régions/territoires d'un même pays) en dimensionnant la surface de chaque pays proportionnellement à sa population.
  La forme et la position relative de chaque pays sont conservées dans la mesure du possible, mais des déformations ou distorsions plus ou moins importantes, apparaissent inévitablement.
- *Cartogramme de distance* (Figure 2.13 b) : Parfois désigné par cartogramme à point central, ce type de cartogramme est généralement utilisé afin de représenter des temps de trajet relatifs et des directions dans un réseau (transport, communication, informatique, etc.).

Denain et Langlois ont défini deux types d'anamorphose [29] :

- Dans un premier cas, l'anamorphose se rapporte à un phénomène associé à des pôles vectoriels. Parmi les rares applications de cette méthode utilisée en géographie, nous pouvons citer celle de l'accessibilité dans un réseau de transport. Nous cherchons à exprimer les disparités de vitesse depuis une ville de référence (le pôle principal) vers les autres villes du réseau (les pôles secondaires), c'est l'anamorphose unipolaire. Sur chaque pôle secondaire, nous définissons un vecteur de déformation qui le rapproche du pôle principal si sa vitesse est supérieure à la moyenne et l'éloigne dans le contraire. Ce sont des vecteurs qui sont utilisés dans l'anamorphose pour réaliser la déformation de tout l'espace, en tout point de la carte. L'anamorphose multipolaire résulte de l'intégration des différents calculs unipolaires sur tous les pôles du réseau, donnant ainsi une vision globale de l'accessibilité de l'ensemble du réseau.
- Le deuxième cas est l'anamorphose scalaire qui permet de cartographier un phénomène décrit par les données scalaires (variable quantitative à une dimension), associées à des pôles comme une population affectée à un centre de zone. Cette déformation repose sur la dilatation ou la contraction de la surface au voisinage d'un pôle en fonction de la valeur qui lui est attribuée.

Il est possible ensuite d'envisager deux types de visualisation statique ou dynamique. Dans le premier cas, seul le résultat final est présenté (avec éventuellement la forme initiale en second plan). En mode dynamique, nous visualisons la déformation de manière continue, par animation ou par une série de cartes intermédiaires depuis la forme initiale jusqu'à la déformation finale.

2.4.2.2 Les algorithmes développés pour les cartogrammes

La première publication scientifique de Tobler proposant un algorithme pour réaliser une anamorphose date de 1973. D'autres travaux postérieurs ont été effectués en vue de produire des algorithmes plus efficaces.

Une des méthodes proposées est celle de Dougenik, Chrisman et Niemeyer en1985 [32]. Elle consiste à exercer des forces partant du centre du polygone (centroïde) vers les points définissant sa bordure. La distance du centroïde de polygone au point le définissant est pris en compte dans la transformation. Ces forces représentent l'écart entre la surface initiale du polygone et la surface qu'il devrait avoir si toutes les surfaces étaient proportionnelles à la quantité à représenter :

- Si la surface d'origine est trop petite par rapport à la quantité à représenter, la force repoussera les points et agrandira l'entité spatiale ;

- Si la surface d'origine est trop grande par rapport à la quantité à représenter, la force attirera les points et réduira l'entité spatiale. La transformation préserve les contiguïtés des entités spatiales et s'effectue par étape ou itération.

Une autre méthode proposée et qui est la plus couramment utilisée est celle de Gastner Newman (2004) [http 33]. Elle est fondée sur le processus physique de la diffusion de la chaleur (diffusion linéaire). Les principales caractéristiques de cette méthode sont les suivantes :

- Calcul de densité dans une grille régulière donnée qui est progressivement déformée ;
- Fonctionnement par itérations ;
- Rapidité, économie en temps de calcul ;
- Préservation correcte de la topologie (sans superpositions fortuites).

2.4.2.3 Outils de création des cartogrammes

Nous citons des nombreux outils permettant la création de cartogrammes. Il est possible de classer ces outils en deux catégories :

- Les outils indépendants par exemple:
  - **Cart** (Computer software for making cartograms) est développé en C par Mark Newman à partir de son propre algorithme. Le logiciel et les sources sont accessibles en ligne [http 10]. Les formats d'entrée sont peu commodes. Il n'est pas facile à utiliser.
  - **Scape Toad** [http 11] est un Logiciel libre pour effectuer des anamorphoses. C'est une application efficace et performante développée en Java. Cette application travaille avec des fichiers `shape` (ESRI) en entrée, et permet de générer du SVG (Scalable Vector Graphics) ainsi que le fichier `shape` déformé.
- Les outils intégrés dans des applications sont par exemple :
  - QGIS cartogram creator qui est un Plugin QGIS permettant de réaliser des anamorphoses.
  - MAPresso est une applet java permettant de générer des anamorphoses. Elle est Facile à mettre en place, publique et Opensource.
  - Cartograms with d3 & TopoJSON est une bibliothèque Javascript facile à implémenter pour un site web. Elle est assez spectaculaire et rapide mais l'algorithme utilisé n'est pas le plus performant.

2.4.2.4 Exemples des cartes en anamorphoses

Ci-dessous des exemples de cartes anamorphoses qui montrent quelques phénomènes de façon originale.

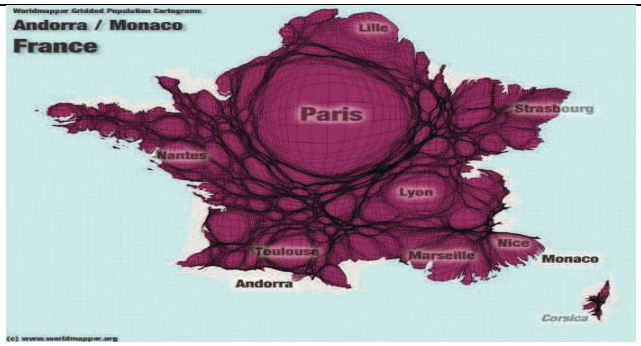
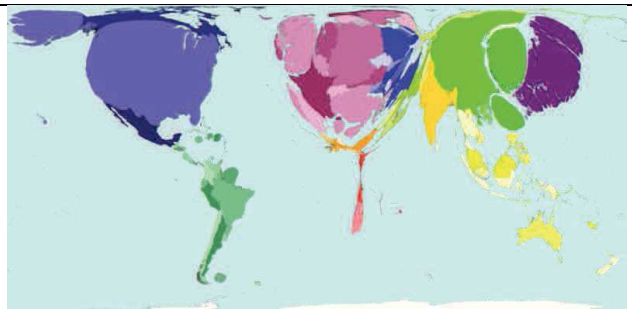
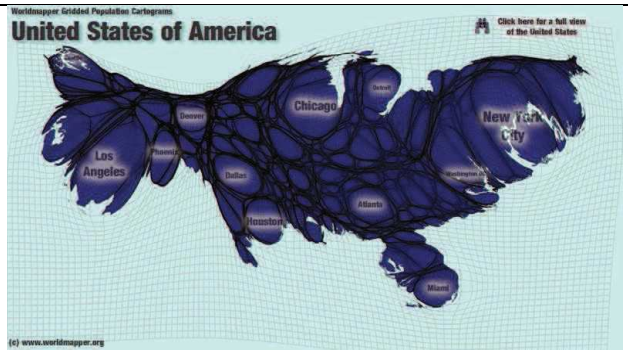
Cartogrammes	Phénomènes décrits
	Répartition de la population française
	Les utilisateurs d'internet en 2002
	La population aux USA

Tableau 2.3 Quelques exemples de cartogrammes [http 12].

2.4.3 Synthèse

Nous avons présenté deux types de résumés visuels : Les chorèmes qui sont des résumés de bases de données géographiques et les cartogrammes qui sont des résumés d'une table de BDG. Le tableau 2.4 présente une comparaison entre ces deux types de résumés visuels.

	Intérêts	Limites
Les Cartogrammes	• C'est une représentation cartographique innovante. • La perception de la quantité est très bonne. • Présente une image très généralisée qui rend bien compte des gradients. • Provoque, suscite l'intérêt et véhicule un message fort. • Présente des cartes de communication qui attirent l'attention sur un phénomène.	• Les cartogrammes dépendent des statistiques définit un obstacle à la génération des images cartographiques. • Les cartogrammes demandent un effort de lecture et ils ne permettent pas de connaître les situations locales. Aussi il y a une perte des repères visuels en regardant les cartogrammes (difficile de retrouver son pays, ou sa région sur la carte). • Les cartogrammes permettent la gestion des données manquantes. • Les cartogrammes selon des variables "sociales" souffrent en effet de trois défauts majeurs : - leur inconstance dans le temps - leur dépendance statistique - la technicité de leur production • Les cartogrammes ne sont pas faciles à produire, et requiert l'usage d'un ordinateur et de logiciels spécialisés, peu diffusés et encore difficiles à utiliser.
Les Chorèmes	• La chorématique part du simple et de l'élémentaire pour construire le complexe et la complexité. Son intérêt réside dans la simplicité logistique • La chorématique vise à mettre en place les règles de lecture des structures spatiales sans chercher un langage très spécialisé qui n'est accessible qu'aux initiés • Les chorèmes peuvent décrire plusieurs phénomènes à la fois, à l'inverse des cartogrammes,	• Les formes géométriques sont trop schématiques • L'accumulation des faits et des mécanismes engendrent des confusions et des ambiguïtés rendant ainsi la lecture de l'espace une tâche difficile • L'absence de localisation précise. • La négligence des influences particulières humaines sur le territoire au profit de forces structurantes extérieures • Le non existence d'outils logiciels ou d'applications qui permettent de créer automatiquement des cartes chorématiques.

Tableau 2.4 Comparaison entre les chorèmes et les cartogrammes

2.5 Conclusion

Dans ce chapitre, nous avons présenté le contexte de notre travail de recherche et les éléments sur lesquels il est basé, nous avons commencé par une étude sur les SIG par la suite les FD et FDS et nous avons terminé par un état de l'art sur les résumés visuels de base de données : les chorèmes et les cartogrammes, dans le chapitre qui suit, nous allons continuer à parler des chorèmes en faisant une étude dequelques cartes chorématiques.

Chapitre 3

Résumés visuels de base des données géographiques

3 Chorèmes : Résumés visuels de base de données géographiques

Dans ce chapitre, nous présentons l'importance des cartes chorématiques et leurs domaines d'application, ainsi qu'une étude sur quelques chorèmes ainsi que le langage qui permet leur description.

3.1 Introduction

Alors que les données géospatiales sont des informations portant sur des objets et événements situés sur la surface terrestre, les chorèmes [11] sont des représentations schématiques des territoires et représentent la structure et l'organisation de ces territoires. Dans ce chapitre nous présentons quelques domaines d'application des chorèmes, nous présentons aussi un langage de description des chorèmes.

Rappelons que les chorèmes constituent un vocabulaire visuel permettant la description des caractéristiques principales d'un territoire; et de notre point de vue, ils sont une base solide et efficace pour la prise de décisions, parce qu'ils font ressortir les aspects les plus significatifs, en laissant de côté des problèmes secondaires.

Disons que lorsqu'il est nécessaire de comprendre la structure d'un territoire, une carte complète à l'échelle n'est pas utile, alors qu'un petit schéma peut être plus utile. Alors les cartes chorématiques sont un outil clé pour schématiser un territoire, et qui permet aux décideurs d'avoir une vue plus claire de la situation.

Parmi les applications, citons la visualisation schématique pour :

- les principaux problèmes politiques, économiques et démographiques,
- les principales caractéristiques de l'environnement et la climatologie,
- l'évolution principale en épidémiologie,
- les risques naturels et technologiques ou de catastrophes, etc.

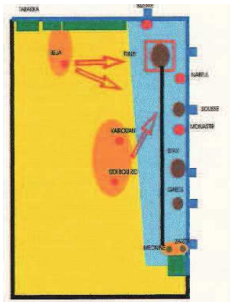
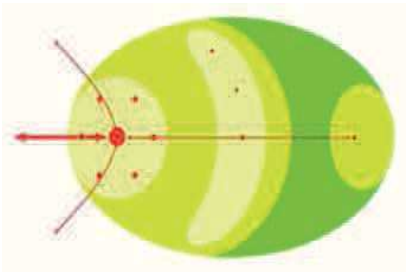
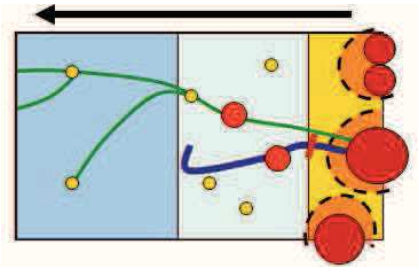
Nous avons remarqué que presque toutes les présentations chorématiques sont inspirées des chorèmes de Brunet [11] et peu de ceux de Ducret [34] et Cheylan [25], en ajoutant quelques modifications pour qu'elles puissent être appliquées chacune dans son domaine. Pour cela nous avons répertorié les chorèmes de Brunet utilisés dans une sélection de quarante cinq cartes chorématiques dans les tableaux ci-dessous.

3.2 Etude de quelques cartes chorématiques

Nous analysons quelques chorèmes dans les tableaux au dessous. Dans le premier tableau, nous avons présenté une vingtaine de cartes chorématiques (Tableau 3.1, A 1.1) que nous avons sélectionnée parmi quarante cinq cartes que nous avons étudiées dans les tableaux A1, A2, A3, A4 et A5 dans l'annexe A, et dans le tableau 3.2, nous identifions les chorèmes les plus utilisés après l'étude.

• Une sélection de cartes chorématiques contenant les chorèmes de Brunet

Nous présentons à travers le tableau 3.1 cinq cartes chorématiques parmi un ensemble de cartes étudiées (cf .Annexe A).

Chorème	Explication
LA TUNISIE [58] 	Cette carte chorématique contient quatre chorèmes qui présentent l'espace tunisien inspiré des chorèmes de Brunet : les villes les plus importantes, les flux de migration, les frontières, les ports et une partition du territoire tunisien en trois grandes zones (Sud, intérieur et littoral).
CARPENTRAS [42] 	Cette carte chorématique présente l'organisation du Ventoux Comtat Venaissin. C'est une représentation schématique (inspirée des chorèmes de Brunet) du bassin d'emploi de Carpentras, qui se compose de trois chorèmes : les chefs lieux (les centres urbains), des axes de communication entre ces derniers et un chorème qui partage Carpentras en trois zones (des espaces périurbains, des espaces agricoles profonds, et des espaces naturels).
LA CHINE [62] 	Cette carte chorématique présente une politique d'aménagement du territoire de "Go West" en chine à partir des différents centres moteurs du littoral vers les périphéries de la "Chine de l'intérieur" et la "Chine de l'Ouest" en s'appuyant sur les villes continentales d'aménagement prioritaire et grâce à la construction d'aménagements ferroviaires et fluviaux spectaculaires.

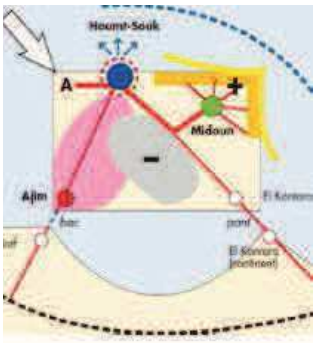
LA FRANCE [61] 	Cette carte chorématique présente l'organisation du territoire français. Elle est composée de trois chorèmes : des chefs lieux (les foyers de puissance et de développement), un axe de mégalopole européenne, des principaux sites et stations touristiques.
D J E R B A - LA TUNISIE 	Cette carte chorématique contient neuf chorèmes : les limites administratives de la ville de Djerba, des chefs-lieux qui représentent des centres importants, l'aéroport, etc., des points de passage aux frontières, des sous-ensemble qui décrivent les activités traditionnelles et la dépression centrale, des liaisons préférentielles entre les centres importants et les frontières, une tête de pont de l'ancien centre maritime, des ruptures qui décrivent l'émigration et l'immigration, une base pour le centre touristique et des évolutions ponctuelles pour décrire le pôle attractif et répulsif.

Tableau 3.1 Exemple de cartes chorématiques utilisant les chorèmes de Roger Brunet

3.2.1 Les chorèmes de Brunet dans les cartes chorématiques

Afin d'identifier les chorèmes les plus souvent utilisés, donc les plus importants, nous avons étudié plusieurs cartes chorématiques et nous avons répertorié pour chaque carte chorématique, les chorèmes de Brunet utilisés, cette étude est présentée dans les tableaux en annexe (Annexe 1).

3.2.2 Les chorèmes de Brunet les plus utilisés

Après une étude de plusieurs cartes chorématiques, nous avons pu identifier les chorèmes de Brunet les plus utilisés par les géographes. Ces chorèmes sont proposés au tableau 3.2.

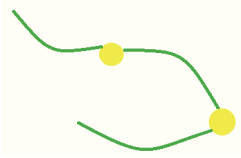
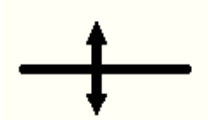
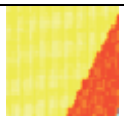
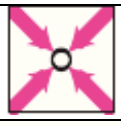
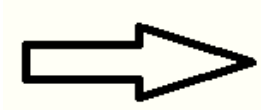
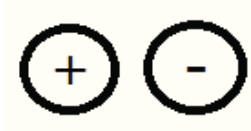
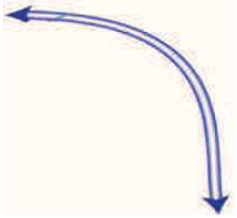
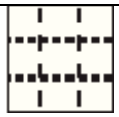
Points			
Limites administratives			
Lignes et points de passage			
Contact			
Tropisme			
Dynamique territoriale			
Hiérarchie Réseau			

Tableau 3.2 Les chorèmes de Brunet les plus utilisés.

Cette étude a aussi montré que les géographes peuvent utiliser les chorèmes de Brunet les plus importants (cf. tableau 3.2), ou créer leurs propres chorèmes selon leurs besoins. C'est pour cette raison que l'on trouve plusieurs types de chorèmes, et qu'il n'existe pas une définition formelle donnant l'ensemble des chorèmes.

De toute façon, nous faisons le constat que les chorèmes sont largement utilisés dans de nombreux domaines puisqu'ils expriment tout ce que l'on veut décrire d'une manière simple et facile à lire.

3.3 Le langage de description des chorèmes

Le Langage ChorML (*Chorem Markup Language*) [81] a été crée par André Coimbra à l'INSA de Lyon. C'est un langage de type XML avec la fonctionnalité de mémoriser l'information des chorèmes et de permettre la communication de telles informations entre les différents modules du système ChEVIS.

C'est la combinaison de XML et de GML [http 27]. Il spécifie les résultats des algorithmes de fouille des données spatiales. Les éléments du langage sont en particulier :

- Les informations de caractère général : l'identificateur de la carte, le nom de la carte, le nom de l'auteur, la date de création, le système de référence, le nom de la base des données originale et la dernière date de mise à jour, etc.
- La liste des chorèmes dans laquelle les données géographiques sont codées en GML.
- Une pré-légende, qui contient une description en format texte de chaque chorème.
- La liste des relations topologiques et non-topologiques parmi les chorèmes.

Le ChorML qui a été conçu pour stocker et échanger les chorèmes.

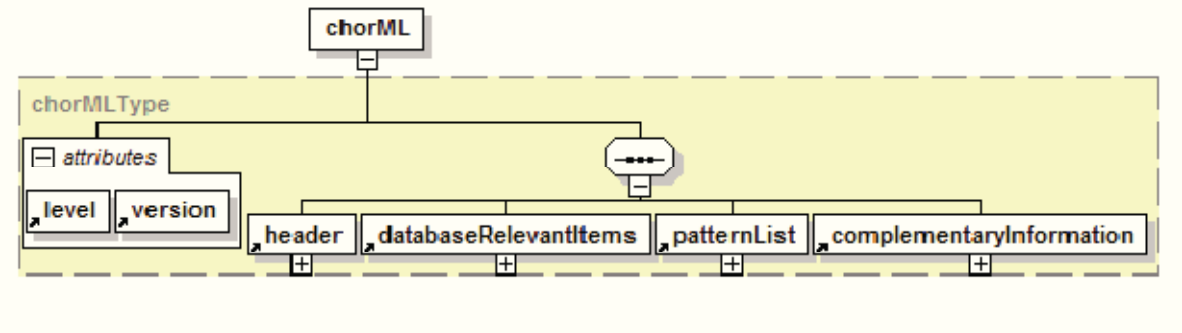


Figure 3.1 Structure d'un document ChorML [81].

Le ChorML est un langage de type XML, multi-niveaux, qui a été défini avec la fonctionnalité de mémoriser l'information des chorèmes et de permettre la communication de telles informations entre les différents modules du système. Ces niveaux sont décrits comme suit : (i) *le niveau 0* correspond à la base des données géographique initiale avec les métadonnées associées, (ii) *le niveau 1* est celui des chorèmes extraits, (iii) *le niveau 2* est celui de la visualisation de chorèmes.
De manière plus précise :

- *Le niveau 0* est composé de XML et GML. Il a été pensé pour stocker les informations sur les proto-chorèmes, l'origine des données et les fonctions appliquées afin d'obtenir les chorèmes.
- *Le niveau 1* de ChorML est également une combinaison de XML et de GML. Il spécifie les résultats des algorithmes de fouille des données spatiales.
- *Le niveau 2* de ChorML est une combinaison de XML, et de SVG.

A titre d'exemple dans le tableau 3.3, on montre comment une ville est représentée à travers ces trois niveaux de langage.

Niveau	Représentation d'une ville
Niveau 0	Surface caractérisée par les coordonnées de son centroïde (longitude/ latitude).
Niveau 1	Point de longitude/latitude avec son importance.
Niveau 2	Point avec des coordonnées en pixels, représenté par un cercle avec son rayon et sa couleur.

Tableau 3.3 Représentation des villes à travers les niveaux de ChorML.

3.3.1 Les structures des motifs dans ChorML

Dans [48] et [86] ont été identifiés plusieurs types de modifications ou de connaissances géographiques : les Faits, les Clusters (regroupements géographiques), les Flux et les Co-localisations.
Dans cette section, nous nous proposons de les examiner :

- **Les faits**

Un fait est considéré comme le résultat d'une ou plusieurs requêtes exécutées sur la base de données. Un ensemble de règles est défini dans le but d'obtenir des informations élémentaires de la base de données. Ces informations vont être représentées comme des faits dans ChorML.

La structure de l'élément « fait », comme le montre la figure 3.2, possède comme éléments enfants la requête exécutée et la liste résultante d'éléments. Chaque élément a un nombre d'association qui sera utilisé pour affecter un degré d'importance à l'élément. Afin de stocker les coordonnées et la géométrie des objets, nous avons défini l'élément `_SHAPE_`.

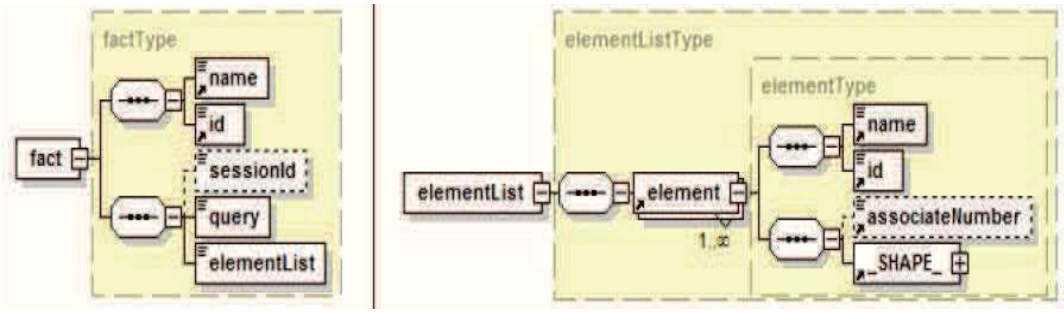


Figure 3.2 Les éléments en ChorML de fact et elementListType [81].

Cet élément est du type `gml:abstractGeometricPrimitive` donc peut être n'importe quel type d'objet géométrique défini dans GML comme des points, des courbes et des polygones. Par exemple, un fait pourrait être constitué d'un ensemble de bureaux de poste dans une ville ou la capitale d'un pays.

- **Les regroupements géographiques (Cluster)**

Le regroupement est la méthode utilisée pour regrouper les données en classes, par conséquent, un objet dans un cluster possède certaines similitudes avec d'autres objets dans le même cluster. Par exemple, on pourrait regrouper des parcelles d'une ville par le

type d'utilisation des sols, ou regrouper des régions par leurs similitudes météorologiques [56].

Les clusters sont de bons candidats pour générer des chorèmes**,** Le principal élément `ClusteringModel`, (cf. figure 3.3), est utilisé pour stocker les résultats du processus de regroupement.

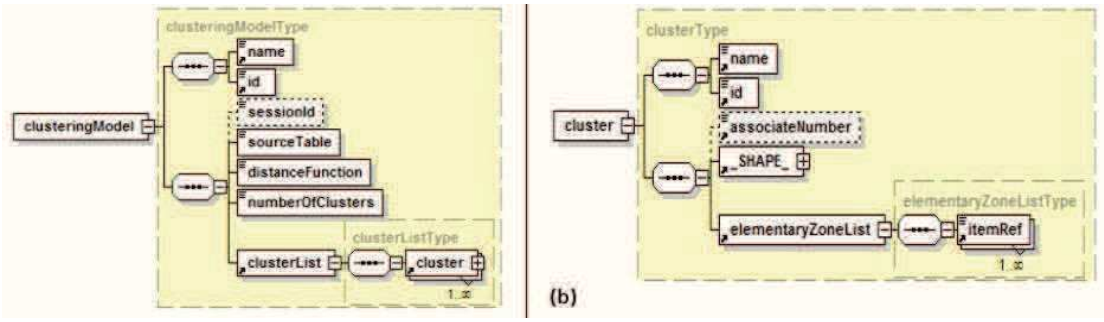


Figure 3.3 : Les éléments en ChorML de `clusteringModelType` et `clusterType` [81]**.**

La représentation des éléments est basée sur celle qui est utilisée dans PMML [81] bien que PMML ne comprenne pas de représentation spatiale.

Afin de s'appuyer sur la représentation spatiale, nous avons inclus l'élément `_SHAPE_` défini dans la section précédente. Chaque cluster est composé par un ensemble de zones élémentaires où chacun est un `itemRef`**.**

L'élément `itemRef` est une référence à une entrée stockée dans `database RelevantItems`, donc la liste des zones élémentaires qui composent un cluster fait référence à des zones telles que définies dans la base de données. Afin de s'assurer de la complétude du modèle, nous devons nous assurer que toutes les zones élémentaires sont associées à un cluster et un seul cluster.

- **Les flux (flow)**

Les flux sont utilisés pour représenter la dynamique spatiale dans un territoire. « Nous considérons comme des flux chaque mouvement matériel ou immatériel, de biens, de personnes, d'informations, entre les différents endroits ».

Les flux sont généralement représentés par des flèches dans la cartographie commune, bien que Brunet ajoute de nouveaux éléments pour les représenter à l'aide des chorèmes, des lignes de contact et le tropisme.

Une étude a montré que trois types de flux sont nécessaires : flux de trajet (`pathFlow`), flux de source divergente (`sourceDivergingFlow`) et flux de puits convergent (`sinkOrientedFlow`) (cf. figure 3.4).

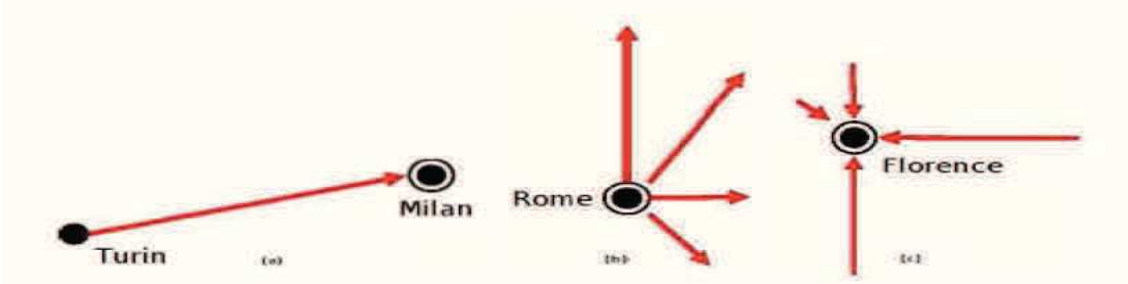


Figure 3.4 Exemples de flux : a. Trajet, b. Source divergente, c. Puits convergent [80].

Le **flux de trajet** représente un flux où l'origine et la destination sont bien définies et il peut éventuellement avoir une forme géométrique (par exemple une grosse flèche). Un flux de source divergente a une origine bien définie, mais la destination est un peu incertaine : la destination est dans ce cas une liste de directions géographiques, comme le Nord ou le Sud-Est; on pourra utiliser ce formalisme par exemple pour illustrer le lien de départ d'émigration de l'Italie vers d'autres pays. Un flux de puits convergent possède une destination bien définie mais l'origine est une liste de directions géographiques convergentes, par exemple l'immigration vers les Etats-Unis.

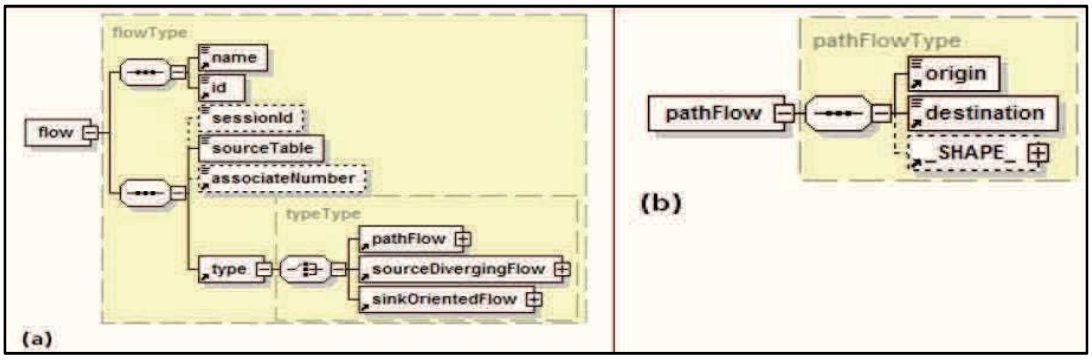


Figure 3.5 Structure du motif de `flowType` et `pathFlowType`
a. Structure de flux de source divergent, b. Flux de puits orienté [81].

Quelles que soient ces catégories, les éléments origine et destination se réfèrent à un élément de chorème ou à un élément de la base de données. En conséquence, la provenance et la destination seront toujours définies comme dit précédemment, il n'est donc pas possible d'avoir un flux sans origine ni destination. Un exemple de flux de trajet peut être le flux de marchandises de Turin à Milan, comme illustré dans la figure 3.4 (a).

- **La Co-localisation**

Les motifs de Co-localisation sont des ensembles de caractéristiques de lieux qui sont présumés être proches les uns des autres avec une certaine probabilité.

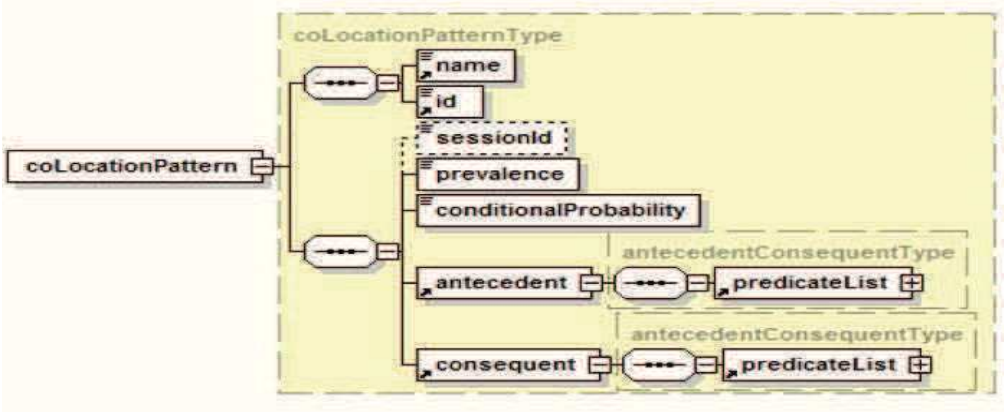


Figure 3.6 Structure du motif de Co-localisation [81].

3.3.2 Le système de génération du langage ChorML : ChorML Generator

Une première version du système ChorML Generator avait été crée par nous mêmes [19, 20]. Rappelons-en les grandes lignes.

Notre système permet de lancer des requêtes SQL ou des procédures PL/SQL afin d'extraire des données géographiques de la base et transformer en ChorML. Il permet aussi de coder des informations extérieures qui ne sont pas enregistrées dans la base comme le nom de la mer, des pays voisins, etc.

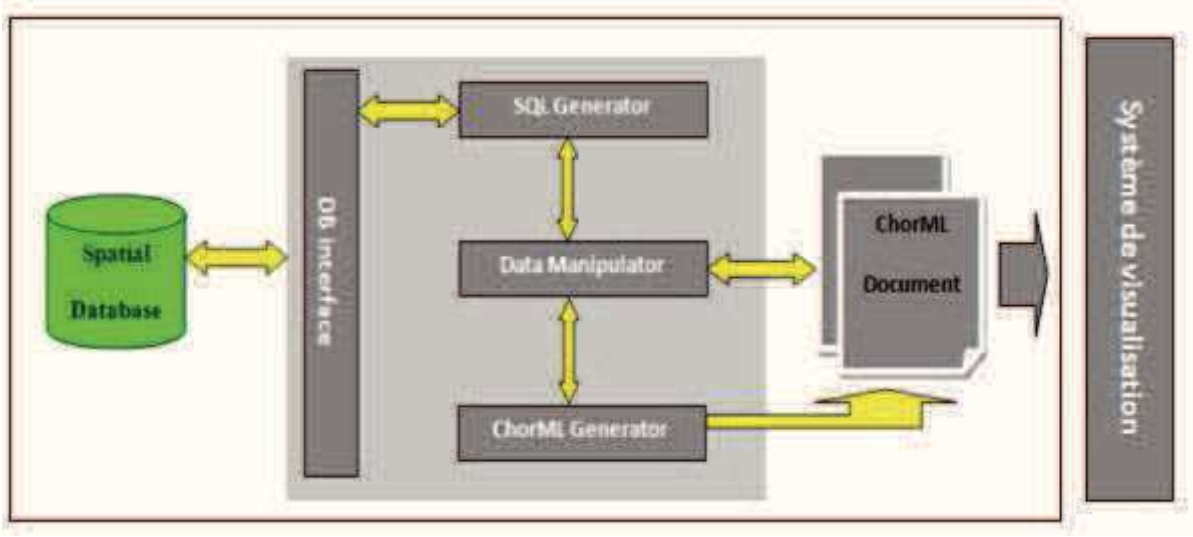


Figure 3.7 Architecture du ChorML Generator [19].

Le ChorML Generator est le premier système qui permet de générer le ChorML vu la difficulté de codage des requêtes lancées c'est pour cela nous avons essayé de créer un système qui permet de coder un nombre plus important de requêtes. Le système offre trois tâches :

- **Génération** : c'est à travers ce module que les requêtes seront lancées et exécutées. En fait SQL Generator assure (1) la préparation et l'exécution de la requête SQL appropriée et (2) la récupération du résultat et stockage des données dans une structure bien déterminée.

- **Manipulation** : la tâche principale de ce module est de récupérer le résultat du SQLGenerator et d'injecter les données dans le module ChorML Generator. Il permet aussi de faire des modifications sur le ChorML déjà généré. Ce dernier est composé de trois méthodes : (1) Une méthode qui permet de récupérer un résultat d'une requête d'un type donné (sql/plsql) afin d'appeler les méthodes nécessaires au module ChorMLGenerator, elle prend comme paramètres une requête, un nom du fichier pour stocker le résultat de génération et un type de requête. (2) Une méthode qui permet de récupérer le résultat d'une requête cluster afin d'appeler les méthodes de génération de résultats correspondantes au type de requête. (3) Une méthode de génération de contraintes qui seront intégrées dans le fichier final ChorML.

- **ChorML Generator** : ce module a pour objectif la construction d'un document XML selon la spécification ChorML. Ce module est un composant essentiel dans l'application développée car il permet la génération de document XML selon la spécification ChorML.

3.4 Conclusion

Nous avons proposé dans ce chapitre une solution s'appuyant sur le concept de chorèmes et sur leur capacité à résumer des situations impliquant des objets statiques et des phénomènes dynamiques, en les associant à des notations visuelles. Cela constitue une synthèse immédiate des données pertinentes et donne aux utilisateurs experts un aperçu global des objets et des phénomènes ainsi qu'une approche dans une situation spécifique.

Une première étude a été menée pour évaluer les différentes solutions. Parmi celles-ci, le chorème semblait être le plus pratique, grâce à son expressivité et sa capacité de synthèse. L'idée principale est de produire interactivement des cartes simplifiées avec les caractéristiques les plus importantes pour l'utilisateur expert. Les cartes sont obtenues par une analyse géographique (comme la fouille de données et la fouille de données spatiales), et restituées par une visualisation inspirée de la définition de Chorème.

Cette solution offre aux utilisateurs experts, les motifs décrivant des positions, des mouvements et des faits faciles à comprendre et à expliquer aux utilisateurs non experts intéressés par les mêmes questions. Dans le chapitre qui suit nous décrivons notre approche proposée qui permet l'extraction des motifs importants et les visualiser en carte chorématique.

Chapitre 4

Méthodologie d'extraction et de visualisation des chorèmes

4 Méthodologie d'extraction et de visualisation des chorèmes

Dans ce chapitre, nous présentons notre méthodologie pour l'extraction des motifs à partir d'une base de données géographiques afin de générer des cartes chorématiques.

4.1 Introduction

Rappelons que l'objectif principal de notre travail de recherche est de mettre au point une méthode qui permet la génération de résumés visuels, inspirés par les chorèmes, à partir d'une base de données géographiques. Sur cette base de données, nous appliquons des algorithmes de fouille de données spatiales pour permettre tout d'abord l'extraction des faits les plus saillants, puis la visualisation de ces faits sous forme de cartes chorématiques.

Après une étude de différents algorithmes de fouille de données, notre choix s'est fixé sur l'algorithme *k*-means et sur celui de l'algorithme Apriori pour développer notre propre algorithme d'extraction des patterns.

4.2 Méthodologie d'extraction des motifs

Notre méthodologie a pour objectif d'extraire les motifs qui servent à construire les résumés visuels de base des données comme illustré dans la figure 4.1. Ces motifs sont les suivants :

- Les *Clusters (regroupements géographiques)*
- Les *Faits*
- Les *Flux*
- Les *Co-localisations.*

Sans oublier deux éléments de nature un peu différente qu'il est nécessaire d'expliciter :

- Les *contraintes topologiques*
- Les *informations extérieures*.

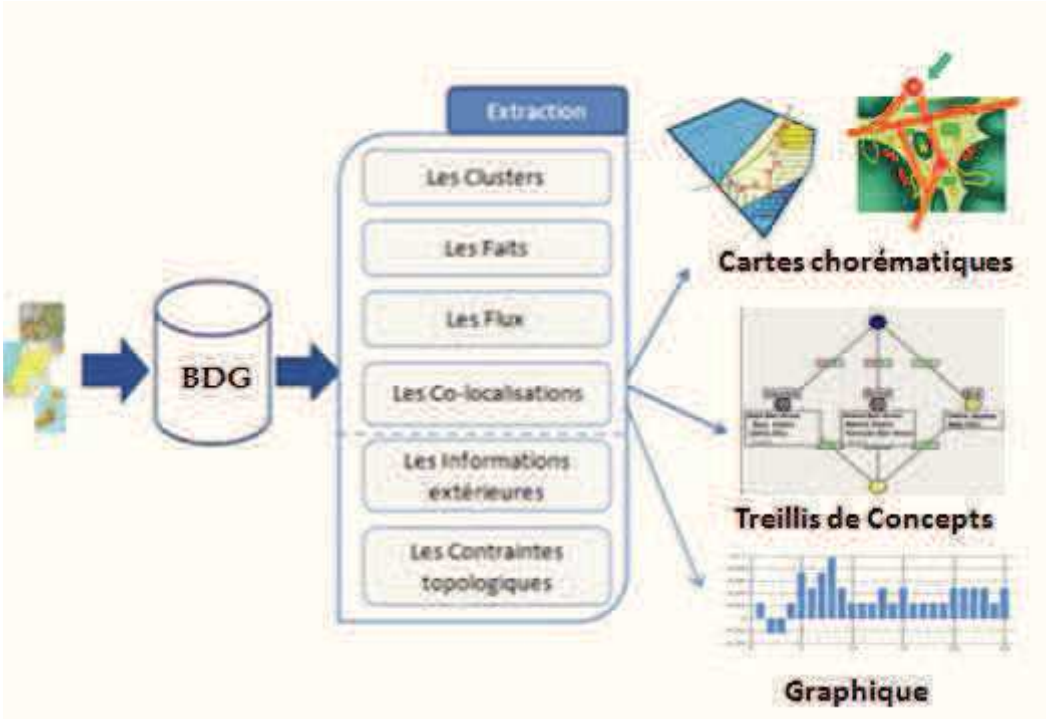


Figure 4.1 Extraction des motifs pour la génération des chorèmes.

Pour extraire ce genre de motifs, il nous faut partir d'une base de données géographiques décrite de manière relationnelle sous forme de n relations ; ces relations pourront soit correspondre à des tables natives ou à des vues construites sur ces tables :

$T_i\ (A_{1i}, A_{i2}, A_{i3},.., A_{ij},.. A_{im})$
…
$T_n\ (A_{1n}, A_{n2}, A_{n3},.., A_{nj},.. A_{nl})$

Pour pouvoir en extraire des motifs géographiques, ces relations devront respecter les conditions suivantes:

- Il doit exister au moins un attribut A_{ij} de type géométrique (le type et les coordonnées de la géométrie, le type pouvant être : Ligne, Multi lignes, Point ou Polygone) noté CGeometry,
- Il existe au moins une relation T_0 qui décrive le territoire complet :

T_0 *(n°territoire, nom, CGeometry, etc.)*

Après avoir lancé les procédures d'extraction des connaissances géographiques, celles-ci pourront être stockées dans de nouvelles tables notées CGi ($A_{1i}, A_{i2}, A_{i3},.., A_{ij},.. A_{im}$) dont on donnera plus loin les structures détaillées.

4.2.1 Prétraitement des données

Les données collectées doivent être " préparées ". Avant tout, elles doivent être nettoyées puisqu'elles peuvent contenir plusieurs types d'anomalies : des données peuvent être omises à cause des erreurs de frappe ou de mesure ou bien alors à cause des erreurs dues au système lui-même ; dans ces cas il faut remplacer ces données ou les éliminer complètement.

Les données peuvent être incohérentes c'est-à-dire qui sortent des intervalles permis et on doit alors les écarter ou les normaliser. Dans notre cas, nous avons normalisé les données, en leur faisant subir une projection dans un intervalle bien précis. Le prétraitement comporte aussi la réduction des données qui permet de réduire le nombre d'attributs pour accélérer les calculs et représenter les données sous un format optimal pour l'exploration.

Une fois les données collectées, nettoyées et prétraitées, nous passons à la deuxième phase d'extraction des motifs.

4.2.2 Extraction des motifs *clusters* (regroupements géographiques)

Le regroupement est la méthode utilisée pour rassembler les données en classes. Par conséquent, un objet dans un cluster possède certaines similitudes avec d'autres objets dans le même cluster. Par exemple, on pourrait regrouper des zones par leurs similitudes météorologiques. Dans notre cas, les regroupements sont construits par voisinage géographique.

Pour extraire des clusters géographiques, il faut tout d'abord que le territoire soit une tessellation et qu'il existe ensuite deux relations :

- Une relation T_1 décrivant la zone

T_1 (n°territoire, n°Zone, CGeometry, etc.)

- Une relation T_m, décrivant le voisinage (relation topologique « touches »)

T_m(n°territoire, n°Zone, n°zone-adjacent, etc.)

Chaque cluster est composé par un ensemble de zones élémentaires. La liste des zones élémentaires qui composent un cluster fait référence à des zones telles que définies dans la base de données.

A l'issue de cette analyse, toutes les zones élémentaires doivent être associées à un cluster et à un seul cluster.

Les regroupements géographiques peuvent être composés d'une ou plusieurs zones et ils peuvent être présentés sous la forme :

CG_1 (n°territoire, n°cluster, CGéométrie, etc.)
CG_2 (n°territoire, n°cluster, n°zone-composante)

Ces clusters doivent former une tessellation et avoir les mêmes attributs.

Comme nous l'avons mentionné dans notre étude dédiée aux diverses méthodes de fouille de données spatiales, la méthode clustering spatiale et plus précisément l'algorithme *k*-means semble le plus approprié pour extraire des connaissances sous forme de motif *cluster.*

Grâce à la méthode *k*-means, nous pouvons sélectionner les zones qui sont très proches (proximité géographique) et qui partagent les mêmes caractéristiques dans la BD (proximité sémantique).

La méthode d'extraction des motifs de type *cluster* consiste à effectuer les étapes suivantes :

1. Considérer chaque zone sélectionnée à travers le filtre, comme une classe *C*
2. (Ré) affecter chaque zone *V* au cluster *Ci* de centre *Mi* tel que *dist(V,Mi)* soit minimal
3. Recalculer *Mi* de chaque cluster (le centre géométrique)
4. Aller à la deuxième étape si on vient de faire une affectation.

4.2.3 Extraction des motifs *Fait*

Un fait sera considéré comme le résultat d'une ou plusieurs requêtes exécutées sur la base de données. Selon Laurini [64]**,** un ensemble de requêtes est défini dans le but d'obtenir des informations élémentaires de la base de données.

Chaque élément *du fait* a un indicateur qui sera utilisé pour affecter un degré d'importance à l'élément. Nous pouvons avoir plusieurs types géométriques de *fait* : des points, des courbes et des polygones. Par exemple, un fait pourrait être constitué d'un ensemble de bureaux de poste dans une ville ou bien désigner la capitale d'un pays.

Pour extraire des faits, il nous faut des relations de cette forme :

T_l *(n°territoire, n°élément, Requête, Niveau-imp, CGeometry, etc.)*

Nous proposons la définition suivante pour un fait :

Un Fait Fa= {Fa1….Fa} / Ensemble des Faits
Fa_x *= {*$Fa_{1,x}$*….*$F_{a,p}$*} 4*$\leq x \leq p$ *: p* attributs d'un fait
Un Fait= {n° territoire, n°élément, Requête, Niveau-imp, CGeometry, etc.}

4.2.4 Extraction des motifs *Flux*

Les flux sont utilisés pour représenter la dynamique spatiale dans un territoire. Nous considérons comme flux, chaque mouvement matériel ou immatériel, de biens, de personnes ou d'informations entre les différents endroits.

L'étude proposée par [27] a montré que trois types de flux sont nécessaires : flux bipolaires, flux divergents et flux convergents, certains pouvant être mono-directionnels ou symétriques.

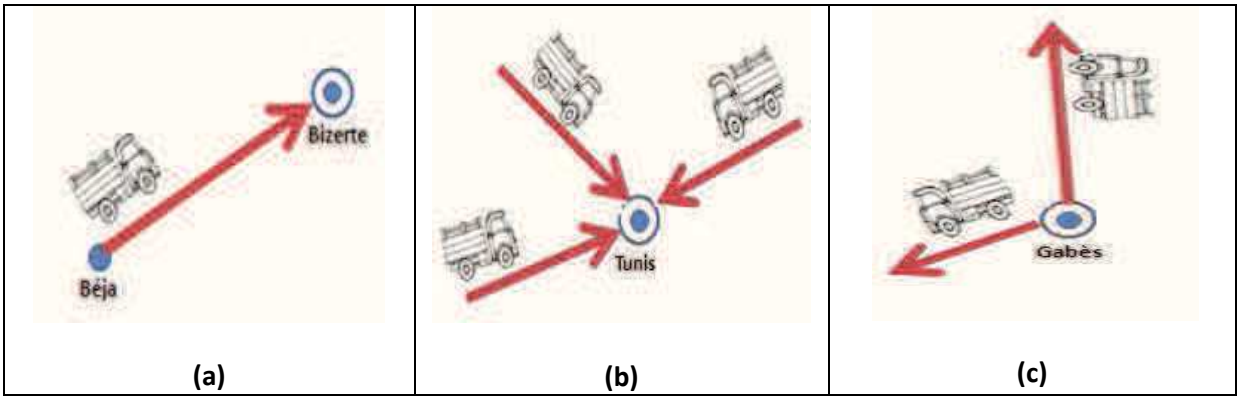


Figure 4.2 Exemples de flux : a. bipolaire, b. convergent, c. divergent.

Un exemple de flux bipolaire non symétrique peut être le flux de marchandises de Béja à Tunis, (figure 4.2 a). Un flux convergent pourrait être l'extension du système de transport public autour de Grand Tunis (figure 4.2b) et un flux divergent pourrait représenter la migration des habitants de la ville de Gabès vers d'autres régions (figure 4.2 c).

4.2.4.1 Les flux symétriques

Le *flux symétrique* représente un flux où l'origine et la destination sont bien définies et il doit avoir une valeur numérique : mathématiquement, on a une matrice carrée $n \times n$. Si les flux sont représentés en termes de pourcentage, il faudra reprendre les valeurs originelles. Il existe deux catégories, l'une où la matrice est complète, c'est-à-dire où l'on ne considère que des flux ayent les mêmes origines et destinations et l'autre où ce cas n'existe pas ; en d'autres termes la diagonale de la matrice est vide.

Si cette relation possède $n \times n$ tuples, alors la matrice est complète. Si l'on n'a que $n \times (n-1)$ tuples alors la diagonale est vide. Dans les autres cas, il s'agira de cas d'erreurs.

Ville O \ Ville D	TUNIS	GABES	SFAX	LE KEF	GAFSA
TUNIS		650	2500	100	120
GABES	1800		3800	54	20
SFAX	1200	1000		23	32
LE KEF	200	80	130		30
GAFSA	620	102	270	28	

Figure 4.3 Exemple de matrice des flux de migrations.

Soit le schéma relationnel suivant :

TabF (n° territoire, n°Zone $_{Orig}$, n°Zone $_{Dest,}$ valeur, autres attributs)

Où
n°Zone $_{Orig}$ est l'identifiant de la zone origine
n°Zone $_{Dest}$ est l'identifiant de la zone destination

Pour extraire les flux de trajet, il faut tout d'abord déterminer le point de départ et le point d'arrivée, qui peuvent être des villes ou des clusters. Nous proposons, ci-dessous, une définition logique d'un flux :

Un Flux F= {F1….Fa} / Ensemble des Flux
$F_x = \{F_{1,x} \ldots F_{a,x}\}$ $4 \leq x \leq a$: ***a attributs d'un flux***
Flux= { n° territoire, n°Zone $_{Orig}$, n°Zone $_{Dest,}$ CGeometry}

Dans le cas des clusters, il faut déterminer le centroïde.

4.2.4.2 Les flux divergents

Le *flux divergent* a une origine bien définie, mais la destination est un peu incertaine ; par exemple plus de 200 000 personnes par an quittent la Région parisienne pour aller s'installer en province. Dans ce type de cas, les destinations précises ne sont pas mentionnées et l'on peut créer une liste de directions géographiques, comme le Nord ou le Sud-Est.

Mathématiquement, on a une matrice $n \times m$; si les flux sont représentés en termes de pourcentage, il faudra reprendre les valeurs originelles.
A titre d'exemple, nous pouvons considérer quatre destinations possibles :

- Le Nord Est
- Le Nord Ouest
- Le Sud Est
- Le Sud Ouest

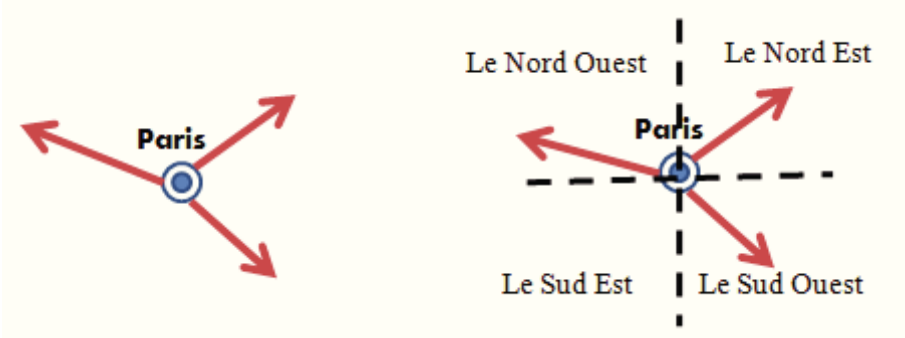


Figure 4. 4 Exemple de flux source divergente

Un autre cas possible est que les sources sont enregistrées dans d'autres tables extérieures ; dans ces deux cas, le flux peut être défini comme suit :

FluxD (n° territoire, n°Zone $_{Orig}$, n°Zone $_{Dest,}$ Destination-type, CGeometry, autres attributs}

D'où *n°Zone $_{dest}$* ou *Direction-type* peuvent être « NULL » au sens BD.

4.2.4.3 Les flux convergents

Le *flux convergent* possède une destination bien définie mais l'origine est une liste de directions géographiques convergentes, par exemple l'immigration vers la France. Dans ce type de cas, les origines précises ne sont pas mentionnées et l'on peut créer une liste d'origines géographiques, comme le Nord ou le Sud-Est.

Mathématiquement, on a une matrice $m \times n$; si les flux sont représentés en termes de pourcentage, il faudra reprendre les valeurs originelles.
A titre d'exemple, nous pouvons considérer quatre origines de flux possibles :

- Le Nord Est
- Le Nord Ouest
- Le Sud Est
- Le Sud Ouest.

Le flux peut être défini comme suit :

FluxC (n° territoire, n°Zone $_{Orig}$, n°Zone $_{Dest,}$ Origine-type, CGeometry, autres attributs}

D'où *n°Zone $_{Orig}$* ou *Direction-type* peuvent être « NULL » au sens BD.

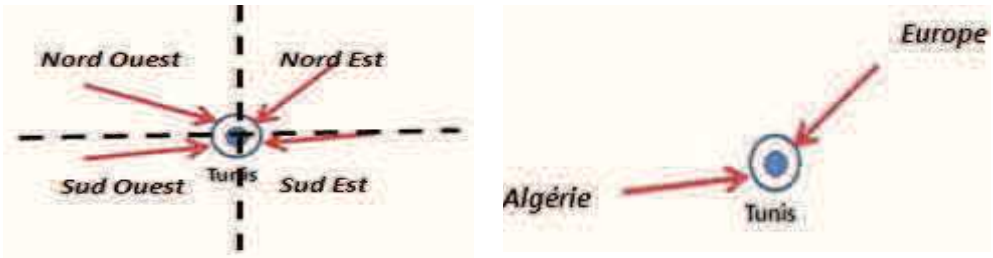


Figure 4. 5 Exemple de flux source convergente.

4.2.5 Extraction des motifs de *colocalisation*

La colocalisation[1] est la présence simultanée de deux ou plusieurs objets spatiaux au même endroit ou à des distances significativement proches.

La colocalisation est associée à deux actions d'intérêts définies: prévalence et probabilité conditionnelle. Les règles peuvent être écrites par un ensemble de prédicats (antécédent) qui mène à une conclusion (conséquent). Les prédicats sont représentés par l'élément *antécédent* et la conclusion comme *conséquent*. Ce qui s'écrit :

Antécédent ➔ *Conséquent*

Comme exemples de règles de colocalisation, on peut citer (les chiffres et les degrés de certitudes sont approximatifs) :

- "à côté de chaque grande ville, il y a un aéroport"
  - $\forall v \in$ villes, *nbreHabit (v)* >100000, $\exists a \in$ *aeroport : Dist (v, a) <50 km (80%)*
    ➔ *Grande ville avec aéroport (v, a) 80%*

- "la plupart des grandes villes au Canada sont à proximité de la frontière Canada-Etats-Unis"
  - $\forall v \in$ *villesCanadiennse, nbreHabit (v) >100000 ,* $\exists f \in$ *frontière : f(v)= «USA»* $\wedge$ *dist (v, f) < 100 km (80%)*
    ➔ *Grande ville canadienne proximité USA(v) 80%*

- "les grands restaurants de pizza pourraient en être colocalisés avec les grands magasins de vidéo"
  - $\forall r \in$ *Restaurants,* $\exists$ *r. spécialité = "PIZZA" ,* $\exists m \in$ *Magasin-vidéo: dist (r, m) > 100 m (70%)*
    ➔ *Grand restaurant de pizza à proximité magasins de vidéo (r,m) 70%*

- "dans les villes balnéaires, plus on se rapproche de la mer, plus les prix augmentent"
  - $\forall v \in$ *villes,* $\exists$ *v. type = "balnéaires"* $\exists$ *surface z1, surface z2*$\in$ *surface(v): dist (surface z1, mer) < dist (surface z2, mer), prix (surface z1)>prix (surface z2) (80%)*

[1] Il existe aussi le cas où deux objets peuvent être au même endroit, mais à deux dates différentes. Ce cas ne rentre pas dans notre problématique.

➔*Dans les villes balnéaires le prix au mètre carré augmente en se rapprochant de la mer 80%*

- "en Tunisie, plus on se rapproche des côtes, plus le nombre d'émigrants augmente"
  - *∀ v1,v2 ∈ villes ∃ , dist (v1,mer)> dist (v2, mer), nombreE (v1) > nombreE (v1) (80%)*

    ➔*Dans les villes tunisiennes le nombre d'émigrants augmente en se rapprochant de la côte 80%*

En définitif, les colocalisations trouvées seront stockées dans la table suivante :

Tab-Co (N° territoire, n°élément 1, n°élément 2, Distance, degré_de_certitude)

Les règles de colocalisation semblent intéressantes dans la création de chorèmes car elles définissent l'organisation des objets sur le territoire avec une précision quantitative, même si l'association des règles de colocalisation de chorèmes est encore une question d'étude à davantage approfondir.

4.2.6 Extraction des informations extérieures

Ce sont des éléments qui ne sont pas stockés originellement dans la base de données mais qui servent à mieux décrire la carte chorématique puisqu'ils donnent des informations complémentaires. A titre d'exemple, les noms des pays voisins ou des mers environnantes. En effet, on constate que la plupart des SIG se cantonnent aux données sous leur propre juridiction et l'extérieur leur est inconnu. Mais, pour comprendre un territoire, il est important de savoir si un morceau de limite correspondant à une frontière avec un autre pay, ou bien à la présence d'une mer ou d'un fleuve.

Dans ce cas-là, il faudra créer des tables supplémentaires comme :

IEi (n° territoire, Cgeometry, nom_du_lieu_exterieur, type)

D'où le type peut être "*mer*" ou "*terre*".

Les tables dans lesquelles on aura les contraintes d'intégrité spatiale suivantes :

- Les attributs de type C-Geometry seront uniquement des lignes ;
- L'ensemble de ces lignes devra correspondre à la totalité du contour ;
- De manière à unifier l'ensemble, par exemple les lignes seront orientées dans le sens trigonométrique, c'est-à-dire que l'extérieur sera systématiquement à droite de ces lignes.

a. Carte initiale de la Tunisie telle que stockée dans la BDG

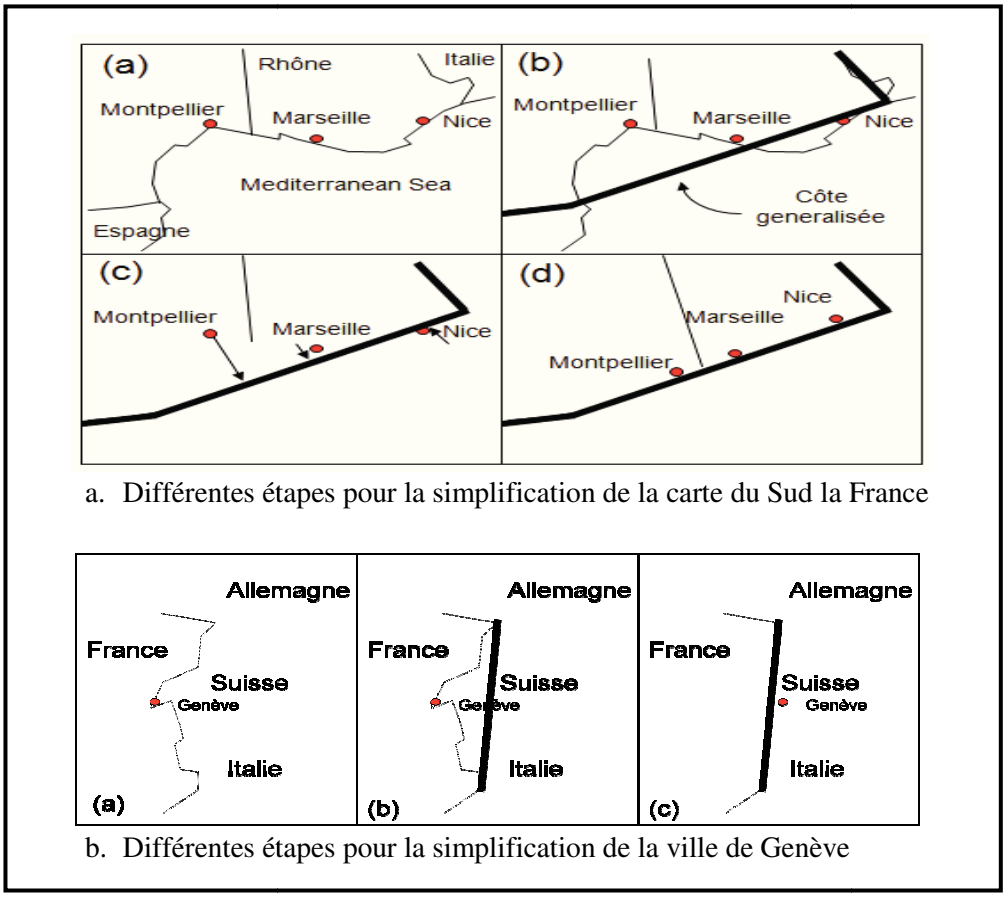
b. Carte enrichie avec des informations extérieures

Figure 4.6 Carte de la Tunisie.

4.2.7 Les contraintes topologiques

La constitution des chorèmes s'appuie sur des algorithmes de généralisation qui notamment simplifient le tracé des lignes. La conséquence est que parfois des éléments au voisinage de ces lignes se trouvent mal disposés. A titre d'exemple, les ports peuvent se retrouver en plein milieu de la mer ou en plein milieu des terres. Pour éviter ce genre d'inconvénients, il faudra imposer des contraintes topologiques. Ceci se fera grâce à une interface dont la structure sera donnée au chapitre 5.

Les contraintes topologiques sont deux types : interne (cf. figure 4.7a) et externe (cf. figure 4.7b).



a. Différentes étapes pour la simplification de la carte du Sud la France

b. Différentes étapes pour la simplification de la ville de Genève

Figure 4.7 Les différentes étapes de simplification (**a.** contraintes topologiques internes) et (**b**. contraintes topologiques externes) [27].

Pour définir les contraintes, il est nécessaire de construire des zones tampons, qui permettent à l'utilisateur de changer la position de certains éléments au sein d'une région définie.

La figure 4.8 montre la zone tampon (buffer) pour un point, une ligne et un polygone.

Nous supposons que les contraintes topologiques s'exécutent seulement dans la zone tampon ; ce qui revient à déduire que l'extérieur de la zone tampon n'est pas concerné par les contraintes. Par ailleurs, c'est seulement dans cette zone que les objets (par exemple, les villes) pourront être déplacés afin de respecter la contrainte.

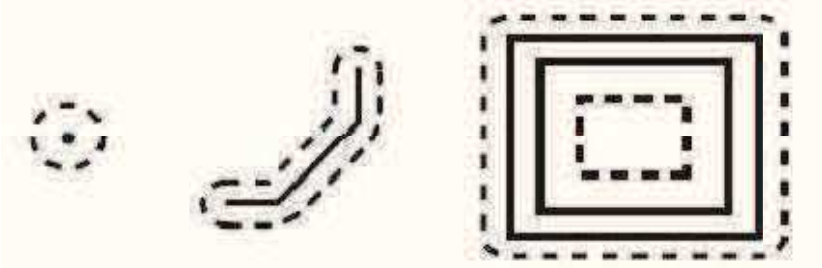


Figure 4.8 Définition de zones tampon autour des entités ponctuelles, linéaires et polygonales [65].

Bien que la définition de la taille des zones tampons soit encore une question d'étude à approfondir, nous considérons que dans notre cas, une zone tampon variable dont la largeur est

fonction de la surface. Pour simplifier le problème, supposons que l'on ait un territoire circulaire afin d'obtenir une règle (thumbrule) expérimentale. Prenons par exemple, la largeur de la zone tampon soit égale à 10 % du rayon R de la surface S de l'objet étudié comme le montre la figure 4.9.

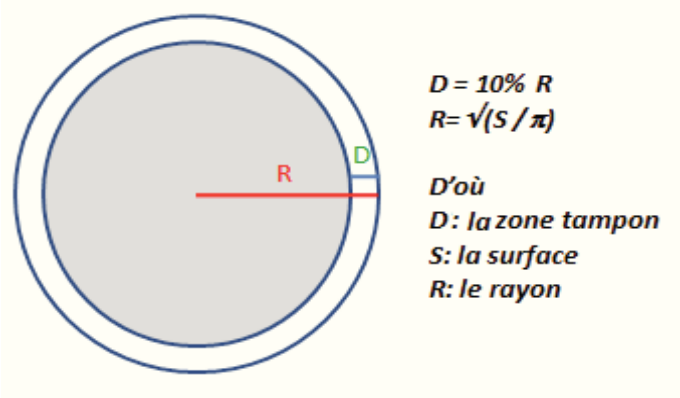


Figure 4.9 Cercle et sa zone tampon.

Si on suppose que le territoire est circulaire, la largeur de la zone tampon serait $D = 0.10\sqrt{(S/\pi)}$, soit $D = \alpha\sqrt{S}$ avec $\alpha = 1/19.2$ soit environ $\alpha = 1/20$.
Basée sur cette considération, nous prenons comme régle la formule expérimentale suivante :

$$D = \sqrt{S}/20$$

Prenons les deux exemples suivants :

- la surface de la Tunisie est égale à 163610 km^2, d'où une zone tampon de 10% du rayon du cercle équivalent de la surface donne 28 km.
- la surface de la France est égale à 543965 km^2, d'où une zone tampon 10% du rayon du cercle équivalent de la surface donne à 42 km.

En d'autres termes, ces valeurs définissent la zone, dans laquelle seuls les objets sous contraintes peuvent être déplacés.
D'un point de vue mathématique, ces contraintes s'écrivent de la façon suivante :

C_{top} *(n° territoire, n°contrainte, n°Zone$_1$, n°Zone$_2$, Relation-topologique , n°contrainte (Foreign-Key), autres attributs)*
D'où
n°Zone est l'identifiant d'une zone
n°Zone 2 : la zone tampon
Relation- topologique est la relation topologique entre les zones

Ces contraintes sont stockées dans des tables complémentaires dont voici la structure :
Tab C_{top} *((n° territoire, n°contrainte, autres attributs)*

Des contraintes peuvent, par exemple, imposer :

- que les ports restent sur la terre,
- que les fleuves se situent dans la mer,
- que des villes frontières restent toujours dans le bon pays,
- etc.

4.2.8 Connaissances plus complexes

Une fois les connaissances élémentaires extraites, on pourra constituer des connaissances plus complexes basées sur les précédentes comme :

- Flux sur des clusters
- Réseaux composés de plusieurs flux
- Etc.

4.3 Méthodologie d'extraction des motifs importants

Puisque le nombre des motifs est important, nous avons essayé de les réduire en se basant sur l'élimination des connaissances inutiles d'un point de vue de l'expert et en même temps exclure ceux qui sont redondants.
Notre approche se base sur l'élimination de la redondance et l'intégration des contraintes de l'utilisateur.

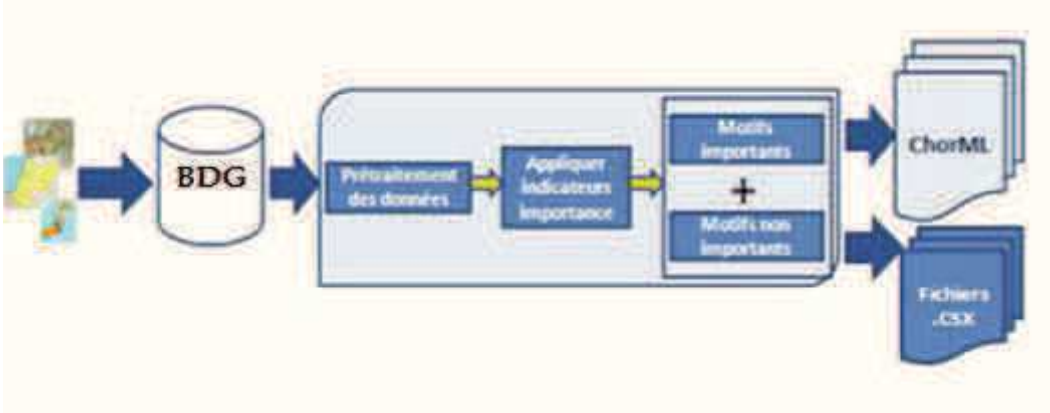


Figure 4.10 Sélection des motifs importants.

Un motif important ou saillant est un motif facile à comprendre et qui contient de nouvelles connaissances avec un certain degré de certitude, le motif est potentiellement utile dans le processus de validation ou l'invalidation des hypothèses-utilisateur.

Un motif important doit également répondre à deux mesures, l'une objective et l'autre subjective :

- *Objective* : basée sur des mesures statistiques comme le poids de redondance et l'ordre de classement.
- *Subjective* : basée sur le point de vue de l'utilisateur sur les connaissances. Par exemple, le fait que cela soit inattendu et nouveau, d'où vient l'idée de jugement concernant la fiabilité d'objet.

Pour extraire les motifs importants, nous devons tout d'abord éliminer la redondance et par la suite appliquer un autre filtrage (indicateurs subjectifs) pour choisir les motifs importants.
Un motif important *Mi* est :

$M_i = PRM_i * N_i$

Où

PRM désigne le poids de redondance du motif M

N_i correspond à la note de l'utilisateur

4.3.1 Elimination de la redondance

Concernant notre critère d'importance, c'est-à-dire la non redondance des motifs, plusieurs pistes peuvent être explorées pour détecter la redondance et la nouveauté lors du filtrage. Déterminons une formule mathématique qui serve à calculer le poids de redondance d'un motif (*PRM*).
Un *poids de redondance* de motif *M:* soit le *PRM* est la somme des poids *P* des motifs redondants *M'* par rapport au motif *M*. nous présentons, ci-dessous, l'algorithme d'élimination de la redondance :

***Entrée** : lire fichier.txt : ensemble des motifs*
***Sortie** : MI : Ensemble des motifs importants*

1. $MI \leftarrow \phi$
2. *Calculer le poids de chaque motif* P_i
3. ***Pour** i de 1 à n **faire***
4. ***Si** T[i] $\subseteq$ T [i+1]*
5. *Pi* $\leftarrow P_i + P_{i+1}$
6. ***Supprimer** T [i+1]*
7. ***Fin si***
8. *MI* $\leftarrow T[i]$
9. ***Fin pour***

Le pseudo-code de l'algorithme montre comment calculer le poids de motif redondant (*PRM*) à partir des motifs fréquents et la possibilité d'extraire les motifs importants en se basant sur l'élimination des connaissances redondantes et les regrouper dans une seule connaissance qui exprime la même information.

4.3.2 Filtrage à l'aide d'une mesure « Subjective »

Dans notre approche, la définition du profil utilisateur influe sur la personnalisation du choix des motifs importants en choisissant les attributs à exclure.
En effet, le but de la personnalisation est de faciliter l'expression des besoins de l'utilisateur et lui permettre d'obtenir des informations pertinentes qui se définissent comme étant un ensemble de critères et de préférences personnalisables spécifiques à chaque utilisateur ou communauté d'utilisateurs.

Ce qui nous amène à dire que l'importance d'un motif est relative à la satisfaction des besoins de l'utilisateur en termes de choix. Selon notre méthode, le jugement de l'expert sert à améliorer l'ordonnancement et l'apparition des motifs importants aux premiers rangs, il sert aussi à une apparition plus graduée, en permettant d'amplifier les détails autour d'un centre d'intérêt tout en gardant le contexte global.

Nous présentons, ci-dessous, l'algorithme d'élimination de la redondance.

***Entrée** : Lire MI_0 (les motifs importants)*
***Sortie** : MI_1 : Ensemble des motifs encore plus importants*

1. ***Ecrire** (''entrée de la variable à analyser'', x)*
2. ***Pour** i de 1 à n faire*
3. ***Si** T[i] ⊆ x **alors***
4. $P_i = P_i * (-1)$
5. ***Fin pour***
6. ***Pour** l de 1 à n-1 **faire***
7. ***Pour** j de 0 à n-l **faire***
8. ***Si** (T[j] > T [j+1]) **alors***
9. Tampon ←T[j]
10. T[j] ←T [j+1]
11. T [j+1] ← Tampon
12. ***Fin si***
13. ***Fin pour***
14. ***Fin pour***
15. ***Retourner** MI_1*

Cet algorithme montre comment l'expert peut intervenir et juger de l'utilité d'une connaissance à partir des objets bien définis au préalable, déterminant si l'objet est utile ou non.

4.4 Méthodologie de visualisation

Nous avons opté pour deux systèmes de visualisation : les treillis de concepts et les chorèmes.

4.4.1 La visualisation sous forme de treillis de concepts

Pour présenter les connaissances sous forme de treillis de concepts, il faut tout d'abord extraire les motifs puis les associer à des ensembles.

Nous sommes en présence de trois valeurs qui présentent un degré d'importance de motifs :

- ***Bruit*** : un PRM motif négatif (-PRM)
- ***Normal*** : un PRM motif égal à 1/*n* (*n* : nombre de motifs)
- ***Important*** : un PRM motif supérieur à 1/*n* (*n* : nombre de motifs)

4.4.1.1 Treillis de concepts

Un treillis de concepts[2] ou treillis de Galois [54] représente les connaissances sous forme d'une hiérarchie de concepts.

Formellement, soit *G* un ensemble d'objets, *M* un ensemble d'attributs et *I* une relation binaire définie sur le produit cartésien $G \times M$, c'est-à-dire $I \subseteq G \times M$. Le triplet (*G, M, I*) est appelé contexte formel : Pour un objet $g \in G$ et un attribut $m \in M$, $(g, m) \in I$ signifie que l'objet g " possède " l'attribut

[2] Pour l'affichage du treillis de concepts, nous avons intégrer le logiciel **ToscanaJ**, c'est une application qui s'intéresse à l'affichage sous la forme de treillis imbriqués, d'un ensemble de données interrogées à partir d'une base de données ou en utilisant des structures de données mappées et mémoire comme dans notre cas nous allons travailler par un fichier **.**CSX.

m. La figure 4.11 montre la table binaire représentant un contexte formel. Chaque ligne (colonne) correspond à un objet de *G* et chaque colonne correspond à l'attribut de *M*. Une case contient une croix si et seulement si l'objet (correspondant à la ligne de cette case) possède l'attribut (correspondant à la colonne de cette case)**.**

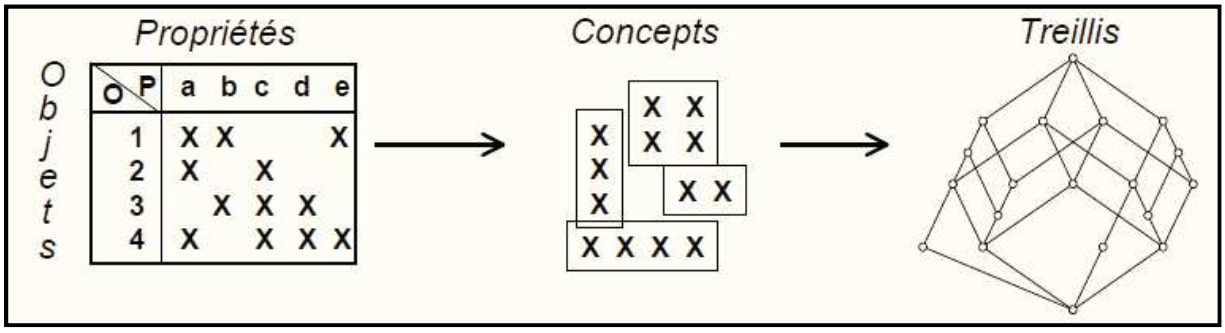


Figure 4.11 Visualisation d'un treillis de concepts [54].

4.4.2 La visualisation sous forme de chorèmes

A l'aide de notre méthodologie nous avons conçu un système avec lequel nous pouvons visualiser les motifs sous forme de résumés visuels de BDG (les chorèmes) [51]. La figure 4.12 présente l'approche proposée qui se décompose en trois phases consécutives : le prétraitement des coordonnées géographiques, la création des chorèmes et puis leur affichage.

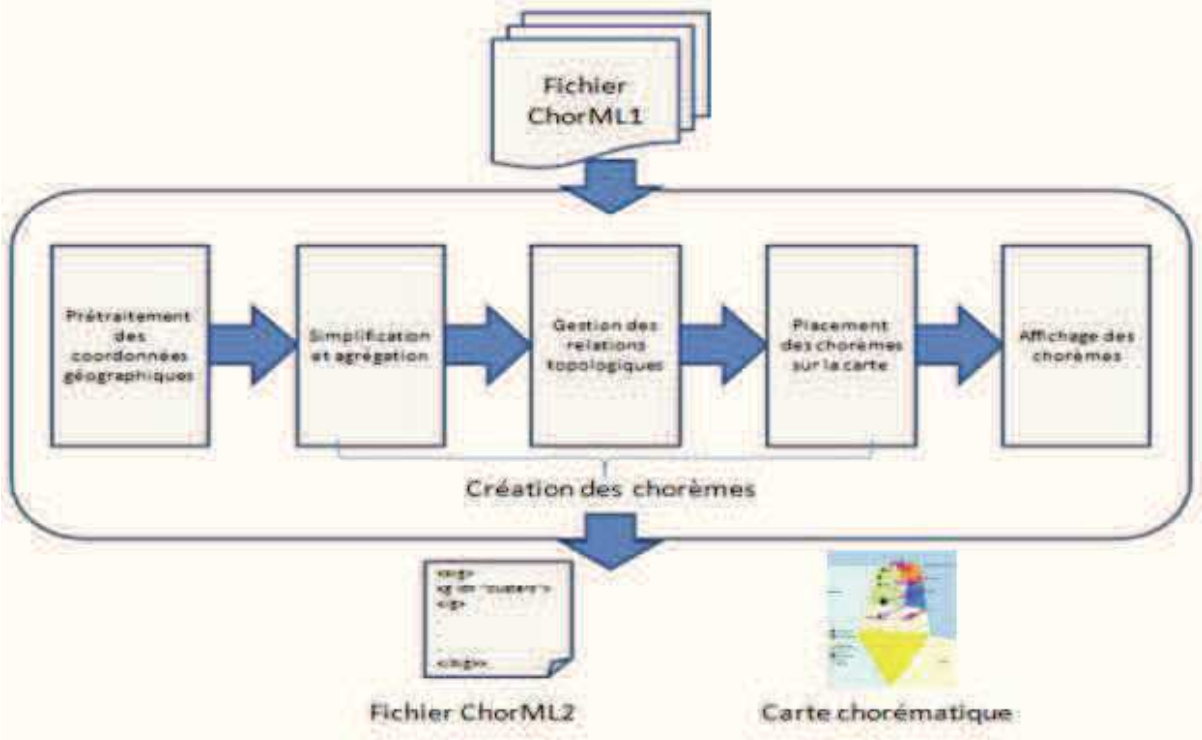


Figure 4.12 Approche de visualisation des chorèmes.

4.4.2.1 Le prétraitement des coordonnées géographiques

Etant donné que la terre est ronde et que les cartes sont plates, la conversion des informations de la surface courbe en une surface plate nécessite une formule mathématique appelée *« la projection cartographique »*. Nous essayons, au cours de cette phase, de convertir les coordonnées longitudes-latitudes des chorèmes stockées dans le fichier ChorML en des coordonnées *x, y*. Cette conversion est basée sur la projection de Mercator.

Rappelons que la projection de Mercator [http19] qui est une projection cylindrique du globe terrestre sur une carte plane a été inventée par « Gerardus Mercator » en 1569.

Figure 4.13 La projection de Mercator [http19].

Les parallèles et les méridiens sont des lignes droites et l'inévitable étirement Est-Ouest en dehors de l'équateur est accompagné par un étirement Nord-Sud correspondant. Dès lors, l'échelle Est-Ouest est partout semblable à l'échelle Nord-Sud. il s'agit d'une projection conforme, c'est-à-dire qu'elle conserve les angles. Toute ligne droite sur une carte de Mercator est une ligne d'azimut constant. Ceci la rend particulièrement utile aux marins, même si le trajet ainsi défini n'est pas généralement sur un grand cercle et n'est pas donc, le chemin le plus court [http19]. Les formules pour calculer le projeté abscisse E et ordonnée N à partir de la latitude sphérique φ et de la longitude λ sont les suivantes:

$$E = FE + R\,(\lambda - \lambda_0)\,;$$

$$N = FN + R\,\ln\,[\tan\,(\pi/4 + \varphi/2)]$$

Où λ_0 est la longitude d'origine naturelle (centre de la carte), FE et FN sont les fausses abscisses et ordonnées fictive, alors que R = 6372.797 représente le rayon de la terre en km. Dans le cas de la projection Mercator sphérique, ces valeurs ne sont pas réellement utilisées. Nous pouvons donc, simplifier la formule comme suit [http20] :

$$x = R\,(\lambda - \lambda_0),\ y = R\,\ln\,[\tan\,(\pi/4 + \varphi/2)]$$

4.4.2.2 La création des chorèmes

En partant des propriétés conceptuelles des chorèmes géographiques, notre système vise à apporter des modifications sur leurs propriétés géométriques. Pour ce faire, nous choisissons d'appliquer deux opérations de généralisation cartographique qui sont la simplification et l'agrégation. Ces dernières sont fondées sur des fonctions spatiales.

Suite à l'application de ces opérations, les relations topologiques entre les chorèmes géographiques et les chorèmes d'annotation sont violées. Pour résoudre ce problème, nous procèdons à une phase de correction par application de deux algorithmes spécialisés.

Pour placer tous les chorèmes sur une carte, nous choisissons le format SVG [http 25]. Nous définissons encore pour les chorèmes phénoménologiques les points de départ et d'arrivée tout en assurant l'harmonie du schéma global.

4.4.2.2.1 La simplification et l'agrégation des chorèmes géographiques

- **L'opération de simplification / généralisation**

La simplification des formes géométriques se fait en réduisant le nombre de sommets composant les objets spatiaux et qui correspondent aux chorèmes dans notre cas, tout en essayant de garder la forme originale.

Du début des recherches sur l'automatisation jusqu'aujourd'hui, différents algorithmes de simplification ont été proposés. Nous pouvons classer ces algorithmes en cinq grandes catégories comme indiqué dans le tableau 4.1 [82].

Catégorie	Description
Independent Point Algorithms	Les algorithmes de cette catégorie ne tiennent compte d'aucune relation entre les points de l'objet.
Local Processing Routines	Les algorithmes de cette catégorie déterminent l'importance d'un point en tenant compte du voisinage immédiat.
Uncontrained Extented Local Processing Routine	En plus du voisinage immédiat, les algorithmes de cette catégorie évaluent une portion de la ligne.
Contrained Extented Local Processing Routine	Les algorithmes appartenant à cette catégorie sont similaires à ceux du groupe précédant sauf qu'ils comportent des restrictions par rapport à la zone de recherche.
Global Routines	Les algorithmes appartenant à cette catégorie considèrent un segment ou une ligne en entier. En plus, ils tiennent compte des points critiques identifiés. Le plus populaire algorithme de cette classe est le RDP 1973.

Tableau 4.1 Les cinq catégories des algorithmes de simplification [82].

Pour effectuer l'opération de simplification, nous avons choisi d'appliquer deux algorithmes : « Radius » [http21] et « RDP » [33]. Les deux algorithmes appartiennent à la dernière catégorie. Ce choix est appuyé par le fait que les formes géométriques à traiter sont des polygones (formés d'une suite cyclique de segments consécutifs et délimitant une portion du plan). Donc, nous devons tenir compte des relations entre les points qui les constituent.

Nous avons choisi de combiner ces deux algorithmes. Nous effectuons tout d'abord, une réduction par Radius avec une petite tolérance, ce qui va permettre de diminuer le nombre de points sans perdre trop d'informations. L'algorithme RDP étant rapide avec peu de points, nous avons tout intérêt à lui en passer le moins possible. Cela donnera donc, une solution rapide et efficace à la fois.

- **L'algorithme Radius**

Le principe de « Radius » [http21] est très simple. Nous partons du point 0, nous définissons un cercle dont le rayon correspond à la tolérance souhaitée et nous éliminons tous les points du polygone original qui sont inscrits dans ce cercle. Nous passons, par la suite, au point suivant (à savoir le point le plus proche du cercle) et nous recommençons. Une description de Radius est fournie dans le tableau 4.2

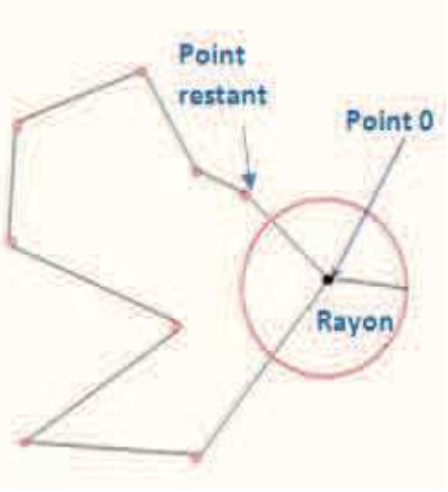
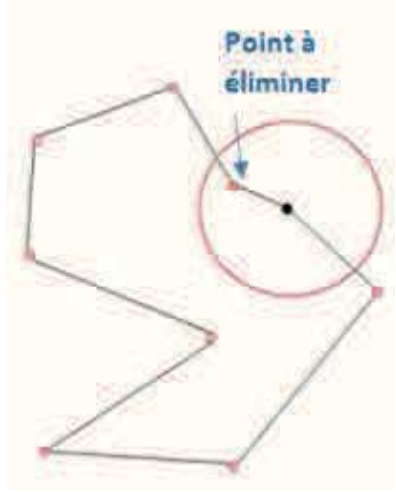
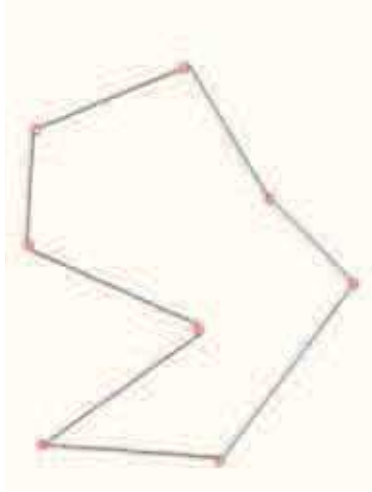
		
Etape 1 : Définition du polygone, du rayon et du point 0	**Etape 2 :** Elimination du point inscrit dans le cercle	**Etape 3 :** Polygone obtenu

Tableau 4.2 Description du principe de l'algorithme Radius.

- **L'algorithme RDP**

Pour l'algorithme RDP (Ramer-Douglas-Peucker) [33], une polyligne (n nœuds) est simplifiable et remplacée par une ligne simple (deux nœuds) si la distance de son nœud le plus éloigné de la droite formée par les extrémités de la polyligne est inférieure à un seuil.

L'algorithme s'exécute d'une manière récursive par la méthode « diviser pour régner ». À l'initialisation, nous sélectionnons le premier et le dernier nœud (cas d'une polyligne) ou un nœud quelconque (cas d'un polygone). Ces nœuds présentent les bornes. À chaque étape, nous parcourons tous les nœuds entre les bornes et nous sélectionnons le nœud le plus éloigné du segment formé par ces bornes :

1. S'il n'y a aucun nœud entre les bornes, l'algorithme se termine.
2. Si la distance est inférieure à un certain seuil, nous supprimons tous les nœuds entre les bornes.
3. Si la distance est supérieure, la polyligne n'est pas directement simplifiable. Nous appelons de manière récursive l'algorithme sur deux sous-parties de la polyligne : de la première borne au nœud distant et du nœud distant à la borne finale.

La figure 4.15 présente les différentes étapes de l'algorithme RDP :

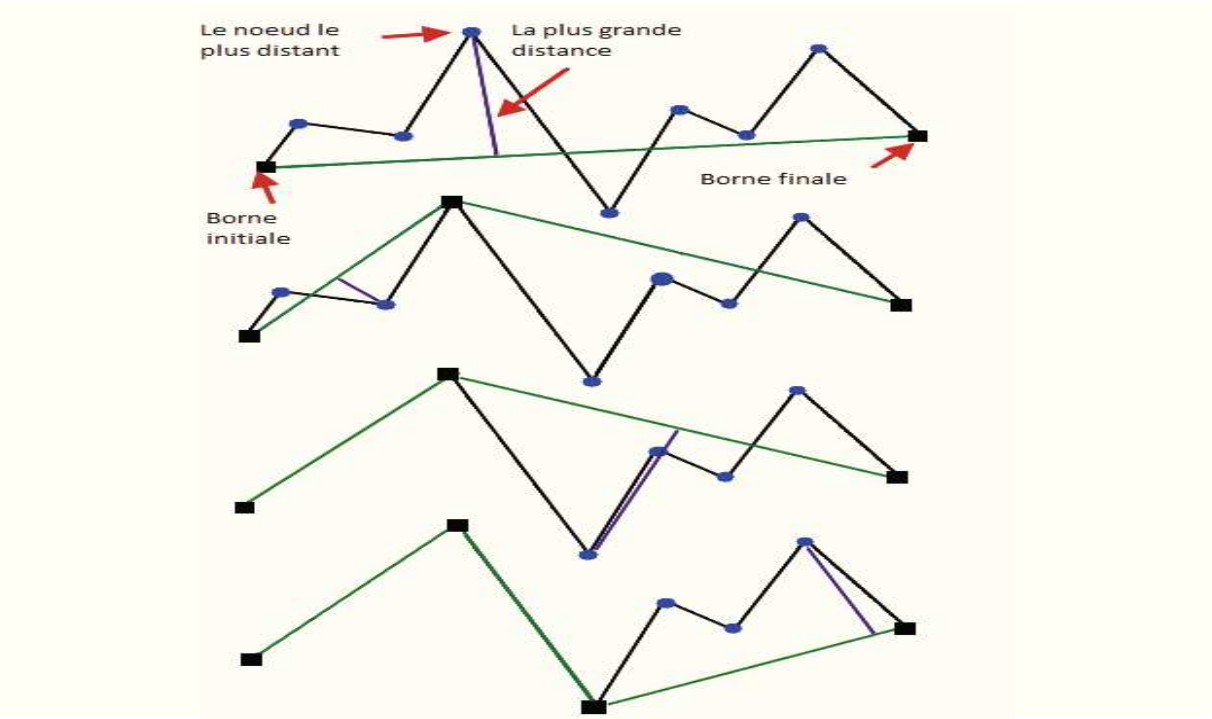


Figure 4.15 Le principe de l'algorithme RDP.

- **L'opération d'agrégation**

L'agrégation génère une représentation géométrique suite au regroupement des caractéristiques de certains éléments partageant quelques propriétés communes. Dans notre cas, nous ne sommes intéressées que par les caractéristiques spatiales des chorèmes géographiques. Les sommets de ces chorèmes, dont la distance est inférieure à la distance spécifiée par rapport aux autres, représentent le même emplacement qui leurs sont attribués une valeur de coordonnées commune.

Une tolérance d'agrégat est utilisée pour intégrer ces sommets. Nous signalons que dans ses limites, tous les sommets risquent d'être légèrement déplacés dans le processus de validation. Selon [http23], la tolérance d'agrégat est par défaut 0,001 mètre dans les unités du monde réel. C'est 10 fois la distance de la résolution *x, y* (qui définit le niveau de précision numérique utilisé pour stocker des coordonnées), ce qui est recommandé pour la plupart des cas. Si les coordonnées sont stockées en degrés de longitude-latitude, la valeur de tolérance *x, y* par défaut est 0,0000000556 degrés.

Nous choisissons une faible valeur de tolérance *x, y* pour garantir l'attribution de la même position uniquement pour les sommets qui sont très proches (situés dans leur tolérance x, y mutuelle). Nous disons que les coordonnées sont dans la tolérance lorsqu'elles sont considérées comme étant coïncidentes et elles sont ajustées pour partager le même emplacement.

Il est important de savoir que la tolérance *x, y* n'est pas fixée pour généraliser des formes géométriques. En fait, elle est destinée à intégrer le réseau linéaire et les limites au cours des opérations topologiques. Ceci signifie qu'elle facilite la découverte des entités coïncidentes dont les sommets figurent dans un même emplacement.

Notre processus d'agrégation est basé sur le théorème de Pythagore [http22]. Nous commençons par le calcul de la distance qui sépare chaque deux sommets d'un chorème. Pour ce faire, nous construisons un triangle rectangle après avoir dégagé son troisième sommet.

Nous passons par la suite, à l'application du théorème de Pythagore pour vérifier si les deux sommets peuvent être agrégés ou non.

- **La construction d'un triangle rectangle**

Pour pouvoir construire un triangle rectangle à partir de chaque deux sommets $s1(x_1, y_1)$ et $s2\ (x_2, y_2)$ d'un chorème donné, nous dégageons un troisième sommet *s3*. Ce dernier est le résultat d'intersection des deux droites issues de s1 et s2.

Supposons deux droites *D1* et *D2,* telles que : *D1* est la projection de *s1* suivant l'axe *x* et *D2* est la projection de *s2* suivant l'axe *y*. L'équation réduite de chacune de ces droites est donnée comme suit :

D1 : $y = m\,x+p$; avec *m* correspondant au coefficient directeur et *p* l'ordonnée à l'origine.

Posons *s1'* $(x_{1'}, y_{1'})$ un sommet appartenant à *D1* obtenu en incrémentant l'abscisse de *s1*. Nous pouvons calculer m par l'application de la formule : $m= (y_{1'} - y_1)/(x_{1'}- x_1)$.

Pour déterminer *p*, il suffit de résoudre l'équation suivante:
$y_1 : m\,x_1+p$ *donc* $p= y_1- m\,x_1$.

Etant donné que *D2* est parallèle à l'axe des ordonnées alors, son équation peut être simplifiée en *D2 :* $x_2= k$

Généralement, les coordonnées du point d'intersection de deux droites sécantes sont les solutions du système formé des deux équations de droites. *D1* et *D2* sont deux droites d'équation $y = m\,x + p$ et $x = k$ donc, elles sont sécantes. Pour déterminer les coordonnées de *s3*, nous remplaçons x_2 par sa valeur dans l'autre équation, d'où :
$x_3 = x_2$ *et* $y3= mx_2+p$

- **Application du théorème de Pythagore**

Nous considérons que la longueur de l'hypoténuse correspond à la valeur de la distance maximale. En se basant sur le théorème de Pythagore, la distance maximale dans laquelle les coordonnées sont agrégées est égale à la racine carrée de deux fois la tolérance *x, y*. Nous considérons que la tolérance suivant l'axe *x* est égale à la tolérance suivant l'axe *y*. De ce fait, le carré de la distance maximale est égal à la somme des carrés des longueurs des deux autres côtés.

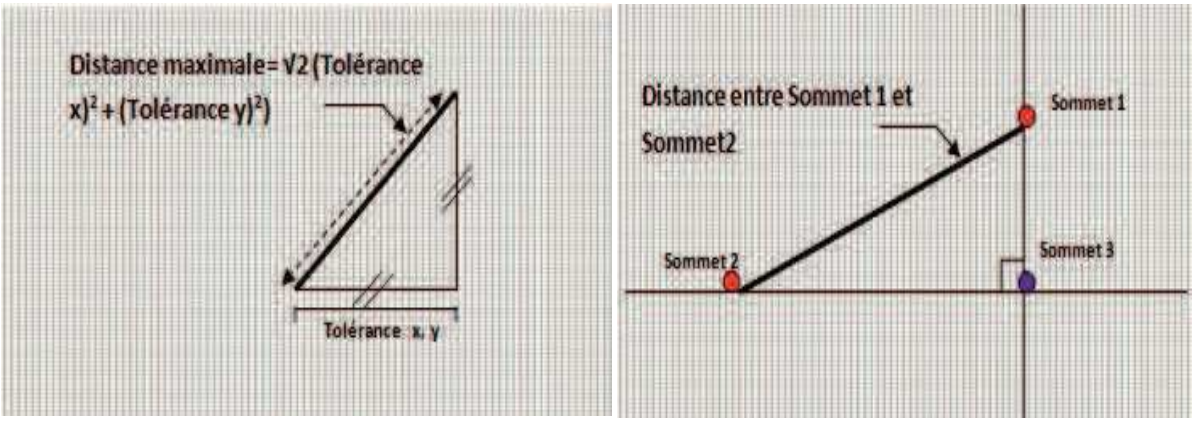


Figure 4.17. Calcul des distances suivant le théorème de Pythagore.

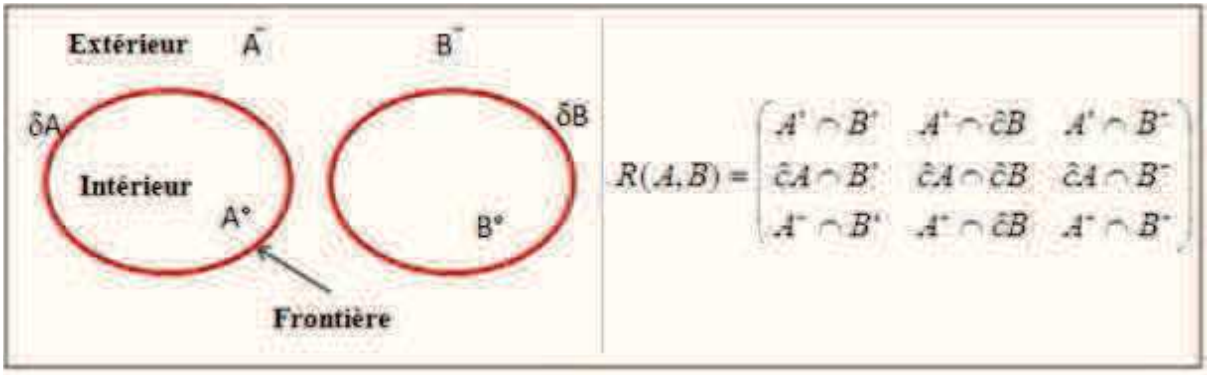
Une fois que la valeur de la distance maximale est obtenue, nous procédons au calcul de la distance séparant chaque deux sommets.
Nous devons tout d'abord, trouver le troisième sommet pour pouvoir construire un triangle rectangle. Ce sommet est le résultat de l'intersection des projetées des deux autres. Les exigences du théorème de Pythagore sont répondues comme nous obtenons un triangle rectangle en *s3*. Il est donc, possible d'appliquer la même formule pour trouver la valeur de la distance.
Nous passons finalement à la comparaison de cette valeur avec celle de la distance maximale :

- Si (distance maximale >= distance séparant les deux sommets *s1* et *s2*) alors, remplacer ces derniers par un unique sommet dont $x = (x_1+x_2)/2$ et son $y = (y_1+y_2)/2$
- Sinon, les deux sommets sont assez éloignés qu'ils ne subissent pas une opération d'agrégation.

4.4.2.2.2 Gestion des relations topologiques

Au niveau de la deuxième phase, nous traitons les relations topologiques entre les chorèmes géographiques[3] et d'annotation[4], décrites dans le fichier ChorML1. Nous utilisons pour cela les relations telles que définies par Egenhofer [36] (cf. Figure 4.18).

[3] Les chorèmes géographiques représentent des objets avec des fonctionnalités spatiales simples, tels que les points, les lignes, les polygones, et les objets composés de leurs combinaisons, telles que les réseaux.
[4] Les chorèmes d'annotations sont les étiquettes des cartes. Des chorèmes d'annotations supplémentaires peuvent être ajoutées par les concepteurs afin de compléter une carte chorématique.



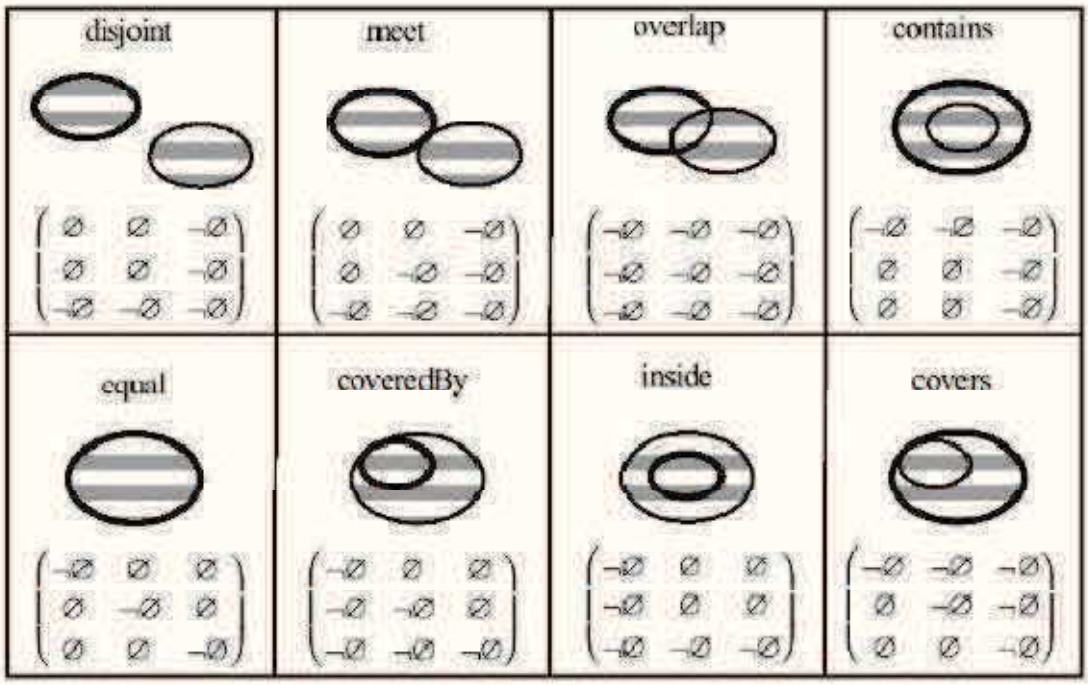


Figure 4.18 Le modèle des 9 intersections d'Egenhofer [36].

Nous implémentons deux algorithmes spécialisés en vue de définir la nature de la relation séparant chaque deux chorèmes.

Chaque relation identifiée est comparée avec la relation d'origine décrite au niveau du fichier ChorML1. Dans le cas où elles sont différentes, nous essayons de corriger la première en effectuant des opérations de déplacement successives.

- **Détermination des relations topologiques**

  Pour préciser la nature des relations topologiques entre les chorèmes, nous proposons un algorithme spécialisé. La structure de cet algorithme est présentée comme suit :

Algorithme 1 : Détermination des Relations Spatiales

Entrées : Liste des chorèmes géographiques, liste de contours des chorèmes géographiques, liste des chorèmes d'annotation, liste de contours des chorèmes d'annotation

Sorties : Liste des relations spatiales

pour i allant de 0 à "taille de la liste des contours - 1" **faire**

si (contour *i* est un contour intérieur) **alors**

le chorème auquel appartient le contour *i* ***contains*** chacun des chorèmes ayant un contour adjacent au contour *i*;

chaque chorème ayant un contour adjacent au contour i ***inside*** le chorème auquel appartient le contour *i*;

sinon

pour (*j* allant de 0 au "nombre des contours adjacents au contour *i*") **faire**

si (la zone de contact entre contour *i* et contour *j* est inférieure au seuil *A*)

alors les deux chorèmes considérés sont reliés par ***meet*** ;

sinon si (la zone du contact est supérieure au seuil *A* et inférieure à seuil *B*)

alors les deux chorèmes considérés ***overlap*** ;

sinon si (zone de contact est supérieure au seuil *B*) **si** (contour *i* est moins long que contour *j*);

alors le chorème auquel appartient le contour *i* ***covers*** le chorème auquel appartient le contour *j*;

sinon le chorème auquel appartient le contour *i* ***covered b*** le chorème auquel appartient le contour *j*;

En plus de la procédure décrite par cet algorithme, il serait mieux de rajouter un post-traitement qui facilite la détermination des relations. Nous vérifions si les deux chorèmes concernés sont disjoints ou pas. Si c'était le cas, nous n'aurons plus besoin de passer par le traitement décrit au niveau de l'algorithme 1. Le post-traitement, dont nous parlons, est décrit dans l'algorithme 2.

Algorithme 2 : Complément des Relations Spatiales

Entrées : Liste des relations spatiales

Sorties : Liste des relations spatiales

Pour chaque couple de chorèmes (*i*, *j*) **faire**

(S'il n'existe pas de relation entre *i* et *j* dans la liste des relations) **alors**

Rajouter une relation "***i disjoint j***" dans la liste;

Rajouter une relation "***j disjoint i''*** dans la liste;

Il convient d'apporter une précision en ce qui concerne l'algorithme 1 et plus précisément au sujet des seuils mentionnés. Prenons l'exemple de la figure 4.19. Pour différentier les trois relations « *covers/ covered by* », « *overlap* » et « *meet* », nous considérons les longueurs :

- du contour de l'objet rouge ;
- du contour de l'objet bleu ;
- du contour commun aux deux objets.

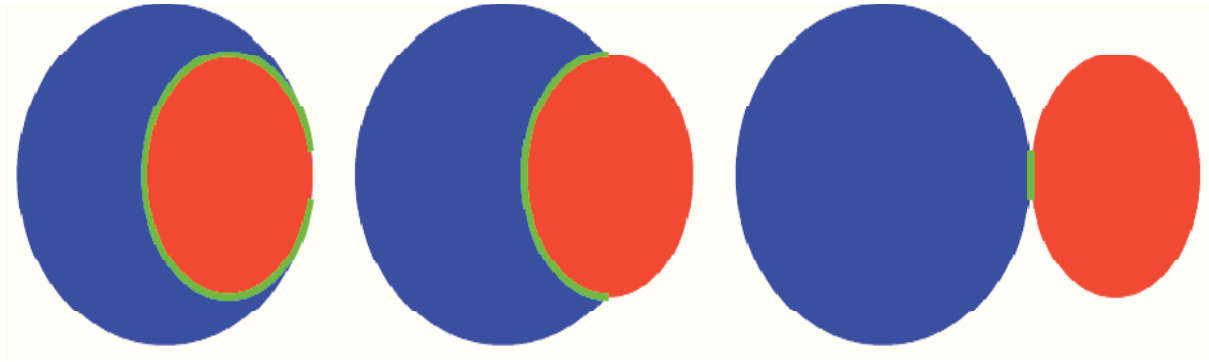


Figure 4.19 Différenciation des relations spatiales à l'aide d'un seuil.

Nous pouvons prendre par exemple un seuil *A* égal à 10% et un seuil *B* égal à 90%. Dans ce cas, nous déclarons que les deux objets se touchent si et seulement si la longueur de leur contour commun (en vert) est inférieure à 10% de la longueur du contour de l'objet bleu et inférieure à 10% de la longueur du contour de l'objet rouge. Nous pouvons de même, déclarer que l'objet rouge recouvre l'objet bleu si et seulement si la longueur du contour commun est supérieure à 90% de la longueur du contour de l'objet rouge. Enfin, les deux objets se chevauchent dans tous les autres cas.

• Déplacement des objets

Après avoir identifié la nature de la relation entre deux chorèmes donnés, nous pouvons être devant deux cas : (i) la relation convient à ce qui est décrit en ChorML1 ou (ii) la relation trouvée est différente de celle existante entre les deux chorèmes donnés. En d'autres termes, la relation trouvée ne convient pas à la relation d'origine décrite en ChorML1.
Dans le deuxième cas, nous devons procéder à une correction de la relation obtenue. Cela se fait en appliquant des opérations de déplacement jusqu'à ce que nous aboutissons à la relation d'origine.
Nous considérons par exemple, le cas d'un chorème géographique et d'un chorème d'annotation. Notre traitement commence par la détermination de la direction de l'un par rapport à l'autre : à gauche, à droite, vers le haut ou vers le bas. Cette détermination s'effectue en comparant les coordonnées du chorème d'annotation par : *Min x, Min y, Max x et Max y* du chorème géographique. Notons que ces coordonnées sont obtenues en insérant le chorème géographique dans un rectangle (le plus petit rectangle englobant), comme le montre la figure 4.20, où nous considérons que *Min x et Min y* partagent la même position.
Nous passons par la suite, au calcul de la distance euclidienne séparant le chorème d'annotation et tous les sommets du chorème géographique. La plus petite distance désigne le partage d'un même emplacement entre les deux chorèmes. Suivant la relation topologique d'origine contenue dans ChorML1, un déplacement, sur les deux axes *x* et *y*, du chorème d'annotation vers le sommet le plus proche est effectué avec une distance bien déterminée. Ce traitement est récursif jusqu'à l'arrivée à la condition d'arrêt qui est dans notre cas, la résolution du conflit topologique identifié.

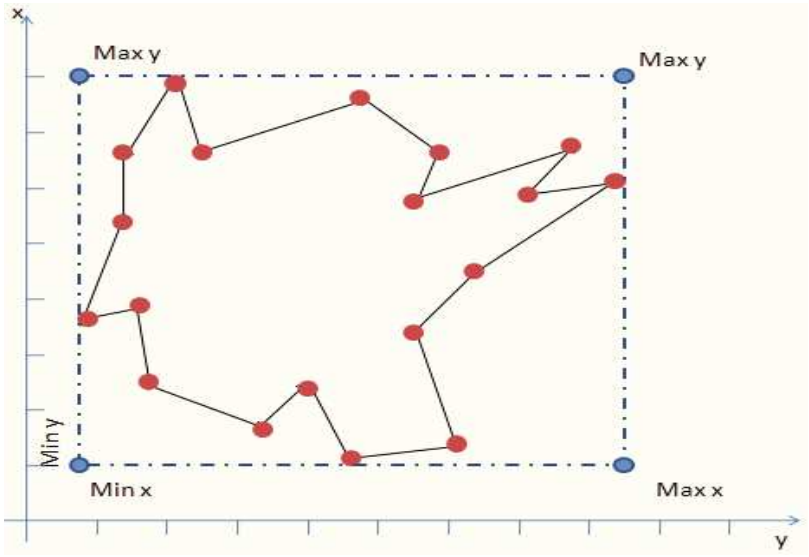


Figure 4.20 Détermination des limites d'un polygone.

Notons que ce traitement se répète jusqu'à la satisfaction des relations d'origines. Notons encore que le diamètre du chaque chorème d'annotation est proportionnel à son importance exprimée dans le fichier ChorML1.

4.4.2.2.3 Placement des chorèmes sur la carte

Les positions des chorèmes géographiques et des chorèmes d'annotation sur la carte sont les résultats de la troisième phase. Concernant les chorèmes phénoménologiques[5], nous proposons un traitement pour définir leurs emplacements. Ce traitement est décrit dans ce qui suit.

Dans le cas des chorèmes de flux, il faux déterminer les centroïdes des zones élémentaires ou des clusters. Une première solution est d'utiliser le centre de gravité de la zone.

a. Calcul des centres de gravité des chorèmes géographiques

Nous procédons, en premier lieu, au calcul des centres de gravité des clusters (chorème géographique) correspondant à la région émettrice et à la région réceptrice. Pour ce faire, nous proposons une méthode simple et rapide par rapport à la méthode existante.

• Méthode existante pour le calcul du centre de gravité d'un polygone

Il existe une méthode prédéfinie permettant le calcul du centre de gravité d'un polygone donné. Considérons, par exemple, la surface présentée dans la figure 4.21(b).

Nous choisissons pour cette surface un axe *x, y* qui simplifie au maximum les calculs. Nous passons à la décomposition de cette surface en un ensemble d'éléments simples. Par la suite, nous procédons au calcul de (i) l'aire de chaque élément simple, (ii) la distance du centre de gravité à l'axe considéré et (iii) le moment statique par rapport à ce même axe [http24].

Notons que le moment statique se calcule en multipliant l'aire par le centre de gravité de l'élément considéré.

[5] Les chorèmes phénoménologiques décrivent des phénomènes spatio-temporels impliquant un ou plusieurs chorèmes géographiques. Le groupe initial des chorèmes phénoménologiques que nous avons identifiés se compose de trois types, à savoir flux, tropisme, et diffusion.

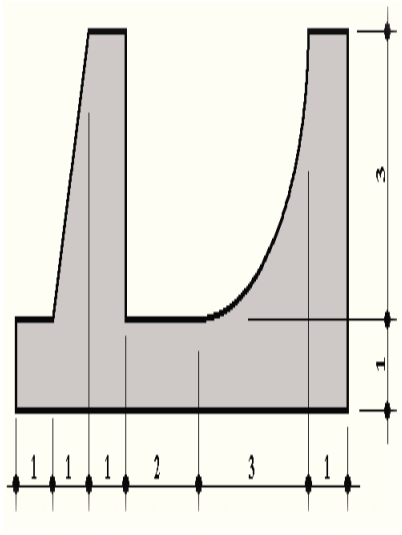


Figure (a). Choisir l'axe ***x, y***

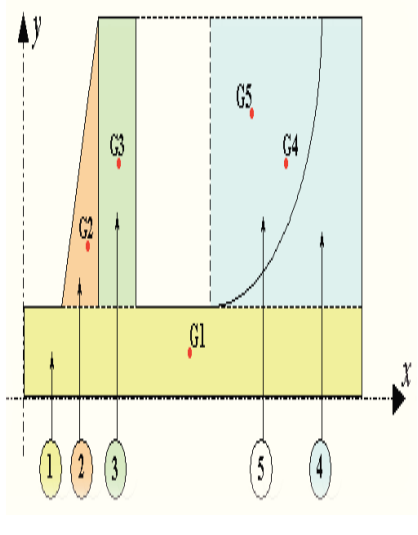


Figure (b). Décomposition de la surface en un ensemble d'éléments simples

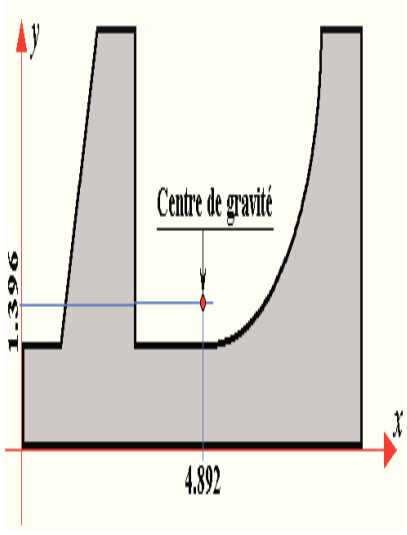


Figure (c). Centre de gravité obtenu

Figure 4.21 Conduite de calcul d'un centre de gravité [http24].

Nous obtenons finalement, la position du centre de gravité du polygone par rapport à l'axe des *x* et l'axe des *y* en appliquant ces deux formules :

Centre de gravité par rapport à l'axe des x : y = ∑ moments statiques des éléments simples / ∑ des aires des éléments simples

De même :

Centre de gravité par rapport à l'axe des y : x = ∑ moments statiques des éléments simples / ∑ des aires des éléments simples

Comme le montre la figure 4.21 (c), nous risquons d'avoir un centre de gravité en dehors du polygone en adoptant cette méthode.

• Méthode proposée pour le calcul du centre de gravité d'un polygone

Etant donné l'importance du temps de réponse et en considérant le nombre des clusters à tenir, nous proposons une solution plus rapide que celle décrite, ci-dessus. Dans ce cas, le centroïde est le centre de rectangle englobant minimum (Cf. Figure 4.22).

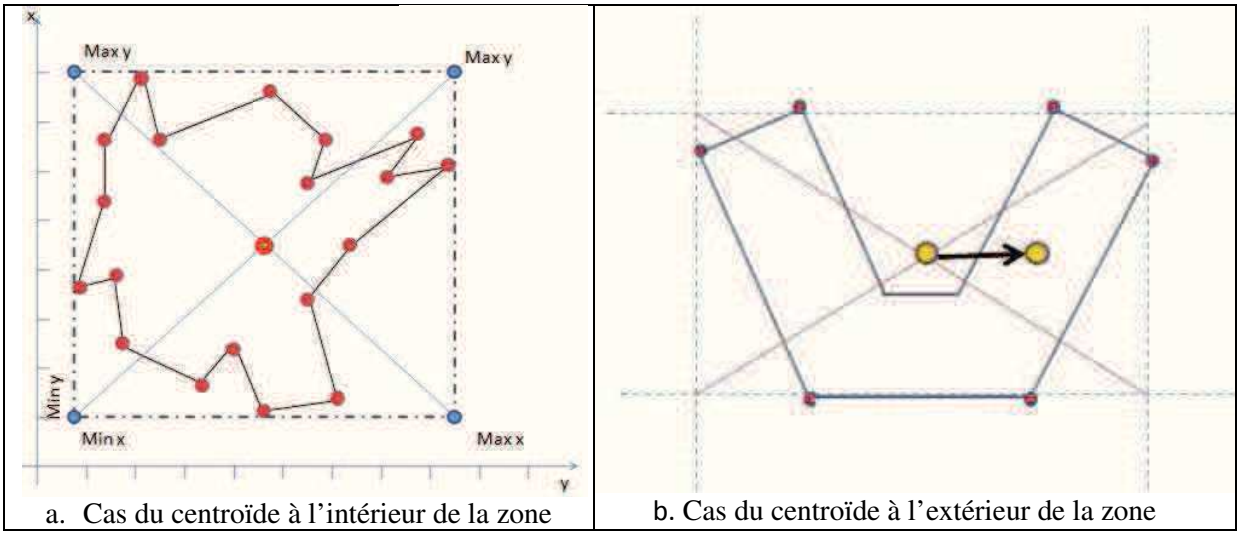
Il serait suffisant d'insérer chaque chorème géographique dans un rectangle défini par les maximum et minimum coordonnées de x et y et avoir son centre, tel que :

$x = Min\ x + (Max\ x - Min\ x)/2,$

$y = Min\ y + (Max\ y - Min\ y)/2$

Une telle solution garantie un temps optimum de calcul du centre de gravité pour chaque cluster décrit dans le fichier ChorML1. En fait, nous n'aurons pas besoin de décomposer une surface donnée en des éléments simples ni d'appliquer consécutivement plusieurs formules. Elle

élimine encore, la possibilité du placement du centre de gravité en dehors du cluster concerné en cas de figure un peu tordue.



a. Cas du centroïde à l'intérieur de la zone b. Cas du centroïde à l'extérieur de la zone

Figure 4.22 Méthode proposée pour calculer le centre de gravité d'un polygone ; (a) cas d'un centroïde à l'intérieur, (b) cas d'un centroïde extérieur qu'il faudra déplacer.

b. Placement des chorèmes phénoménologiques

Pour les chorèmes phénoménologiques, nous ne traitons que les flux. Nous définissons pour chacun d'eux un point de départ et un point d'arrivée. Ces points correspondent aux centres de gravité des régions en question. L'épaisseur des lignes varie selon son importance décrite dans le fichier ChorML1.

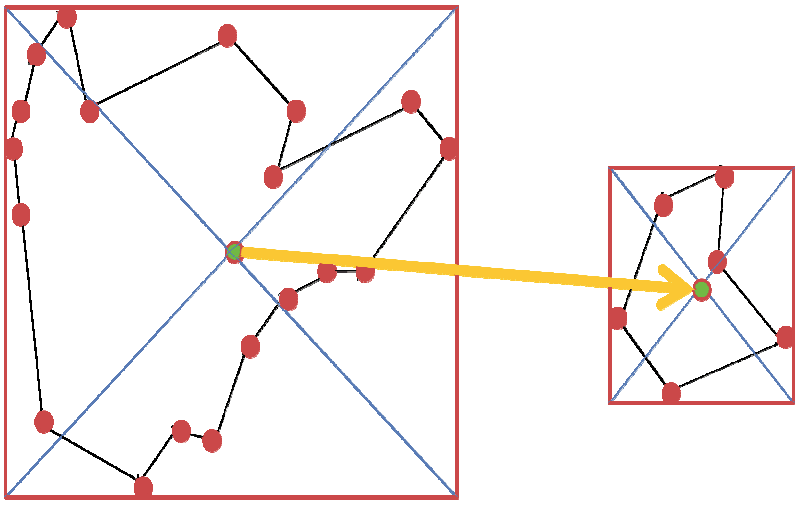


Figure 4.23 Placement des chorèmes phénoménologiques.

4.4.2.3 La visualisation des chorèmes

Une fois avoir subi un ensemble de traitements au cours des phases précédentes, tous les chorèmes sont placés suivant le schéma de visualisation qui est dans notre cas, le format SVG. Pour ce faire, la transformation du fichier ChorML obtenu vers une nouvelle extension SVG est nécessaire.

En vue d'assurer l'interaction avec l'utilisateur, nous offrons la possibilité de modifier la carte chorématique produite à travers un ensemble d'opérations :

- Déplacement ;
- Translation ;
- Changement de couleur ;
- Simplification de la forme ;
- Modification de la taille d'un élément ;
- Zoom in ;
- Zoom out ;
- etc.

Ces opérations sont possibles à travers un éditeur graphique supportant le format SVG. Dès lors, l'utilisation d'un tel outil permet de mieux répondre aux besoins des utilisateurs lorsqu'ils demandent plus de raffinements supplémentaires en ce qui concerne les propriétés sémantiques et graphiques des chorèmes.

4.5 Conclusion

Dans ce chapitre, nous avons présenté notre méthodologie permettant, tout d'abord, l'extraction des motifs à partir d'une base de données géographiques, ensuite la visualisation des patterns extraits sous forme de graphes de treillis ou cartes chorématiques.

A partir d'une base de données géographiques, le système effectue le nettoyage des données et ensuite, il lance les algorithmes de fouille de données pour extraire les motifs de Clusters, Faits, Flux et Colocalisations. Les motifs les plus intéressants sont identifiés à l'aide de deux mesures subjectives et objectives, puis ils passent par un processus de visualisation, sous forme de treillis de concepts ou de cartes chorèmatiques. Par la suite, il sont visualisés sous forme des chorèmes à l'aide du sous système des visualisation des résumés visuel.

Chapitre 5

Architecture, implémentation et expérimentation

5 Architecture, implémentation et expérimentation

Ce chapitre est dédié à la phase d'implémentation d'un prototype devant valider notre approche décrite au chapitre 4. Nous allons, dans ce chapitre, tout d'abord rappeler d'une manière brève notre approche. Nous avons baptisé notre prototype ChoreMAP. Nous avons, enfin, monté une expérimentation grandeur réelle concernant le cas des migrations internes et le cas de migration de marchandises en Tunisie.

5.1 Introduction

Après une étude de différents algorithmes de fouille de données, nous avons implémenté la fouille de données spatiales avec Postgresql /PostGIS. C'est grâce à cet algorithme que nous obtenons des motifs, qui seront filtrés par un sous système d'extraction de patterns importants. Enfin, visualisés soit comme chorèmes dans une carte chorématique, soit comme treillis de concepts. Les sections suivantes décrivent l'architecture du prototype réalisé.

5.2 Rappel de l'approche proposée

Comme dit précédemment, nous proposons une méthodologie pour générer les résumés visuels à partir d'une base de données géographiques. Les données de notre base de données passent d'abord par une phase de prétraitement, puis nous appliquons la fouille de données spatiales pour sélectionner les motifs de type clusters. Ensuite, nous déterminons les flux. Enfin nous appliquons notre algorithme pour dégager les relations topologiques entre les clusters et les villes enregistrées dans la base,

Les motifs extraits dans la première phase vont subir un filtrage grâce à des mesures objectives et subjectives. Enfin ils seront présentés sous deux formats : graphe à l'aide de la technique des treillis de concepts ou chorèmes à l'aide des phases de simplification et d'agrégations.

Les sections qui suivent présentent le prototype *ChoreMAP* que nous avons réalisé pour valider notre approche. Afin de mettre en œuvre l'application objet de ce travail, certains choix techniques doivent être explicités en vue de garantir la faisabilité ainsi que la qualité du projet, les outils de base sont décrits dans l'annexe C.

5.3 ChoreMAP : Outil d'extraction et de visualisation des cartes chorématiques

Notre système d'extraction et de visualisation des résumés visuels (cf. figure 5.1) est composé de trois sous-systèmes :

- Le sous-système *EPS* (*Extract Patterns Sub-system*) qui permet l'extraction des motifs d'une base de données géographiques

- Le sous-système *ESPS* (*Extract Salient Patterns Subsystem*) qui permet de filtrer les motifs extraits du premier sous-système pour générer les patterns importants qui seront visualisés sous forme de treillis de concepts
- Le sous- système *VS* (*Visualization Subsystem*) qui permet de visualiser les motifs.

A partir d'une base de données géographiques, le système effectue un prétraitement des données, ensuite grâce au sous-système *EPS* (*Extract Patterns Sub-system*), nous effectuons la fouille de données et la fouille de données spatiales, nous obtenons les motifs qui vont subir par la suite une action de filtrage par l'utilisateur, puis ils seront codés en ChorML1 v0. Les motifs les plus intéressants sont identifiés à l'aide d'un deuxième sous-système *ESPS* (*Extract Salient Paterns Subsystem*), puis ils sont codés en ChorML1 pour être visualisés. Ils passent par un processus de visualisation grâce à un troisième sous- système *VS* (*Visualization Subsystem*).

La figure 5.1 montre l'architecture du système proposé, qui se compose de trois niveaux principaux, à savoir les deux sous-systèmes d'extraction et un sous système de visualisation.

EPS est destiné à obtenir et manipuler de l'information à partir de données disponibles, *ESPS* manipule les données générées par le premier sous-système, il été renforcé par l'intégration des techniques spatiales de la fouille de données, le filtrage des données basées sur SQL ; quant à la procédure pour réduire le nombre de motifs, elle se base sur l'intervention de l'expert en tenant compte de ce qu'il veut montrer sur sa carte. De plus, *VS* gère cette information en lui assignant une représentation visuelle en terme de graphe, treillis de concepts ou des cartes chorématiques.

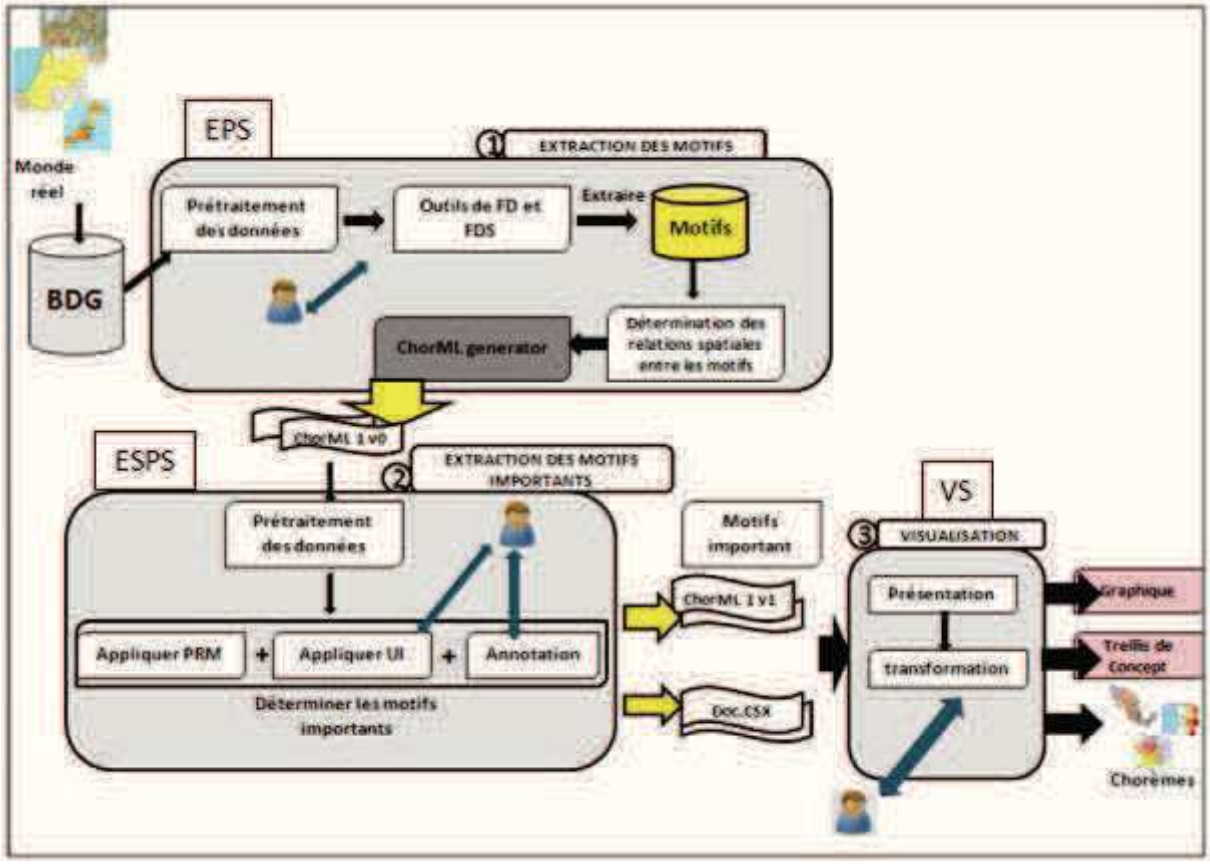


Figure 5.1 Architecture de notre prototype ChoreMAP.

5.3.1 Base de données spatio-temporelles

Nous avons créé notre base de données sur Postgresql/PostGIS (cf. figure 5.2). Sous forme d'entrepôt de données avec un ensemble de données spatio-temporelles concernant la population en Tunisie provenant de l'INS (Institut supérieur des statistiques en Tunisie) et étudiées selon deux axes de dimension : Dimensions Temps Années et Dimension Espace Villes (cf. figure 5. 3)**.**

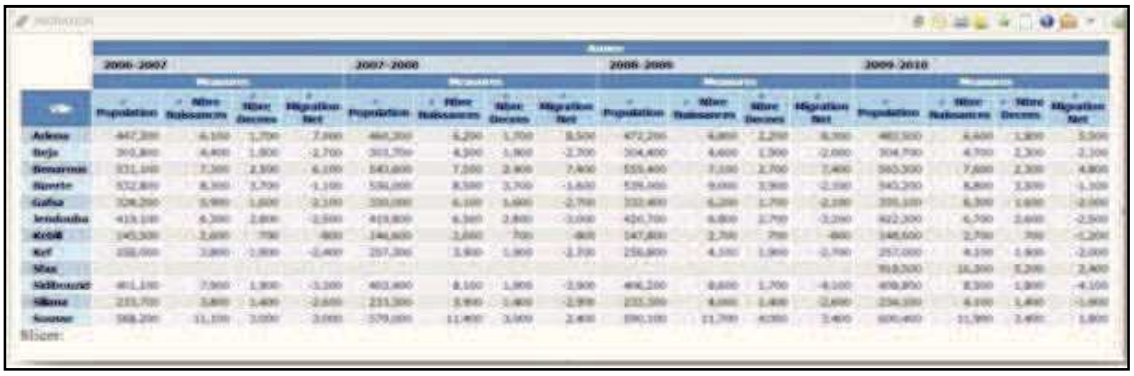


Figure 5.2 Base des données spatio-temporelles.

La transformation de notre base de données en base de données cubiques se fait via l'ETL (*Extract Transform Load*) Talend qui intervient essentiellement dans la chaîne décisionnelle, lors du processus d'intégration des données.

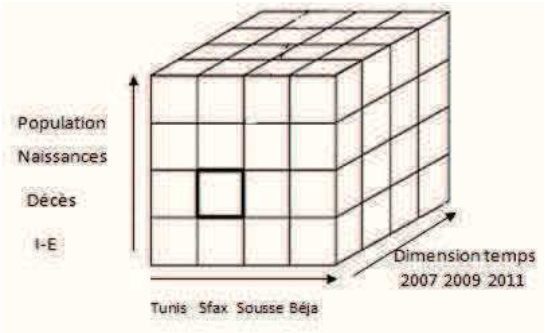


Figure 5.3 Structure de la base des données.

5.3.2 Sous- système d'extraction des motifs *EPS*

Le processus commence par la prise en compte d'une base de données géographiques et les technologies de fouilles des données. Puis le sous-système d'extraction intègre le SGBD. Les développements se font dans l'environnement *Java 1.7* et SGBD *PostgreSql 9.2*. En outre, notre système fournit aux utilisateurs des informations pertinentes sur les résultats partiels, c'est-à-dire, les utilisateurs peuvent interagir avec ce sous-système afin d'affiner les requêtes ou les fonctions, améliorant ainsi la qualité des résultats.

Nous allons extraire quatre types de motifs : les flux, les regroupements, les faits et les co-localisations.

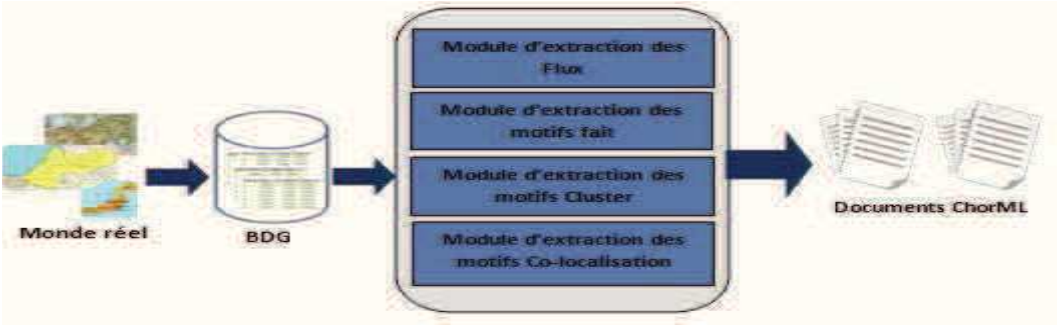


Figure 5.4 Architecture du sous-système EPS.

5.3.2.1 Sous système d'extraction des flux

Dans notre étude, nous allons nous intéresser aux flux migratoires et aux flux de marchandises. Dans les paragraphes ci-dessous nous présentons nos formules d'extraction de ces types de flux.

- **Cas n°1 : Extraction des flux de migration**

Pour extraire des connaissances liées au flux de migration entre les villes, nous avons mis au point une méthode qui permet d'étudier la migration à partir du lieu de naissance ; à travers des chiffres du recensement sur la province nous pouvons obtenir une bonne approximation sur les mouvements migratoires. Cette méthode [22] se fait par la comparaison entre le nombre de personnes nées dans une province par rapport à leur province de résidence actuelle.

Pour extraire les motifs relatifs aux flux de migration, il faut passer par trois étapes :

1. *(Intégration des données*) est chargée de stocker les données dans la base de données. En effet, les données descriptives sont indépendantes par rapport aux formes géométriques situées dans les couches thématiques.
2. *(Sélection et représentation)* est équipée des fonctionnalités de bases d'un SIG. Elle permet d'extraire et d'afficher les couches thématiques à partir de la base de données stockées. Ainsi, d'autres options avancées comme l'ajout ou l'élimination des couches, la modification de leurs apparitions, l'affichage des labels, etc. Ce sous-module permet également d'analyser les données.
3. *(Stockage)* permet de stocker les connaissances dans un document ChorML.

Pour déterminer les motifs flux de migration, il nous faut tout d'abord :

1. Calculer la migration nette de chaque ville,
2. Déterminer un seuil à l'aide d'un expert,
3. Déterminer les villes réceptrices qui ont une valeur de migration nette supérieure à ce seuil
4. Enfin, dans chaque ville réceptrice, nous comparons le nombre de personnes nées dans une province par rapport à leurs provinces de résidence actuelles puis à laide d'un seuil nous pouvons déduire les départs et les arrivées de flux migratoires entre les villes.

- **Cas n°2 : Extraction des flux de marchandise**

Ce deuxième module est destiné à extraire des connaissances liées aux flux de marchandises entre les villes (entre les clusters). Notre méthode [22] permet d'étudier les sources d'importations des produits choisis, à travers les chiffres de recensement sur la ville, les quantités consommées et produites des produits par ville, nous pouvons obtenir les sources possibles du produit choisi.

Cette méthode se fait par la comparaison entre la quantité produite et la quantité consommée par une ville pour un produit choisi par l'expert.

Dans notre système, nous travaillons sur les bases de données géographiques, où les données sont hétérogènes. En effet, les données d'attributs sont indépendantes des formes géométriques stockées dans les couches thématiques distinctes. Le résultat est stocké dans une base de données cubique par l'ETL (Extract, Transform, Load).

Après le stockage des données dans la base de données cubique, nous appliquons la méthode d'extraction des flux des marchandises. Cette méthode consiste à :

1. Calculer les dépenses de marchandises de chaque ville.
2. Déterminer la ville réceptrice et le produit à étudier à l'aide d'un expert.
3. Déterminer les villes qui ont une valeur de dépense supérieure au seuil saisi. Ces villes sont considérées comme des villes émettrices du produit choisi.
4. Finalement, les flux extraits sont stockés dans un fichier ChorML.

5.3.2.2 Sous système d'extraction des regroupements

Comme nous l'avons mentionné dans notre étude dédiée aux diverses méthodes de fouille de données spatiales, la méthode clustering spatiale et plus précisément l'algorithme *k*-means semble la plus appropriée pour extraire des connaissances sous forme de motifs de type cluster.

Via la méthode *k*-means, la distance euclidienne et les différents attributs de la base de données sont déjà choisis par l'utilisateur. Nous pouvons sélectionner les villes les plus proches d'entres elles et qui partagent les mêmes caractéristiques dans la BD.

L'architecture du système d'extraction du motif cluster est composée de trois principaux sous-modules:

- *Le sous-module de filtrage de données qui* permet de filtrer les données selon le choix de l'expert soit manuellement par la saisie de requête soit automatique par la sélection des listes.
- *Le sous-module de sélection et d'analyse* qui est équipé pour extraire et afficher des clusters contenant des villes qui sont proche et ont des caractéristiques communes. Le déroulement de ce module est explicité comme suit : (i) Utiliser la base de données selon le filtre de données (ii) Appliquer l'algorithme *k*-means en passant par quatre étapes :
  1. Chaque ville sélectionnée à travers le filtre sera considérée comme une classe *C*.
  2. (Ré) affecter chaque ville *V* au cluster C_i de centre *Mi* tel que $dist(V,M_i)$ soit minimale
  3. Recalculer M_i de chaque cluster (le centre géométrique)
  4. Aller au deuxième étape si on vient de faire une affectation.
- *Le sous-module de stockage* qui permet de stocker les connaissances extraites dans un fichier ChorML.

5.3.2.3 Module d'extraction des motifs faits

Dans ce module, nous allons extraire les villes importantes puis nous allons les représenter dans un document ChorML. L'architecture de notre méthode d'extraction de motifs de type fait est composée de deux sous-modules :

- *Le sous-module de sélection et d'analyse* qui est équipé pour extraire et afficher les villes que nous considérons comme des villes importantes.
- *Le sous-module de stockage* qui permet de stocker les connaissances extraites dans un fichier ChorML.

5.3.2.4 Module d'extraction les motifs de co-localisation

Après l'extraction des faits et les clusters qui regroupent un ensemble de villes proches et qui partagent les mêmes caractéristiques, nous avons élaboré un algorithme qui nous permet d'extraire les relations topologiques (*meet, contains, disjoint*) entre les clusters et les villes qui seront représentées dans la même carte chorématique.

L'architecture de notre méthode d'extraction de motif type co-localisation est composée de deux modules :

- *Le module d'extraction de relation topologique* qui permet d'extraire les relations topologiques entre les clusters extraite qui sont stockées dans le document ChorML et les villes sont stockées dans la base de données.
- *Le module du stockage* qui permet de stocker les connaissances extraites dans un document ChorML.

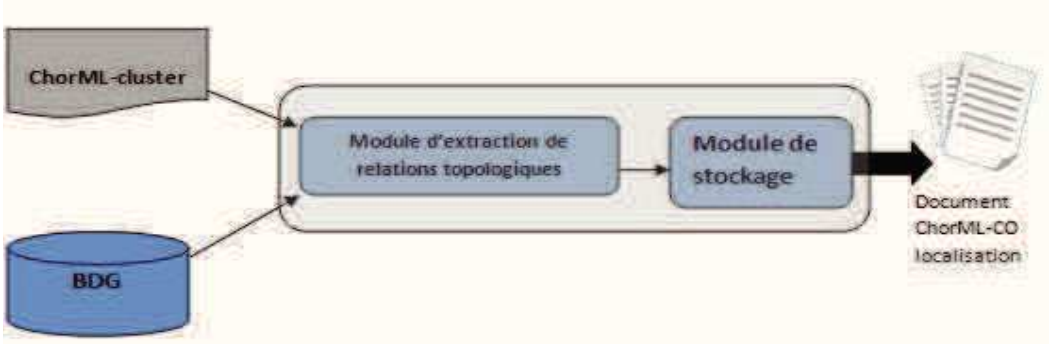


Figure 5.5 Architecture du sous-module d'extraction de motifs de type co-localisation.

Le sous-système d'extraction des motifs permet d'extraire le motif de type cluster, fait, flux et co-localisation et genère un document ChorML 1v0 à l'aide d'un générateur ChorML.

Rappelons que le ChorML Generator [19, 20] a pour finalité la génération d'un document XML selon la spécification du langage ChorML. Ce document fait le lien entre le sous-système d'extraction de motif et le sous-système d'extraction des motifs importants. Ce générateur est composé de deux modules :

- *Le module de manipulation des données* qui récupère les informations extraites et les injecter dans le module du générateur du ChorML, il permet aussi de faire des modifications sur le ChorML déjà généré.
- *Le module de la génération de ChorML* qui a pour objectif la construction d'un document XML selon la spécification du langage ChorML, Il permet aussi la génération

complète d'un document ChorML, en fusionnant les documents générés " l'entête " et " les informations complémentaires ou extra-données " aussi la génération des frontières ajoutées aux cartes chorématiques des informations extérieures comme les noms des pays voisins, des mers, etc.

5.3.3 Sous-système d'extraction des motifs importants *ESPS*

L'approche que nous proposons a pour objectif la découverte des motifs saillants. L'étude sur l'état de l'art a montré qu'une partie des contributions actuelles visait à introduire la subjectivité nécessaire à la découverte de ces motifs, par le biais de contraintes spécifiées par l'utilisateur.
La figure 5.6 présente l'architecture générale de l'approche proposée.

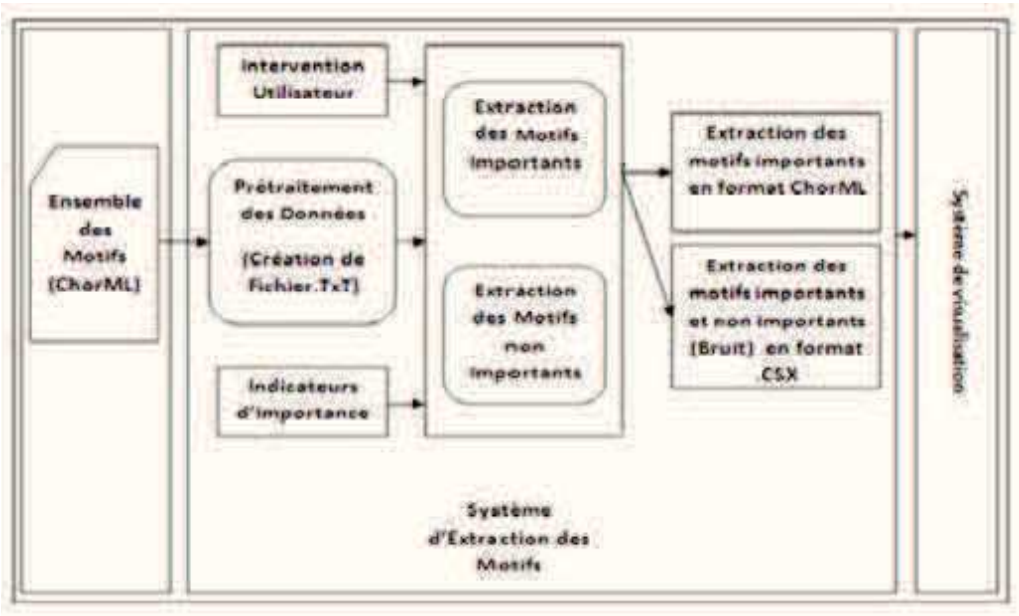


Figure 5.6 Système d'extraction des motifs saillants.

Les problématiques de modélisation de la connaissance évidemment de nombreuses difficultés. Il faut être en mesure de trouver un accord entre les différents experts du domaine et mobiliser des équipes pour la formalisation et la construction du modèle. Il s'agit d'une activité qui demande généralement un investissement important en terme du temps et d'énergie dépensée. De plus, les bénéfices réels que l'on pourra retirer de l'utilisation du modèle s'avèrent difficile à évaluer a priori. Face aux problématiques de recherches demeurant « ouvertes » nous avons choisi d'aborder les axes suivants:

- ***Élimination de redondance***, plusieurs pistes peuvent être explorées pour détecter la redondance et la nouveauté lors du filtrage.
  La redondance et l'importance sont deux concepts difficiles à définir clairement. La redondance peut être vue selon les différents angles :

  - un motif peut être une copie conforme d'un autre motif,
  - un motif peut être une copie d'un autre motif écrit dans une autre langue,
  - le contenu d'un motif peut être une partie d'un autre motif.

Plusieures pistes peuvent être explorées pour détecter la redondance et la nouveauté lors du filtrage. Notre réflexion est de proposer une formule mathématique qui sert à calculer le poids de redondance d'un motif (PRM).

- ***Intégration des contraintes utilisateur.*** Cette intégration constitue un pas vers l'utilisation des connaissances, en intégrant le jugement de l'expert au processus d'extraction des motifs. Dans notre approche, la définition du profil d'utilisateur influe sur la personnalisation du motif important en choisissant l'attribut à exclure. Le but de la personnalisation est de faciliter l'expression du besoin de l'utilisateur et de lui permettre d'obtenir des informations pertinentes lors de ses accès à un système d'information. La pertinence de l'information se définit par un ensemble de critères et de préférences personnalisables et spécifiques à chaque utilisateur ou communauté d'utilisateurs. Les données décrivant les utilisateurs sont souvent regroupées sous forme de profils.
Parmi les données qui constituent un profil utilisateur, on trouve une dimension relative à la qualité. Elle permet d'exprimer des préférences extrinsèques comme l'origine de l'information, sa précision, sa fraîcheur, sa durée de validité, le temps nécessaire pour la produire ou la crédibilité de sa source. Il est, par exemple, possible de poser une requête en spécifiant des préférences en termes de qualité comme une réponse rapide ou une information fraîche. Donc, la définition d'un motif important est relative à l'utilisateur. Plus précisément, à la satisfaction de ses besoins en termes de choix.

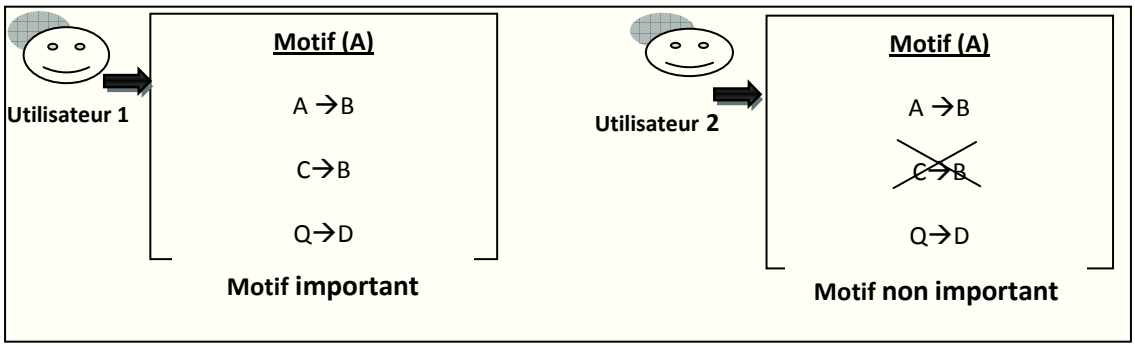


Figure 5.7 Intervention de deux utilisateurs où, devant le même motif, l'un le déclarant important et l'autre non.

- ***Système d'annotation supervisé.*** La tâche de l'expert est d'analyser et d'interpréter les connaissances extraites. Pour cela, nous allons mettre à sa disposition un système d'annotation qui va lui permettre de porter un jugement sur les connaissances et en particulier sur les motifs qui la composent.
En pratique, il n'est pas rare de retrouver un motif déjà connu ou inintéressant dans un grand nombre de connaissances (par exemple, une forêt est remplie d'arbres, mais une forêt sans arbre attire l'attention !). L'expert sait reconnaître ces motifs et il nous est apparu nécessaire de lui donner les moyens de les indiquer par le biais d'une syntaxe bien définie. Ces annotations vont avoir un intérêt double : d'une part, le moteur d'affichage des connaissances va pouvoir rendre compte de la nature des différents motifs qui composent chaque connaissance. D'autre part, ces annotations sont utilisées pour faciliter la mise à jour de l'ensemble des motifs importants. Les annotations peuvent être de trois types :

- *Non valide (Nv) :* Le motif contient une connaissance non valide de l'expert. Il est alors possible d'intégrer la notion de ce motif à la base de données. À l'itération suivante du processus, l'utilisateur ne verra plus apparaître ce type de motifs car ils seront jugées non valides.
- *Non Important (NImp) :* Ces annotations décrivent une relation valide mais non pertinente par rapport au contexte dans lequel se situe l'expert.
- *Important (Imp) :* L'annotation est intéressante. C'est-à-dire, qu'elle surprend l'expert du domaine et exige une analyse approfondie.

Lorsque l'expert rédige les annotations, il a pour objectif qu'un maximum de règles contenant uniquement des motifs de type *Nv* et *NImp*, soient éliminées lors de la prochaine itération du processus. Ce filtrage peut intervenir par le biais de la mesure d'intérêt subjective (une modification du l'ensemble de motifs en fonction des annotations collectées doit diminuer l'intérêt de ces motifs).
En effet, nous souhaitons que les connaissances affichées intègrent un maximum d'informations relatives aux annotations, dans le but de faciliter les étapes ultérieures de l'analyse. L'idée est d'affiner progressivement le modèle de connaissance utilisé pour le filtrage des connaissances en y intégrant les jugements récemment découvertes. Cependant, l'interprétation des motifs extraits et le choix des modifications à apporter à la base d'annotations et à l'ensemble de motifs sous la forme de fichier ChorML ne sont pas des tâches faciles à réaliser.

En définitive, étant donnée l'importance de l'utilisateur, nous avons affaire à un système d'annotations supervisées.

5.3.4 Sous- système de visualisation *VS*

La visualisation de l'information aide l'utilisateur à acquérir et à accroître ses connaissances (reconnaissance rapide de motifs, couleurs, formes et textures). Les utilisateurs sont dotés d'une capacité à visualiser l'information très développée. Nous utilisons des méthodes graphiques afin de mieux accueillir des notions abstraites ou pour représenter ce qui l'entoure.
Notre système donne la possibilité de visualisation des connaissances sous trois formes :

- *Les treillis de concepts* [54], pour représenter les connaissances sous forme d'une hiérarchie de concepts. Nous avons fait appel aux techniques de visualisation permettant d'afficher les résultats d'une manière compréhensible par l'être humain. Notre système permet l'affichage sous la forme de treillis imbriqués, des structures de données mappées et mémoire : les fichiers .CSX.

- *Les graphiques.* Comme ils sont montré dans la figure 5.8, nous présentons un graphique des motifs et leurs degrés d'importance.

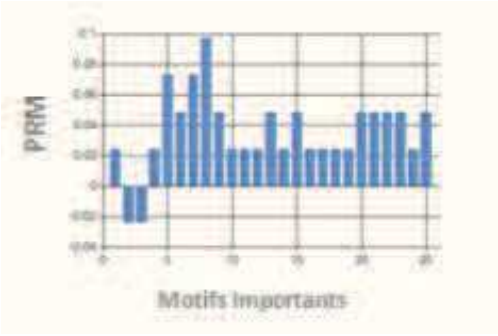


Figure 5.8 Les motifs saillants sous forme de graphique.

- *Les cartes chorématiques* qui représentent schématiquement l'espace choisi selon les méthodes décrites au chapitre 4. Nous avons proposé un nouveau système pour la visualisation des résumés visuels qui tourne autour de trois phases principales : consécutives : le prétraitement des coordonnées géographiques, la création des chorèmes et puis leur affichage. Nous rappelons que l'interaction entre les phases de notre système est assurée par un niveau approprié du langage ChorML.

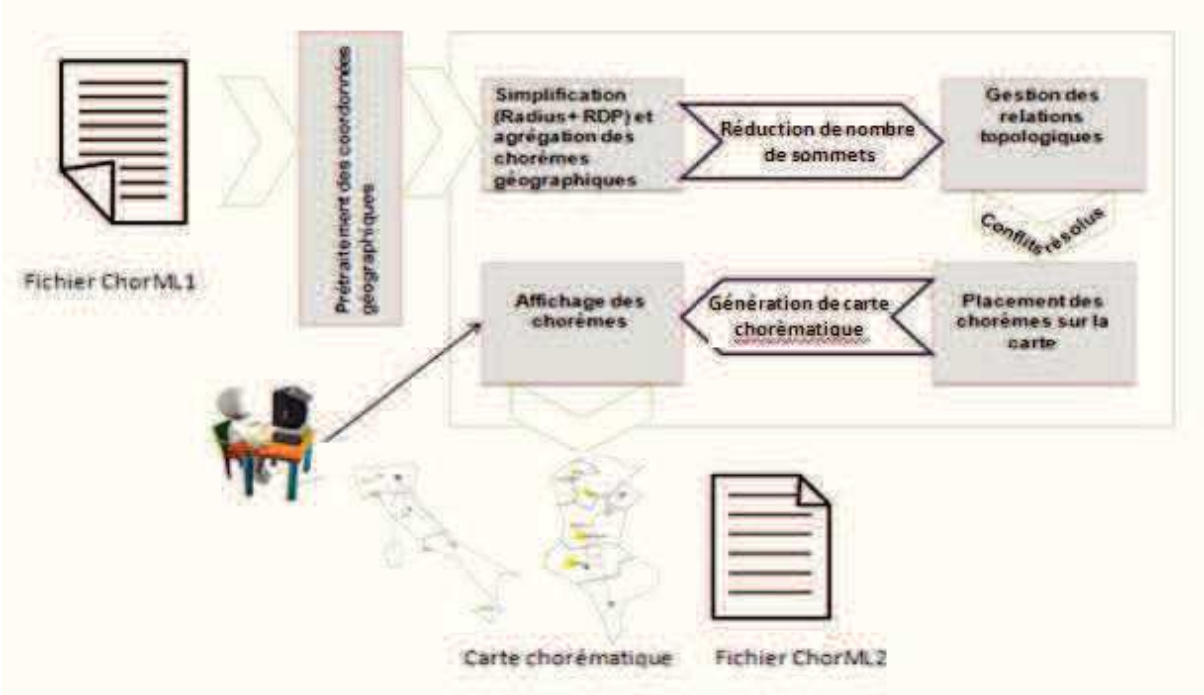


Figure 5.9 Architecture détaillée de notre approche de visualisation.

Il s'agit de fournir à l'utilisateur une compréhension qualitative du contenu de l'information. Cette visualisation doit permettre à l'utilisateur final de faire des découvertes, proposer des explications ou prendre des décisions. Ces actions peuvent se faire aussi sur des motifs (clusters, tendances, émergences, anomalies) ou sur des ensembles d'éléments ou encore sur des éléments isolés.

.

5.4 Expérimentation

5.4.1 Etude de cas n°1 : Les migrations internes en Tunisie

Afin de montrer les potentialités de l'approche que nous venons de proposer dans cette recherche et des résultats obtenus, nous avons mis en place une application informatique. En effet, nous avons mis en œuvre notre proposition théorique au sein d'une application dédiée au l'extraction des motifs importants et les visualiser.

Nous présentons le déroulement de différents processus de notre application à travers une étude de cas décrivons les migrations internes dans la Tunisie un rôle assez important dans la répartition spatiale de la population.

- *La migration interne en Tunisie:*

La migration joue un rôle assez important dans la répartition spatiale de la population aussi bien au niveau national que régional de par l'impact qu'elle peut avoir sur la croissance de la population d'une zone géographique donnée.
Le recensement de 1984 [69], donne une évaluation des flux migratoires internes dans chaque région et dans chaque milieu.
On retrouve deux types de migration:

1. les migrations " traditionnelles " et les migrations " modernes ". Chacun de ces deux types de migrations est déterminé par des facteurs spécifiques étroitement liés aux caractéristiques économiques et démographiques de la région considérée.
2. Les migrations traditionnelles sont des mouvements très anciens qui se manifestent principalement par l'émigration de populations rurales de la Tunisie du Sud saharienne et présaharienne vers le Nord. Elles se portent d'une manière privilégiée, mais non exclusive, vers la capitale. Ces mouvements semblent être la conséquence de la crise économique de la région du Sud.

En ce qui concerne les migrations modernes [69], ces mouvements proviennent particulièrement des régions du Nord-Ouest. Ils semblent être le résultat de deux types de causes : les causes structurelles (il s'agit de la crise des structures agraires notamment dans la Vallée de la Medjerda et dans les plaines du Kef, caractérisées par une forte concentration de la propriété et de l'exploitation, où s'opèrent une mécanisation rapide et une modification du système de culture) et les causes secondaires liées essentiellement de la crise de l'emploi. Au niveau régional, la corrélation entre l'émigration et le sous-emploi semble être vérifiée : les régions de faible sous-emploi (Tunisie centrale, en particulier l'arrière-pays de Sfax) sont aussi des régions de faibles départs ; de même, les régions où les taux de sous-emploi sont en général supérieurs à la moyenne (Sud-Est, Sahel, Nord-Ouest) sont les principales zones d'exode.

- Extraction des Motifs Flux

L'extraction des motifs c'est la détermination des villes réceptrices et des villes émettrices, à l'aide de la formule décrite au chapitre 4, le système peut identifier les flux de migration.
Au début l'utilisateur accède à l'interface *Flux* montré dans la figure 5.8, il choisit la date et le seuil de migration nette qui a été déterminé par un expert de domaine.
Lorsque la saisie de date et de seuil est terminée une liste des villes les plus réceptrices d'immigrants sera affichée (cf figure 5.15), ensuite il faut déterminer les villes émettrices d'émigrants qui sera extraite selon un autre seuil déterminé aussi via un expert (le nombre de personnes qui ont une ville de résidence parmi les villes trouvées dans la liste de villes réceptrices et qui ont une ville de naissance différente de celle-ci.

Figure 5.15 Phase de sélection de l'année et seuil
pour l'extraction des flux de migration.

A fin de déterminer les flux de migration entre les villes, un document ChorML sera créé qui contient six parties:

- Le nom de flux
- Le nom de base de données source et la date de création de fichier.
- Le type de motif
- L'origine de notre flux (la ville émettrice)
- La destination de flux (la ville réceptrice)
- Le type d'objet et les coordonnées spatiales de flux (extraite via un outil s'appel OpenJUMP)

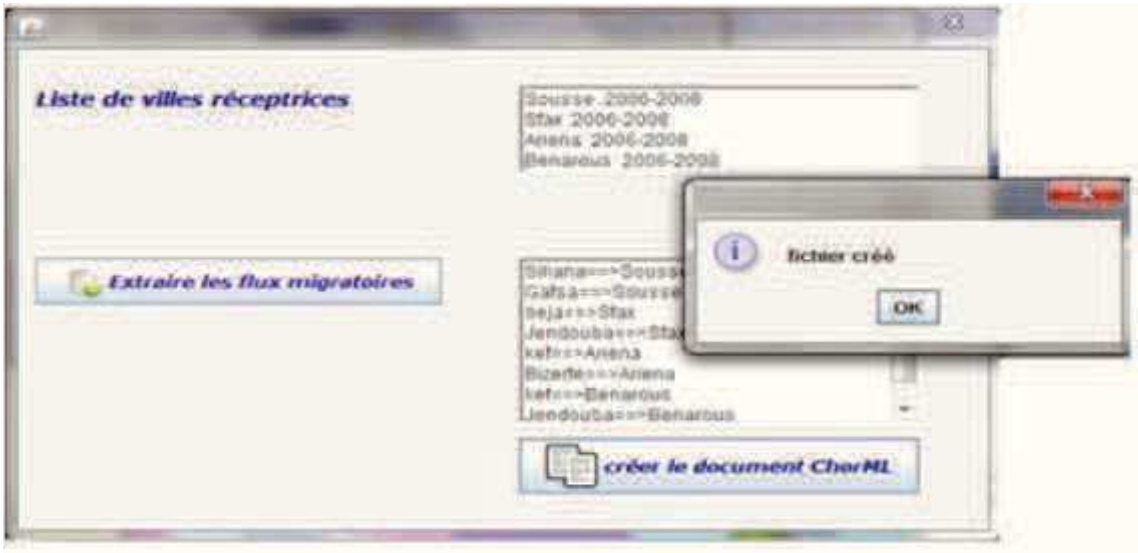


Figure 5.16 Extraction des flux migratoires et création du fichier ChorMLflux.

- Extraction des Motifs cluster:

Après l'extraction des différents flux migratoires dans la phase précédente, l'utilisateur peut regrouper les différentes villes réceptrices et émettrices dans différents regroupements (Zones) à travers l'interface " Cluster " présentée dans la figure 5.17. Tout d'abord, le système offre à l'utilisateur la possibilité de filtrer les données. Ici le système nous offre deux choix :

- Soit le filtrage par requête : l'objectif est de permettre à l'utilisateur de lancer des requêtes spatiales et non-spatiales sur la base de données. Cette interface permet à l'utilisateur l'interrogation de la base de données en utilisant une approche de type SQL. Les requêtes sont créées en temps réel et ensuite soumises à la base de données pour répondre à la demande de l'utilisateur.
- Soit par plusieurs options sont possibles :

1) Sélectionner le nom de base dans la liste de base de données, le nom de table, les attributs et les opérateurs mathématiques à utiliser.
2) Préciser la date et le seuil.
3) Déterminer le nombre de groupes à extraire ; par la suite, le système construit la requête et interroge la base des données.

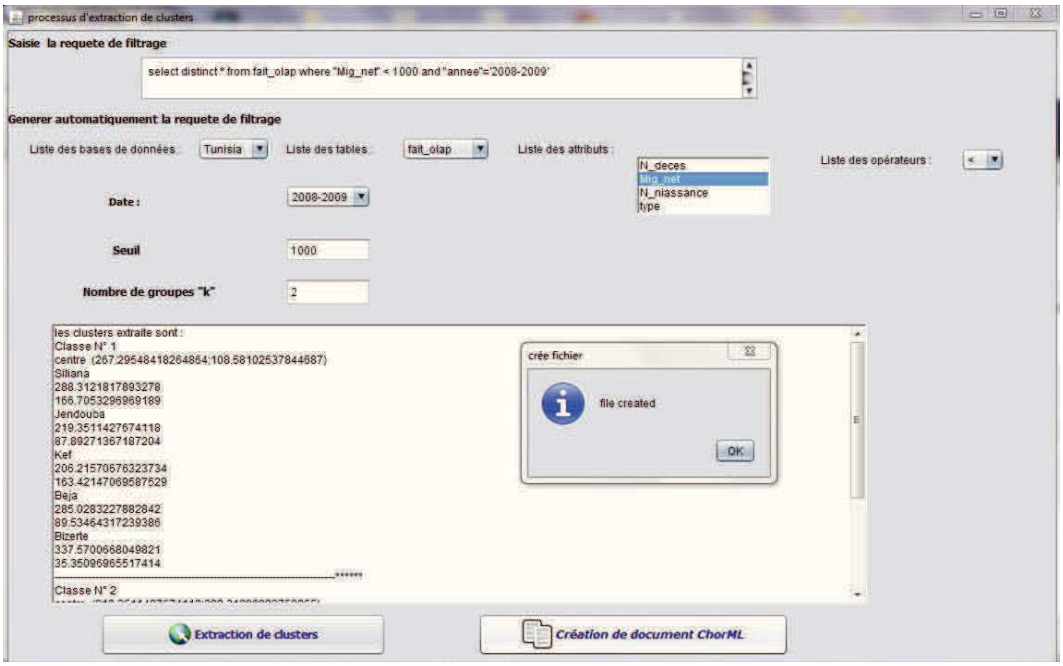


Figure 5. 17 Extraction des clusters.

Après le filtrage, l'utilisateur extrait les classes (regroupements), les résultats seront affichés dans une liste qui contient le numéro de classe, le centre de chaque classe de coordonnées (*x,y*), les villes qui appartiennent à chaque classe et leurs centres de coordonnées *(x, y).*

Après le processus d'extraction de clusters Les résultats sont stockés dans un document ChorMLCluster qui contient :

- Le nombre de classes
- Nom de la base de données, la date de création de fichiers
- Type de distance utilisée dans le regroupement
- Numéro de cluster
- Formes et coordonnées géographiques de clusters
- Les éléments de clusters.

- Extraction des co-localisations

C'est l'extraction des relations topologiques entre les clusters et les Faits, une interface affiche à l'utilisateur les numéros de classes et les types des relations entre ces derniers avec les restes de villes qui seront affichées dans la même carte chorématique.

Après le processus d'extraction des relations entre les clusters et les villes le système stocke les résultats dans un document *ChorMLco-localisation* qui contient :

- Le type de relation
- Les numéros de clusters
- Les numéros de villes.

Après ces différentes étapes le système fusionne les différent ChorML en un seul document ChorML1 qui sera une entrée au sous système d'extraction des motifs importants.

- Extraction des motifs importants

L'utilisateur peut appliquer une autre fois un filtrage sur les motifs déjà extraits à l'aide de notre formule d'importance par la suite il peut visualiser les résultats comme présenté dans la figure 5.18.
Il peut voir les motifs de départ et les motifs importants aussi forment de graphique.
La présence d'un graphique des motifs importants permet d'aider l'utilisateur à identifier les résultats.

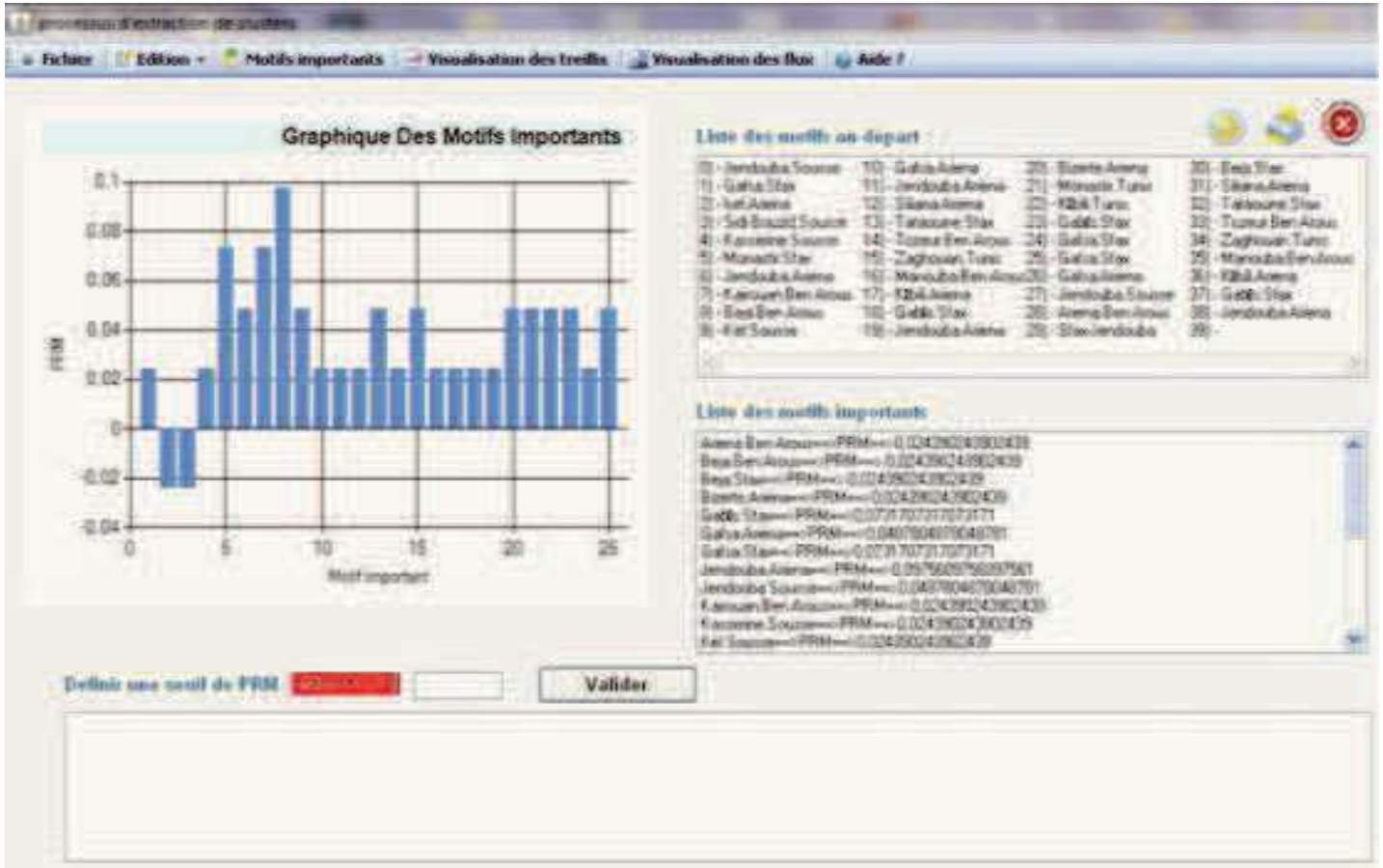


Figure 5.18 Présentation des motifs importants.

- La visualisation des résultats

La visualisation des résultats est basée sur la technique des treillis de concepts pour présenter les connaissances sous forme d'une hiérarchie de concepts comme montré dans la figure 5. 20.

- Un ensemble d'objets qui présente par des motifs importants extraite a l'aide de notre approche.
- Un ensemble d'attributs ou on trois attributs qui présentent une degré d'importance de motif :
  - ✓ **Bruit** : un *PRM* motif négatif (*-PRM*)
  - ✓ **Normal** : un *PRM* motif égale 1/*n* (*n* : nombre de motif)
  - ✓ **Important** : un *PRM* motif supérieur à 1/*n* (*n* : nombre de motif)

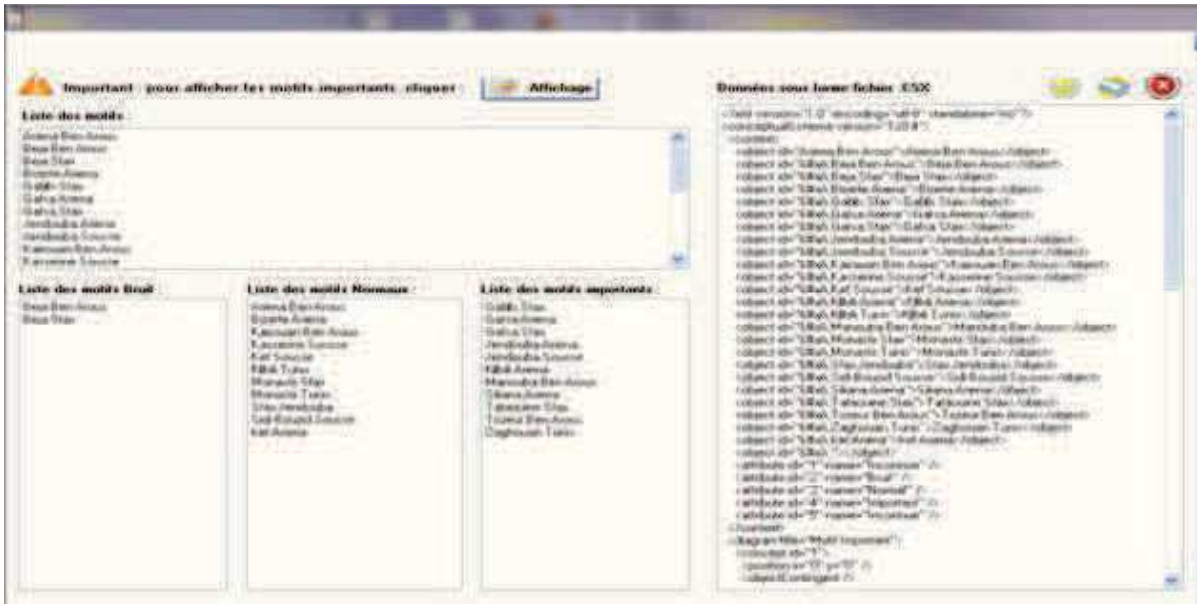
Pour l'affichage du treillis de concepts, nous avons intégrer le logiciel " ToscanaJ " [http 34] vu sa simplicité d'utilisation. C'est une application qui s'intéresse à l'affichage sous la forme de treillis imbriqués, d'un ensemble de données interrogées à partir d'une base de données ou en utilisant des structures de données mappées et mémoire comme dans notre cas nous allons travailler par un fichier .CSX (Cf. Fig 5.19) Elle permet aussi de montrer les nœuds ayant certains liens d'intérêt avec un nœud choisi.

motifimportant.csx

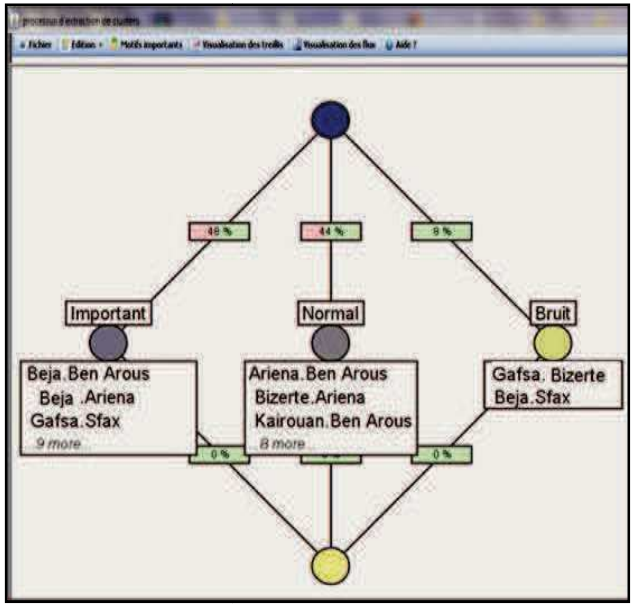
```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<conceptualSchema version="TJ0.4">
  <context>
      <object id="Ariena.Ben Arous">Ariena.Ben Arous</object>
      <object id="&#xA;Beja.Ben Arous">
          Beja.Ben Arous</object>
      <object id="&#xA;Beja.Sfax">
          Beja.Sfax</object>
      <object id="&#xA;Bizerte.Ariena">
          Bizerte.Ariena</object>
      <object id="&#xA;Gab□s.Sfax">
          Gabes.Sfax</object>
      <object id="&#xA;Gafsa.Ariena">
          Gafsa.Ariena</object>
      <object id="&#xA;Gafsa.Sfax">
          Gafsa.Sfax</object>
      <object id="&#xA;Jendouba.Ariena">
          Jendouba.Ariena</object>
    <attribute id="1" name="Inconnue" />
    <attribute id="2" name="Bruit" />
    <attribute id="3" name="Normal" />
    <attribute id="4" name="Important" />
    <attribute id="5" name="Inconnue" />
  </context>
  <diagram title="Motif Important">
    <concept id="1">
      <position x="0" y="0" />
      <objectContingent />
      <attributeContingent />
```

Figure 5.19 Visualisation des clusters au format CSX.

Aussi, on peut les visualiser à l'aide de notre système de visualisation des chorèmes comme montré dans la figure 5.21.



(a) La visualisation des motifs importants et non importants



(b) La visualisation du treillis de concept

Figure 5. 20 La visualisation des motifs.

La figure 5.21 présente la carte chorématique qui a été générée grâce aux résultats obtenus à partir de la phase de génération de ChorML qui décrit le nombre des migrants d'une ville tunisienne à l'autre, en tenant compte du seuil proposé. Cette carte propose une vue d'ensemble des motifs saillants du processus de migration et devient alors un support pour la prise de décision.

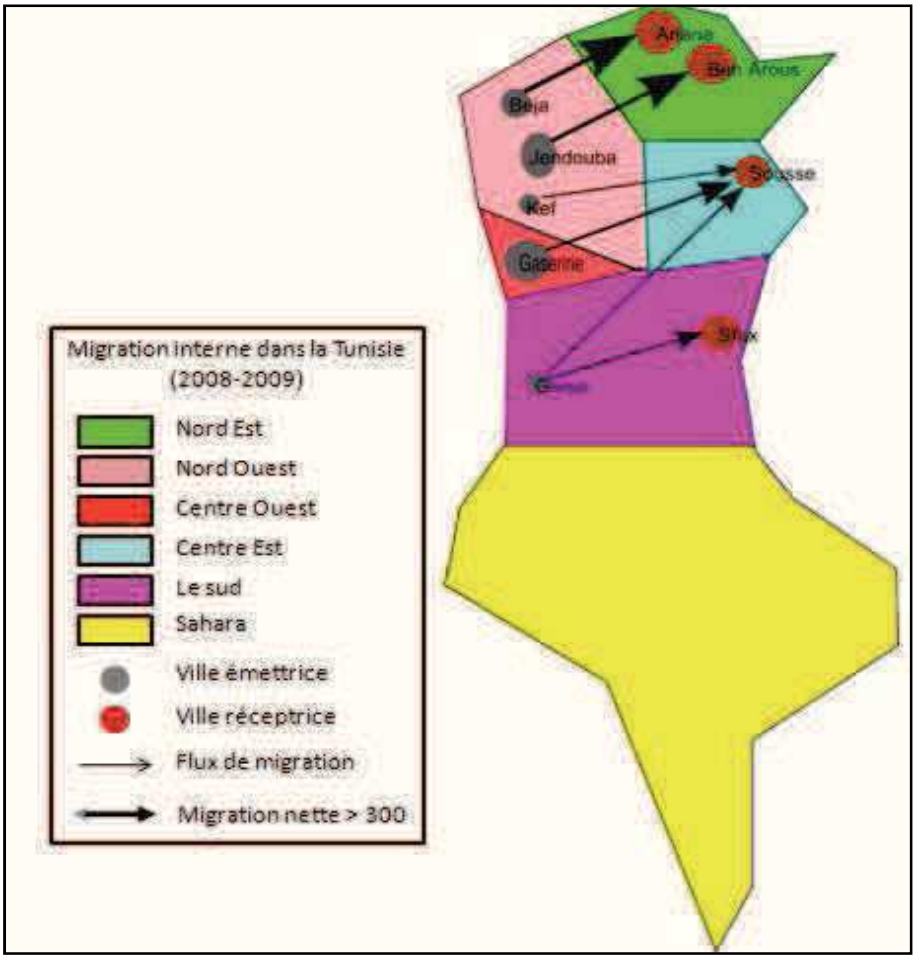


Figure 5. 21 Carte chorématique de la migration interne dans la Tunisie.

Cette carte chorématique (figure 5.21) créée par notre système de visualisation des chorèmes, montre la migration interne en Tunisie, qui est divisée en six grandes régions: Nord-Ouest, le Midwest, du Centre-Est, du Sud, le Sahara et la région de Tunis. Les villes réceptrices et émettrices sont représentées par des cercles, leurs tailles varient selon le nombre des immigrés ou les émigrants et les flèches d'épaisseur variable représentent les flux de personnes.
Ainsi, on peut voir :

- Un pourcentage élevé d'individus sortants du gouvernorat de Jendouba, Kef et Béja qui se déplacent vers l'est.
- De Kassrine et Gafsa, les migrations sont ciblées en particulier vers Sousse, Monastir et Sfax,
- Le nord-ouest envoie un grand nombre de migrants d'autres régions tout en recevant le nombre presque constant des arrivées.
- Un pourcentage élevé des personnes sortant de la région du Nord-Ouest choisit le gouvernorat de Tunis pour la destination.

5.4.2 Etude de cas n°2 : Les flux de marchandises en Tunisie

Le transport intérieur des marchandises correspond aux transports effectués par modes routier, ferroviaire ou fluvial. D'après les définitions internationales, le transport désigne un flux de marchandises déplacées sur une distance donnée et se mesure en tonne-kilomètre. Suivant [1] et compte tenu de la grande quantité d'informations numériques disponibles dans le monde, les statisticiens ont la tâche difficile de faire en sorte que les analystes des entreprises aient rapidement accès à des données commerciales exactes.

En tant que source basique de l'économie tunisienne, les échanges intra-régionaux des produits agro-alimentaires se présentent comme des principaux flux de marchandises. Vue leur importance et étant donné la difficulté de leur analyse par les experts, nous trouvons intéressant de combler cette limite et de proposer une méthode plus facile. En fait, il serait intéressant de représenter ces flux sur une carte chorématique. Ceci offre une vision synthétique et aisée comme le taux massif des données, sur le territoire et les marchandises, va être remplacé par des formes et des symboles faciles à comprendre.

Dans ce qui suit, nous passons à la phase de test où nous utilisons en entrée du système notre un fichier provenant du sous système d'extraction contenant les flux des céréales entre les régions tunisiennes. L'ensemble des données concernant ce phénomène est tiré de l'INS (Institut National de des Statistiques) de la Tunisie et date de l'année 2010. Cet ensemble est filtré par un sous-système d'extraction des motifs les plus intéressants en appliquant un ensemble des techniques de fouille de données spatiales. Le résultat consiste en un fichier ChorML1. Ce fichier contient six clusters dont chacun présente un ensemble de régions partageant les mêmes caractéristiques. Certaines régions ont un excès de production de céréales. En effet, elles produisent plus que 100000 tonnes en 2010 ce qui satisfait énormément leurs besoins. D'autres régions ont, par contre, un manque de production (production inférieure à 100000 tonnes en 2010). En vue de réduire ces écarts, des échanges entre ces régions auront lieu. Nous parlons donc des flux interrégionaux de céréales.

- **Mise à jour du fichier ChorML1**

  Chaque composante de ce fichier possède un identifiant et ses coordonnées sont exprimées consécutivement en longitude-latitude et en *x, y* suite à l'application de la formule de Mercator.
  Nous pensons qu'il serait avantageux de fournir à l'utilisateur un accès aux données contenues dans le fichier ChorML1. Pour ce faire, nous stockons ces données dans une base SQL.

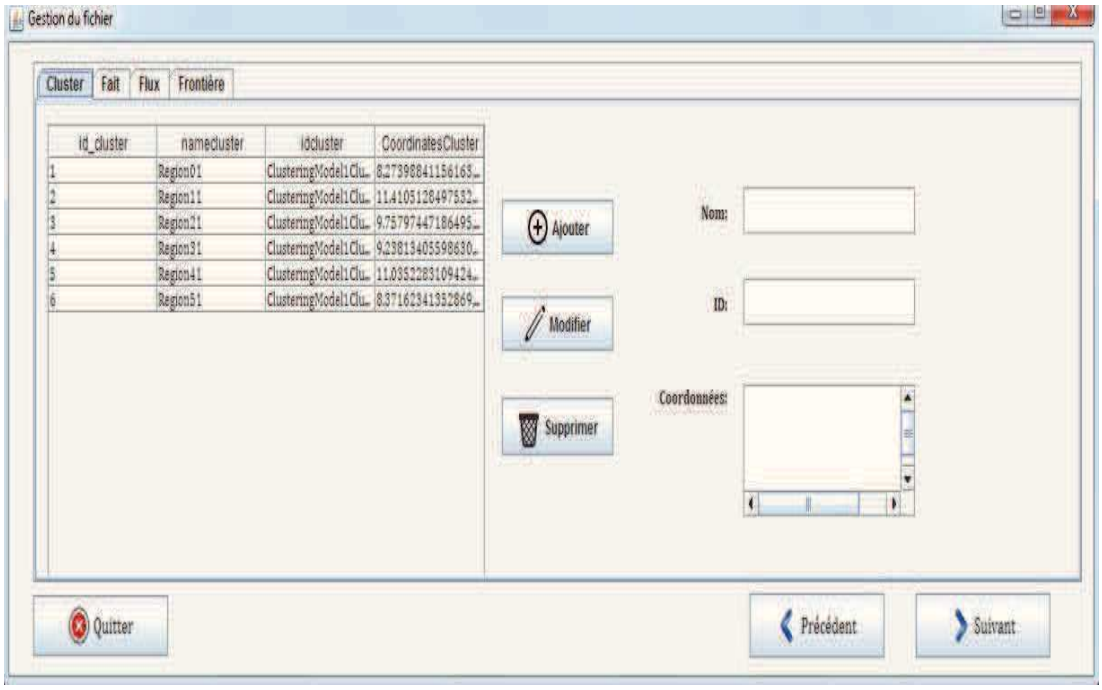


Figure 5.22 Mise à jour du fichier ChorML1.

L'accès à ces données est fourni, par notre système, comme le montre la figure 5.22 Les composantes du fichier sont réparties selon leurs types. Dès lors, la distinction des clusters, des faits et des flux devient plus aisée que dans un fichier ChorML1.
Via cette interface, nous offrons à l'utilisateur la possibilité de gérer son fichier d'origine. Il peut donc, le mettre à jour en ajoutant, modifiant ou supprimant un ou plusieurs éléments (chorème(s) géographique(s), chorème(s) d'annotation et/ou chorème(s) phénoménologique(s)).

- **Application des opérations de simplification et d'agrégation**

Après avoir géré le fichier, l'application des opérations de généralisation cartographique: la simplification et l'agrégation sur les formes géométriques des clusters s'effectuent. L'utilisateur peut intervenir dans ce processus en modifiant les valeurs de tolérance par défaut.
L'entrée de grandes valeurs implique des formes géométriques plus simples de l'ensemble des clusters.

Figure 5.23 Interface des opérations de simplification et d'agrégation.

- **Génération des résultats**

Après avoir résoudre l'ensemble des conflits topologiques et définir les emplacements finaux de tous les chorèmes, notre système génère les résultats. Ces derniers sont fournis à travers l'interface présentée dans la figure 5.24.

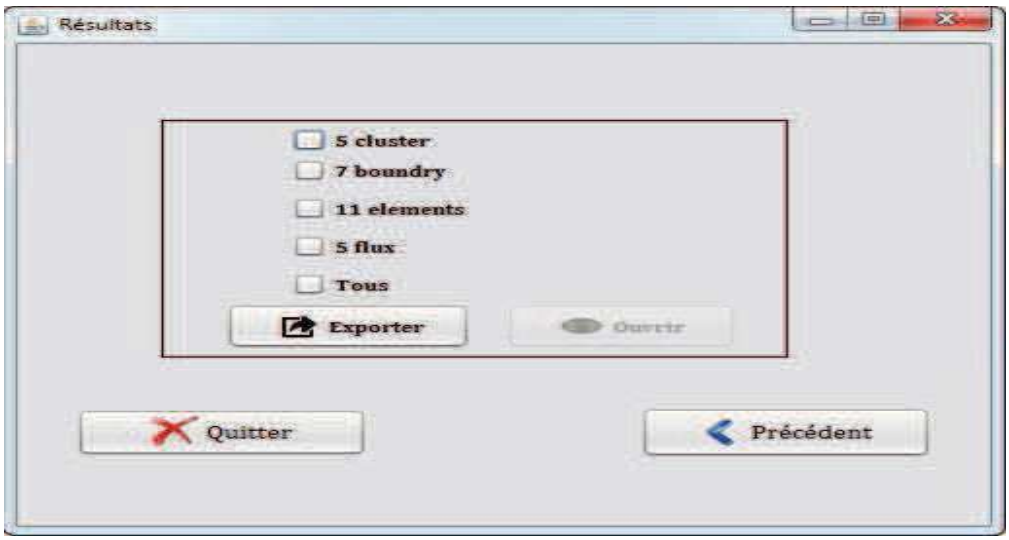


Figure 5.24 Exportation du fichier ChorML2 et génération de la carte chorématique.

Les résultats consistent en une carte chorématique affichée à travers l'éditeur Inkscape en cliquant sur le bouton « Ouvrir ». Un fichier ChorML2 est également produit.
Ce fichier est accessible en cliquant sur le bouton « Exporter ». Il est composé, comme indiqué précédemment, des tags XML et SVG. L'ensemble des éléments décrits par ChorML2 sont les suivants :

– Des métadonnées ;
– Une liste simplifiée des chorèmes géographiques qui résulte de la phase de simplification et d'agrégation ;

– Une liste décrivant les emplacements des chorèmes d'annotation obtenus suite à la correction des relations topologiques ;
– Une liste définissant les emplacements des chorèmes phénoménologiques ;

L'ensemble des modifications effectuées par l'utilisateur sur le fichier résultant.

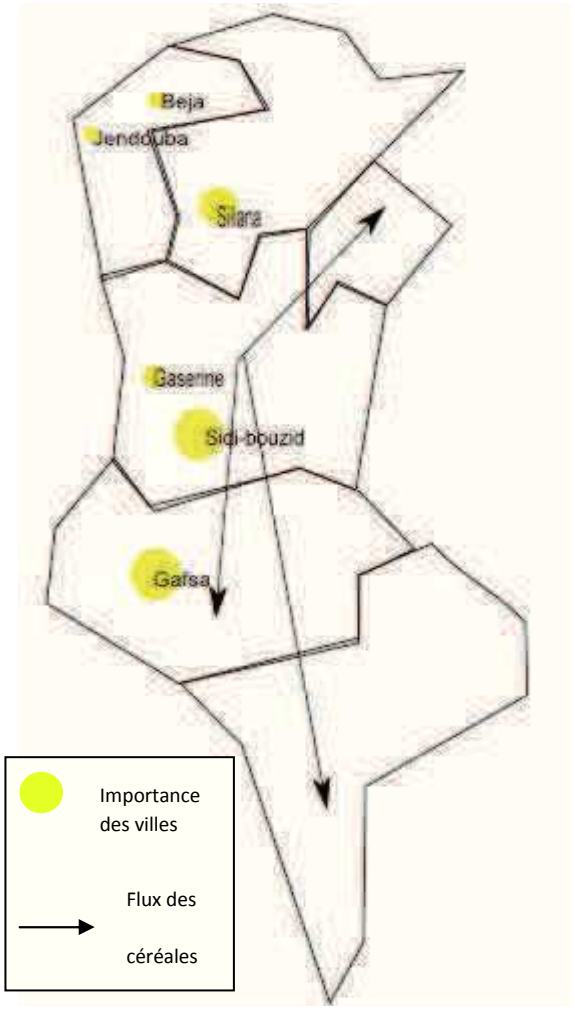


Figure (a). Carte chorématique produite avant affinement

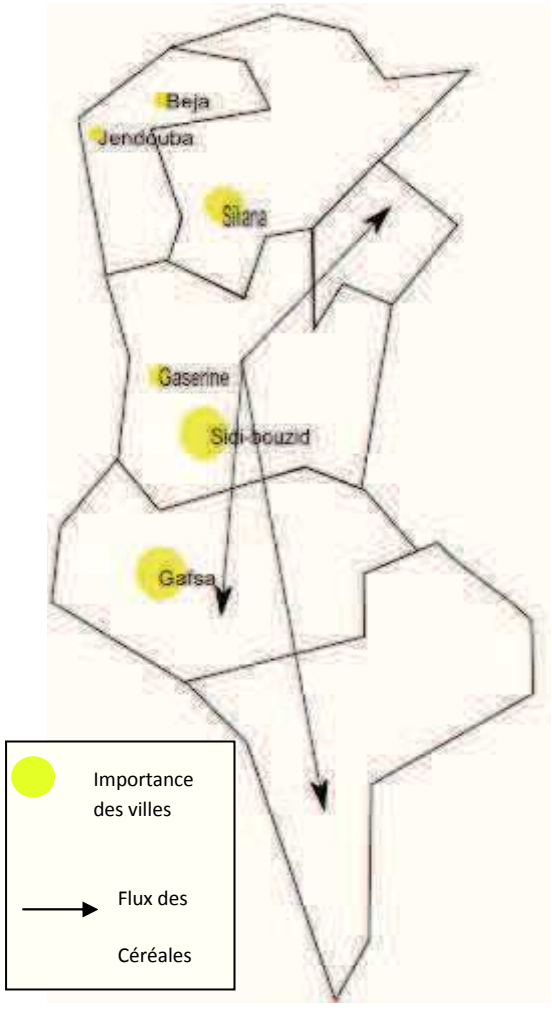


Figure (b). Carte chorématique produite après affinement

Figure 5.25 La carte chorématique produite avant affinement et après affinement.

La carte chorématique correspondante de la Tunisie et représentant les flux de céréales pendant l'année 2010 est présentée dans la figure 5.25. Nous distinguons, à travers cette carte, cinq macro-villes dont les quantités de céréales produites excèdent leurs consommations. Ces villes sont respectivement: Sidi Bouzid, Gafsa Siliana, Gasserine, Béja et Jendouba. Leur importance est exprimée à travers les diamètres des ellipses sur la carte (Cf. Fig 5.27).
Le surplus de production est distribué sur les régions qui en sont dépourvues (Sousse, Monastir, Mednine, Tataouine, etc.). Nous distinguons trois flux massifs dont les

épaisseurs sont proportionnelles aux quantités de céréales transmises d'une région à une autre**.** Ces flux ont lieu entre : Le centre et le Nord-est, le centre Sud-est et le centre et le Sud-ouest;

Nous présentons, dans la figure 5.26, un extrait du fichier ChorML2 généré par notre système.

```
<g
   id="Flows"
   transform="matrix(2.0658361,0,0,2.2949809,-442.80044,681.74989)">
  <path
     d="m 5324.2392,1462.6051 450.3753,-436.9979"
     id="Flow1"
     style="fill:none;stroke:#000000;stroke-width:1px;stroke-
linecap:butt;stroke-linejoin:miter;stroke-opacity:1;marker-end:url
(#Arrow1Lend1)" />
  <marker
     refX="0"
     refY="0"
     orient="auto"
     id="Arrow1Lend1"
     style="overflow:visible">
    <path
       d="M 0,0 5,-5 -12.5,0 5,5 0,0 z"
       transform="matrix(-1.1,0,0,-1.1,-1.1,0)"
       style="fill-rule:evenodd;stroke:#000000;stroke-width:1pt"
       id="path4409" />
  </marker>
  <path
     d="m 5324.2392,1467.0642 -40.1325,646.5785"
     id="Flow2"
     style="fill:none;stroke:#000000;stroke-width:2px;stroke-
linecap:butt;stroke-linejoin:miter;stroke-opacity:1;marker-end:url
(#Arrow1Lend2)" />
  <marker
     refX="0"
     refY="0"
     orient="auto"
     id="Arrow1Lend2"
     style="overflow:visible">
    <path
       d="M 0,0 5,-5 -12.5,0 5,5 0,0 z"
       transform="matrix(-1.1,0,0,-1.1,-1.1,0)"
       style="fill-rule:evenodd;stroke:#000000;stroke-width:1pt"
       id="path4413" />
  </marker>
  <path
     d="m 5333.1575,1458.1459 298.7638,1310.9936"
     id="Flow3"
     style="fill:none;stroke:#000000;stroke-width:2px;stroke-
```

Figure 5.26 Un aperçu du fichier TunisiaChorML2 produit.

- **Interaction de l'utilisateur sur la carte produite**

Cette phase est assurée par l'éditeur graphique « Inkscape » [http 26]. L'utilisateur peut, à travers une variété d'opérations, adapter la carte chorématique produite par notre système à ses besoins et désirs.
Un tel éditeur permet encore, la génération d'un fichier SVG comportant l'ensemble des modifications apportées sur notre fichier ChorML2.
Nous présentons, dans la figure 5.27, notre carte chorématique suite au changement des couleurs des clusters à travers Inkscape.

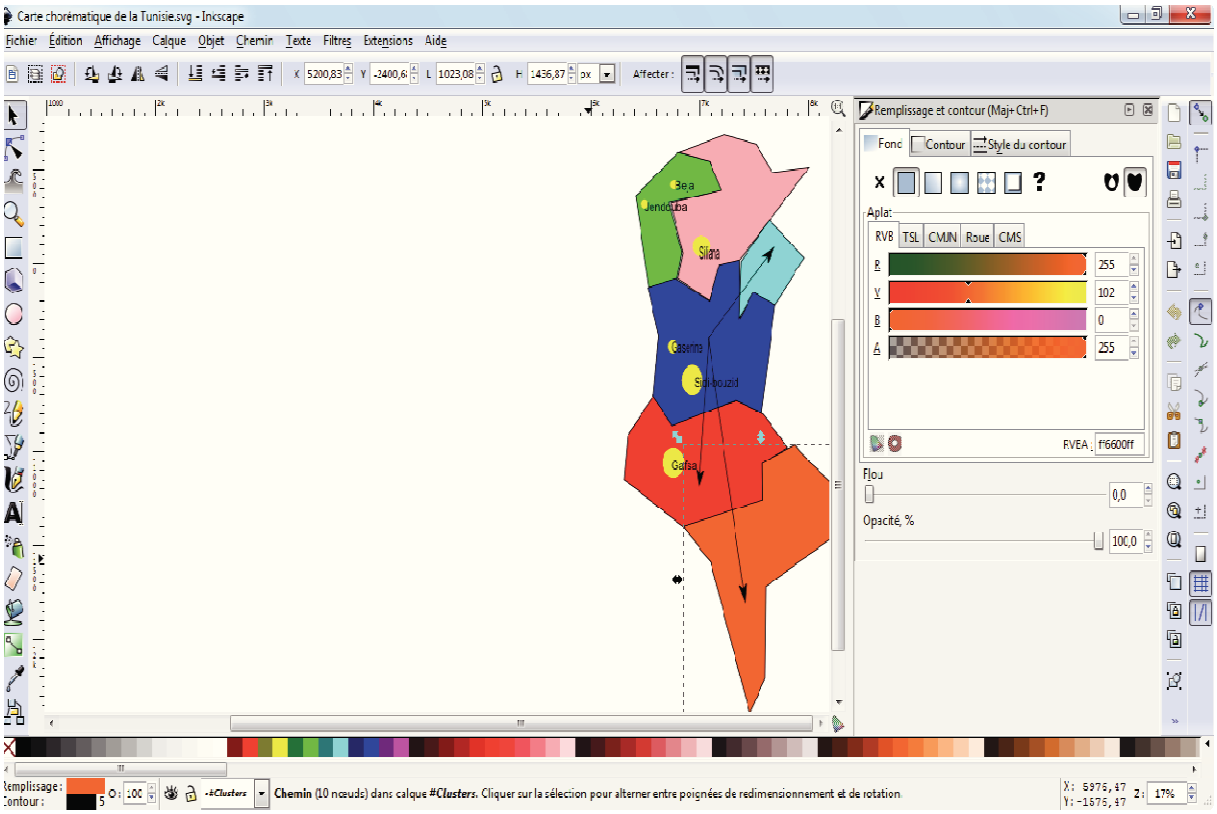


Figure 5.27 Interaction de l'utilisateur sur la carte produite.

5.5 Conclusion

L'objectif de ce travail de recherche est d'étudier une solution cartographique innovante capable de représenter la dynamique, le mouvement et les changements qui sont à l'origine de ces situations complexes. La solution proposée est basée sur le concept de Chorème défini comme représentation visuelle schématisée d'un territoire.

Grâce à notre prototype, des expérimentations sur une base de données géographiques tunisienne ont été effectuées pour générer des motifs important et les visualiser sous forme de treillis et carte chorématique.

Chapitre 6

Conclusion générale et perspectives

6.1 Conclusion générale

Afin d'accomplir régulièrement des tâches liées au travail d'exploration scientifique de la complexité de notre monde, les systèmes d'information géographiques offrent aux utilisateurs la possibilité de devenir plus productifs et conscients et plus sensibles à notre planète Terre.

Notre recherche s'inscrit dans la construction d'une «Terre numérique» – un accès global à toutes les données possibles sur les lieux du globe terrestre – cadre dans lequel des chercheurs et des praticiens sont confrontés à de nombreux défis, notamment le développement des systèmes de visualisation avec des interfaces conviviales qui permettent l'analyse, la modélisation et la simulation des données, au delà de la simple vision de ces dernières.

Les bases de données géographiques contiennent les informations nécessaires à la compréhension de notre environnement, qui nous aident à prendre des décisions par rapport à notre entourage. Bien sûr, ces informations ont besoin d'une représentation claire, facile à comprendre pour pouvoir aider à la prise de décisions. Dès lors, nous avons besoin de cartes qui donnent une vision synthétique d'ensemble et qui intègrent des résumés visuels décrivant facilement l'information importante à montrer et expliquer.

La motivation de notre thèse est basée essentiellement sur le concept de « Chorème ». Un chorème est la représentation schématique d'un espace géographique, avec des caractéristiques importantes que nous souhaitons représenter sur une carte afin de les mettre en évidence pour une étude ou obtenir une meilleure compréhension. Chaque chorème est un dessin qui a sa propre forme et sa propre signification. La signification peut correspondre à un processus ou bien à la représentation de la dynamique d'un certain lieu. Par conséquent, un chorème est un outil puissant qui permet de représenter la connaissance que l'on possède sur un certain lieu, et ceci grâce à sa capacité à symboliser et à encapsuler une méthodologie et son interprétation correspondante. Nous pouvons ainsi montrer des situations climatiques, géographiques, économiques, sociologiques, géologiques, agronomiques, etc. basées sur leur contexte spatial, statistique et temporel, et ce, grâce à la combinaison de plusieurs chorèmes pour constituer une carte chorématique.

Pour obtenir les représentations souhaitées, il est nécessaire de faire une analyse en profondeur de la structure et des aspects les plus importants du lieu à représenter. Pour ce faire, nous utilisons les techniques de la fouille de données géographiques (ou spatial datamining). Notre travail vise à définir des solutions cartographiques afin de mieux représenter les informations géographiques extraites à partir du contenu de bases de données, qui se réfèrent à la fois aux objets statiques et phénomènes dynamiques. L'analyse profonde d'une base de données géographiques génère souvent des résultats caractérisés par un gros volume difficiles à exploiter. La représentation visuelle dans une carte simplifiée des informations extraites de cette analyse devient une solution pour résoudre le problème d'une complexité encore plus grande, surtout lorsqu'il s'agit de domaines comme la politique, l'économie ou la démographie.

• Nos Contributions

Nos travaux de recherche ont permis de proposer une solution innovante basée sur le concept de chorème et sur sa capacité à résumer les scénarios impliquant des objets statiques et des phénomènes dynamiques, en les associant avec des notations schématiques visuelles.

Après une étude de différents algorithmes de fouille de données, notre choix s'est fixé sur l'algorithme *k*-means et celui de l'algorithme Apriori pour développer notre propre algorithme d'extraction de patterns.

D'une manière générale, notre approche comprend trois phases. La première concerne l'extraction de patterns à partir de la fouille de données et notamment la fouille de données spatiales, alors que la seconde est dédiée à l'identification des patterns les plus importants. La dernière phase est allouée à la visualisation de résumés visuels.

Notre méthodologie a pour objectif d'extraire les motifs qui servent à construire les résumés visuels de base des données géographiques. Ces motifs sont les suivants : Les clusters (regroupements géographiques), les faits, les flux, les co-localisations, les contraintes topologiques et les informations extérieures. Dans ce but, nous avons explicité la structuration des tables de BDG afin de pouvoir extraire les connaissances géographiques.

Puisque le nombre de motifs extraits de la première phase est important, nous procédons à une réduction en se basant sur l'élimination des connaissances inutiles d'un point de vue de l'expert et en même temps exclure les connaissances redondantes.

Ainsi, notre approche proposée dans la deuxième phase se base sur l'élimination de la redondance et l'intégration des contraintes d'utilisateur.

Finalement, nous passons à la phase de visualisation des résultats basée sur la technique des treillis de concepts et les chorèmes. Nous avons créé un sous système de visualisation des chorèmes après des phases de simplification et d'agrégation.

Des expérimentations ont été conduites sur des données en Tunisie et en Italie.

6.2 Perspectives

Notre solution dédiée à la génération de chorèmes est supervisée. Elle offre aux utilisateurs experts un modèle qui décrit des positions, des évolutions et des faits, faciles à comprendre et à expliquer aux utilisateurs non experts.

Malgré cet état de fait, différentes questions peuvent être posées :

- Est-il envisageable d'extraire des résumés visuels à partir des bases de données géographiques de manière automatique sans faire appel à un expert ?
- Parmi les multiples motifs découverts par la fouille de données, comment choisir les plus importants ?
- A plus long terme et dans l'optique du mantra de Ben Shneiderman (Overview, zoom and filter, details en demand), est-il envisageable que des chorèmes hiérarchisés puissent constituer un modèle d'accès aux bases de données géographiques ?
- Comment aborder les données continues du temps réel provenant de capteurs, (par exemple, en météorologie) ?

- **Travaux futurs à court terme**

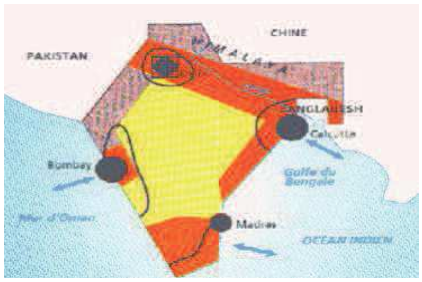
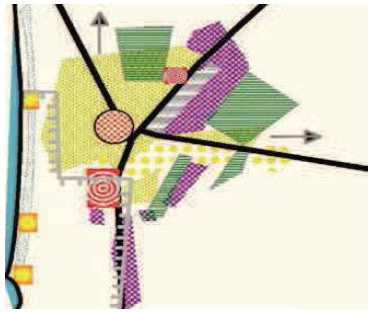
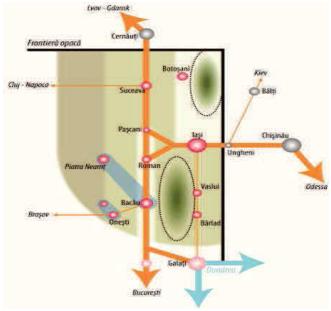
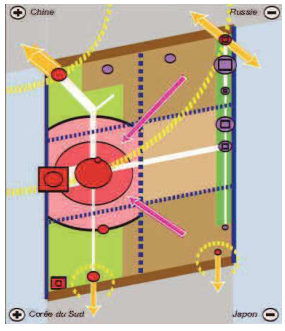
Comme extensions à notre système, nous envisageons :

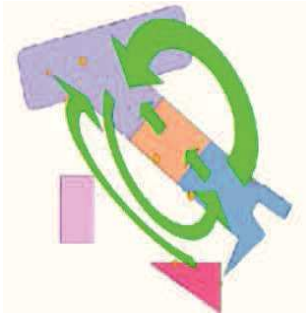
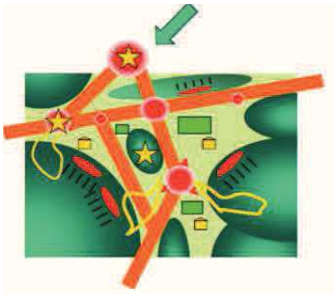
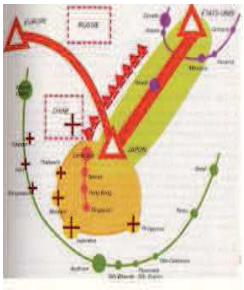
- d'améliorer notre système pour pouvoir extraire les motifs importants, et ce, de manière automatique,
- d'intégrer le système de visualisation avec le sous-système d'extraction,
- de traiter d'autres types de connaissances géométriques comme les gradients, les autres chorèmes de Brunet et Ducruet ainsi que les autres connaissances liées aux champs continus.

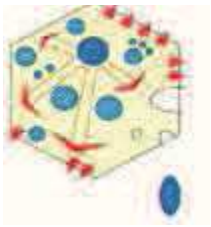
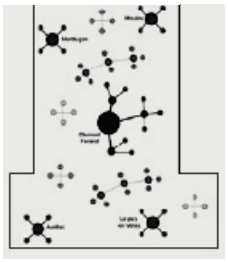
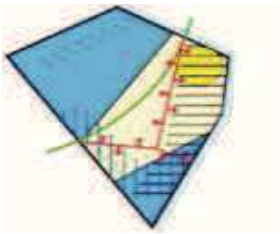
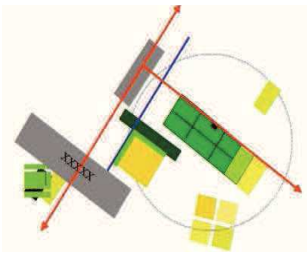
Annexe A

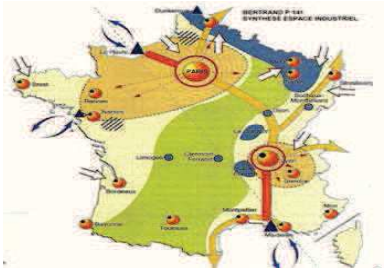
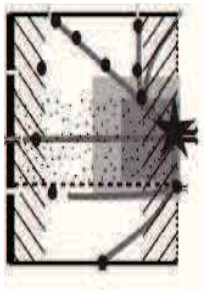
Des cartes chorématiques contenant les chorèmes de Brunet

Nous présentons, ci-dessous, une vingtaine de cartes chorématiques contenant les chorèmes de Brunet parmi un ensemble de cartes étudiées.

Chorème	Explication
L' I N D E [18] 	Cette carte chorématique présente les contrastes de peuplement et le développement économique et ses facteurs. Elle se compose de quatre chorèmes: Les chefs lieux (les villes les plus importantes), une partition des zones selon la population, une partition des zones selon leurs développement économique et les frontières.
NOUAKCHOUTT [92] 	Cette carte chorématique présente la dynamique spatiale de la ville de Nouakchott. Elle est composée de six chorèmes: les régions selon les habitants(les habitats spontané, les habitats de classe et de classe modeste), les aéroports, les pôles industriels, les pôles économique principales et enfin les infrastructures portuaires.
L A R O U M A N I E [47] 	cette carte chorématique présente la Roumanie elle est composée de quatre chorèmes : Les zones et les axes industriels, les axes de transports principaux et secondaires et l'infrastructure (les montagnes, les collines et les hauts plateaux).
L'archipel nord-coréen [34] 	C'est la carte chorématique de l'Archipel nord-coréen, Il est possible de distinguer six régions principales à partir des dix provinces administratives. Les périphéries Sud-est, Sud-ouest, Nord-est, Nord-ouest dépendent des régions centrales. Aussi il est possible de distinguer mouvement de la population récent vers les plaines et villes de l'ouest. La carte présente aussi les voisinages de l'espace nord-coréen.

L'ITALIE [27] 	Les migrations en Italie est représenté. Le territoire est divisé en cinq grandes régions : Les grandes villes sont présentées par des points, les flèches représentent les flux de personnes entre les régions. Les chefs lieux les villes les plus importantes (grande population)
VALDONNEZ [61] 	C'est la carte chorématique du Valdonnez qui présente le développement des villages, surtout à la croisée des routes principales, qui sont considérablement élargies Aussi elle présente les activités touristiques développées entre les principaux chefs-lieux (les sites touristiques).
CORSE [28] 	Cette carte chorématique contient six chorèmes : la limite administrative de la Corse, des sous ensembles pour décrire des régions intérieures, des chefs-lieux pour décrire Ajaccio et Bastia, des ruptures entre la zone d'Ajaccio, Bastia et Balagne avec les zones intérieures et des têtes de pont à Balagne, Bastia, Sartenais, etc.
JAPON [30] 	Cette carte chorématique contient cinq chorèmes : des limites administratives du Japon, de la Russie, de Chine, etc., des chefs-lieux qui représentent des villes importantes, des liaisons préférentielles qui représentent les îles, des aires en contact qui sont les aires en commun entre le Japon et la Corée et des points de passage vers la Corée, les îles, etc.

LA FRANCE [12] 	Cette carte chorématique contient trois chorèmes de Brunet : des chefs-lieux qui représentent des villes importantes, une limite administrative de la France, des axes de propagation qui sortent de Paris et des points de passage vers les différentes frontières.
AUVERGNE [56] 	Cette carte chorématique contient cinq chorèmes : des têtes de réseau et de graphes qui décrivent la hiérarchie des réseaux des pôles urbains et ruraux et la limite administrative de l'Auvergne.
LA BRESIL [58] 	Cette carte chorématique présente le problème de l'au au Brésil, elle est composé de 4 macro régions : deux zones humides, une zone sèche et une zone de désertification en présentant les limites des trois principaux bassins fluviaux et les limites méridionale de la Foret tropicale.
LA FRANCE [64] 	C'est une carte chorématique de la France, elle est composée de six chorèmes : les chefs-lieux (les villes les plus importantes) les flux de migration, les montagnes, le zones industrielles et les zones rurale.
LOU [60] 	C'est la carte chorématique du Lou (les états unis), elle est construite par la combinaison de cinq chorèmes élémentaire : Parc vache laitières, les rivières, les bâtiments d'exploitation, les axes de communication entre ces derniers et enfin les clôtures (mobiles et fixes)

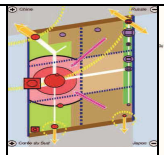
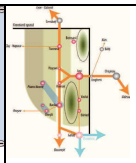
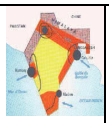
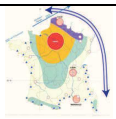
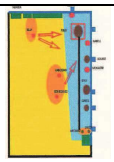
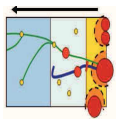
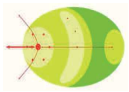
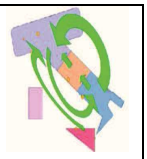
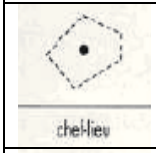
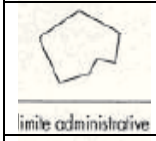
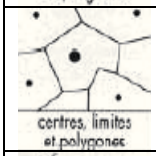
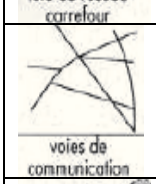
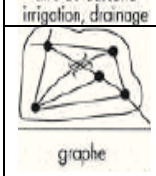
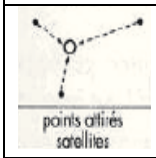
L A F R A N C E [65] 	C'est une carte chorématique de la France, elle présente des chefs-lieux : les pôles industriels majeurs, les centres isolés et les espaces sous industrialisés. Elle présente les axes terrestres majeurs, les couloirs industriels et les limites de France avec les investissements étrangères.
L'A R G E N T I N E [85] 	Cette carte chorématique contient sept chorèmes : une limite administrative qui décrit l'Argentine, des bandes qui décrivent les frontières, des graphes pour les villes importantes en communication constante, des chefs-lieux des villes importantes, des aires en contact, des points de passage aux frontières et des sous-ensembles pour les zones en croissance.

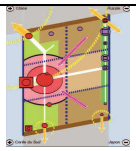
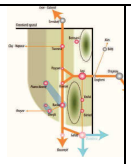
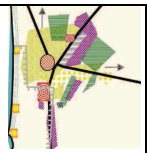
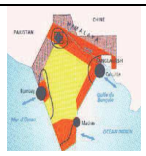
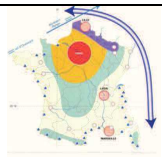
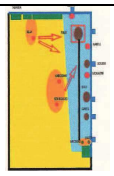
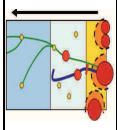
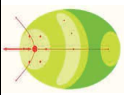
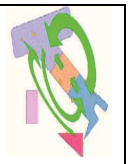
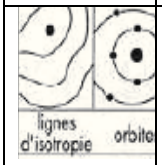
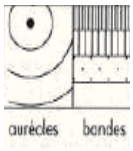
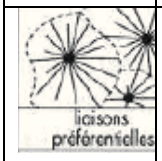
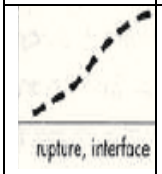
A1.1 Cartes chorématiques utilisant les chorèmes de Roger Brunet

Annexe B

Etude des cartes chorématiques

Cette partie consiste à repèrer les chorèmes les plus utilisés.

Cartes Chorémati-ques / Chorèmes De Brunet									
 chef-lieu	X	X	X	X	X	X	X	X	X
 limite administrative	X	X	X	X	X	X	X	X	X
 État, région...									
 centres, limites et polygones									
 tête de réseau carrefour									
 voies de communication					X	X	X		
 aire de desserte irrigation, drainage									
 graphe							X		
 points attirés satellites									

Cartes Chorémati-ques / Chorèmes De Brunet									
									
			X						
									
		X	X				X	X	
	X					X	X		
	X	X	X	X	X		X		X
									
	X	X						X	
									

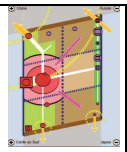
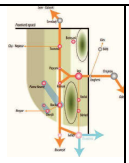
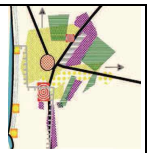
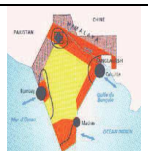
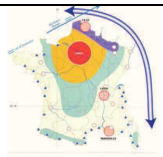
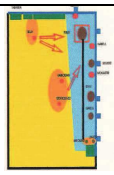
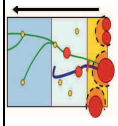
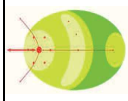
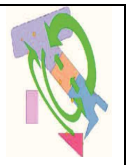
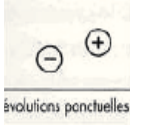
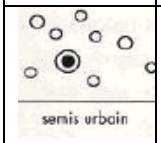
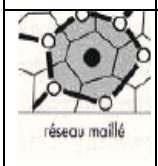
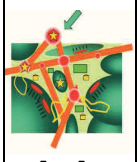
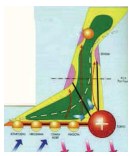
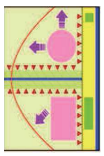
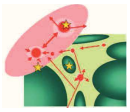
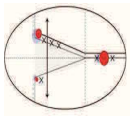
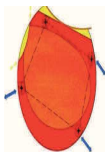
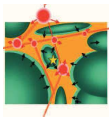
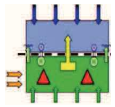
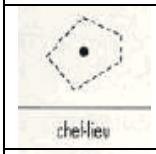
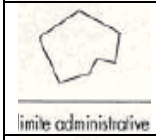
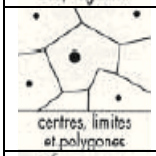
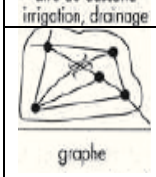
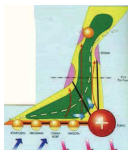
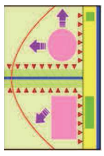
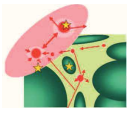
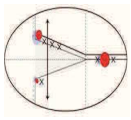
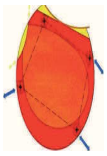
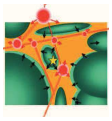
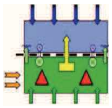
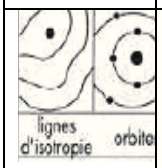
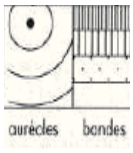
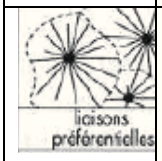
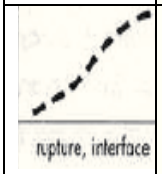
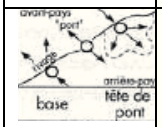
Cartes Chorémati-ques / Chorèmes De Brunet									
 surfaces de tendance									
 dissymétrie									
 évolutions ponctuelles	X								
 axes de propagation		X	X	X	X	X	X	X	X
 aires d'extension ou de régression									
 tissu du changement									
 semis urbain									
 relation de dépendance limites administratives									
 sous-ensemble									
 réseau maillé	X	X					X		

Tableau A2.1 Les chorèmes de Brunet répertoriés dans les cartes chorématiques

Cartes Chorémati-ques / Chorèmes De Brunet	 [54]	 [86]	 [http2]	 [86]	 [http2]	 [http7]	 [http7]	 [86]	 [http1]
 chef-lieu	X	X		X	X		X	X	
 limite administrative	X		X	X	X			X	
 État, région...									
 centres, limites et polygones									
 tête de réseau carrefour									
 voies de communication	X					X			
 aire de desserte irrigation, drainage									
 graphe									
 points attirés satellites		X	X					X	

Cartes Chorémati-ques / Chorèmes De Brunet									
									
		X							X
									
		X	X	X			X		
		X				X	X		X
	X		X						
									
	X	X		X	X			X	
						X			

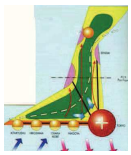
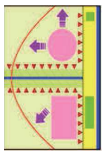
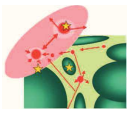
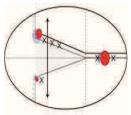
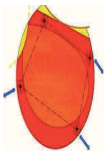
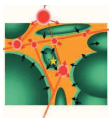
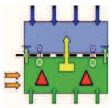
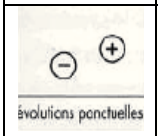
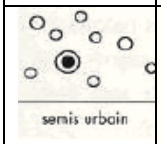
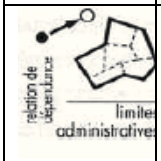
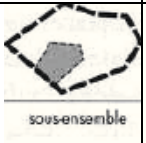
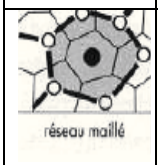
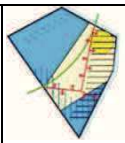
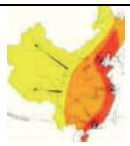
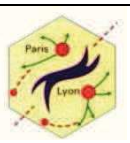
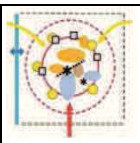
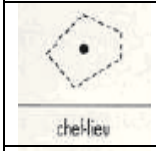
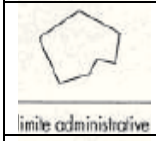
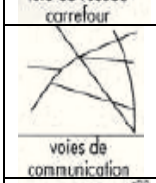
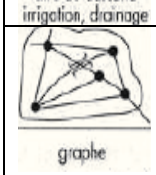
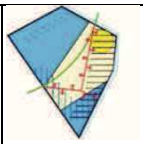
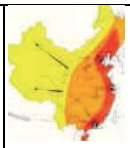
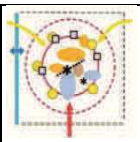
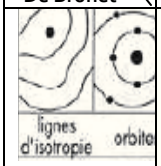
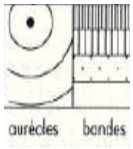
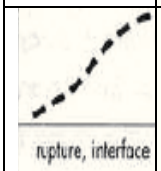
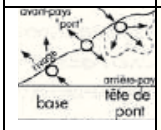
Cartes Chorémati-ques / Chorèmes De Brunet									
 surfaces de tendance									
 dissymétrie									
 évolutions ponctuelles		X				X	X		
 axes de propagation	X	X	X	X	X		X	X	X
 aires d'extension ou de régression									
 tissu du changement									
 semis urbain									
 relation de dépendance limites administratives									
 sous-ensemble					X				
 réseau maillé							X		

Tableau A2.2 Les chorèmes de Brunet répertoriés dans les cartes chorématiques

Cartes Chorématiques / Chorèmes De Brunet	 [13]	 [58]	 [http6]	 [12]	 [17]	 [67]	 [http5]	 [64]	 [http4]
 chef-lieu	X		X	X	X	X	X	X	X
 limite administrative	X	X	X	X	X		X	X	X
 État, région...									
 centres, limites et polygones									
 tête de réseau carrefour									
 voies de communication							X		X
 aire de desserte irrigation, drainage									
 graphe						X			
 points attirés satellites									

Cartes Chorémati-ques / Chorèmes De Brunet									
									
		X			X				
									
		X		X		X	X		X
	X			X		X			
		X	X						
									
				X			X		
									

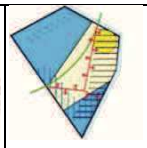
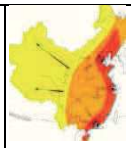
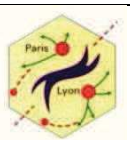
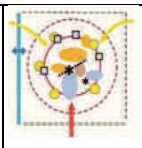
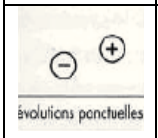
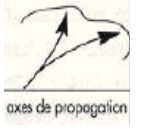
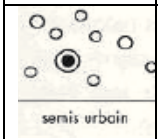
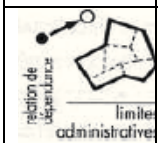
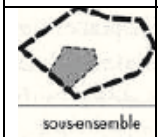
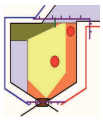
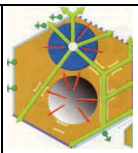
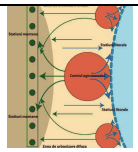
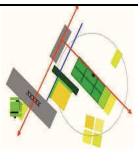
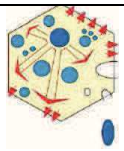
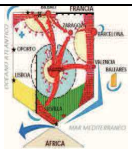
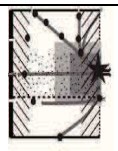
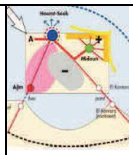
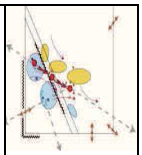
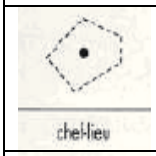
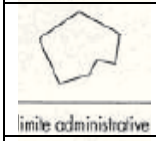
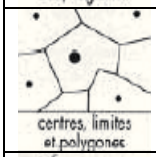
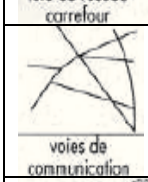
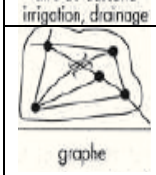
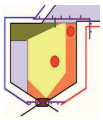
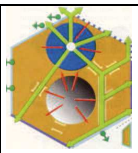
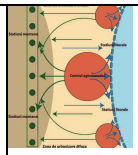
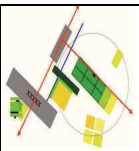
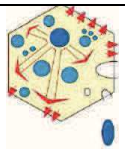
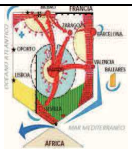
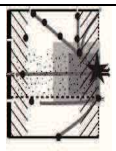
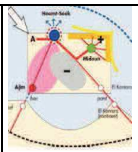
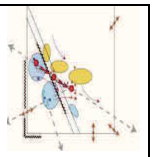
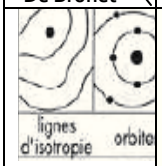
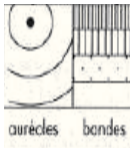
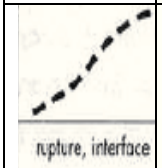
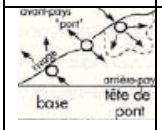
Cartes Chorémati-ques / Chorèmes De Brunet									
									
									
									
			X	X		X	X	X	
						X			
									
									
									
									
									

Tableau A2.3 Les chorèmes de Brunet répertoriés dans les cartes chorématiques

Cartes Chorémati-ques / Chorèmes De Brunet	 [http 9]	 [http 4]	 [http 8]	 [54]	 [84]	 [84]	 [84]	 [53]	 [15]
 chef-lieu	X	X	X		X	X	X	X	X
 limite administrative	X		X	X	X	X	X	X	X
 État, région...									
 centres, limites et polygones									
 tête de réseau carrefour									
 voies de communication									
 aire de desserte irrigation, drainage									
 graphe									
 points attirés satellites		X							

Cartes Chorémati-ques / Chorèmes De Brunet									
						X			
									
								X	
		X	X		X	X	X	X	X
		X	X			X		X	X
						X	X		
		X						X	
		X	X			X			
								X	

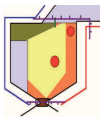
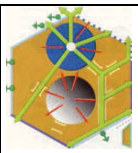
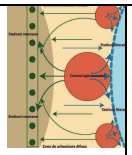
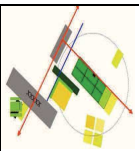
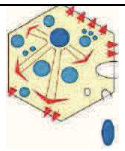
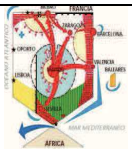
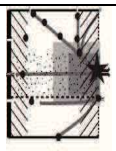
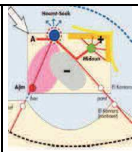
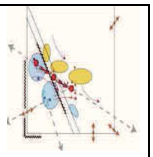
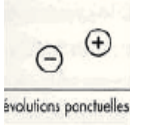
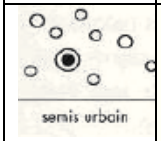
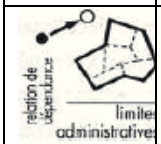
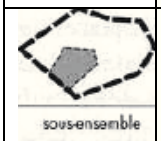
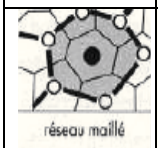
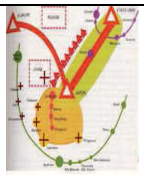
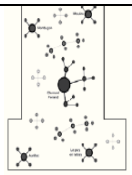
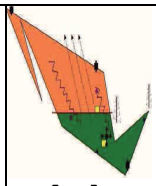
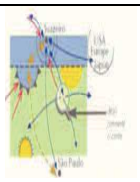
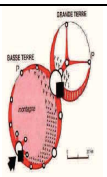
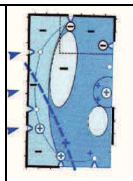
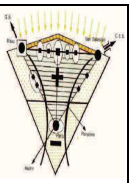
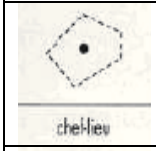
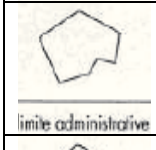
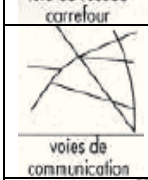
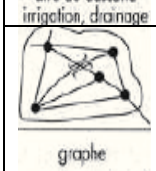
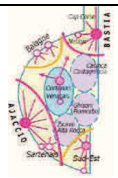
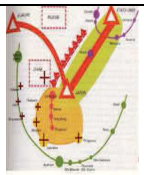
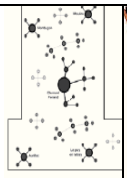
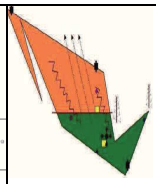
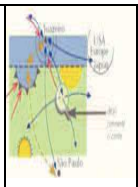
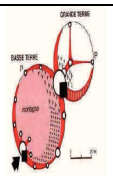
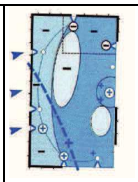
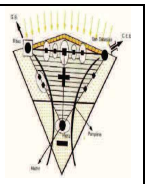
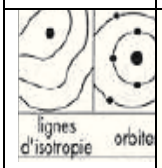
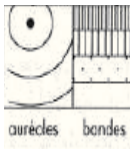
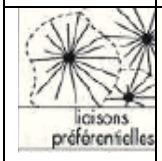
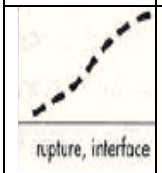
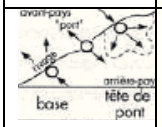
Cartes Chorèmati-ques / Chorèmes De Brunet									
									
									
								X	
		X	X	X	X	X			X
									
									
									
									X
								X	X
				X					

Tableau A2.4 Les chorèmes de Brunet répertoriés dans les cartes chorématiques

Cartes Chorématiques / Chorèmes De Brunet	 [28]	 [30]	 [30]	 [73]	 [7]	 [74]	 [14]	 [79]	 [61]
 chef-lieu	X	X		X	X	X	X		X
 limite administrative	X	X	X	X	X	X		X	X
 État, région...							X		
 centres, limites et polygones									
 tête de réseau carrefour			X						
 voies de communication									
 aire de desserte irrigation, drainage									
 graphe									
 points attirés satellites									

Cartes Chorémati-ques / Chorèmes De Brunet									
					X				
									X
		X				X			
		X		X	X	X		X	
					X	X		X	
		X		X					X
	X					X			
									
									X

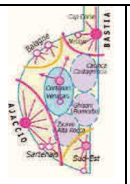
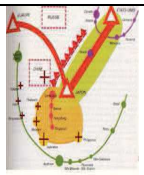
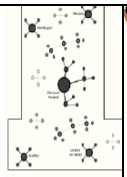
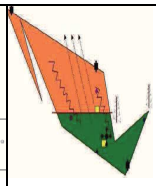
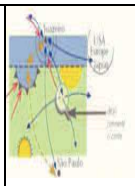
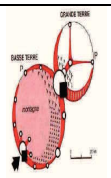
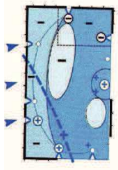
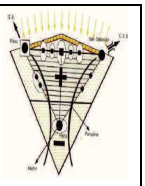
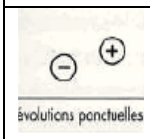
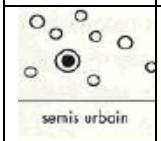
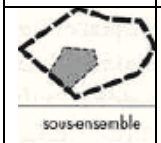
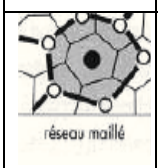
Cartes Chorémati-ques / Chorèmes De Brunet									
 surfaces de tendance									
 dissymétrie									
 évolutions ponctuelles								X	X
 axes de propagation					X	X	X		X
 aires d'extension ou de régression									
 tissu du changement									
 semis urbain									
 relation de dépendance limites administratives									
 sous-ensemble	X					X			
 réseau maillé									

Tableau A2.5 Les chorèmes de Brunet répertoriés dans les cartes chorématiques

Annexe C

Les outils de développement de base

Les outils de base

Afin de mettre en œuvre l'application objet de ce travail, certains choix techniques doivent être explicités en vue de garantir la faisabilité ainsi que la qualité du projet. L'ensemble des outils exploités sont décrits ci-dessous:

- **Java** [http 13]

Nous avons privilégié Java dans sa version 1.7 de son JDK, comme étant le principal langage de programmation du système à développer. La caractéristique majeure de ce langage est sa portabilité. De plus, dans le cas du développement d'une interface comme celle de ce projet, la bibliothèque Swing est une des plus importantes. Elle propose une série d'objets graphiques tels que des fenêtres, des boutons, etc. Pour toutes ces raisons nous choisissons le langage Java et nous avons utilisé Netbeans 7.0 [http 14].1comme un environnement de développement.

- **JDOM** [http 15]

Une riche bibliothèque de classes a été développée par le groupe Apache pour Java spécifiquement dédiée à la lecture du document XML et la réparation de ces différents composants. Cette bibliothèque contient deux principales classes pour « parser » un document qui est DOMParser et SAXParser [http 16]. On note que SAX et DOM sont deux API qui facilitent le travail avec un document XML au sein d'un programme. La grande différence entre ces deux API apparaît dans la seconde opération qui consiste à travailler avec un document XML dans un programme et avec Java en particulier. En fait, SAX ne charge pas le document en mémoire alors que DOM construit en mémoire une représentation arborescente du document. L'API SAX est donc particulièrement adaptée aux gros documents. Par contre, elle offre des facilités de traitement plus réduites. Le fonctionnement par événements rend difficiles des traitements non linéaires du document. Au contraire, l'API DOM rend plus faciles des parcours de l'arbre.

Dans le cadre du notre projet, nous avons travaillé avec JDOM 2.0.5. Notons que JDOM est une API « open source » et que c'est une forme dérivée de DOM.

- **Le SIG OpenJUMP** [http 17]

OpenJUMP est né d'un regroupement du projet JUMP GIS « Java Unified Mapping Platform », développé par Vivid Solutions et ouvert au monde du libre en 2003. Open Jump peut être considéré comme un outil relativement complet par rapport à l'offre actuelle des logiciels libres. Il a une structure en couche et en tables. Il permet en effet de lire et de créer des fichiers vecteur au format shapefile ou GML, de prendre en charge des données raster (ecw, png, tiff), de montrer des données extraites de services web WFS ou WMS ou d'exporter des données au format SVG. Le logiciel est en lien avec une base de données PostGIS. Ces données peuvent être analysées à l'aide d'outils de géométrie et d'attributs, considérés comme la force particulière du logiciel et enrichis d'un nombre croissant d'outils d'analyse vectorielle que ce soit en topologie ou en superposition.

OpenJUMP offre plusieurs fonctionnalités comme :

- Affichage d'une image multicanal
- Gestion de données multi-sources

- Connexion aux bases de données : ARCSDE, ORACLE, POSTGIS, MYSQL
- Création de graphiques sur la base des données attributaires
- Import et export de standards (dxf, gml, shp, postgis, geoconcept, mapinfo)
- Ouverture de fichiers texte type txt ou excel
- Lecture de données GPS
- Outil conversion de la géométrie
- Sélection attributaire
- Requêtes attributaires, spatiales et de calcul
- Calcul de distance
- Editions attributaires et édition des objets géographiques
- Editions de données (construction ou modification d'objets vectoriels)
- Outils de dessin (point, ligne, polygone).

Nous avons utilisé ce logiciel pour importer les coordonnées géographiques des objets spatiaux comme les coordonnées des villes (point) ou de flux (deux points) ou de zones (polygones).

- **Talend Open Studio (TOS)** [http 18]

Talend Open Studio (TOS) est une plate-forme d'intégration de données Open Source, basée sur le langage Java. C'est un ETL du type « générateur de code ». Pour chaque traitement d'intégration de données, un code spécifique est généré, ce dernier pouvant être en Java ou en Perl. Les données traitées et les traitements effectués sont donc intimement liés.
TOS permet de répondre à toutes les problématiques liées au traitement des données dans la chaîne décisionnelle :

- ETL : Extraction, Transformation, et Chargement des données
- EAI : Echange de données Inter-Application
- Synchronisation des données

Dans notre travail nous avons utilisé ce logiciel pour créer une base de données cubique.

Annexe D

Convention de la thèse en cotutelle

INSA LYON

CONVENTION DE THESE

ENTRE LES SOUSSIGNES :

L '**«Institut Supérieur de Gestion, relevant de l'Université de Tunis»**
Sis au «41, Rue de la Liberté, Le Bardo, Tunis 2000»(Tunisie)
Tél. « (+216) 71 58 85 14 » / « (+216) 71 58 85 53 »
Fax. « (+216) 71 58 84 87 »
Dénommée ci-après «ISG»
Représenté aux fins des présentes par son directeur, Monsieur **«Mehrez CHAHER»,**
ET :
L' **« Institut National des sciences appliquées de Lyon »**
Sis au « 20 avenue Albret Einstein 69621, Villeurbanne Cedex » (France)
Tél. « (+33)4 72 43 60 55 »
Fax. « (+33)4 72 43 83 13 »
Dénommé ci-après INSA de Lyon
Représentée aux fins des présentes par son directeur, Monsieur « **Alain STORK** ».qui a donné délégation au Professeur Daniel BARBIER pour les études doctorales.

IL EST D'ABORD EXPOSE CE QUI SUIT :

Vu le décret n°97-1801 du 03 septembre 1997, modifiant et complétant le décret n° 93-1823 du 6 septembre 1993, fixant les conditions d'obtention des diplômes nationaux (tunisiens) sanctionnant les études doctorales.

Vu la loi n° 84-52 du 26 janvier 1984 modifiée sur l'enseignement supérieur, vu l'arrêté du 25 avril 2002 relatif aux études de troisième cycle, vu l'arrêté du 06 janvier 2005 relatif à la création d'une procédure de cotutelle de thèse, prévus par la législation en vigueur en France,

Vu le décret n°93-1823 du 6 septembre 1993, fixant les conditions d'obtention des diplômes nationaux sanctionnant les études doctorales tel que modifié et complété par le décret n°97-1801 du 3 Septembre 1997, et le décret n°1665 du 4 Août 2003 en Tunisie.

Les deux parties, animées par la volonté de favoriser les échanges de doctorants entre elles et de renforcer, ainsi, la coopération scientifique et universitaire entre la Tunisie et la France, décident d'un commun accord, dans le cadre de la législation dans leurs pays respectifs, d'utiliser la procédure de cotutelle concernant :

Mme Ibtissem CHERNI ép. MISSAOUI,
Née le : 03 septembre 1984 au Kef, Tunisie,
Nationalité : Tunisienne.
Adresse dans le pays d'origine : « L'Institut Supérieur de Gestion de Tunis, 41, Rue de la Liberté, Cité Bouchoucha 2000 Le Bardo, Tunis, Tunisie ».
Adresse dans le Pays d'accueil :
« Bâtiment Blaise Pascal, LIRIS, INSA de Lyon, 7 Avenue Capelle F-69621 Villeurbanne Cedex, France ».

1/4

CECI ETANT EXPOSE, IL A ETE CONVENU ET ARRETE CE QUI SUIT

Article 1

La Doctorante doit être inscrite dans les deux établissements, à l'INSA-Lyon et à ISG-Tunis ; Elle acquitte, chaque année, ses droits d'inscription dans un des établissements. Elle est exemptée de ces droits dans l'autre établissement. Sur la durée de la thèse, la doctorante doit obligatoirement s'acquitter de ses droits d'inscription au moins une fois dans chaque établissement.

Dans ce contexte, les deux parties prennent acte et enregistrent les données suivantes :

1. Date de l'inscription en thèse de la doctorante, sous le régime de cotutelle : rentrée universitaire 2010/2011.

2. Durée prévisionnelle des travaux de recherche dans le cadre de la thèse en accord avec la législation en vigueur trois ans. Cette durée peut être prolongée d'une année au maximum après accord spécifique établi entre les deux établissements signataires de la convention, dans le respect des conditions prévues par les législations en vigueur dans les pays respectifs et après accord du Président de l'Université de Tunis :

   - Périodes prévues à l'établissement de recherche en Tunisie : 9 à 10 mois par an.
   - Périodes prévues à l'établissement de recherche en France : 2 à 3 mois par an.

Article 2

Lors de son séjour en France, la couverture sociale de la doctorante est assurée par elle-même conformément à la législation en vigueur.

Article 3

Dans chacun des établissements concernés, la doctorante effectuera ses travaux de recherche sous la direction et la responsabilité des directeurs de thèse suivants :

- Monsieur Sami FAIZ, Maître de conférences habilité, Université de Jendouba ;

- Monsieur Robert LAURINI, Professeur, Institut National des Sciences Appliquées de Lyon.

Les directeurs de thèse s'engagent à exercer pleinement et conjointement, auprès de la doctorante, les compétences qui leur sont attribuées par la réglementation en vigueur et les traditions universitaires dans leurs pays respectifs.

Article 4

La composition du jury de soutenance obéit à la réglementation en vigueur dans chacun des pays impliqué dans la cotutelle. Le jury de soutenance, désigné par les deux établissements partenaires, sera composé à parité par des représentants scientifiques des deux pays. En tout état de cause, le jury doit comprendre obligatoirement les directeurs de thèse.

Article 5

La thèse, préparée en cotutelle, sera rédigée en **français,** sera soutenue en français et complétée par un résumé d'une page écrite en **anglais**.

Article 6

La thèse donnera lieu à une soutenance unique en Tunisie (à l'ISG de Tunis) et à un rapport de soutenance unique obéissant à la réglementation en vigueur dans chacun des pays impliqués dans la cotutelle.

ISG-Tunis, s'engage à délivrer à la doctorante, le titre de docteur, et à transmettre une copie du dossier complet de soutenance à l'institution partenaire, **l'INSA-Lyon**, qui s'engage à délivrer au doctorant, à son tour, le titre de Docteur.

Article 7

Les modalités de dépôt, signalement et reproduction de la thèse ainsi que l'autorisation de la soutenir obéissent à la réglementation en vigueur dans le pays où a lieu la soutenance.

La date et le lieu de soutenance sont fixés d'un commun accord et notifiés par écrit par les codirecteurs de thèse aux chefs des établissements concernés.

Article 8

La protection du sujet de thèse de la doctorante ainsi que la publication, l'exploitation et la protection des résultats de recherche issus de ses travaux dans les deux établissements sont assujetties à la réglementation en vigueur et assurées conformément aux procédures spécifiques à chaque pays impliqué dans la cotutelle.

Les résultats obtenus au cours de la préparation de cette thèse pourront être publiés, mais resteront la propriété conjointe des deux établissements ISG-Tunis et l'INSA-Lyon.

Article 9

La conclusion de la présente convention a été préalablement, autorisée par la cotutelle du **Président de l'Université de Tunis** en date du.....sous le n°.....et

Par Monsieur le **Directeur de l'INSA-Lyon.**

Article 10

Soucieux de l'intérêt des doctorants et du développement de la coopération entre leurs pays respectifs, les établissements d'enseignement supérieur et de recherche sus-indiqués s'engagent à respecter les dispositions ci-dessus et à faire tout ce qui est nécessaire pour l'application de la présente convention dans les meilleures conditions.

En cas de litige, les parties à la présente convention s'engagent à rechercher toute solution amiable avant d'en décider la résolution.

Article 11

Au cas où le régime de cotutelle viendrait à être dénoncé par une des parties concernées, celle-ci devra le notifier par écrit à son établissement d'origine (ISG-Tunis) en indiquant les raisons de sa décision.

L'établissement d'origine (ISG-Tunis) devra en informer l'établissement d'accueil (INSA-Lyon) et l'Université de Tunis dans un délai d'un mois.

Fait en neuf (09) exemplaires originaux.

Le Président de l'Université de Tunis Abderraouf MAHBOULI	**Directeur de l'INSA-Lyon par délégation** Daniel BARBIER
Le Directeur de l'Institut Supérieur de Gestion de Tunis Mehrez CHAHER	**Le Directeur du Laboratoire LIRIS** Attila BASKURT Directeur LIRIS UMR 5205 - CNRS Attila BASKURT
Le directeur de l'école Doctorale Abdelwahed OMRI	**Le directeur de l'école Doctorale** le 28.06.11 Johannes KELLENDONK Alain MILLE
Le directeur de recherche Sami FAIZ	**Le directeur de recherche** Robert LAURINI
La doctorante. Lu et approuvé Ibtissem CHERNI	

4/4

Bibliographie

[1] ABABACAR S. Déterminants, caractéristiques et enjeux de la migration sénégalaise, REVUE Asylon(s) (3) Migrations et Sénégal, 2008.

[2] AGRAWAL R., IMIELINSKI T., SWAMI A. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD, Conference on management of data, ACM press, New York, 1993, pp. 207-216

[3] AGRAWAL R., MANNILA H., SRIKANT R., TOIVONEN H., VERKAMO A. Fast discovery of association rule. In Fayyad U., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. editiors, Advances in knowledge discovery and data mining, MIT express, Cambridge, MA, 1996, pp.307-328.

[4] ANSELIN L. What is special about spatial data? Alternative perspectives on spatial data analysis. Technical paper 89-4. Santa Barbara, NCGIA 1989, pp. 63-77.

[5] BEN HAFSA N. La chorématique : Une grammaire spatiale, disponible sur : http://epigeo.voila.net/chorematique.htm. (Consulté en décembre 2010).

[6] BERRY MICHAEL J.A., LINOFF G. Data mining techniques for marketing, sales and customer support [en ligne]. Third Edition, Edition John Wiley & Sons, 861p, 1997. Disponible sur : http://www.amazon.fr/Data-Mining-Techniques-Relationship-Management/dp/0470650931. (consulté en mars 2011)

[7] BONIN M., CARON P. Territoire, zonage et modélisation graphique: recherche action et apprentissage. GEOCARREFOUR, Vol. 76-3, 2001, pp. 241-252.

[8] BREIMAN L., FRIEDMAN J.H., OLSHEN R., STONE C.J. Classification and Regression Trees, Wadsworth, Belmont, 1984.

[9] BREIMAN L., FRIEDMAN J.H., OLSHEN R.A., STONE C.J. Classification and regression trees. Edition Wadsworth & Brooks. Montery, California, 1984.

[10] BRIN S., MOTWANI R., ULLMAN J.D., TSUR S. Dynamic itemset counting and implication rules for market basket data. In proceeding ACM SIGMOD, international conference on management of data. Tucson, Arizinia, USA, 1997. pp. 355-264.

[11] BRUNET R. La carte-modèle et les chorèmes. Mappemonde 4, Montpellier, GIPReclus, 1986, pp. 2-6.

[12] BRUNET R. La Corse, région d'Europe, Mappemonde 76. 2004. pp. 1-16.

[13] BRUNET R. Le Languedoc-Roussillon en modèle. Mappemonde 3, 1994, pp. 1-4.

[14] BRUNET R. Une épure de la Guadeloupe. In : Chorèmes et modèles, Mappemonde 4, 1986, pp. 24-25.

[15] BRUNET R., SALLOIS J. La France : les dynamiques du territoire. DATARReclus, 1986, 256p.

[16] BUNEL H. Etat de l'art des systèmes d'information géographique. mémoire d'ingénieur

en système d'information, CNAM Paris, 2005.

[17] CAPITAINE M., LARDON S. Chorèmes et graphes pourmodéliser les interactions entre organisation spatiale et fonctionnement des exploitations agricoles. In : Géomatique et espace rural (Journées de la recherche CASSINI Montpellier), Libourel T. & Maurel P. (eds), Éditions CIRAD, 2001, pp. 145-163.

[18] CAREMEL J F. Chorème : l'organisation spatiale de l'Inde, manuel HATIER 5EME. Disponible sur : http://www.histoire-geo.org/docperso/inde.html (consulté en juin 2011).

[19] CHERNI I. Résumés visuels de bases de données géographiques : Transformation de requêtes spatiales en ChorML. Master en Informatique. Université de Jendouba, Tunisie, 2009, 73p.

[20] CHERNI I., LOPEZ K., FAIZ S., LAURINI R. Un langage et un générateur pour représenter les résumés visuels de bases de données géographiques. Revue des Nouvelles technologies de l'information, RNTI-E-19, Extraction et Gestion des Connaissance, EGC'2010, 2010, pp. 691-692.

[21] CHERNI I., LOPEZ K., LAURINI R., FAIZ S. Un langage et un générateur pour représenter les résumés visuels de bases de données géographique. RNTI la Revue des Nouvelles Technologies de l'Information (Cépaduès). Université Francois-Rabelais de Tours, France, 2009.

[22] CHERNI I., OUERGHI M., LAURINI R., FAIZ S. Extraction system for the automatic generation of visual summaries inspired by chorems, ICDIM 2014 - Hong Kong September. September 2014. (Article accepté).

[23] CHERNI I., OUERTANI S., FAIZ S., SEERVIGNE. S., LAURINI R. Chorems: A New Tool for Territorial Intelligence, UDMS: Urban Data Management Society – LONDON, 2013, pp. 29-31.

[24] CHEVRIN V., COUTURIER O., NGUIFO E.M., ROUILLARD.J., User-driven association rules mining to decisions supports systems : Recherche anthropocentrée de règles d'association pour l'aide à la décision, Revue d'Interaction Homme-Machine Vol 8 N°2, 2007.

[25] CHEYLAN J., DEFFONTAINES J., LARDON S., THERY H. Les chorèmes : un outil pour l'étude de l'activité agricole dans l'espace rural, Mappemonde, 1990. n° 4, pp. 2-4.

[26] CHRISTELLE S., Règles d'association [Rapport]. - [s.l.] : IFI, 2004.

[27] DEL FATTO V. Visual summaries of Geographic Databases by Chorems. Thèse de doctorat : INSA de Lyon et Universita di Salerno, 2009.

[28] DEL FATTO V., LAURINI R., LOPEZ K., SEBILLO M., VITIELLO G. A Chorem-based Approach for Visually Synthesizing Complex Phenomena. Review Information Visualization, Palgrave Macmillan, 7, 2008, pp. 253-264.

[29] DENAIN J., L ANGLOIS P. Cartographie en anamorphose, Mappemonde n°49, vol 1, 1998, pp.16-19.

[30] DOMINGO J. Le Japon dans le système mondial des échanges de marchandises et de capitaux, Mappemonde, n°3, 1993, pp. 7-9

[31] DORLING D. the Visualization of Spatial Structure. PhD thesis, United Kingdom: Department of Geography, University of Newcastle upon Tyne. 1991.

[32] DOUGENIK J.A., CHRISMAN N.R., NIEMEYER D.R. An algorithm to construct continuous area cartograms, Professional Geographer 37, 1985. pp. 75-81.

[33] DOUGLAS D., PEUCKER T. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature, The Canadian Cartographer 10(2), PP 112-122, 1973.

[34] DUCRET C. Benchmarking urban networking strategies in Europe An application of chorems to France and Great Britain. The Korea Spatial Planning Review 49(1). 2006. pp. 3-24.

[35] DUCRET C., STANISTAS R. L'archipel nord-coréen transition économique et blocage territoriaux. Mappemonde n°87, 2007.

[36] EGENHOFER M., Herring J., Categorizing Binary Topological Relationships Between Regions, *1991.*

[37] ESPINASSE B. Introduction aux méthodes de Fouille de données, 2009. Disponible sur : http://www.lsis.org/espinasseb/Supports/DWDM-2013/8-IntroFouille-2009.pdf

[38] ESTER M., KRIEGEL H.P., SANDER J. Spatial Data Mining: A Database Approach, in proceedings of 5th Symposium on Spatial Databases. Berlin, Germany, 1997, pp. 67-82.

[39] ESTER M., KRIEGEL H.P., SANDER J., XU X. A density-Based algorithm for discovering clusters in lager spatial databases with noise. In proceeding of second international conference on knowledge discovery and data mining, Portlad, 1996, pp. 226-231.

[40] ESTER M., KRIEGEL H.-P., SANDER J., XU X. Clustering for Mining in Large Spatial Databases. Special Issue on Data Mining, KI-Journal, ScienTec Publishing, 1998, pp. 67-82.

[41] FAYYAD USAMA M., DJORGOVSKI S. G., WEIR N. Advances in Knowledge Discovery and Data Mining, AAAI Press / MIT Press, 1996.

[42] FAYYAD USAMA M., Grinstein G., WIERSE A. Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufman Publishers, 2001.

[43] FRAWLEY W.J., PIATESKY-SHAPIRO G., MATHEUS J. Knowledge Discovery in Databases: An Overview. AI Magazine, 1992, pp. 57-70.

[44] GARDARIN G. Bases de données, 1999.

[45] GARDARIN G. Internet / Intranet et bases de données: Data Web, Data Media, Data Warehouse, Data Mining. Editions Eyrolles. 1999, 246 p.

[46] GARDARIN G., PUCHERAL Ph., WU F. Bitmap Based Algorithms for mining association rules. In proceeding of 15th workshop on bases de données avancées, Tunis, October 1998.

[47] GEORGE T., Încercând choreme [en ligne]. Disponible en ligne sur: http://tzurca.weblog.ro/2009/02/06/ncercnd-choreme-de-criticat. (Consulté en décembre 2009).

[48] GUO Z., ZHOU S., XU Z., ZHOU A. G2ST: a novel method to transform GML to SVG. In ACM international symposium on Advances in Geographic Information Systems, Association for Computing Machinery, 2003, pp. 161-168.

[49] HAN J., KAMBER M. Data Mining. Concepts and Techniques. Morgan Kaufman Publisher, 2006, 703p.

[50] HAN J., PEI J., YIN Y. Mining frequent patterns without candidate generation. ACM-SIGMOD International conference on Management of Data, 2000, pp. 1-12.

[51] HASNI S. Système de génération des résumés visuels basés sur les Chorèmes. Master en Informatique. Université de Jendouba, Tunisie, 2014, 71p.

[52] JAMBU M. Introduction au data mining. Analyse intelligente des données. Edition Eyrolles, 1ère édition, 1998, 136 p.

[53] JOQN A. Organisation du territoire CBE Pays du Ventoux Comtat Venaissin. Disponible sur : http://joanalpini.e-monsite.com/album-cat-1-246404.html. (consulté en juin 2011).

[54] KAYTOUE M. NAPOLI A. Classification de données numériques par treillis de concepts et structures de patrons. Journées Nationales de l'Intelligence Artificielle Fondamentale, 2009.

[55] KOLATCH, E. Clustering Algorithms for Spatial Databases: A Survey, 2001

[56] KOPERSKI K. and HAN J. Discovery of Spatial Association Rules in Geographic Information Databases, In Advances in Spatial Databases (SSD'95). Portland, ME, 1995, pp. 47-66.

[57] 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). Vancouver, Canada, 1998, pp. 45-54.

[58] LAFON B., CODEMARD C., LAFON F. Essai de chorème sur la thématique de l'eau au Brésil [en ligne]. 2005. Disponible sur : http://histoire-geographie.acbordeaux.fr/espaceeleve/bresil/eau/eau.htm (consulté en juin 2011).

[59] LARDON S. Chorèmes et graphes pour modéliser les interactions entre organisation spatiale et fonctionnement des exploitations agricoles, Journées SIGMA-CASSINI, 2001, pp. 25-28.

[60] LARDON S. Usage des chorèmes, graphes et jeux dans le diagnostic de territoire, in Debarbieux B. et Lardon S. (dir.), Les figures du projet territorial, Paris, Editions de l'Aube, Datar, Collection Bibliothèque des territoires, 2003, pp. 109-129

[61] LARDON S. Usage raisonné des représentations spatiales comme objets intermédiaires dans les projets de développement participation [en ligne], INRA-ENGREF, 2006. Disponible sur: http://www.agroparistech.fr/IMG/pdf/rapport_Joystic.pdf (consulté en mars 2011).

[62] LAURINI R., SEBILLO M., VITIELLO G., SOL MARTINEZ D., RAFFORT F.Computer-generated Visual Summaries of Spatial Databases: Chorems or not Chorems?. SA.P.I.EN.S, 2009, Vol 2. Disponible sur : http://sapiens.revues.org/795. (consulté en Décembre 2010).

[63] LEFEBURE R., VENTURI G. Le Data Mining, Eyrolles Edition, 1998, 330 p.

[64] LEFEBURE R., VENTURI G. Le Data Mining, gestion de la relation client, personnalisation de site Web. Eyrolles Edition. 2001. 455p.

[65] LOPEZ K. Contributions aux résumés visuels des bases de données géographiques basés sur les chorèmes. PhD Thesis. INSA-Lyon, 2010. 204 p.

[66] LU W., HAN J., OOI B. C. Discovery of General Knowledge in Large Spatial Databases. In Proc. of 1993 Far East Workshop on Geographic Information Systems (FEGIS'93). Singapore, 1993. pp. 275-289.

[67] MANGIN C. D'Angelinopolis à Postmetropolis, ou l'exception devenant paradigme : un modèle pour la ville mondiale ?. Mappemonde 61, 2001, pp. 5-8.

[68] MARCHAND J-P. L'organisation de l'espace irlandais. Mappemonde 4, 1986, pp. 35-37.

[69] MZALI H. Marché du travail, migrations internes et internationales en Tunisie. Revue Région et Développement, N°6, 1997, pp. 151-183.

[70] NADJIM C. Fouille de données spatiales : un problème de fouille de données multi-tables. Thèse de doctorat de l'Université Versailles Saint-Quentin-En-Yvelines U.F.R DE SCIENCES. 2004.

[71] NG R., HAN J. Efficient and effective clustering method for spatial data mining. In proceeding of international conference on very large database. Santiago, Chile, 1994, pp. 144-155.

[72] NOUACEUR Z. Essor économique et crise environnementale d'une capitale sahélienne : Nouakchott, In : sécheresse, Vol 21, no1, 2010, pp. 63-70.

[73] OPENSHAW S. Developing automated and smart spatial pattern exploration tools for geographical information systems applications, The Statistician, Vol. 44, n° 1, 1995, pp. 3-16

[74] OPENSHAW S., CHARLTON M., WYMER C., CRAFT A. A mark 1 geographical analysismachine for the automated analysis of point data sets. International Journal of Geographical Information Systems, Vol. 1 (4), 1987, pp. 335-358.

[75] PARENT O , EUSTACHE J. Les Réseaux bayésiens A la recherche de la vérité. Master de recherche : Université Claude Bernard Lyon 1, 2006.

[76] PELLISTRANDI. Internet, un nouveau territoire de la Défense ?. La seconde journée d'approfondissement du Trinôme académique de la Défense d'Alsace, 2009.

[77] PORTUGAL J-A. Le modèle basque. Mappemonde 4, 1984, pp. 43-45.

[78] PREUX P., Fouille de données : Notes de cours, (Consulté le 26 mai 2011).

[79] QUINLAN J.R. C4.5: Programs for machine learning, Morgan Kaufmann, 1993.

[80] QUINLAN J.R. Induction of Decision Trees, Machine Learning (1). 1986, pp. 82-106.

[81] ROCHA COIMBRA A. ChorML: XML Extension for Modeling Visual Summaries of Geographic Databases Based on Chorems. Master en Informatique, INSA de Lyon. 2008.

[82] SABO M.N. Intégration des algorithmes de généralisation et des patrons géométriques pour la création des objets auto-généralisants (SGO) afin d'améliorer la généralisation cartographique à la volée, 2007.

[83] SALLEB A. Recherche de motifs fréquents pour l'extraction de règles d'association et de caractérisation. Thèse de doctorat de l'Université d'Orléans, 2003.

[84] SAVASETE A., OMSEINSKI F., NAVATGR S. An efficient Algorithm for mining association rules in large database. In proceeding of 21st international conference on very large databases. Zurich, Switzerland, 1995, pp. 432-444.

[85] SHANNON C.E, WARREN W. The mathematical theory of communication. University of Illinois press, 1949, 126 p.

[86] SHEKHAR S., ZHANG P., HUANG Y., VATSAVAI R. Trends in Spatial Data Mining. In: Data Mining: Next Generation Challenges and Future Directions. Association for the Advancement of Artificial Intelligence Press, 2004.

[87] SHNEIDERMAN B. Designing the User Interface. Third edition, Addison-Wesley Publishing Company, 1997, pp. 66-72.

[88] VELUT S. Argentine, modèle à monter. L'espace géographique, No 3, 2001, pp. 231-244.

[89] WANG W., YANG J., MUNTZ R. STING: A statistical information grid approach to spatial data mining. Technical report CSD- 97006, computer science department. University of California, Los Angeles, 1997.

[90] ZEITOUNI K. Habilitation à Diriger des Recherches Spécialité Informatique Analyse et extraction de connaissances des bases de données spatiotemporelles, Université de Versailles Saint-Quentinen- Yvelines, 2006.

[91] ZEITOUNI K., YEH L. Les bases de données spatiales et le data mining spatial. Revue internationale de géomatique, Numéro spécial Data mining spatial, Vol. 9, N°4, 1999, pp 389-423.

[92] ZIGHED A., RICCO R. Graphes d'induction - Apprentissage et Data Mining. Edition Hermès Sciences, 2000.

Netographie

[http1] http://clioweb.free.fr/carto/carto.htm - Décembre 2010
[http2] http://clioweb.free.fr/carto/carto.htm#manuels - Décembre 2010
[http3] http://confins.revues.org/docannexe/image/3473/img-5.png - Décembre 2010
[http4] http://lettres.histoire.free.fr/lhg/geo/geo_france/geo_france_03.htm - Décembre 2010
[http5] http://lettres.histoire.free.fr/lhg/geo/geo_france/Organisation_territoire - Décembre 2010
[http6] http://lettres-histoire.info/lhg/geo/geo_japon_ase/cartes_japon_ase/Chine_Peuplement.jpg - Décembre 2010
[http7] http://lettreshistoire.info/lhg/geo/geo_japon_ase/cartes_japon_ase/ Indechoreme.jpg
[http8] http://tzurca.weblog.ro/2008/10/21/de-choreme - Décembre 2010
[http9] http://www.acnancymetz.fr/enseign/histgeo/EspacePeda/LYCEE/Christophe/Espagne -- Décembre 2010
[http10] http://www-personal.umich.edu/~mejn/cart/ - Mars 2011
[http11] http://scapetoad.choros.ch/- Mars 2011
[http12] http://www.worldmapper.org/ - Avril 2011
[http13] http://www.java.com/fr/ - Janvier 2013
[http14] http://fr.netbeans.org/ - Janvier 2013
[http15] http://www.jdom.org/- Janvier 2013
[http16] http://docs.oracle.com/javase/7/docs/api/javax/xml/parsers/SAXParser.html - Janvier 2013
[http17] http://www.openjump.org/ - Mars 2012
[http18] http://fr.talend.com/products/talend-open-studio - Janvier 2013
[http19] http://www.techno-science.net/?onglet=glossaire&definition=4942- Juillet 2013
[http20] http://en.wikipedia.org/wiki/Mercator_projection#Derivation_of_the_Mercator_projection- Juillet 2013
[http21] http://blog.phiphou.com/index.php/?2011/02/19/188-simplification-de-polygones - Juin 2013
[http22] http://www.maxicours.com/se/fiche/5/6/277156.html - Juin 2013
[http23] http://help.arcgis.com/fr/arcgisdesktop/10.0/help/index.html#//006200000003000000 - Août 2013
[http24] http://static3.teamdev.com/downloads/jexcel/docs/JExcel-PGuide.html - Août 2013
[http25] http://www.w3.org/Graphics/SVG/ - Juillet 2013
[http26] http://www.inkscape.org/fr/ - Mars 2014
[http27] http://schemas.opengis.net/gml - Janvier 2011
[http28] http://help.arcgis.com/fr/arcgisdesktop/10.0/help/index.html#//006200000003000000 - Août 2013
[http29] http://editionsrnti.fr/?inprocid=1000365&PHPSESSID=p3d3ef8lmmbeerfic5l64k5j51&lg=en&PHPSESSID=p3d3ef8lmmbeerfic5l64k5j51 – Janvier 2012
[http30] http://help.arcgis.com/fr/arcgisdesktop/10.0/help/index.html#//00080000000q000000 - Août 2013
[http31] http://perso.univ-rennes1.fr/thierry.guillaudeux/Biopuces.pdf - juin 2012
[http32] http://www.geog.ucsb.edu/~tobler/publications/pdf_docs/cartography/Analytic_1.pdf- Mars 2011
[http33] http://en.wikipedia.org/wiki/Cartogram - Mars 2012
[http34] http://pdf.aminer.org/000/225/176/toscana_a_graphical_tool_for_analyzing_and_exploring_data.pdf - Juin 2012

FOLIO ADMINISTRATIF

THESE SOUTENUE DEVANT
L'INSTITUT NATIONAL DES SCIENCES APPLIQUEES DE LYON
L'INSTITUT SUPERIEUR DE GESTION DE TUNIS

NOM : CHERNI DATE de SOUTENANCE : **/**/**

Prénoms : Ibtissem

TITRE : Découverte de chorèmes par fouille de données spatiales

NATURE : Doctorat Numéro d'ordre : AAAAISALXXXX

Ecole doctorale :

L'école doctorale InfoMaths – INSA de LYON
L'école doctorale Sciences de gestion – ISG de TUNIS

Spécialité : Informatique

RESUME :

La motivation de notre thèse est basée essentiellement sur le concept de « Chorème ». Ce néologisme inventé par Roger Brunet désigne à la fois un élément et une structure d'un système. Les chorèmes partent d'éléments géométriques simples comme le point, la ligne, le vecteur, le réseau pour former des sémantiques plus complexes. Ces combinaisons créent des représentations de tout type : de la représentation simple d'un lieu important à travers un point, jusqu'aux flux d'échange qui existent entre des zones à l'aide de cercles, lignes, couleurs, textures, flèches, etc.

Notre travail vise à définir des solutions cartographiques afin de mieux représenter les informations géographiques extraites à partir du contenu de bases de données géographiques, qui se réfèrent à la fois aux objets statiques et aux phénomènes dynamiques. La représentation visuelle dans une carte simplifiée des informations extraites de cette analyse devient une solution pour résoudre le problème d'une complexité encore plus grande, surtout lorsqu'il s'agit de domaines comme la politique, l'économie et la démographie. Nous proposons une solution basée sur le concept de chorème et sur sa capacité à résumer les scénarios impliquant des objets statiques et des phénomènes dynamiques en les associant avec des notations schématiques visuelles.

Notre méthodologie a pour objectif d'extraire les motifs qui servent à construire les résumés visuels de base des données géographiques. Ces motifs sont les suivants : Les clusters (regroupements géographiques), les faits, les flux, les co-localisations, les contraintes topologiques et les informations extérieures.

Il se trouve, cependant, que le nombre des motifs extraits de la première phase est souvent important, nous procédons alors à le réduire en se basant sur l'élimination des connaissances inutiles d'un point de vue de l'expert et en même temps, exclure ceux qui sont redondants.

Pour la phase de visualisation, nous proposons deux choix : la visualisation des résultats basée sur la technique des treillis de concepts et la visualisation sous forme de chorèmes.

D'une manière générale, notre approche comprend trois phases :
- La première phase concerne l'extraction de patterns à partir de la fouille de données et notamment la fouille de données spatiales,
- La seconde est dédiée à l'identification des patterns les plus importants,
- La dernière phase est allouée à la visualisation de résumés visuels.

MOTS-CLES : Extraction, résumés visuels, fouille des données géographiques, chorèmes, visualisation, motifs.

Laboratoire (s) de recherche :

Laboratoire d'Informatique en Image et Systèmes d'Information d'Information UMR CNRS 5205, Lyon, France
Laboratoire de Télédétection et des Systèmes d'Informations à Références Spatiales, Tunis, Tunisie

Directeurs de thèse:

Robert LAURINI - Professeur à l'INSA de Lyon, France
Sami FAIZ - Professeur à l'ISAMM de La Manouba, Tunisie

Composition du jury :

Robert LAURINI - Professeur à l'INSA de Lyon, France - Directeur de Thèse
Sami FAIZ - Professeur à l'ISAMM de La Manouba, Tunisie - Directeur de Thèse
Sylvie SERVIGNE - MdC à l'INSA de Lyon, France - Co-directeur
Ahmed LBATH- Professeur à l'Univ. Joseph Fourier de Grenoble, France - Rapporteur
Faiez GARGOUR I- Professeur à l'ISIM de Sfax, Tunisie - Rapporteur
Rached BOUSSEMA - Professeur à l'ENIT, Tunisie - Examinateur
Thomas DEVOGELE - Professeur à l'Univ. de Tours, France - Examinateur
Imed Riadh FARAH - MdC à l'ISAMM de La Manouba, Tunisie - Examinateur