بسم الله الرحمن الرحيم

Ministère de l'enseignement supérieur et de la recherche scientifique
Université Abdelhamid Ibn Badis Mostaganem

Faculté des Sciences Exactes et d'Informatique
Département de Mathématiques et d'Informatique Filière Informatique

Mémoire de Fin d'études pour l'obtention de Master en Informatique
Option : **Systèmes d'Information géographique**

Thème : **Visual Spatial Data mining (VSDM) appliqué à l'épidémiologie**

Etudiants :

- Belhandouz Abdelhak
- Bensahli Belkacem

Encadrant : Mr.Midoun Mohamed

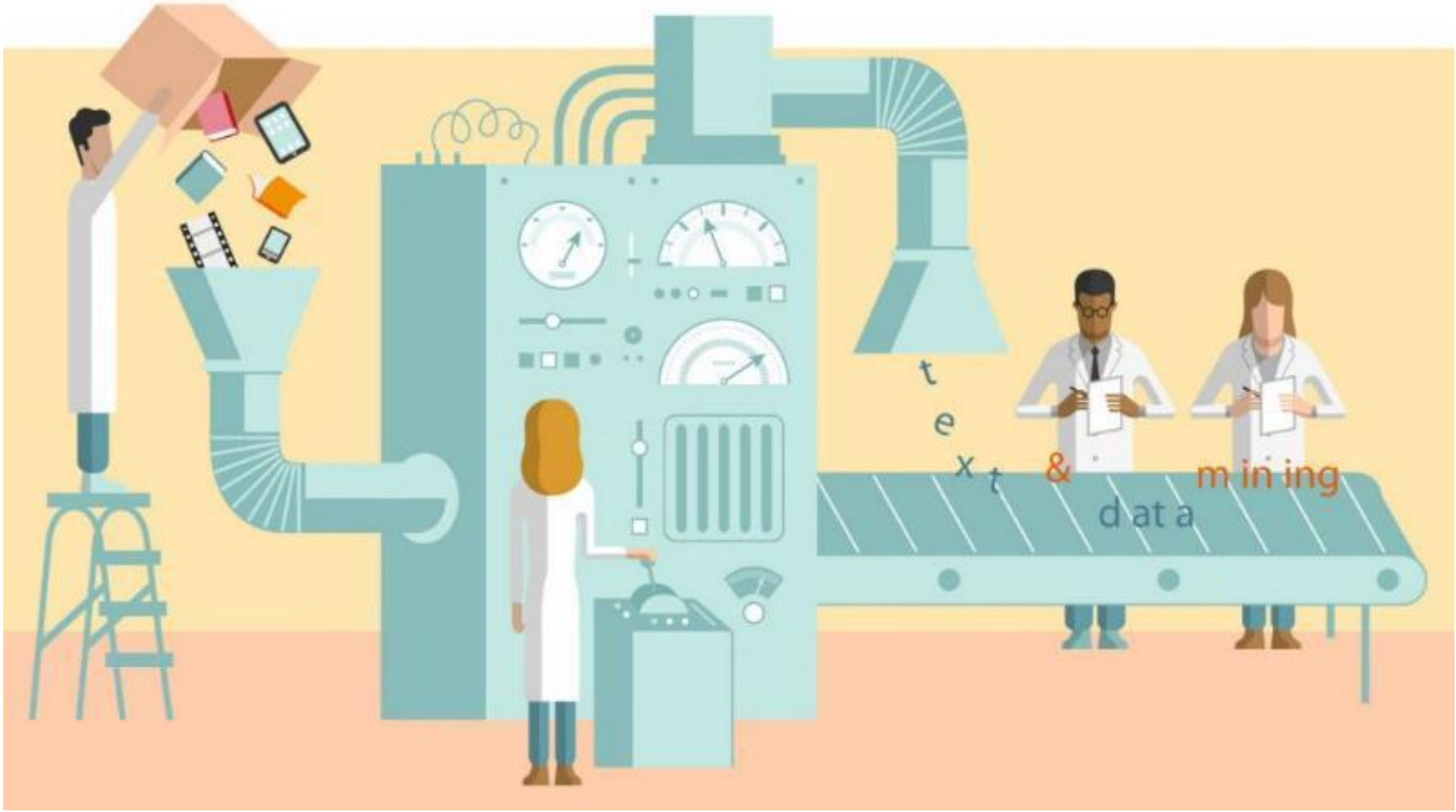
2016 -2017

Plan de la présentation:

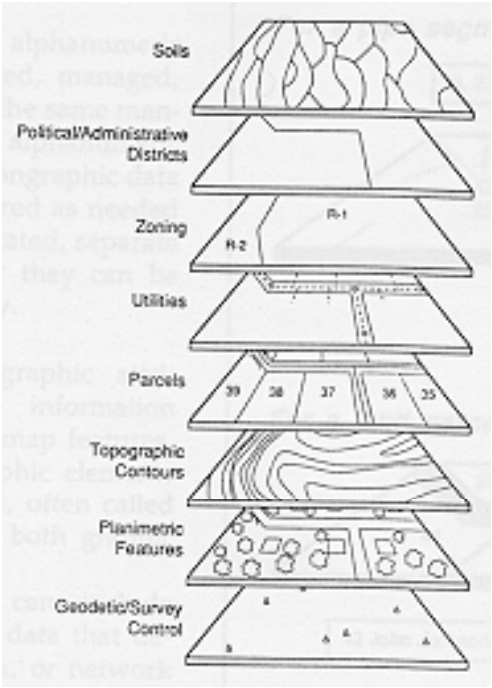
1. Spatial data mining
2. Visuel data mining
3. VSDM et épidémiologie
4. Méthodologie
5. Application
6. Conclusion

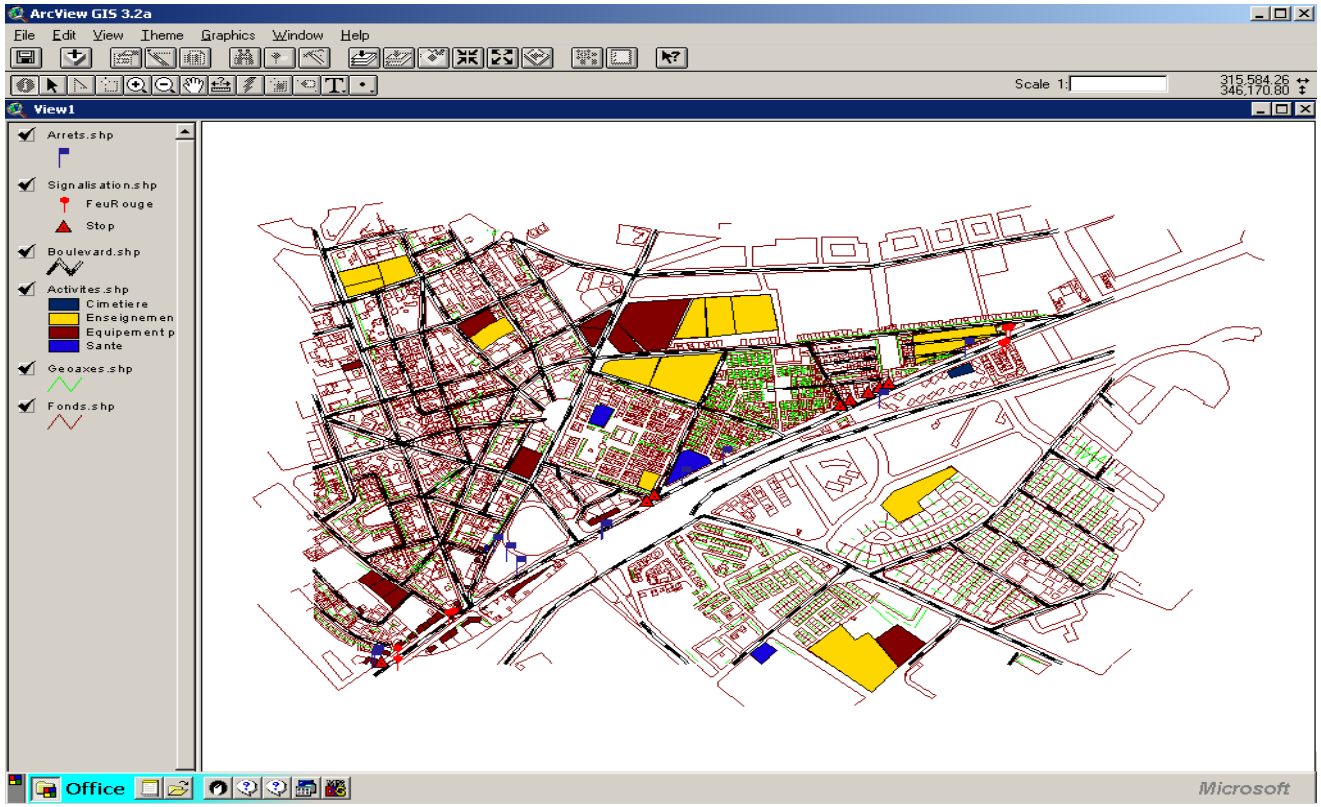
1.Spatial data mining

Data mining

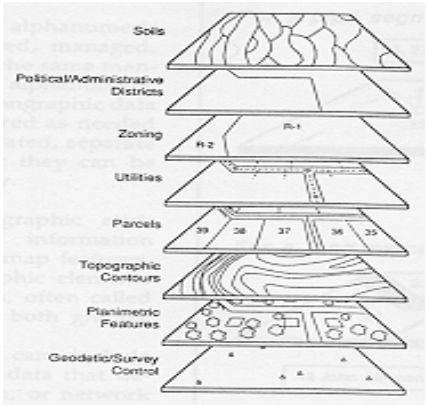


Bases de données spatiales





Spatial data mining



BDS

DM classique

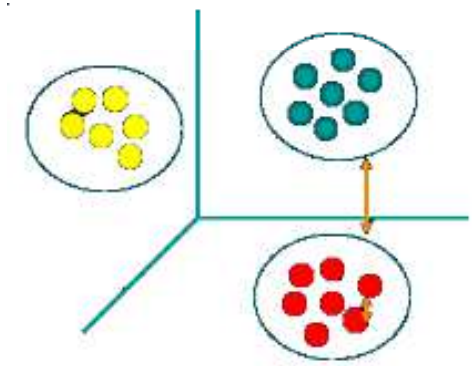
Spatial data mining

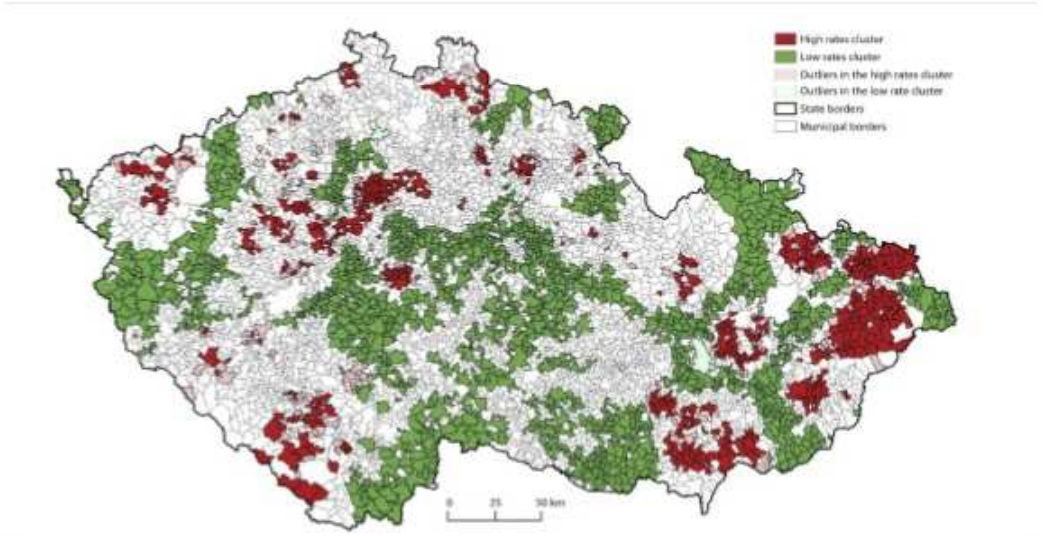
Les méthodes principales du spatial data mining

- Clustering
- Classification
- Prédiction
- Règles d’association
- Hotspot
- outlier

Clustering

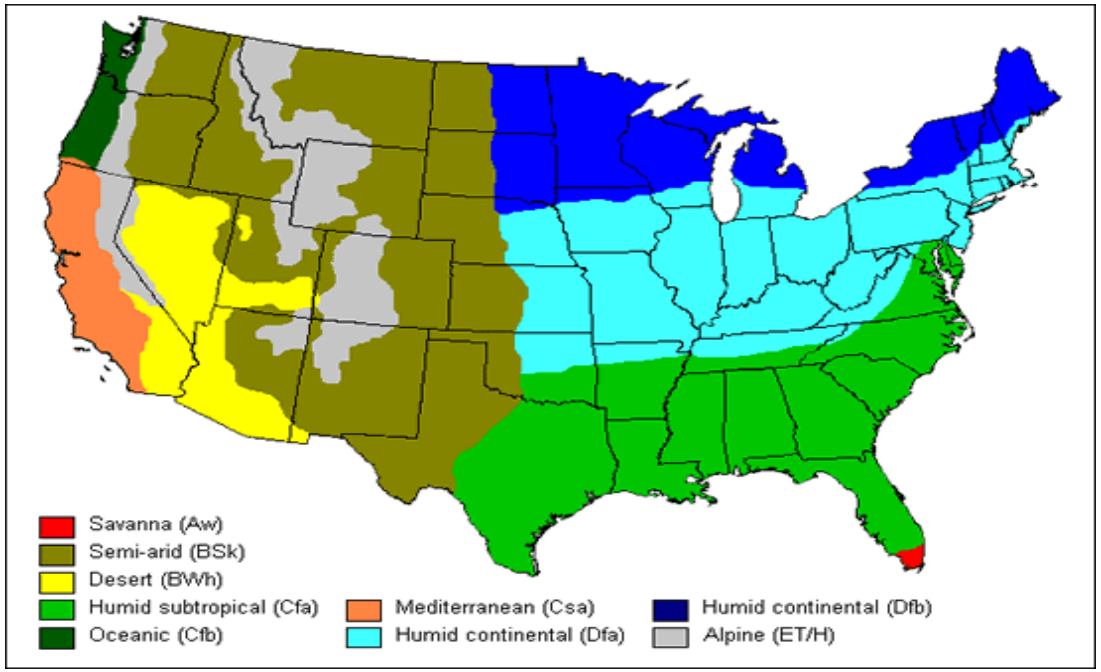
- Groupage d'objets similaires / séparation dissimilaires
- Méthode non supervisé
- Utilisé moins pour classer que pour découvrir des concentrations ou des points chauds
- Exemple : criminologie, épidémiologie, accidents





Classification

- La classification est la tâche de trouver un modèle qui classe chaque cas dans l'une des nombreuses classes prédéfinies.
- Méthode supervisé
- Exemple: Classification climatique ..



Prédiction

- modélise des données numériques pour prédire des valeurs inconnues ou manquantes et pas nécessairement des événements futurs
- Tâche d'apprentissage supervisée
- Exemple: prédire les risques engendrés par les changements climatiques

Règle d'association

- Identifie les relation entre les données spatiales

idem* + *Rel° spatiales* => *idem* + *Rel° spatiales

Avec (S,C) avec *S* comme support et *C* la confiance.

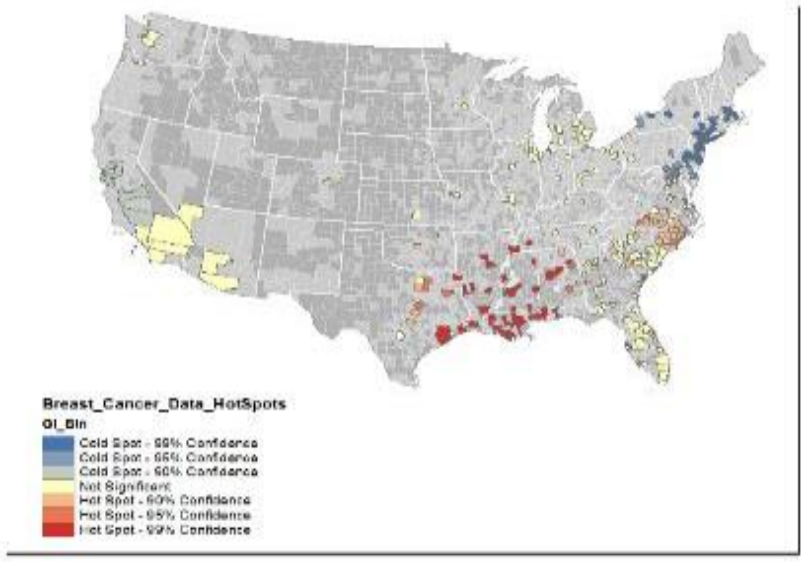
Exemple :

Exemple:

- *station_service ^ dans (zone_rurale) -> proche (autoroute) (25%, 80%)*

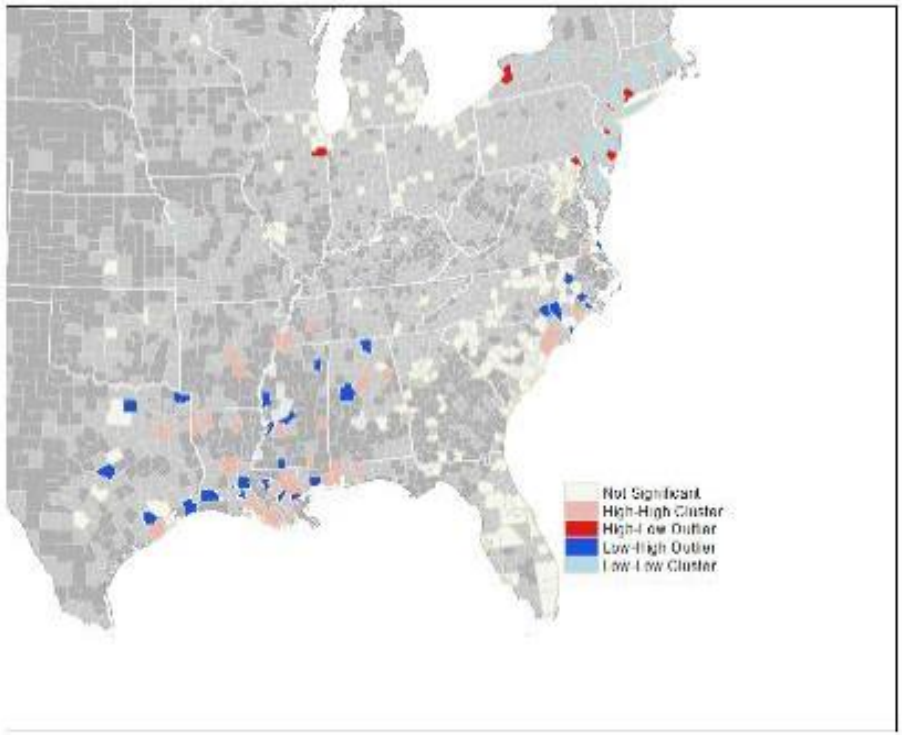
Hotspot

- HotSpot apprend un ensemble de règles qui maximisent (point chaud) ou minimisent (les point froid) une variable ou une valeur par rapport a une cible d'intérêt
- Méthode non supervisé
- Exemple : criminologie, épidémiologie, accidents



outlier

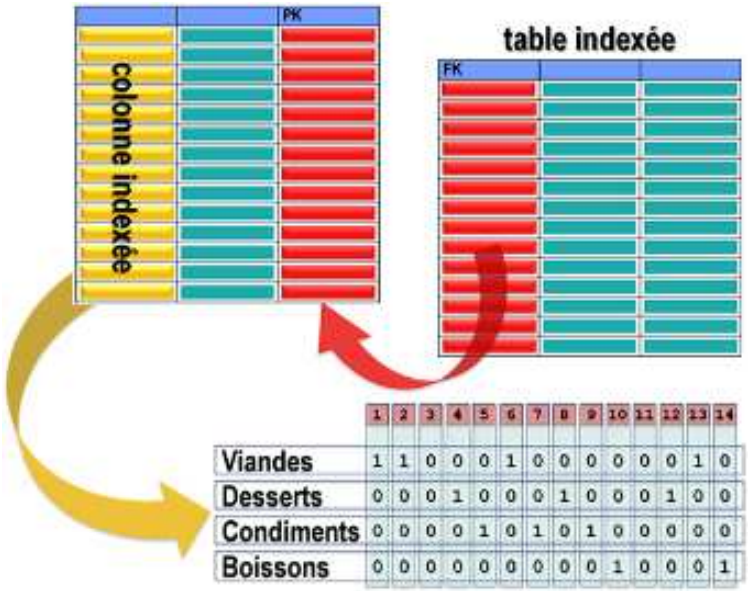
- identifie des clusters de valeurs élevées ou faibles ainsi que des valeurs aberrantes spatiales d’un ensemble pondérées.



Approche pour le SDM

- Il existe deux approches pour l'analyse et l'extraction de connaissances d'une base de données spatiales :
  - Une approche statistique
  - Une approche base de données spatial

L’approche Base de données



indexe de jointure

	1	2	3	4	5	6	7	8	9
1	0	1	0	1	1	0	0	0	0
2	1	0	1	1	1	1	0	0	0
3	0	1	0	0	1	1	0	0	0
4	1	1	0	0	1	0	1	1	0
5	1	1	1	1	0	1	1	1	1
6	0	1	1	0	1	0	0	1	1
7	0	0	0	1	1	0	0	1	0
8	0	0	0	1	1	1	1	0	1
9	0	0	0	0	1	1	0	1	0

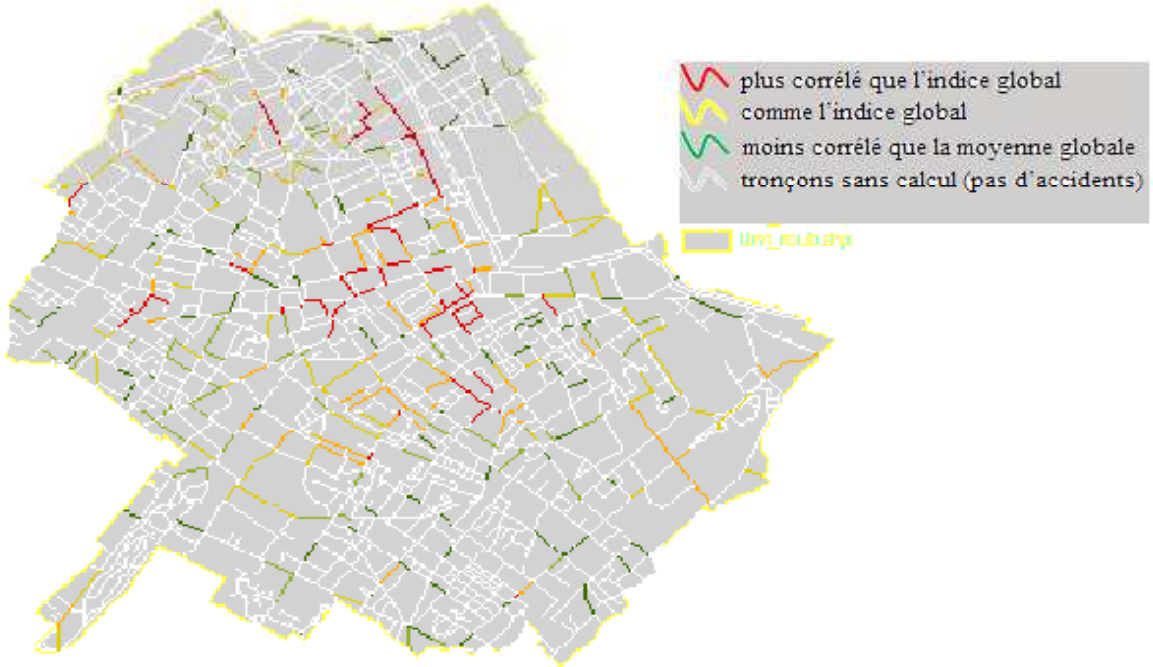
matrice de contiguïté

L’approche Statistiques spatiales

- Consiste a dégager des structures, des caractères, des invariants, des lois de comportement, en mettant en évidence :
  - des distributions de lieux ou des structures spatiales (linéaires, ponctuelles), des structures temporelles et spatio-temporelles, des relations fonctionnelles ($y=f(x_1,x_2,\ldots, x_n)$) permettant d’élaborer des modèles et des lois de comportement
  - Base mathématique solide : mesures, indicateurs

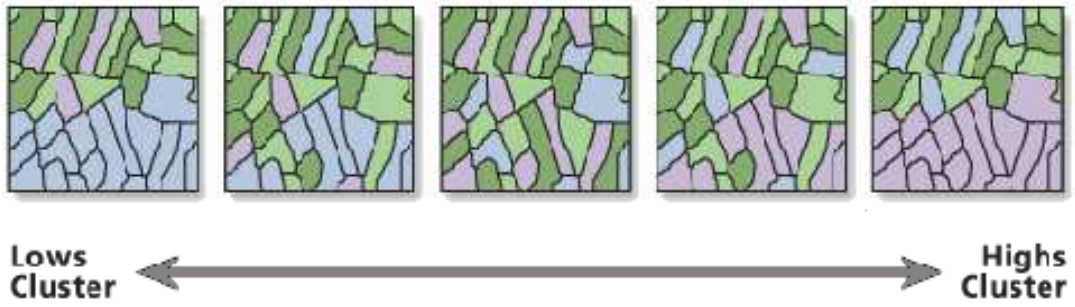
Analyse globale

- **Analyse globale** - Mesure d'autocorrélation spatiale d'une variable
- mesure les relation entre la variable et les autre variable voisins



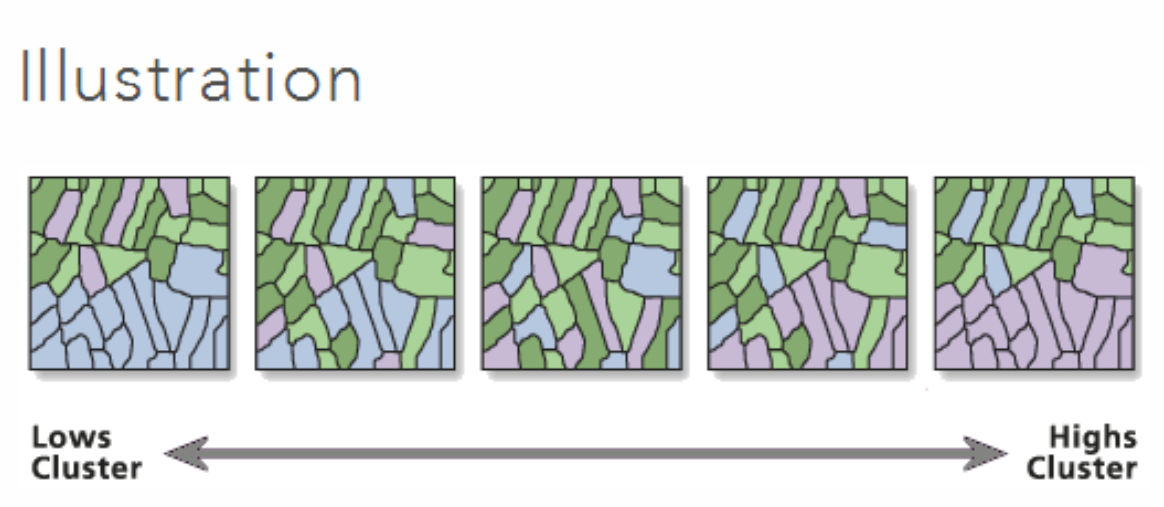
Analyse globale

- **Autocorrélation spatial (Moran I) :** permet de calculer l' indice de l'autocorrélation spatial et représenter le résultat du test sous forme schématique
- l'outil identifie l'emplacement où les valeurs élevées ou faibles sont regroupées dans l'espace, ainsi que les entités ayant des valeurs qui sont très différentes des valeurs d'entités environnantes



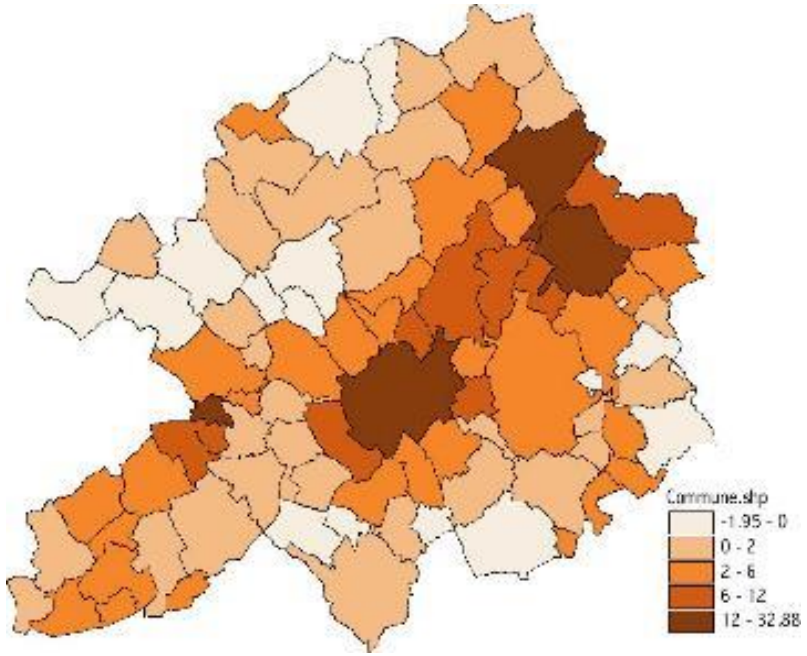
Analyse global

- **Clustering (Getis Ord G)**
  - Mesure le degré d'agrégation des valeurs élevées ou des valeurs faibles à l'aide de la statistique Getis-Ord General G.



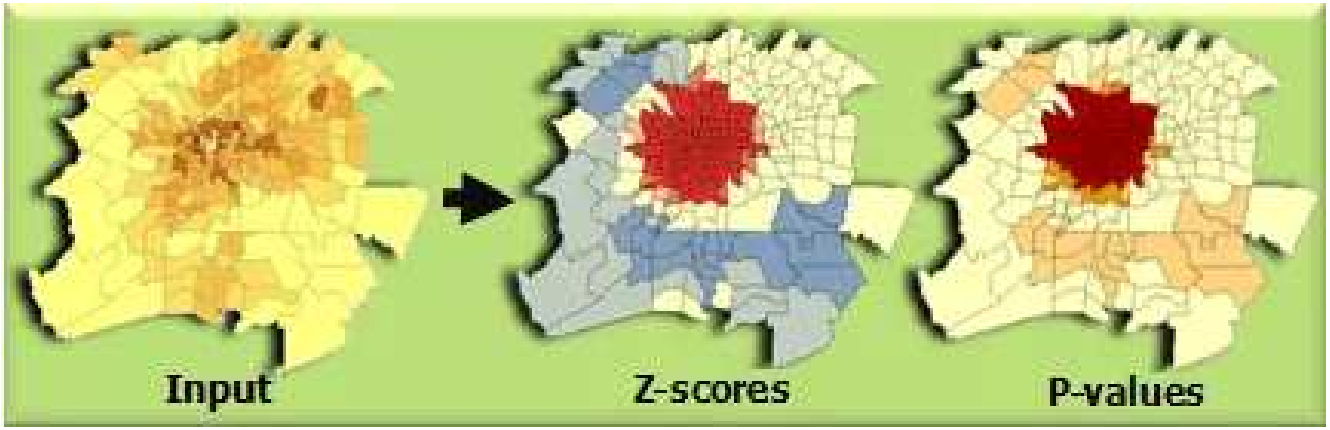
Analyse local

- **Analyse locale** - Indice local d'associations spatiales
  - met en évidence les données atypiques
  - quantifie la contribution individuelle de chaque lieu à l'indice global



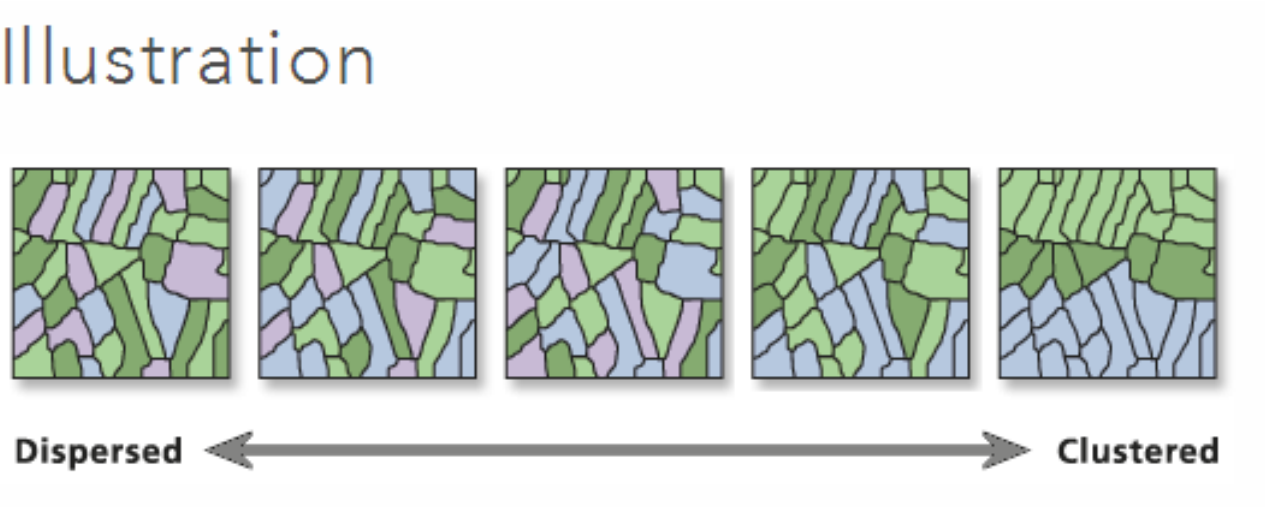
Analyse local

- **Hot Spot (Getis-Ord Gi*)**
  - A partir d'un ensemble d'entités pondérées, identifie les points chauds et les points froids statistiquement significatifs à l'aide de la statistique Getis-Ord Gi*.



Analyse local

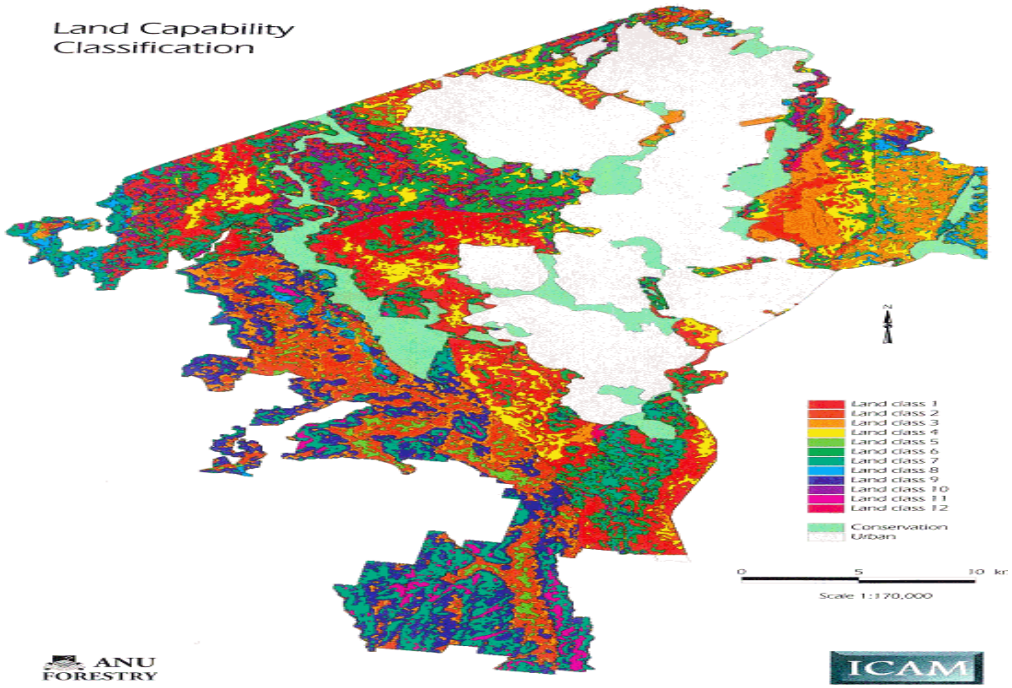
- **Outlier (Anseline Moran LISA)**
  - Mesure l'auto-corrélation spatiale selon l'emplacement des entités et leurs valeurs attributaires à l'aide de la statistique de l'indice global de Moran.



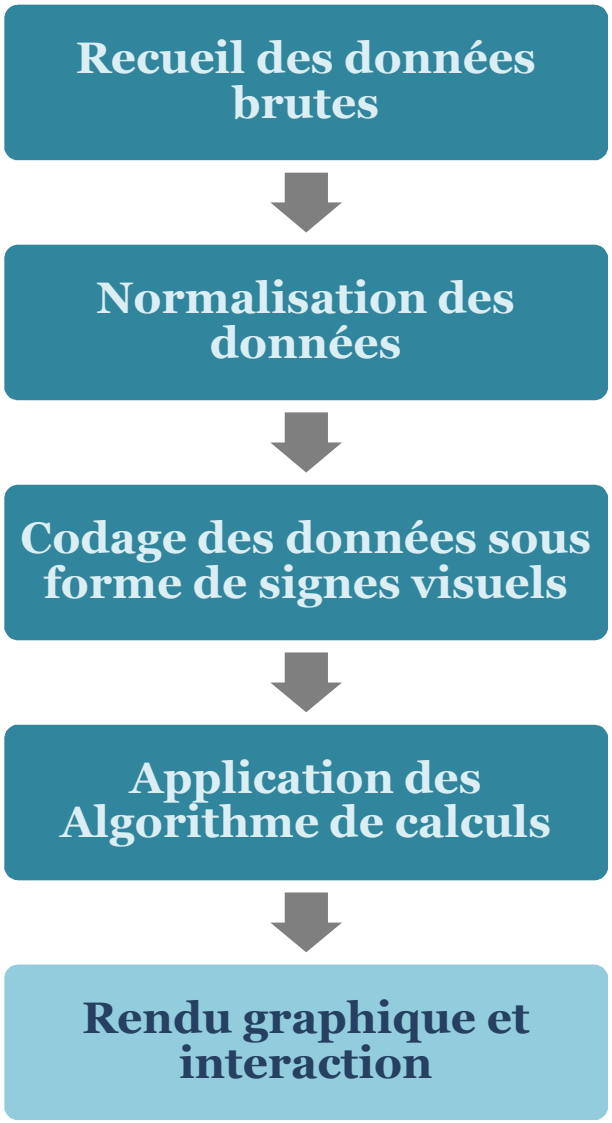
2.Visuel data mining

Visuel data mining

- **Visuel data mining :** c'est la combinaison des techniques usuelles du Data mining avec les méthodes de visualisation de l'information



Processus du visuel data mining



Couplage du SDM et du VDM

❖ **SDM**

- Prend en compte les relations spatiales
- Faible dans les grandes BDS

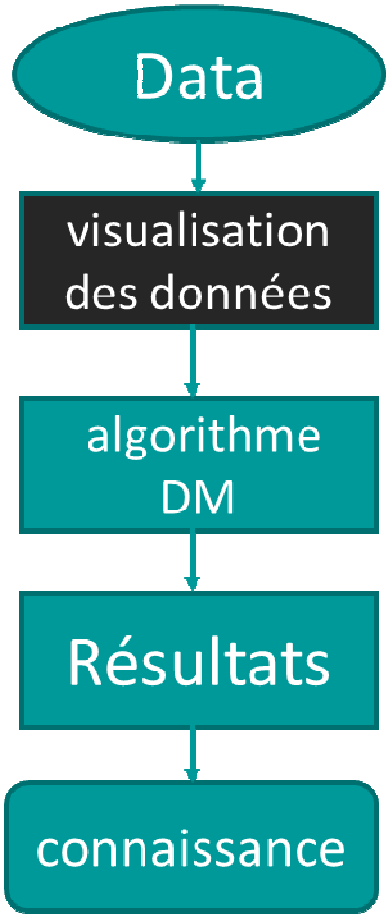
❖ **VDM**

- Permet de découvrir visuellement certains modèles profondément enfuit
- La composante spatiale est difficile à visualiser efficacement dans les BDS

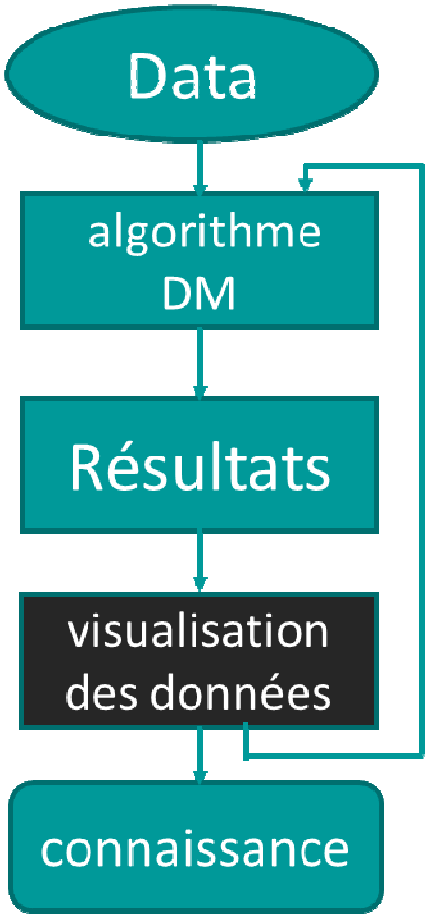
❖ **VSDM**

- Découvrir des modelés spatiaux de manière visuelles
- Montrer visuellement les résultats des algorithmes complexes de SDM afin de mieux les interpréter.

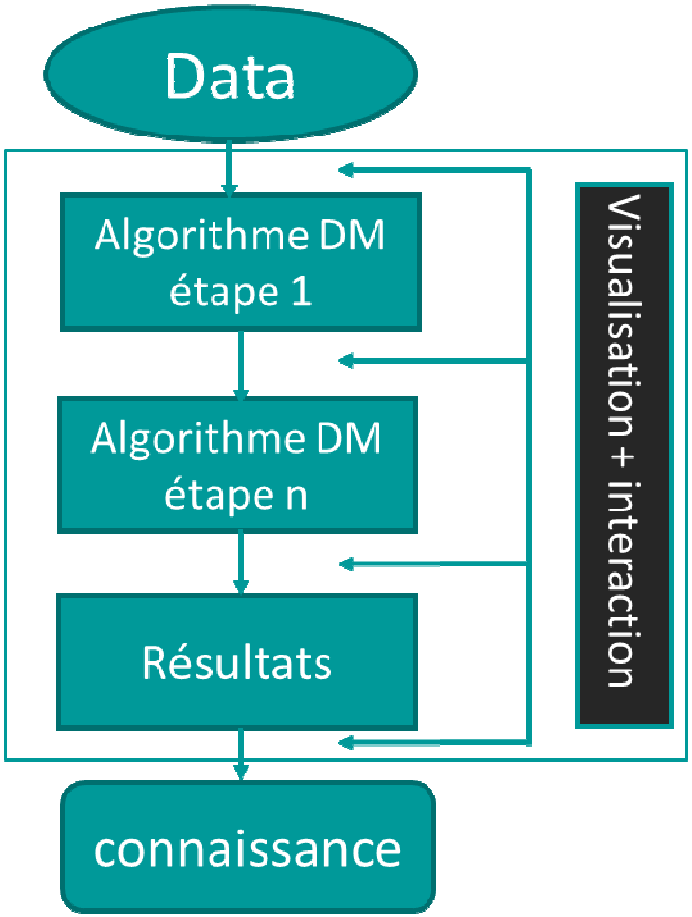
Approches du VSDM



Preceding Visualization (PV)



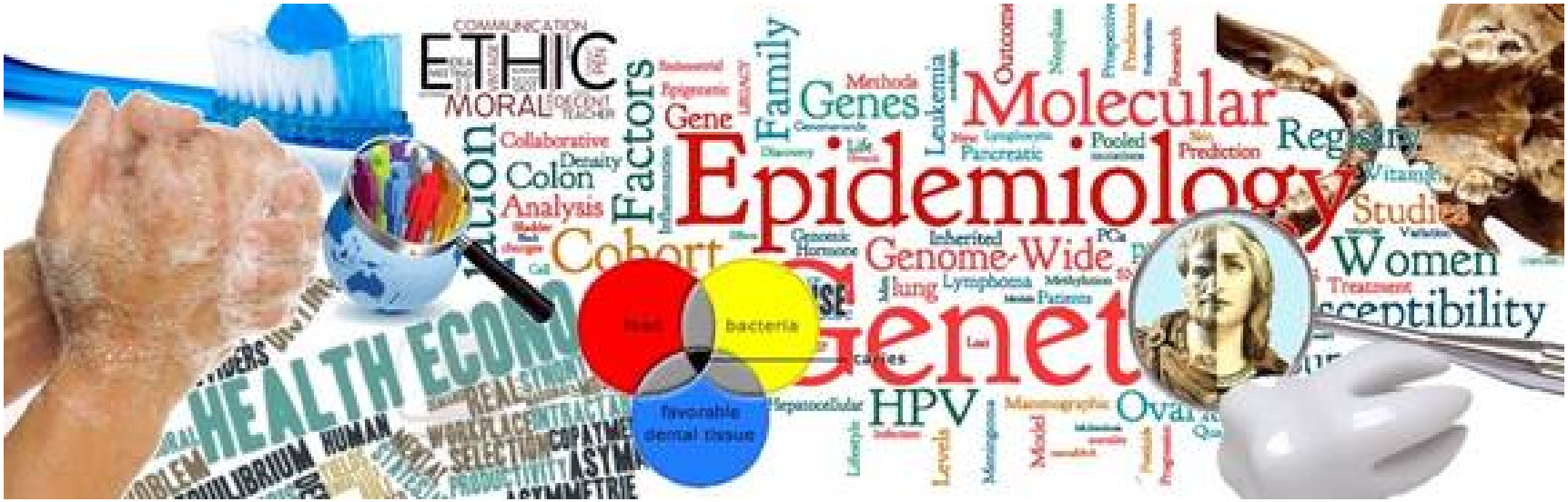
Subsequent Vizualisation (SV)



Tightly Integrated Visualization (TIV)

3.VSDM et épidémiologie

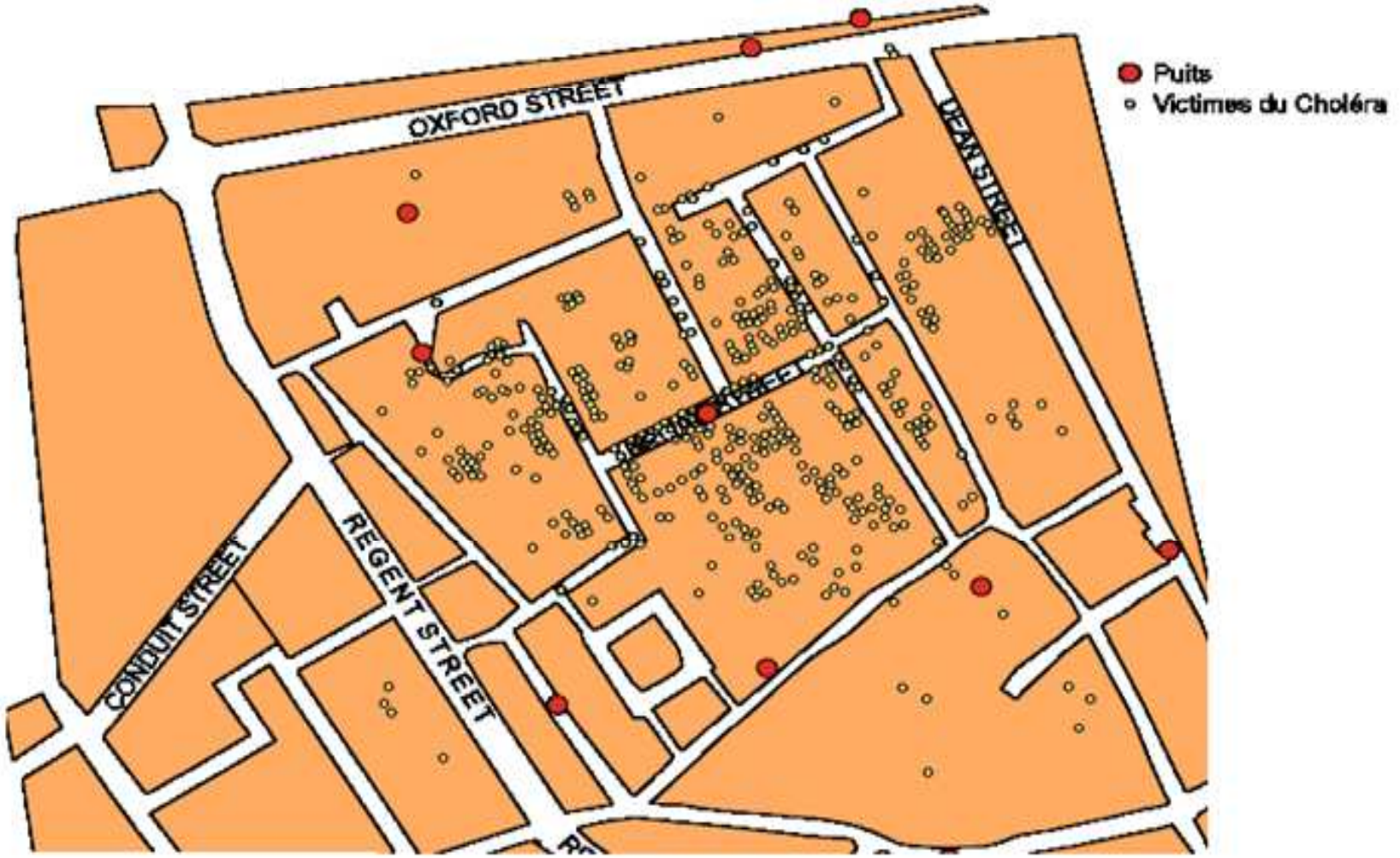
épidémiologie



L’utilisation du VSDM en épidémiologie :

- L’utilisation des méthodes du Data Mining en épidémiologie et santé publique est en forte croissance. C’est la disponibilité de vastes bases de données historiques qui incite à les valoriser, et a les étudié afin de mieux comprendre les modèles et les tendances de la propagation des maladies et à explorer les relations entre les maladie et l’environnement , le climat et autre facteur de risque.

Les statistique spatial et l'épidémiologie



4.Méthodologie

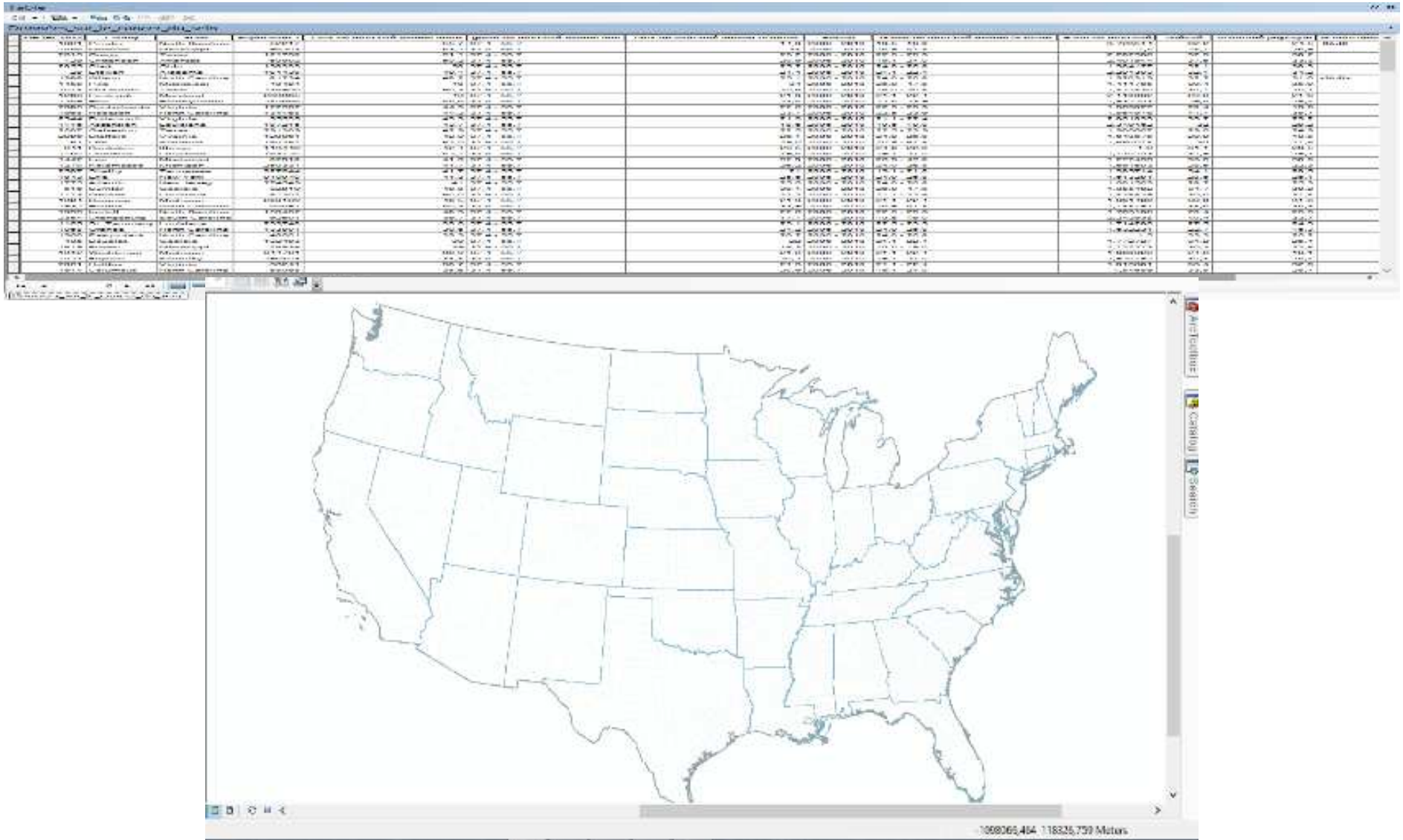
Objective de l’étude :

Appliqué les méthode du SDM (méthode statistique) sur une base de données spatiales épidémiologique(cancer du sein) et interpréter les résultats visuellement

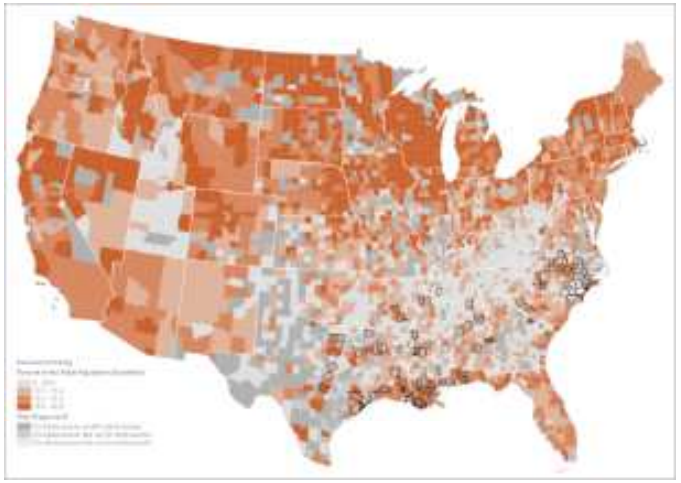
Zone d’étude

- Les états unis d'Amérique

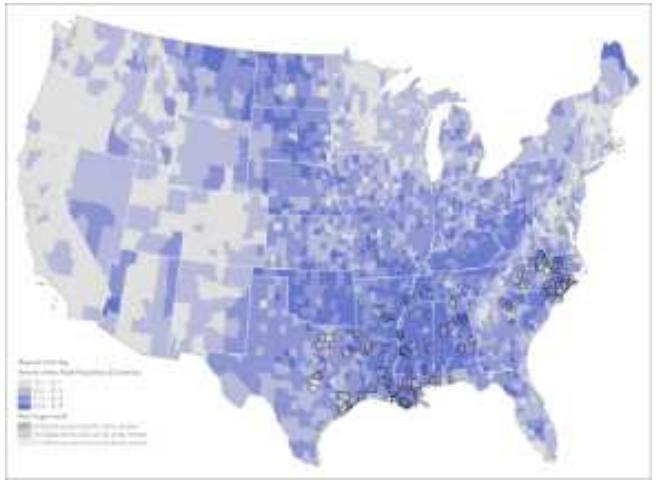
Base de données



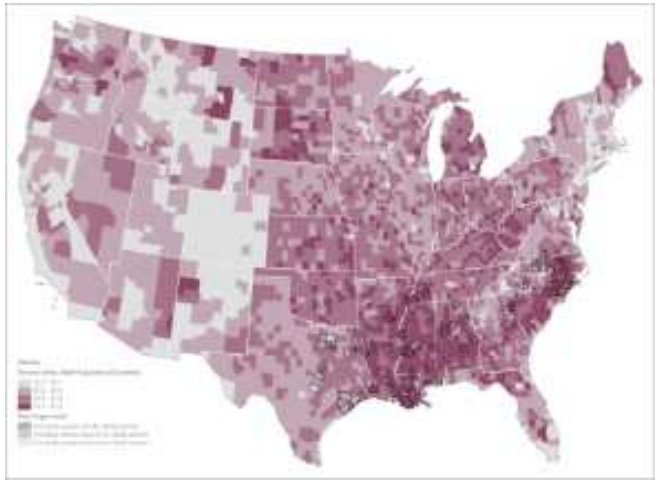
Facteur de risques



Abus d'alcool

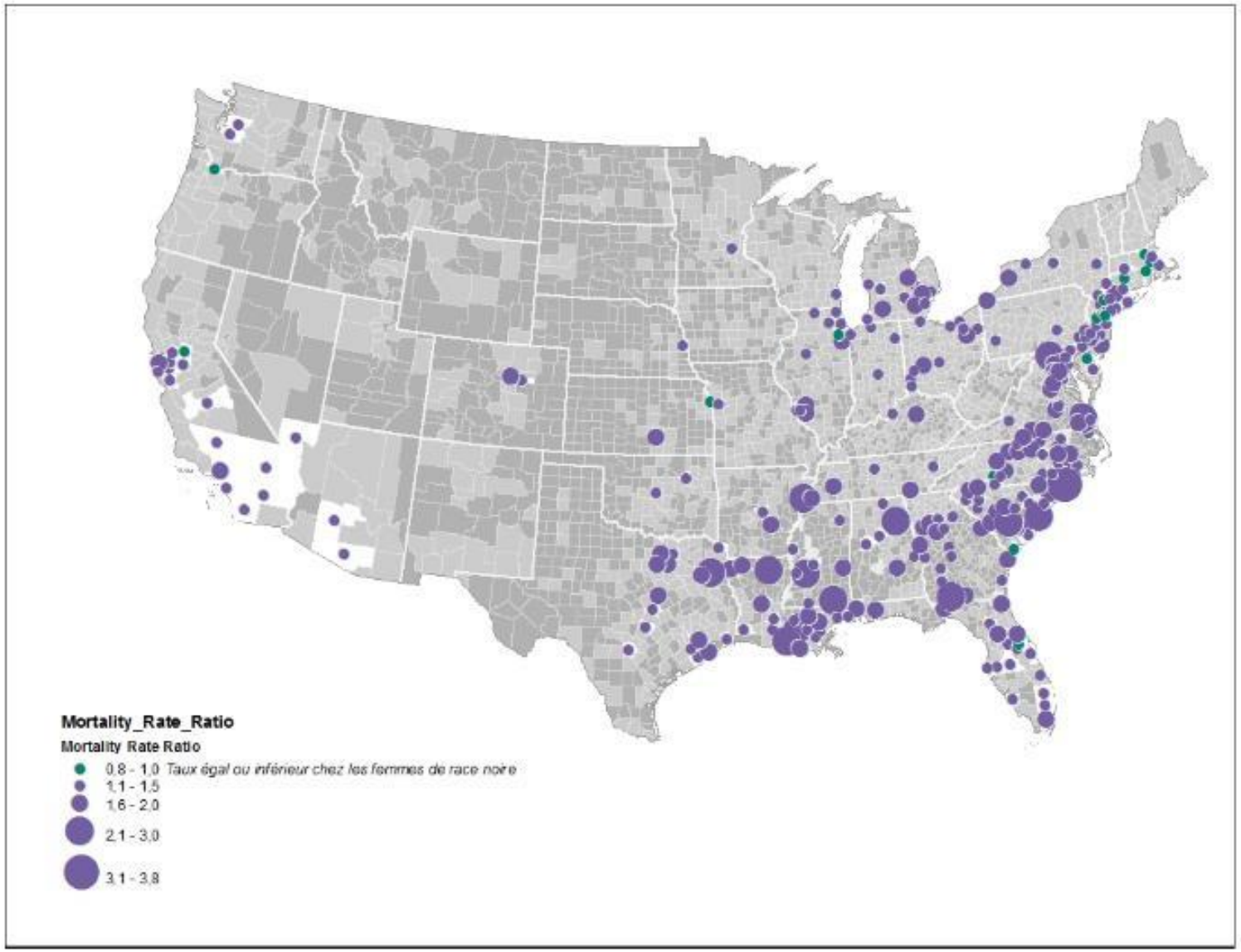


Inactivité physique



Obésité

Taux de mortalité



5.Application

Environnement de développement

- **ArcGIS**

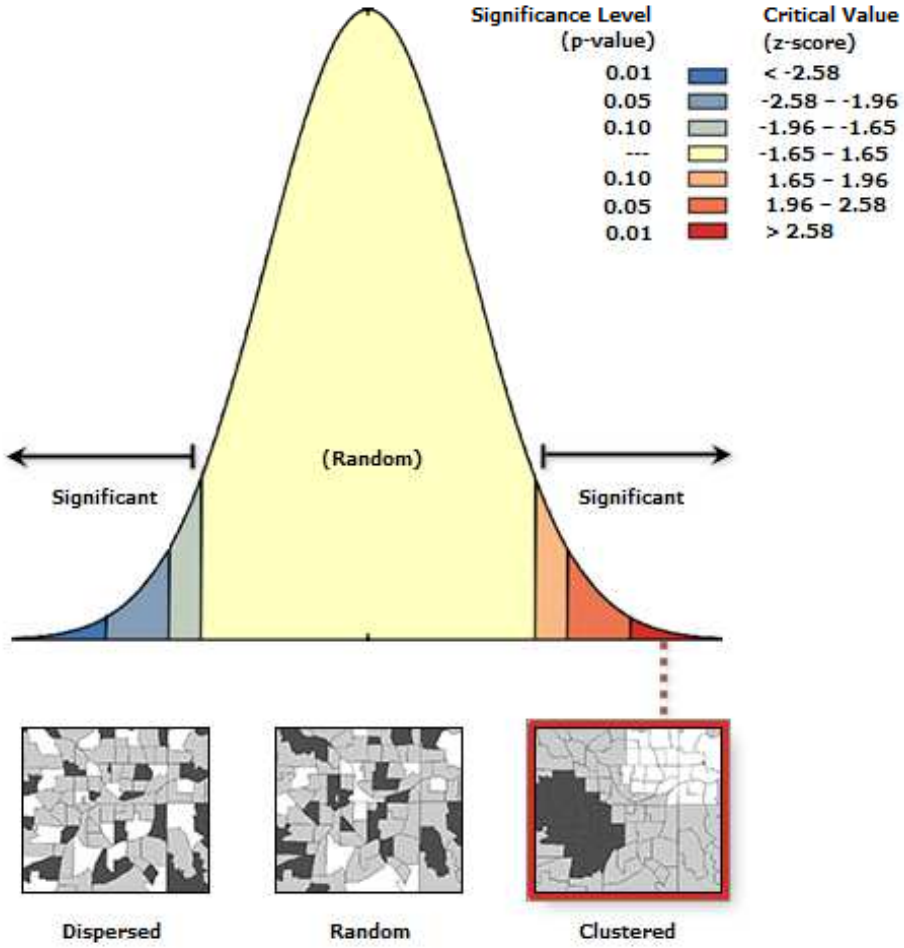
- **Visuel studio**

Méthodes SDM utilisé

- On as utilisé une approche basé sur les statistiques spatiales
  - Analyse global :
    - Autocorrélation spatial (Morane I)
    - Clustering (Getis ORD G)
  - Analyse local :
    - Hotsopt (Getis Ord G*)
    - Cluster and outlier (Anselin Local Moran's I ,LISA)

Analyse Globale

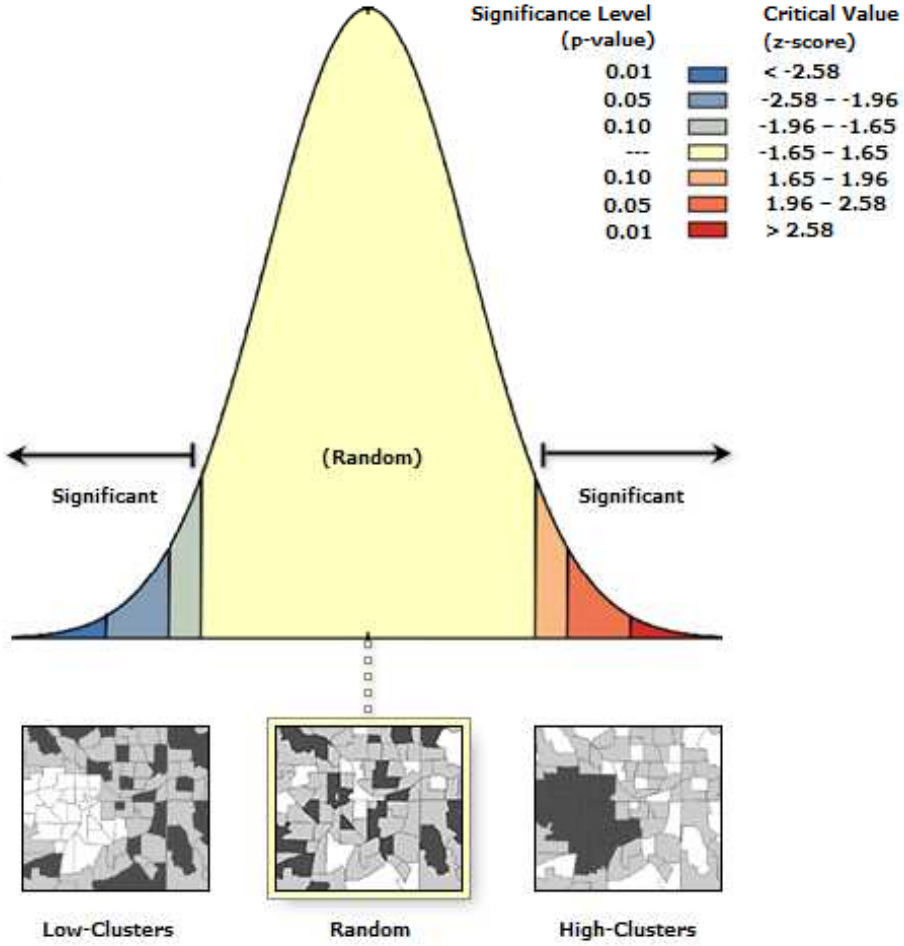
Autocorrélation spatial global (Morane I)



Global Moran's I Summary	
Moran's Index:	0,150409
Expected Index:	-0,003460
Variance:	0,000301
z-score:	8,862506
p-value:	0,000000

Analyse Globale

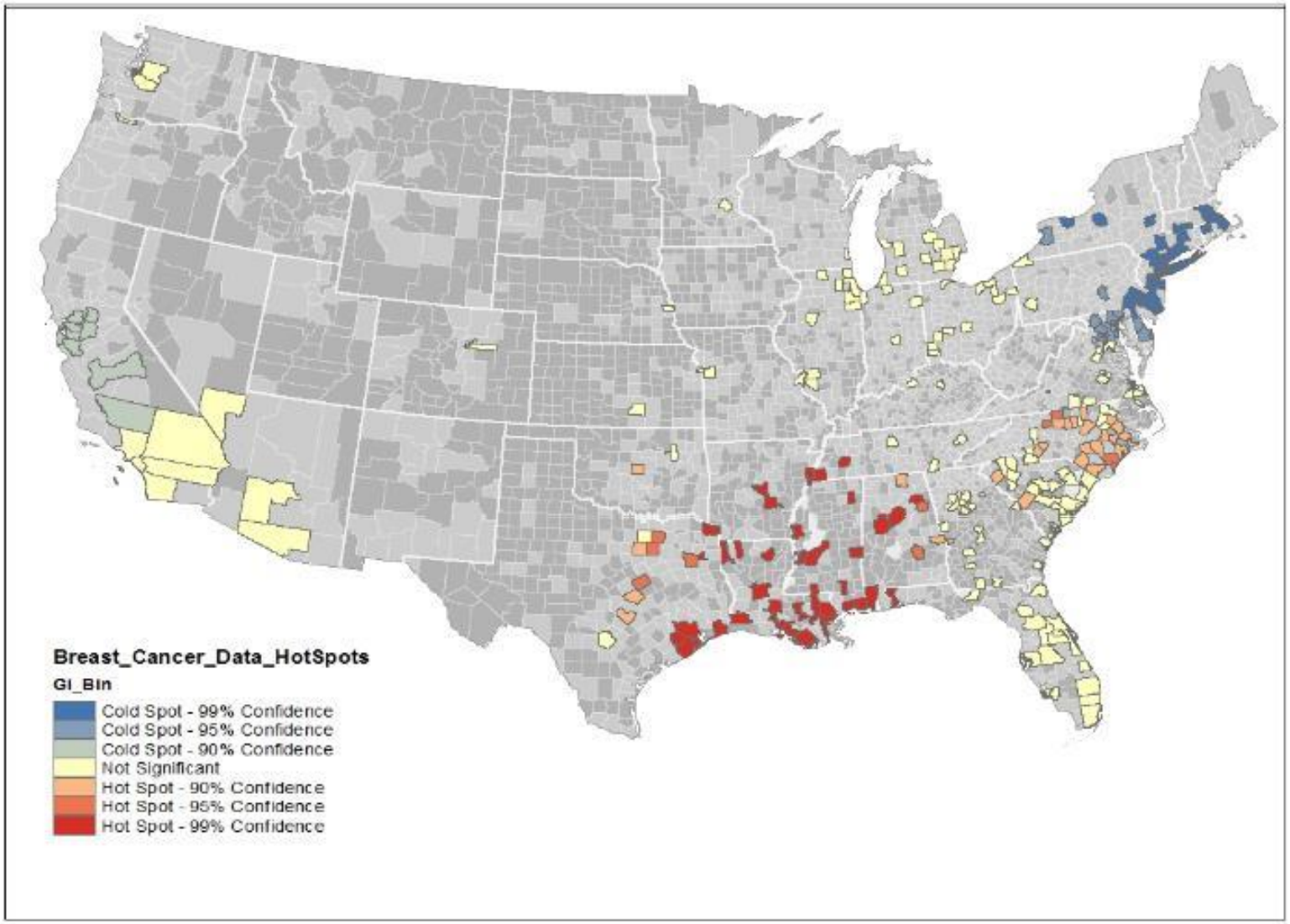
High-Low Clustering (Getis Ord G)



General G Summary	
Observed General G:	0,000001
Expected General G:	0,000001
Variance:	0,000000
z-score:	-1,231656
p-value:	0,218078

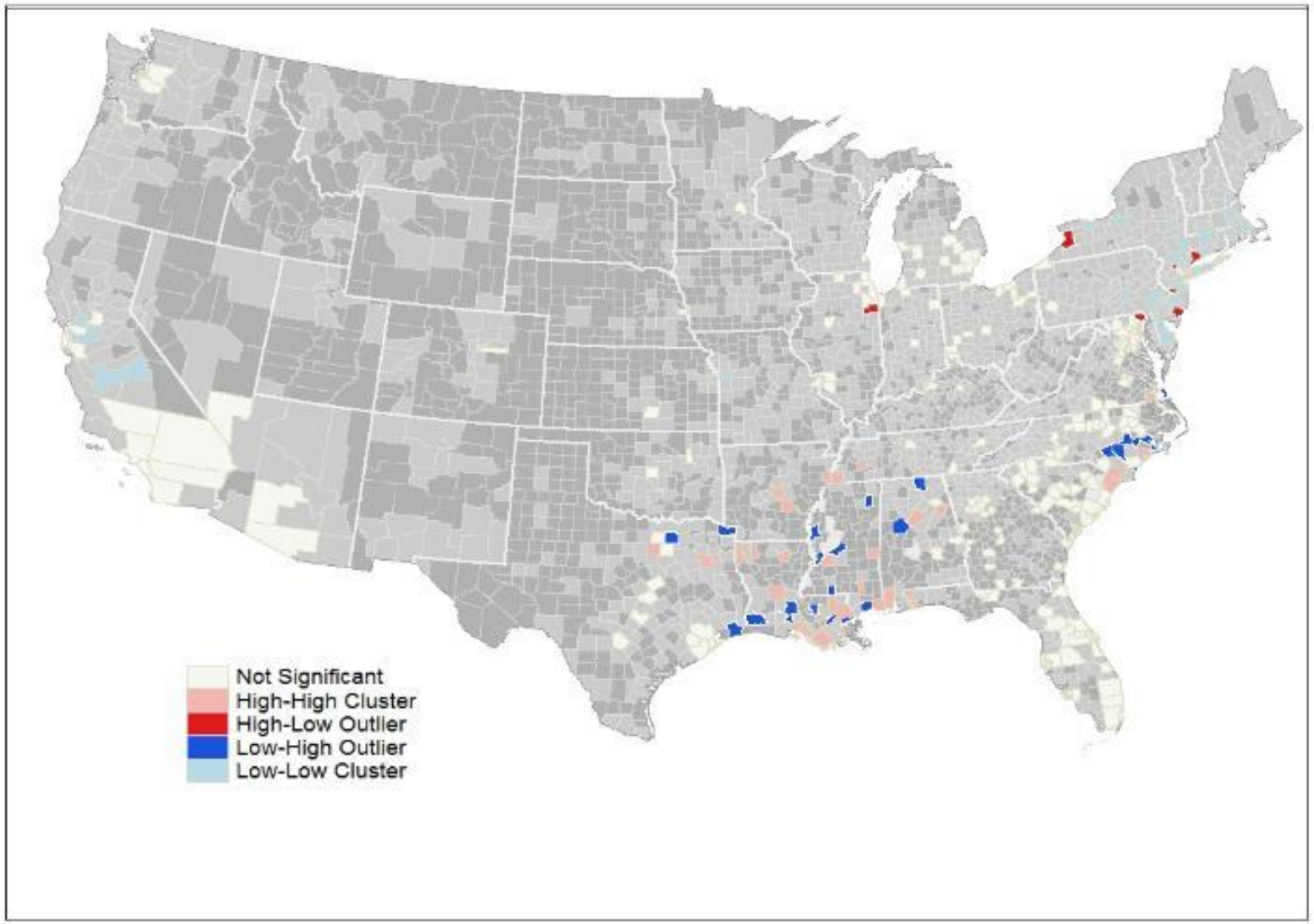
Analyse locale

Visualisation du résultat du hot spot (Getis-Ord *i G*∗)

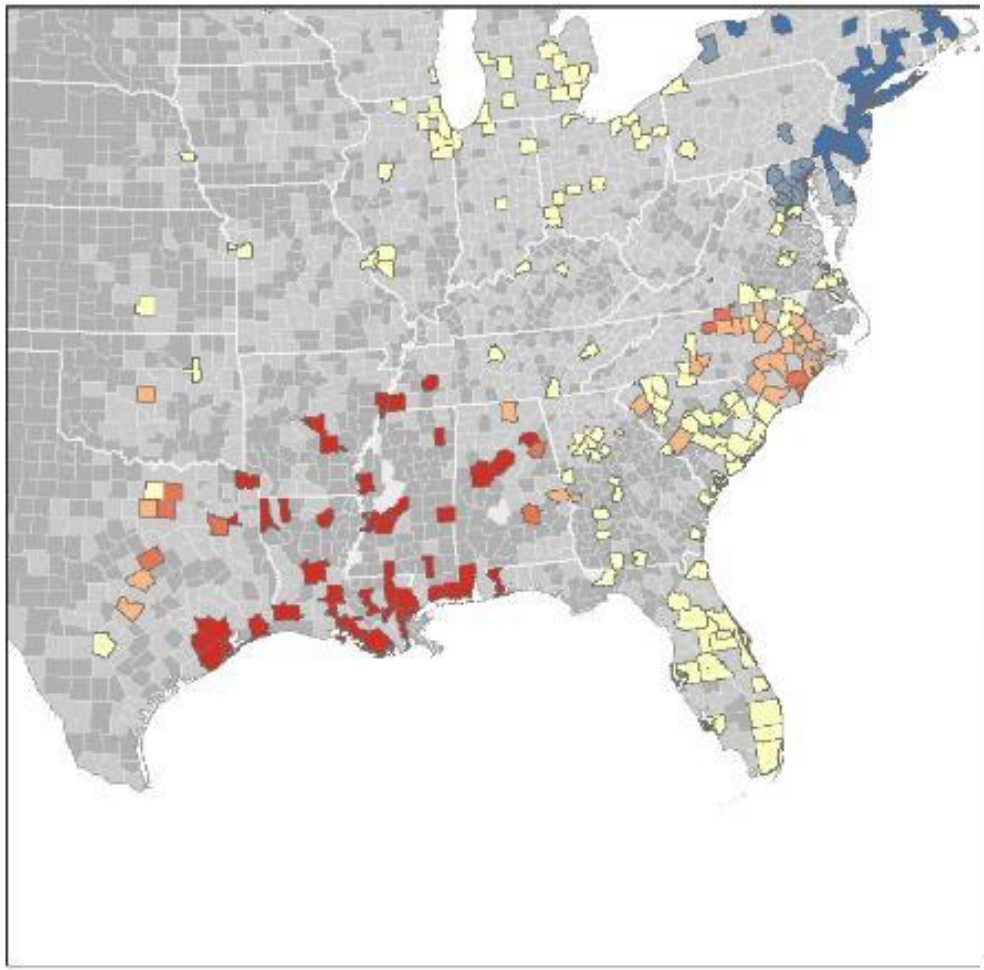


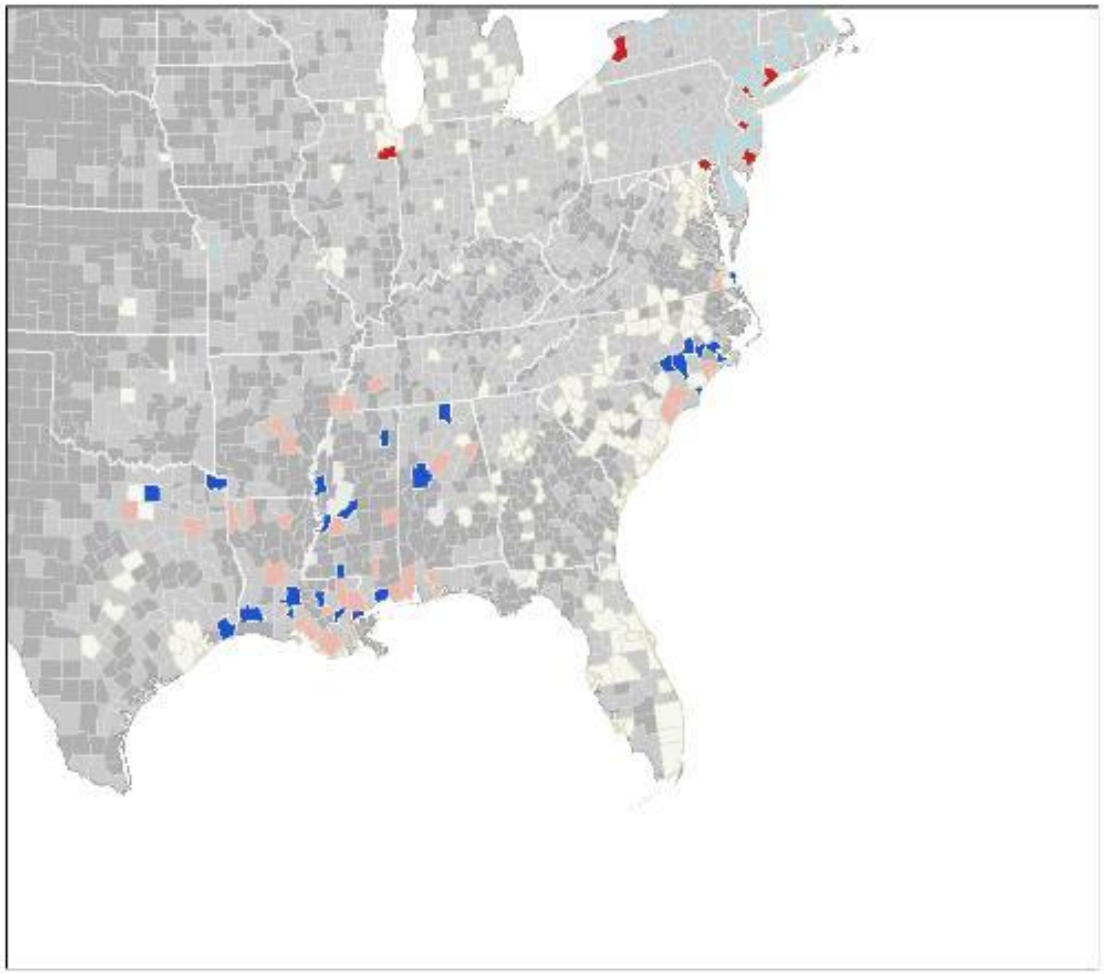
Analyse Locale

Cluster and outlier (Anselin Local Moran's I ,LISA)

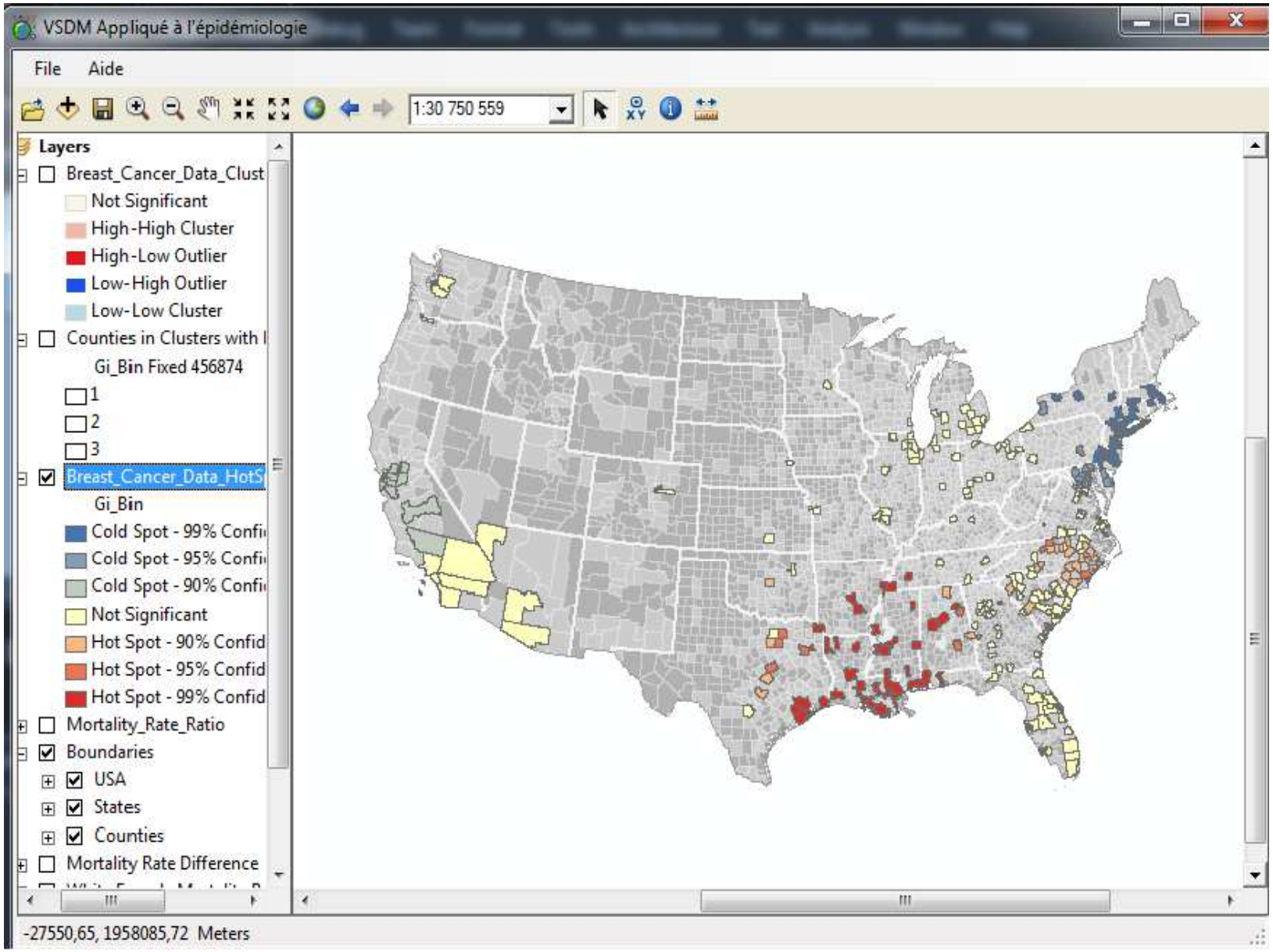


Interprétation plus approfondit





l’interface créé dans visuel studio.



6.Conclusion

- SDM dérive du data mining classique et prend en compte les relation spatiales.
- En intégrant la visualisation dans ce processus on as facilité le processus du SDM pour avoir une meilleure extraction des connaissance des relation spatial complexe.
- Notre travail a u pour but l'application du VSDM et certaines connaissances concernant les techniques d'analyse de données dans le domaine de l'étude des épidémiologies.

.

Merci pour votre intention

Table des matières

Partie II : VSDM et épidémiologie

Liste des figures

Résumé

La plupart des grandes bases de données actuellement disponibles ont une forte composante spatiale et contiennent des informations potentiellement utiles qui pourraient être de valeur. La discipline chargée de l'extraction de ces informations et connaissances est le data mining. La découverte de connaissances est effectuée en appliquant des algorithmes automatiques qui reconnaissent des modèles dans les données.
Les algorithmes d'exploration de données classiques supposent que les données sont générées de façon indépendante et identiquement distribuées. Les données spatiales sont multidimensionnelles, spatialement autocorrélées et hétérogènes.
Ces propriétés font en sorte que les algorithmes de data mining classique soient inappropriées pour les données spatiales, et que leurs hypothèses de base cessent d'être valables. L'extraction de connaissances à partir de données spatiales nécessite donc des approches particulières. Une façon de le faire est d'utiliser l'exploration visuelle des données. Lorsque le data mining visuel est appliquée aux données spatiales, il fait partie de la discipline appelée Visual Spatial Data Mining (VSDM).
Les deux types de data mining : automatique et visuel, ont leurs avantages respectifs. Les ordinateurs peuvent traiter de grandes quantités de données beaucoup plus rapidement que les humains, alors que les humains sont capables de reconnaître des objets et d'explorer visuellement les données beaucoup plus efficacement que les ordinateurs. Une combinaison de l'exploration de données visuelle et automatique rassemble les compétences humaines cognitives et informatiques pour une découverte de connaissance efficace.

Ce projet propose l'utilisation du VSDM pour la découverte de connaissance dans des données spatiales dans le domaine épidémiologique.

Introduction générale

Jamais auparavant dans l'histoire, les données n'ont été générées à des volumes aussi importants qu'aujourd'hui. L'exploration et l'analyse des vastes volumes de données sont de plus en plus difficiles notamment, depuis le développement d'outils de géocodage permettant la localisation par l'adresse. C'est le cas en géomarketing, dans l'analyse de la criminalité ou dans l'analyse de risques d'accidents ou d'épidémies. Cependant, la nature et le volume de données de base dépassent les capacités humaines d'analyse. D'où l'intérêt d'appliquer des techniques d'extraction automatique de connaissances telles que le Data Mining aux bases de données géographiques.

Les techniques du Data Mining classique ne prend pas en compte les relations spatiales qui contienne des informations potentiellement utiles, certains modèles intéressent restent enfuîtes, et ne peuvent pas être découvert.

Le spatial Data Mining est né du besoin d'exploitation dans un but décisionnel de données à caractère spatial produites, importées ou accumulées, susceptibles de délivrer des informations ou des connaissances par le moyen d'outils exploratoires. Il constitue un domaine à part, car il considère les interactions des objets dans l'espace. Ce domaine intègre des techniques provenant à la fois des bases de données spatiales et des SIG, du Data Mining et des statistiques spatiales. Cependant les techniques du SDM montrent leurs limites quand les bases de données sont trop volumineuses.

L'extraction de connaissances à partir de données spatiales nécessite donc des démarches particulières. Une façon de le faire est d'utiliser l'exploration visuelle des données. Lorsque le data mining visuel est appliquée aux données spatiales, il fait partie de la discipline appelée Visual Spatial data Mining (VSDM).

L'objectif de ce mémoire est d'une part comprendre est de maitriser les notions théoriques du data mining spatial et visuel et faire ressortir les plus importantes méthodes, et d'autre part d'découvrir le VSDM et l'appliqué aux problèmes d'épidémiologie.

Notre mémoire est structurée en trois chapitres, organisés comme suit :

Le premier chapitre sera consacré à l'étude du spatial data mining, nous commencerons par définir le data mining cité ses diffèrent méthodes, nous aborderons ensuite à la définition du spatial data mining, puis on fera un panorama sur ses méthodes et ses approches.

Le second chapitre sera divisé en deux grandes parties, la première sera consacrée pour l'étude du visuel spatial data mining ses types de visualisation ses méthodes et point fort, la seconde partie c'est pour le VSDM appliqué à l'épidémiologie.

Dans le troisième chapitre on va étudier une approche de spatiale data mining basée sur les statistiques spatiales, nous appliquerons comme méthodes de SDM le clustering comme

analyse global et les Hostpots comme analyse locale pour étudier des données relatives au cancer aux Etats unis d'Améique.

Chapitre 1 Spatial data mining

Introduction

Jusqu'à 80% des données d'une organisation ont une composante spatiale. Les données spatiales sont de plus en plus nombreuses grâce à l'évolution des outils d'acquisition de données (ex. GPS, images satellites, photos aériennes, etc.) et des méthodes de structuration (ex. raster, vecteur) et de représentation (ex. représentations 2D, 3D). De plus, des outils et des méthodes de représentation des données spatiales (ex. des outils de visualisation) ont été développés pour mettre en évidence les caractéristiques spatiales des données (position, forme, taille, orientation, etc.) et les relations qui existent entre elles (ex. intersection, adjacence, etc.) afin de faciliter leur interprétation dans ce chapitre nous allons définir le data mining cité ses diffèrent méthodes, nous aborderons ensuite le spatial data mining, ses méthodes sons application et ses approches.

1.Le Data Mining

Le Data Mining (traduit en fouille de données) est né dans le contexte où des données de production se sont accumulées au fil du temps et où l'on s'est posé la question de leur devenir.

Le DM est couramment défini comme l'extraction de connaissances intéressantes intelligibles (règles, régularités, patterns, contraintes) cachées dans les bases de données.[1]

1.2 Apprentissage supervisé et apprentissage non supervisé

Les algorithmes et techniques du data mining peuvent être regroupés en méthodes supervisées ou non, selon que l'algorithme ait besoin d'une connaissance préalable des données manipulées.

- **Apprentissage supervisé**

Il consiste à élaborer un apprentissage automatique à partir d'un ensemble de données de référence appartenant à des classes déterminées. Pour cela, les attributs de chaque donnée sont analysés et servent à établir une description des classes. À partir de cette description, il devient possible de trouver la classe d'appartenance de toute autre donnée. Ce type d'apprentissage concerne par exemple les arbres de décision et les réseaux de neurones. Les premiers présentent une situation sous la forme d'un arbre dont les nœuds correspondent à des résultats de décisions. Les seconds s'inspirent du comportement des neurones lors de la transmission d'informations entre eux, chaque neurone étant connecté à plusieurs autres en amont et en aval. La transmission ou non de l'information résultante d'un neurone vers les suivants dépend d'une fonction d'activation.[2]

- **Apprentissage non supervisé**

Dans ce type d'apprentissage, les classes ne sont pas connues d'avance et les informations sont extraites directement des données. Parmi les méthodes usuelles se trouvent le Clustering et la recherche de motifs fréquents[2]. Le Clustering consiste à partitionner les données, et à les regrouper en classes, à partir de critères d'homogénéité qui permettent de définir une distance. Grâce à celle-ci, des groupes de données sont établis, en minimisant les distances intragroupes et en maximisant les distances intergroupes. Les motifs fréquents ou *itemsets* caractérisent des associations entre des attributs de données. Leur recherche sera abordée dans le paragraphe.

1.3 Panorama des méthodes du data mining :[2]

Les méthodes de DM peuvent être classées en deux catégories : les méthodes utilisées dans une phase exploratoire (non supervisé) et les méthodes à caractère plus décisionnel qui cherchent à prédire une donnée particulière (supervisé).

1.3.1 Méthode descriptive:

- **La description :**

La description consiste à mettre au jour :

- Pour une variable donnée : la répartition de ses valeurs (tri, histogramme, moyenne, minimum, maximum, etc.).

- Pour deux ou trois variables données : des liens entre les répartitions des valeurs des variables. Ces liens s'appellent des « tendances ».

- **Le clustering**

Aussi appelée segmentation : consiste à créer des classes (sous-ensembles) de données similaires entre elles et différentes des données d'une autre classe

- **L'association :**

Consiste à trouver quelles valeurs des variables sont corrélées ensemble. Par exemple, telle valeur d'une variable intervient avec telle valeur d'une autre variable.

1.3.2 Méthode prédictive :

- **L'estimation :**

L'estimation consiste à définir le lien entre un ensemble de prédicteurs et une variable cible. Ce lien est défini à partir de données « complètes », c'est-à-dire dont les valeurs sont connues tant pour les prédicteurs que pour la variable cible. Ensuite, on peut déduire une variable cible inconnue de la connaissance des prédicteurs. À la différence de la segmentation (technique prédictive suivante) qui travaille sur une variable cible catégorielle, l'estimation travaille sur une variable cible numérique**.**

- **La prévision :**

La prévision est similaire à l'estimation et à la segmentation mise à part que pour la prévision, les résultats portent sur le futur. les méthodes étudiées sont représentées dans la figure 1 :

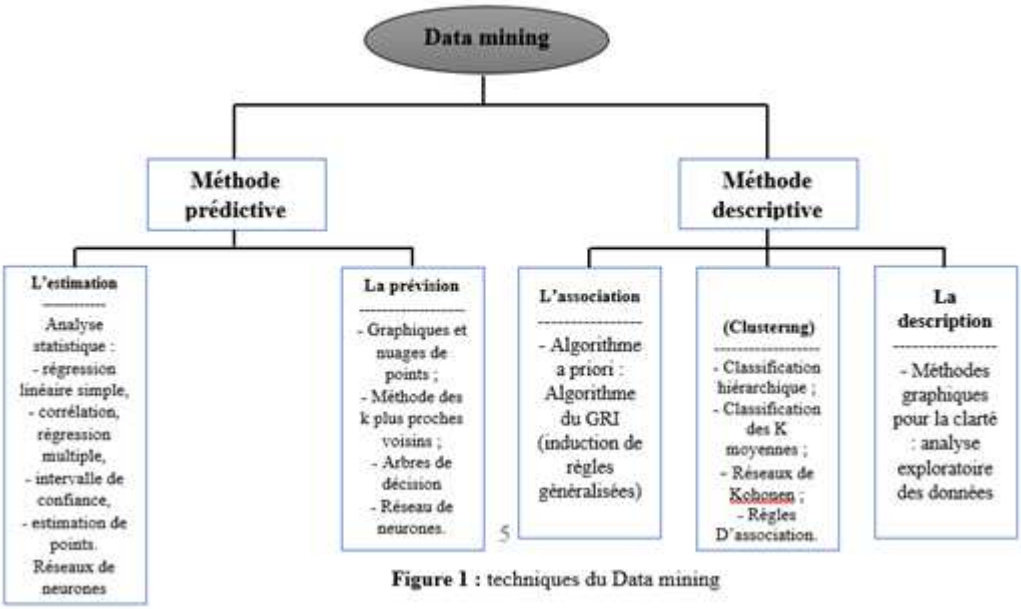


Figure 1 : techniques du Data mining

2. Le spatial Data mining :

C'est le processus de découverte de modèles intéressants et précédemment inconnus, mais potentiellement utiles à partir de bases de données spatiales. La complexité des données spatiales et des relations spatiales limite l'utilité des techniques classiques d'extraction de données pour l'extraction de motifs spatiaux. [3]

Base de données spatial : C'est une base de données optimisée pour stocker et requêter des données reliées à des objets référencés géographiquement, (des découpage administratif, Réseaux routier, Cadastre, POS, Topographie (courbes de niveau)[4]

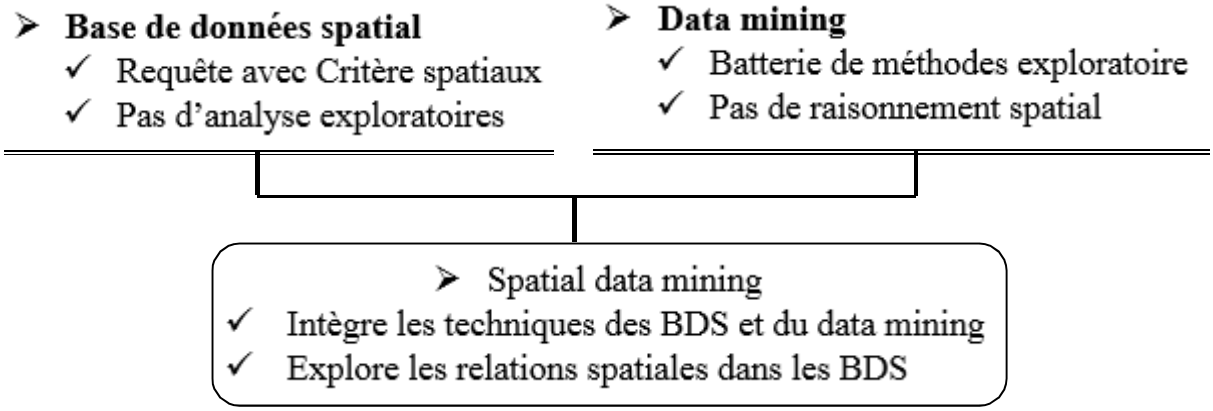


Figure 2 : Spatial data mining

Les méthodes mises en œuvre pour le data mining spatial utilisent de manière intensive les relations spatiales. C'est ce qui distingue ces méthodes de celles appliquées dans le cas de données de type alphanumérique. Ces relations spatiales jouent donc un rôle primordial dans l'analyse de données spatiales et la découverte de connaissances.

2.1 Application du SDM

Le data mining spatial peut être utilisée pour la compréhension des données spatiales, la découverte des relations entre les données spatiales et non spatiales, la construction de bases de connaissances spatiales, l'optimisation des requêtes, la réorganisation des données dans des bases de données spatiales, la saisie des caractéristiques générales de manière simple et concise. Ainsi, la technique peut être appliquée pour la détection, la cartographie et la prédiction de tout phénomène qui manifeste une composante spatiale.

Le SDM trouve son application aussi bien dans le domaine public, dans le domaine scientifique que dans le secteur privé. Mais les objectifs ne sont pas identiques. En exploitant les données géographiques, les administrations recherchent plutôt des modèles concernant la population et son bien-être, tandis que l'industrie a des objectifs de rentabilité dans l'implantation d'usines, d'antennes de télécommunication, de panneaux publicitaires, etc. Dans le domaine des sciences, l'exploration des données spatiales sert à la recherche. En astronomie et en astrophysique, Le SDM sert à la classification automatique d'objets spatiaux, ou bien à découvrir des régions dignes d'intérêt, ou des objets rares dans l'immensité de notre univers. En archéologie, les données géographiques et la fouille de données spatiales sont exploitées pour trouver de nouveaux sites.[5] Le data mining spatial est utilisée en épidémiologie pour

prévoir la propagation des maladies Les Sciences de la vie et de la Terre ont aussi recours à cette technique pour évaluer les tendances au cours du temps des modifications de la végétation dans des zones sensibles.

2.2 Panorama des méthodes du Spatial data mining

Les méthodes types sont un prolongement des tâches de fouille de données intégrant les données et les critères spatiaux. Ainsi, une première phase exploratoire permet une description synthétique (indice d'auto corrélation globale, généralisation, densité, lissage), de découvrir les écarts donnant les spécificités locales ou de chercher des regroupements de données (clusters). Cette première phase permet de guider la phase décisionnelle, où l'on procède à une analyse plus fine afin d'expliquer les écarts ou de caractériser les groupes (caractérisation, règles de classement ou d'associations)[5]. Nous allons maintenant décrire ces différentes méthodes.

2.2.1 Phase exploratoire

Dans le cas où les données seraient corrélées, il faut les simplifier afin de faire apparaître une tendance générale. Pour réaliser cela, il existe différentes approches dont celles basées sur la densité, l'analyse multidimensionnelle lissée ou la généralisation.

- **Généralisation spatiale**

Elle consiste à substituer les valeurs estimées trop détaillées par des valeurs moins détaillées jusqu'au niveau de détail souhaité, puis à agréger et compter les attributs identiques ainsi obtenus. Cette méthode permet de résumer les données et constitue une première étape pour induire des règles d'associations. [6].elle est Le découpage administratif en pays, régions, département, communes. Comme le montre la figure 3.

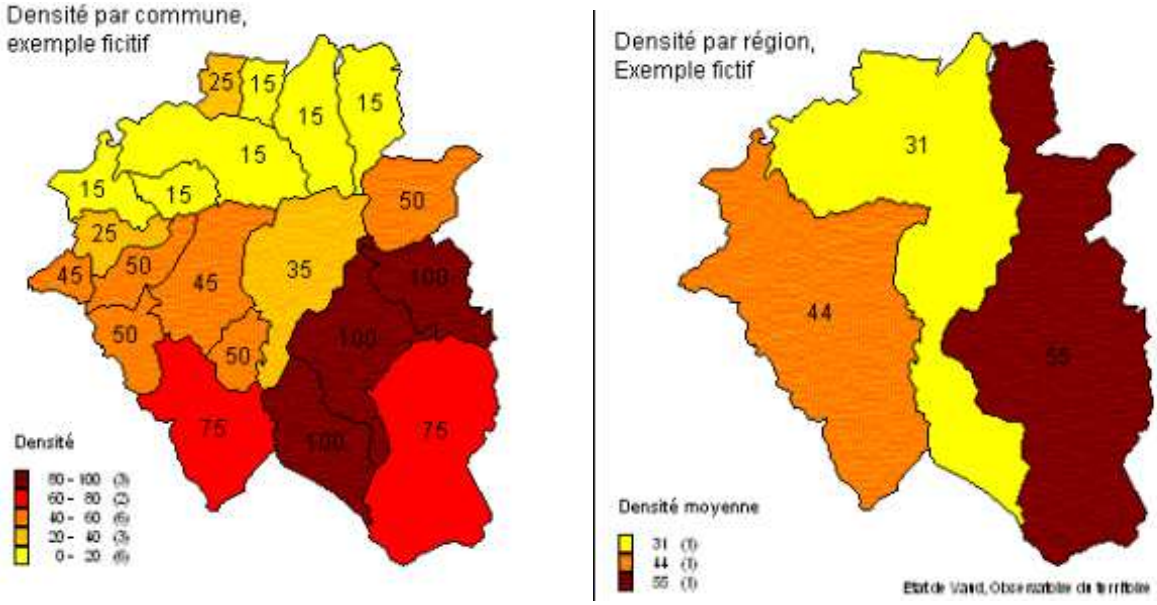


Figure 3: Application de la généralisation spatial pour obtenir la densité par région ou commune [7]

- **Statistiques spatiales**
  - **Analyse globale** permet de mesurer l'autocorrélation spatiale d'une variable et la ressemblance entre voisin avec l'indice globale [7]

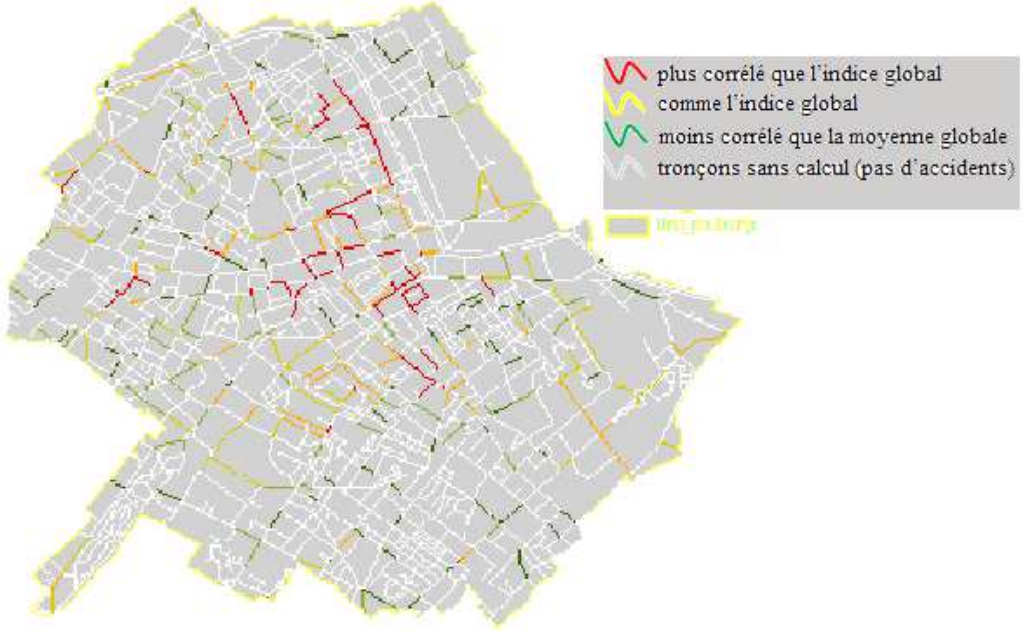


Figure 4 : Tançons des corrélations globales[7]

  - **Analyse locale**

Deux particularités caractérisent indicateur local d'association spatiale. Au contraire de l'indice global dont les valeurs sont comprises entre -1 et 1, les valeurs de l'analyse locale varient sans limites autour de 0. Une valeur négative indique une association spatiale locale de valeurs différentes, d'autant plus importante que le résultat est fort. Une valeur positive montre, au contraire, que localement les unités spatiales ont tendance à se ressembler. La seconde caractéristique est que la somme des indices locaux est proportionnelle à l'indice global calculé sur le même espace.[7]

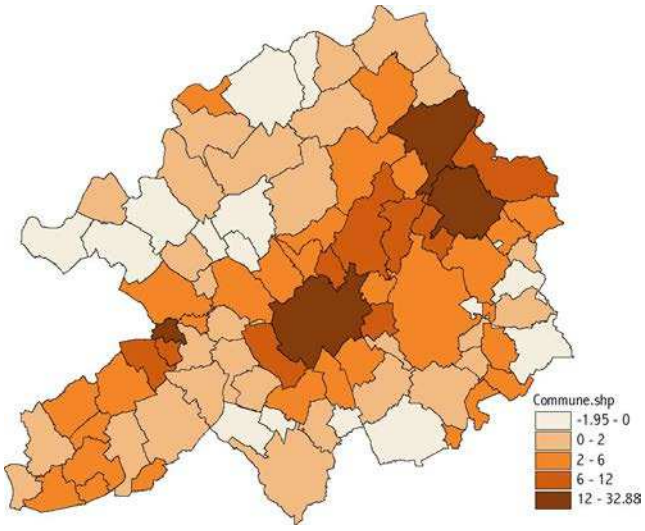


Figure 5 : Analyse locale (Autocorrélation) [7]

- **Clustering**

Le clustering est une méthode **non supervisée** qui regroupe des objets dans des classes. Son but est de maximiser la similarité intra-classes et de minimiser la similarité inter-classes. Elle est couramment utilisée en fouille de données et est bien connue dans le domaine des statistiques [8]**.**

La transposition au domaine spatial des méthodes de clustering s'appuie sur une mesure de similarité d'objets localisés suivant leur distance métrique. Néanmoins, la finalité du clustering en spatial n'est pas tant de former des classes que de détecter des concentrations anormales (par exemple, détecter un point chaud dans l'étude de criminalité, ou des zones à risque en epidiomologie). Cette étape est souvent utilisée en amont d'autres tâches de type décisionnelles comme la recherche d'associations entre groupes et d'autres entités géographiques ou la caractérisation au sein d'un groupe. Un exemple d'application est de former des clusters d'habitations puis de rechercher des caractéristiques communes par cluster.

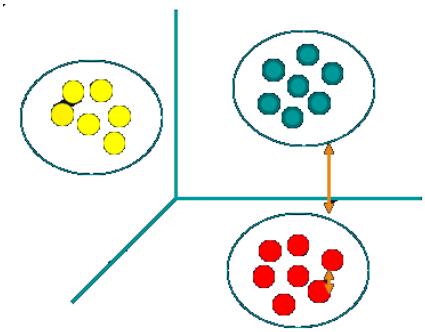


Figure 6 : Clustering maximiser la similarité intra-classes et de minimiser la similarité inter-classes.[8]

2.2.1 Phase décisionnelle

Le terme explicatif ici est lié à une intervention de l'analyste qui à la suite d'une découverte de clusters ou de valeurs atypiques par rapport à une tendance, focalise son analyse sur un sous-ensemble d'objets, sur une partie des variables ou encore sur une zone géographique. Cette partie des données est ensuite expliquer sa particularité par des liens avec certaines valeurs ou par des règles caractéristiques. Ces méthodes, à l'inverse des méthodes précédentes, opèrent sur plusieurs couches thématiques pour permettre d'expliquer un phénomène suivant les propriétés de son entourage. Nous décrivons les méthodes de caractérisation, de règles d'association et de classification.

- **Caractérisation**

Définis la caractérisation comme l'induction des propriétés caractéristiques d'un sous-ensemble de données. Une règle caractéristique est une assertion qui décrit un concept satisfait par tous ou une grande partie des objets sélectionnés.[9] Appliquée à des bases de données spatiales, la caractérisation découvre en plus le niveau d'extension de ces propriétés aux "voisins".

- **Règles d'association**

L'extension de la découverte de règles d'association des données spatiales permet de générer des règles de type : X → Y (s, c) avec s comme support et c la confiance

Support X →Y est le nombre de transaction contenant à la fois tous les items de X et tous les item de Y par rapport au nombre total de transaction

$$Conf(\mathrm{X} \rightarrow Y) = \frac{Supp(\mathrm{X} \cap Y)}{Supp(\mathrm{X})}$$

Telle que X et Y sont des ensembles de prédicats spatiaux et non spatiaux. En d'autres termes, ceci revient à trouver des associations entre des propriétés des objets et celles de leur voisinage.

- **Classification**

La recherche de règles de classement vise à structurer un ensemble d'objets en classes d'objets ayant des propriétés communes. Cette tâche est réalisée par apprentissage supervisé qui, à partir de classes fournies partiellement en extension (un échantillon de la base de données), induit une description en intention permettant de classer les prochaines données.[9]

Ainsi, 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%)

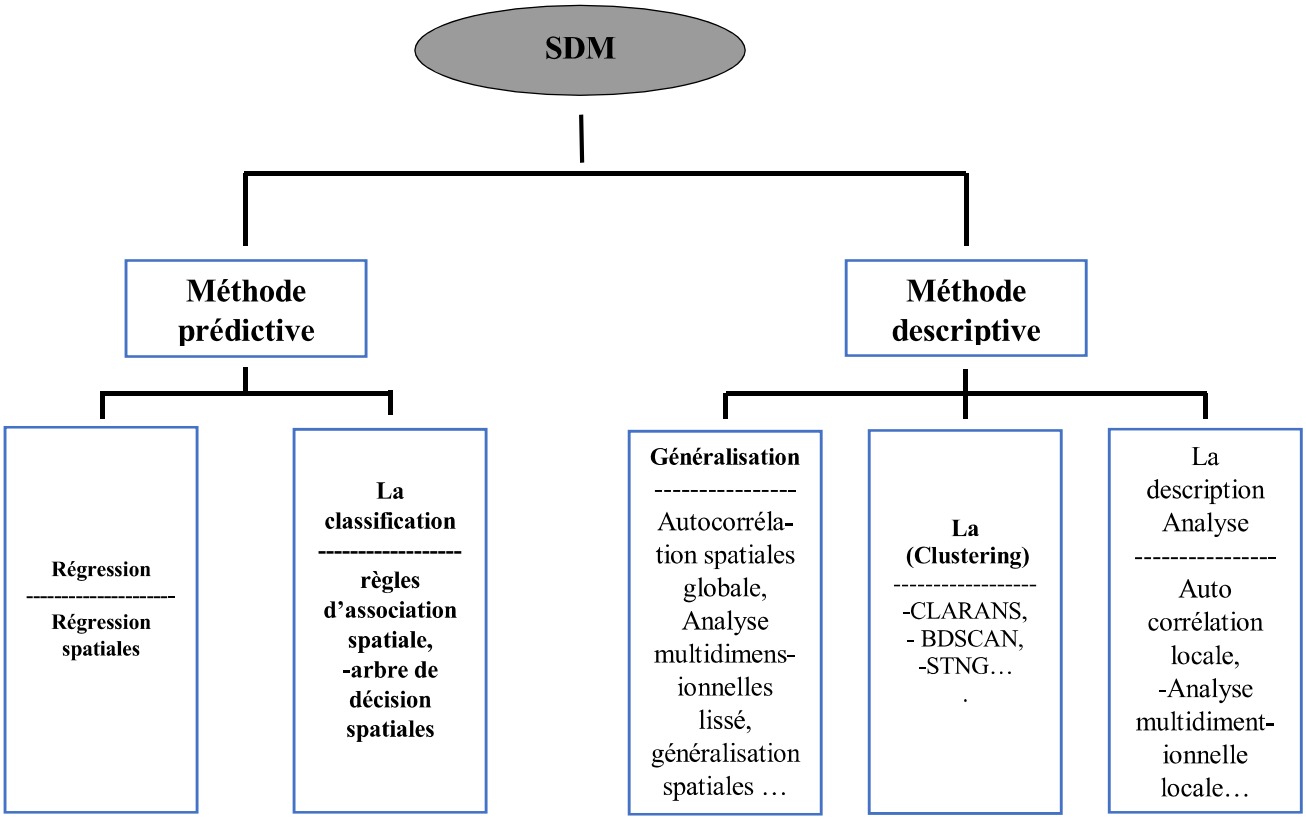


Figure 7 : Techniques spatial Data mining

2.3. Approches du DMS

Il existe deux approches pour l'analyse et l'extraction de connaissances d'une base de données spatiales. La première est issue des statistiques spatiales et la seconde du domaine des bases de données (approche BD). Très peu de liens existent aujourd'hui entre ces deux types de recherches. Malgré cela, elles permettent parfois de résoudre les mêmes tâches d'analyse et ont certains points en commun. Cette section dresse une synthèse de ces similarités tout en soulignant les différences de ces approches. Elle montre aussi les forces et les faiblesses de chaque approche qui sont récapitulées dans le tableau ci-dessous.

	Point fort	Point faible
APPROCHE STATISTIQUE	• Base mathématique solide : mesures, indicateurs • Visuelle report sur des cartes (ESDA) • De multiples possibilités (acquis des proba/stat) délivre des infos précises et quantifiées à l'analyste	• Monocouche thématique comme dans l'autocorrélation ou l'analyse de distribution données uniquement ponctuelles ou zonales • Pas ou peu d'attributs certaines méthodes n'utilisent d'attributs qu'en phase de préparation (ex : cluster) d'autres portent sur une seule mesure seule l'AD contiguë porte sur plusieurs attributs • Plus difficile à interpréter aux néophytes en stat./AD • Ne découvre pas explicitement des règles spatiales
APPROCHE BASE DE DONNEES SPATIALES	• multithème : exploite/découvre les relations spatiales tout type de relations (connexité, distance, ∩, …) toute forme d'objets (points, surfaces, lignes) • Multi-attribut tout type d'attribut (mesure, qualitatif) • Basé sur les techniques BDS (jointures, requêtes) • Facilement interprétable induit directement des règles É des relations spatiales • Utilise des connaissances sémantiques d'experts hiérarchies de concepts, analyse multiniveau	• Moins d'interactivité avec la carte (sauf OLAP spatial) • Validité, robustesse moins sûre et moins mesurable pas assez de modèles • Résultat généré parfois en nombre et difficile à filtrer • Effet boîte noire

Tableau 1. Comparaison des approches au SDM [7]

Conclusion :

Le spatial data mining dérive du data mining classique sauf qu'il présente une spécificité importante pour la prise en compte des relations spatiales.

L'objectif du SDM est d'automatiser partiellement la découverte de connaissances, c'est-à-dire la recherche de relation et information intégrées dans des bases de données spatiales. Sauf que ces Algorithmes ne sont pas très performants lorsque le volume de données spatial traité, soit très grand et complexe, pour surpassé cette faiblesse nous abordons dans chapitre suivant une autre approche qui réglera ce problème,

Chapitre 2 VSDM et épidémiologie

Partie I : VSDM

Introduction

La représentation de données est un élément important pour leur analyse, là où une mauvaise représentation noie l'information dans la masse des données, une représentation adéquate facilite l'identification d'éléments intéressants et rend possible leur interprétation. Ce problème de la représentation des données se pose aussi bien pour leur analyse automatique (par des algorithmes d'extraction de connaissance) que pour une analyse par un expert qui visualise les données (visual data mining). Dans le premier cas, il s'agit d'une question de représentation informatique de données, dans le second cas, il s'agit d'une question de représentation visuelle. L'analyse visuelle des données est facilitée par la capacité des outils de visualisation à exprimer le maximum de contenu informationnel des données.

1.Définition

- **Visuel Data mining : VDM**

La fouie de données visuelles est la combinaison des techniques usuelles du Data mining avec les méthodes de visualisation de l'information[10].

- **Visual spatial data mining :**

C'est l'intégration des approches de visualisation dans le processus du data mining pour fournir des outils et des atouts pour représenter les données géospatiales.

- **Les données géospatiales :**

Les données géospatiales fournissent de l'information sur la forme et la localisation d'objets et d'événements sur la surface terrestre. Elles comprennent l'ensemble des données géométriques (position et forme des objets), des attributs (caractéristiques des objets) et des métadonnées (information sur la nature des données).[11]

2. Le rôle de la visualisation dans le Data mining :

La plupart des bases de données contemporaines contiennent de grandes quantités de données multidimensionnelles, ce qui rend la recherche des informations précieuses une tâche difficile. Avec les systèmes d'exploration de données automatiques d'aujourd'hui, il est seulement possible d'examiner des portions relativement petites de données. N'ayant aucune possibilité d'explorer les grandes quantités de données recueillies les rendent inutiles et les bases de données deviennent des décharges de données. C'est là que l'analyse des données visuelles peut devenir utile**[12]**. La visualisation peut contribuer au processus d'exploration de données de deux façons. Tout d'abord, il peut fournir un affichage visuel des résultats des algorithmes de calcul complexes. Deuxièmement, il peut être utilisé pour découvrir des modèles complexes dans des données qui ne sont pas détectables par les méthodes de calcul actuelles, mais qui peuvent être identifiés par le système visuel humain. La première approche consiste à visualiser les résultats des algorithmes automatiques de data mining. La deuxième approche est le Visual data mining.

3. Outils de visualisation des données : Les outils de visualisation de données dépendent de la nature et la structure des données traitée. Et peuvent être classifiés dans deux catégories principales :

- Visualisations multidimensionnelles
- Visualisations hiérarchiques spécialisées et de paysage

3.1 Outils de la visualisation multidimensionnels :

Les outils de visualisation de données les plus utilisés généralement sont ceux qui représentent graphiquement les ensembles de données multidimensionnels. Les outils de visualisation de données multidimensionnelles permettent aux utilisateurs de comparer visuellement les dimensions des données (valeurs de colonne) avec d'autres dimensions de données à l'aide d'un système de coordonnées spatiales[10]. La figure 8 présente des exemples de types de graphes de visualisation les plus courants. Les autres types de graphes multidimensionnels courants qui ne sont pas représentés dans la figure 1 comprennent les histogrammes de contours, les erreurs, les graphiques de Westinghouse et de boîtes.

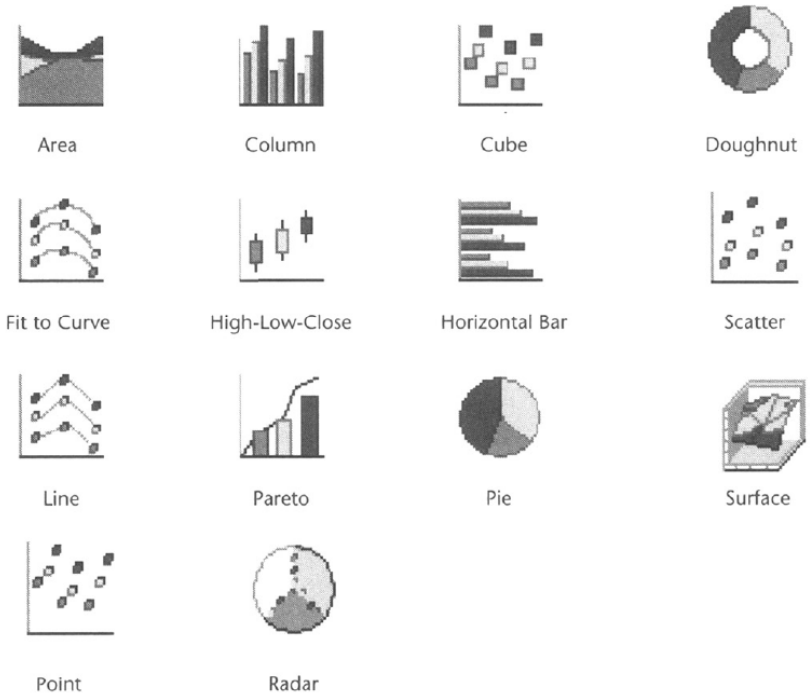


Figure 8 : Types de graphiques de visualisation de données multidimensionnelles.[10]

- **Colonne et barres analogiques :**

Comparent des dimensions de données continues à travers des dimensions de données discrètes dans un x et y-coordonnent le système. Les graphiques de colonne tracent des dimensions de données tout comme graphe linéaire, sauf qu'une colonne verticale est tirée de l'axe des abscisses à l'axe des ordonnées pour la valeur de la dimension de données.

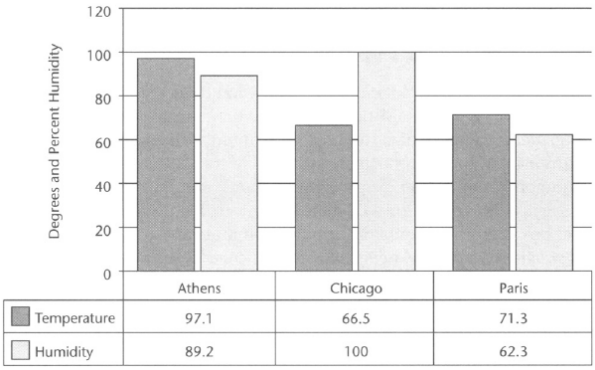


Figure 9 : Graphique de colonne comparant la température et l'humidité par ville.[10]

- **Graphiques de distribution et d'histogramme :**

Une technique analytique extrêmement utile est d'utiliser le bar de base et des graphiques de colonne pour montrer la distribution de valeurs pour une dimension de données (la colonne). La distribution et des graphiques d'histogramme montrent la proportion des valeurs pour des colonnes (numériques) (non-numériques) et continues discrètes comme le bar spécialisé et des graphiques de colonne[10], il est utilisé pour montrer des déséquilibres dans les données. Un histogramme, aussi mentionné comme un graphique de fréquence, trace le nombre de présence de même ou des valeurs distinctes dans l'ensemble de données. Ils sont aussi habitués à révéler des déséquilibres dans les données.

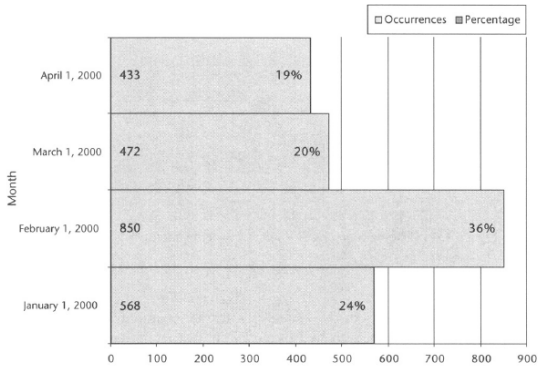


Figure 10 : Graphique de distribution

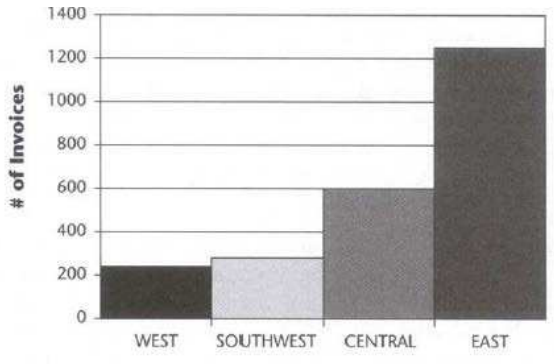


Figure 11 : histogramme

- **Graphiques de Boîte :**

Une variation sur le graphique d'histogramme est le graphique de complot de boîte. Il montre visuellement la statistique d'une colonne continue, En ajoutant les mesures de tendance centrale (comme moyen, médian et mode), les mesures de variabilité et les mesures de distribution.

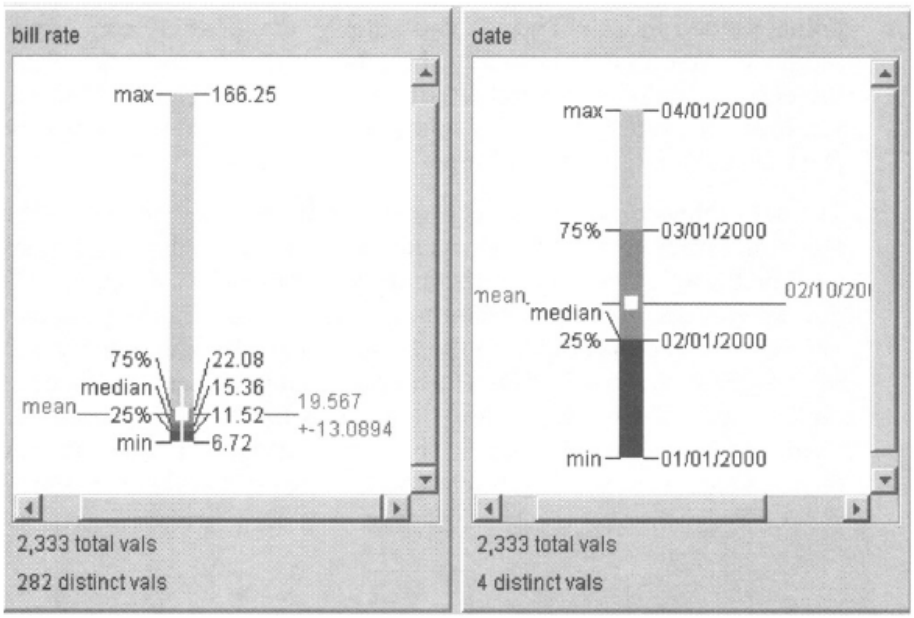


Figure 12 : Graphique en boîte[10]

- **Graphiques linéaires :**

En sa forme la plus simple, un graphique en courbe (le diagramme) est rien de plus qu'un ensemble de points de données tracés dans un x- et une coordonnée Y, connecté par segments de ligne. Les graphiques en courbe montrent comment les valeurs d'une colonne sont comparées à une autre colonne dans un système de coordonnée Y et un x-.

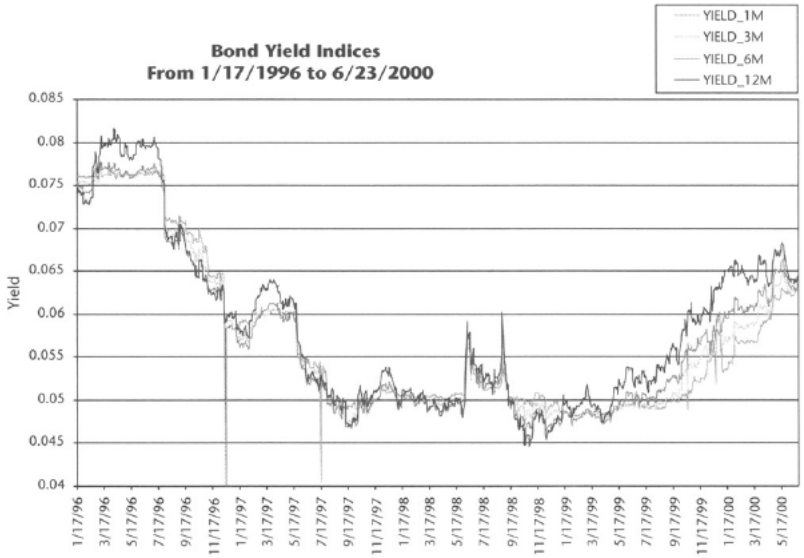


Figure 13 : graphe linéaire

3.2 Outils de visualisation de données hiérarchiques et paysagères :

Les outils de visualisation de données hiérarchiques, paysagères et autres outils spécialisés de visualisation de données sont diffèrent des outils multidimensionnels normaux en ce qu'ils exploitent ou améliorent la structure de soulignement de l'ensemble de données représenté.

Les visualisations d'arbres peuvent être utiles pour explorer les relations entre les niveaux hiérarchiques Les ensembles de données d'autre secteur ont une structure géographique ou spatiale inhérente. Par exemple, les ensembles de données qui contiennent des adresses ont une composante géographique de structure. La visualisation de carte peut être utile pour explorer les rapports géographiques dans l'ensemble de données.

- **Visualisations d'arbre [10]**

Le graphe arbre représente un ensemble de données sous la forme d'un arbre. Chaque niveau de l'arbre se divise en fonction des valeurs d'un attribut différent (hiérarchie dans l'ensemble de données). Chaque nœud de l'arbre affiche un graphe représentant toutes les données dans le sous-arbre ci-dessous. Le graphique arbre affiche les caractéristiques quantitatives et relationnelles d'un ensemble de données en les montrant comme nœuds connectés hiérarchiquement. Chaque nœud contient des informations généralement sous forme de barres ou de disques dont la hauteur et la couleur correspondent à des agrégations de valeurs de données (généralement des sommes, des moyennes ou des comptes). Les lignes (appelées arêtes) relient les nœuds ensemble et montrent la relation d'un ensemble de données à ses sous-ensembles.

La figure 14 illustre le nombre de familles sur Medicaid (un médicament pour les problèmes orthopédiques) à partir d'un ensemble de données en utilisant un graphique en arbre. Le nœud « racine » ou le début de l'arbre montre le nombre total de familles sur Medicaid (la petite colonne de couleur plus foncée à droite) et non sur Medicaid (la colonne plus grande et de couleur plus claire sur la gauche). On peut voir que le nombre de familles sur Medicaid est très faible.

Le deuxième niveau de l'arbre représente le nombre de familles sur Medicaid par les différents types de famille. En visualisant les données de cette façon, on peut être en mesure de trouver des attributs de combinaison et des valeurs qui sont indicatives des familles ayant une chance supérieure à la normale d'être sur Medicaid. Comme vous pouvez le voir à partir de la visualisation d'arbres, certains types de familles ont une chance significativement plus élevée d'être sur Medicaid que d'autres (sous-famille apparentée et les deuxièmes types de famille individuels contre les ménages non familiaux).

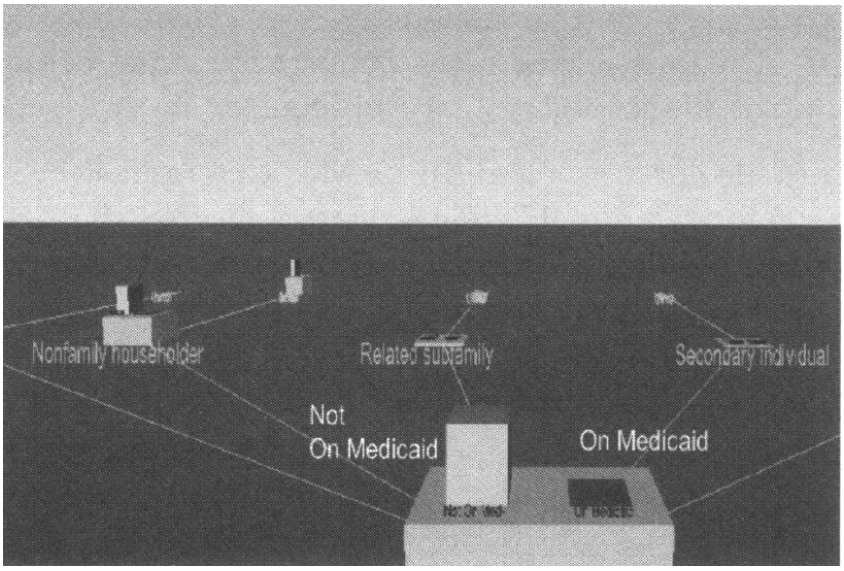


Figure 14 : Visualisation en arbre de la proportion de familles sur Medicaid .[10]

- **Visualisation sur carte** Pour explorer des ensembles de données métier pour des relations spatiales fortes (typiquement géographiques), on peut utiliser une visualisation sur carte. Les valeurs de colonne correspondantes sont affichées sous la forme d'éléments graphiques sur une carte visuelle basée sur une légende spatiale.

La figure 15 est une visualisation de carte d'un ensemble de données métier qui contient des informations sur le nombre d'enregistrements de nouveaux comptes par état. En utilisant une touche de couleur correspondante, les états sont colorés en fonction du nombre d'enregistrements par état. On peut rapidement déterminer à partir de la carte quels lieux de vente (états et régions) sont en train d'inscrire plus de nouveaux clients que d'autres. On peut également voir la signification géographique de l'état ou des régions les plus productrices par rapport à d'autres.[10]

Figure 15 : Visualisation de la carte des enregistrements de nouveaux comptes par état.

4. Couplage du spatial data mining et du Visual data mining

Les techniques de data mining spatial montrent leurs limites quand les bases de données sont trop volumineuses, certaines données restent enfuies à jamais dans les bases de données et beaucoup de modèles ne sont jamais découvert. Le visuel spatial data mining permet de découvrir visuellement certains de ces modèles et explorer ces données profondément enfuies.

Un inconvénient de l'exploration visuelle des données géospatiales est que la composante spatiale des données est difficile à visualiser efficacement. La représentation graphique la plus courante de la composante spatiale dans un système d'exploration de données visuelle est une carte. Cependant, étant donné que les données ne sont souvent pas uniformément réparties sur l'espace, certaines zones d'une telle carte pourraient être peu peuplées alors que dans d'autres régions un taux élevé de surimpression se produit. La surimpression pourrait également être causée par la taille de l'ensemble de données. Le fait que jusqu'à quatre dimensions fournissent un cadre de mesure pour toutes les autres dimensions implique également que la représentation graphique des phénomènes géospatiaux offre moins de degrés de liberté que ce qui est disponible pour la représentation graphique de données non spatiales arbitraires[13] .Lorsque de telles difficultés se produisent, il est impossible de s'appuyer uniquement sur la vision humaine tout en explorant les données. Une découverte efficace des connaissances à partir de données géospatiales est donc très probable si les avantages des méthodes d'exploration visuelle et informatique sont combinés. L'objectif de cette intégration est de construire des systèmes de découverte de connaissances visuellement activés qui pourraient faciliter le processus automatique de reconnaissance des modèles et des relations dans des données complexes et l'interprétation ultérieure des modèles et des relations découverts. De tels systèmes permettent à l'analyste d'explorer visuellement les données avec une manipulation directe des composants de visualisation d'informations et d'appliquer des outils de calcul quand quelque chose d'intéressant apparaît. Le data mining peut être utilisée comme un premier passage et les résultats peuvent ensuite être examinés visuellement [11, 14].

5. La visualisation revêt trois aspects :

La fouille visuelle des données peut être réalisée selon trois types d'approche permettant d'intégrer l'homme dans la boucle [15]:

Visualisation préalable (Preceding Visualization = PV) : Les données sont explorées visuellement avant la mise en oeuvre d'un algorithme. Cela permet de découvrir des motifs intéressants ;

Visualisation a posteriori (Subsequent Vizualisation = SV) : Un algorithme est préalablement utilisé afin d'extraire des motifs. Ceux-ci sont ensuite visualisés pour être analysés par l'utilisateur. En fonction de cette visualisation, l'utilisateur peut être amené à modifier le paramétrage de l'algorithme pour l'exécuter une nouvelle fois ;

Visualisation étroitement intégrée (Tightly Integrated Visualization = TIV) : Un algorithme analyse les données, mais ne donne pas le résultat final. Les résultats intermédiaires

sont cependant visualisés et permettent à l'utilisateur de détecter des motifs intéressants, en fonction de son domaine de connaissance. Comme un algorithme ne peut pas convenir à toutes les situations, son choix est réalisé par l'utilisateur, et les résultats sont ainsi adaptés à son domaine. Ce processus peut être réitéré jusqu'à obtention d'un résultat.

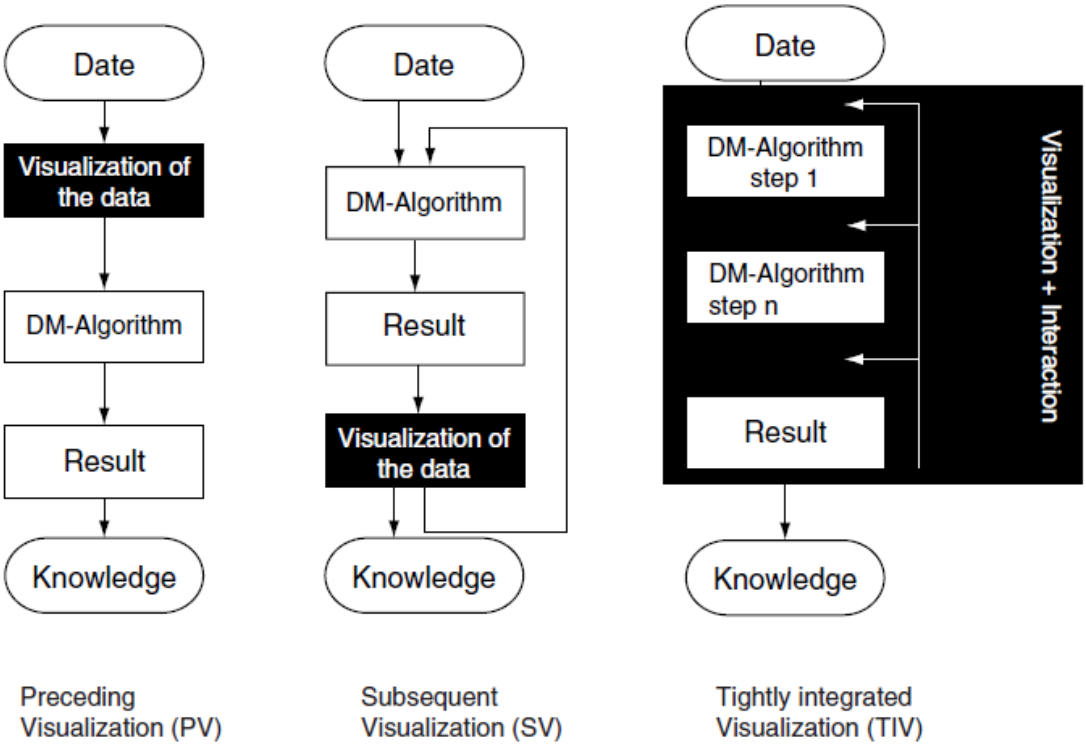


Figure 16: Trois types d'approche pour la fouille visuelle des données [15].

En plus du raisonnement direct de l'humain, les principaux avantages de l'exploration de données visuelles par rapport aux techniques automatiques du data mining sont les suivants :

- L'exploration de données visuelles peut facilement traiter des données hautement homogènes et bruyantes.
- L'exploration de données visuelles est intuitive et ne nécessite aucune compréhension d'algorithmes ou de paramètres complexes mathématiques ou statistiques.
- La visualisation peut fournir un aperçu qualitatif des données, permettant d'isoler les phénomènes de données pour une analyse quantitative supplémentaire.

Les techniques de l'exploration de données visuelles se sont révélées d'une grande valeur dans l'analyse des données exploratoires et ont un grand potentiel pour explorer de grandes bases de données. L'exploration de données visuelles est particulièrement utile lorsque peu de connaissances sur les données et les objectifs d'exploration sont vagues. Étant donné que l'analyste de données est directement impliqué dans le processus d'exploration, le déplacement et l'ajustement des objectifs d'exploration sont automatiquement effectués si nécessaire.

L'intégration de l'être humain dans le processus d'exploration de données et l'application des capacités humaines pour l'analyse des grands ensembles de données peuvent aider à fournir des

résultats plus efficaces dans des domaines importants d'application de données, pour les règles d'association, le clustering, la classification et la récupération de texte[16].

6. Visualisation :

6.1. Règles d'associations :

L'objectif de la génération de règles d'association est de trouver des modèles et des tendances intéressants dans les bases de données. Les règles d'association sont des relations statistiques entre deux ou plusieurs éléments dans l'ensemble de données.

Les règles de l'association nous disent que la présence de certains éléments dans une transaction implique la présence d'autres éléments dans la même transaction avec une certaine probabilité, appelée ***confiance***. Un deuxième paramètre important est ***le support*** d'une règle d'association, qui est définie comme un indicateur la de fiabilité de la règle.

Cependant, les règles d'association sont généralement très faibles au niveau du support et de confiance. L'utilisation d'un niveau plus élevé de support et de confiance peut ne pas être efficace, car des règles utiles peuvent alors être négligées.

Des techniques de visualisation ont été utilisées pour surmonter ce problème et permettre une sélection interactive de bons niveaux de support et de confiance. l'une de ces technique SGI MineSets Rule Visualizer [17] qui mappe les côtés gauche et droit des règles aux axes x et y de l'intrigue et montre la confiance que la hauteur des barres et le support à la hauteur de la hauteur Disques. La couleur des barres montre l'intérêt de la règle. Comme le montre la figure 17

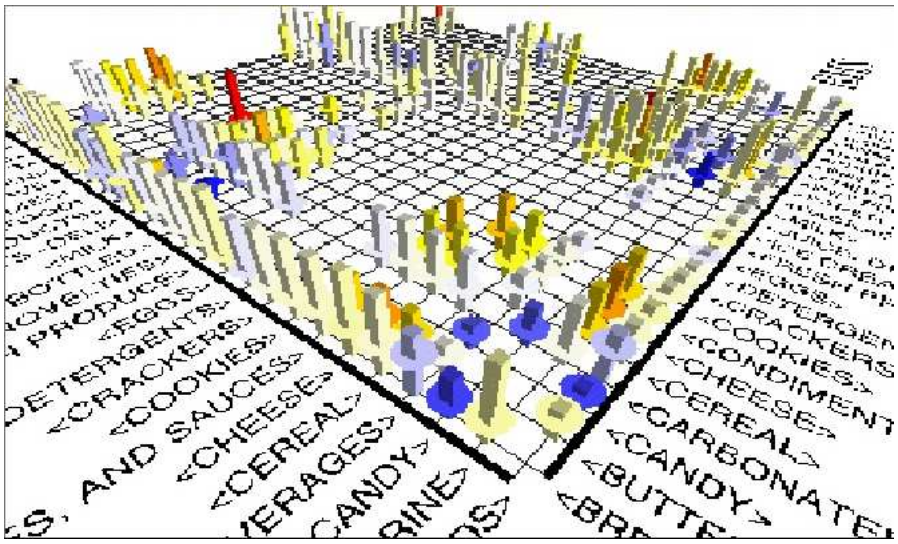


Figure 17 : visualisation des règles d'association[18]

6.2. Classification :

La classification est le processus d'élaboration d'un modèle basé sur un ensemble de données de classe connues. Pour construire le modèle de classification, les attributs de l'ensemble de données de formation sont analysés et une description ou un modèle précis des classes en fonction des attributs disponibles dans l'ensemble de données. La classification est parfois appelée apprentissage supervisé, car l'ensemble de formation est utilisé pour enseigner au système comment classer les données. Il existe de nombreux algorithmes pour résoudre les discussions sur la classification.

Les approches les plus populaires sont les algorithmes qui construisent méthodiquement, des arbres de décision. En outre, il existe des approches qui utilisent des réseaux de neurones, des algorithmes génétiques ou des réseaux bayésiens pour résoudre le problème de classification. Étant donné que la plupart des algorithmes fonctionnent comme des approches, il est souvent difficile de comprendre et d'optimiser le modèle de décision. Des problèmes tels que l'ajustement excessif ou l'élagage des arbres sont difficiles à aborder.

Les techniques de visualisation peuvent aider à surmonter ces problèmes. L'analyste d'arbre de décision dans le système MineSet de SGI [17]**la figure 18** présente un aperçu de l'arbre de décision avec des paramètres importants tels que les distributions de valeurs d'attributs. Le système permet une sélection interactive des attributs affichés et aide l'utilisateur à comprendre l'arbre de décision. Une approche plus sophistiquée qui contribue également à la construction d'arbres de décision est la classification visuelle, proposée par Ankerst et al. [18]. L'idée de base est de montrer chaque valeur d'attribut par un pixel coloré et de les répartir en barres. Les pixels de chaque barre d'attributs sont triés séparément et l'attribut avec la distribution de la valeur la plus pure est sélectionné comme l'attribut fractionné de l'arbre de décision.

La procédure est répétée jusqu'à ce que toutes les feuilles correspondent aux classes pures. Un exemple de l'arbre de décision résultant de ce processus est illustré à la figure18 Par rapport à une visualisation standard d'un arbre de décision, des informations supplémentaires sont utiles pour expliquer et analyser l'arbre de décision, à savoir :

- Taille des nœuds (nombre d'enregistrements d'entraînement correspondant au nœud)
- Qualité de la division (pureté des partitions résultantes)
- Distribution de classe (fréquence et emplacement des instances de formation de toutes les classes).

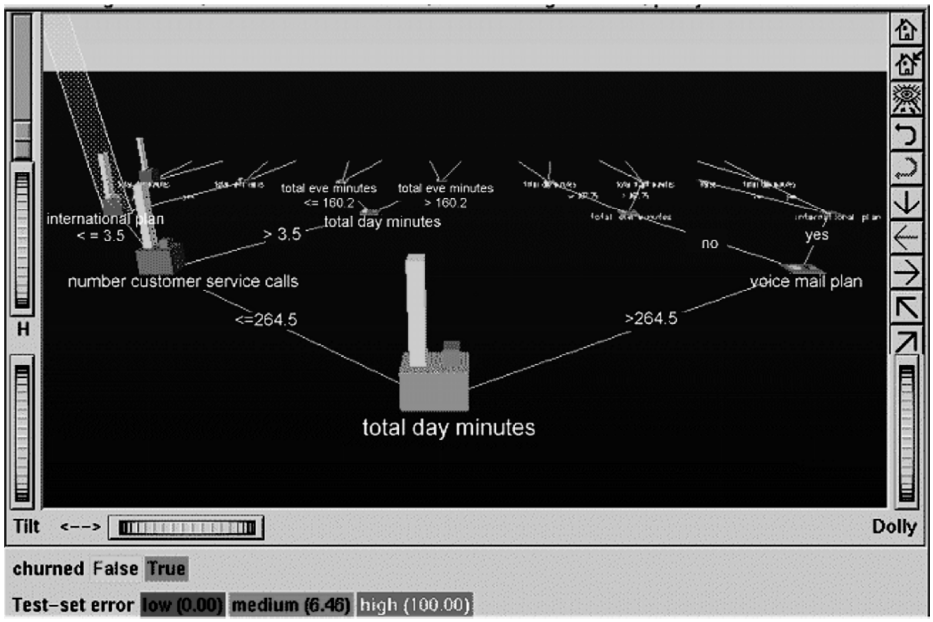


Figure 18 :visualisation de la classification [17]

6.3. Clustering :

Le clustering est le processus consistant à trouver un partitionnement de l'ensemble de données en sous-ensembles homogènes appelés clusters. Contrairement à la classification, le regroupement est un apprentissage non supervisé. Cela signifie que les classes sont inconnues et qu'aucun jeu d'entrainement avec les étiquettes de classe n'est disponible.

La plupart des algorithmes utilisent des hypothèses sur les propriétés des clusters qui sont soit utilisées comme valeurs par défaut, soit doivent être données en tant que paramètres d'entrée.

L'utilisateur obtient des résultats de clustering différents. Dans l'espace 2D ou 3D, l'impact de différents algorithmes et paramètres peut être facilement exploré à l'aide de simples visualisations des grappes. Certaines techniques de dimension supérieure tentent de déterminer les projections 2D ou 3D des données qui conservent autant que possible les propriétés des clusters à grande dimension [19]. Figue 19 avec une projection 3D d'un jeu de données composé de cinq grappes.

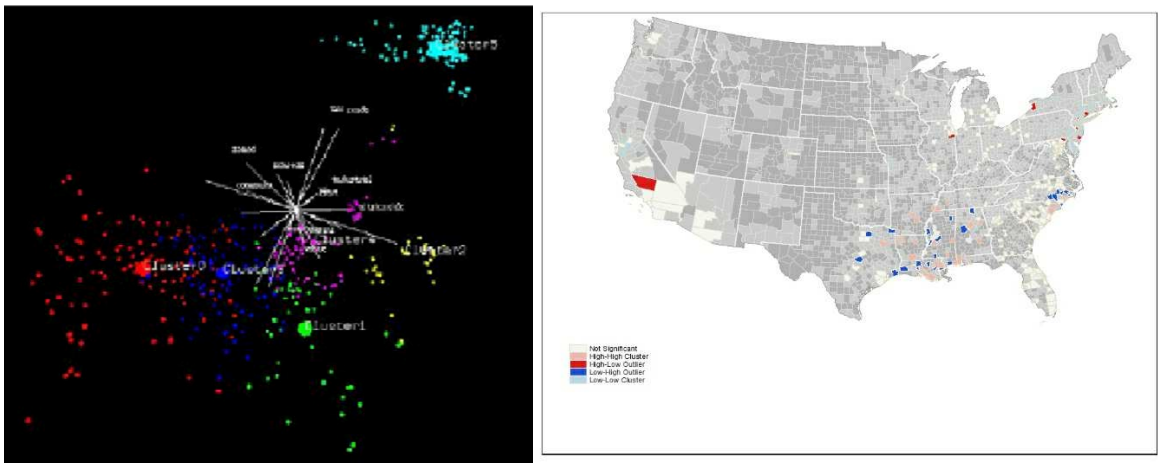


Figure 19 : Visualisation de clustering 3D , 2D

7. Les point fort du VSDM :

Si les avantages des méthodes du data mining visuel et spatial sont combinés, l'extraction des connaissances à partir des données géospatiales serait plus efficace, l'objectif de cette intégration est de construire des systèmes de découverte de connaissances visuellement activée qui pourraient :

- ✓ Faciliter le processus automatique de reconnaissances des modèles et des relations spatiales enfuies dans les bases de données spatiales complexes
- ✓ Faciliter l'interprétation des modèles et des relations découverts.
- ✓ Faciliter à l'analyste d'explorer et de manipuler visuellement les données et d'appliquer des outils de calculs quand quelque chose d'intéressant apparaît.

Partie II : VSDM et épidémiologie

L'utilisation des méthodes du Data mining en épidémiologie et en santé publique est en forte croissance. Comme dans d'autres domaines, c'est la disponibilité de vastes bases de données historiques qui incite à les mettre en valeur, pour les bases de données spatial le data mining classique ne prend pas en compte la composante spatiale qui restent enfuîtes dans ce cas on utilise le VSDM pour une découverte satisfaisante

8. L'épidémiologie :

De nombreuses définitions ont été proposées, mais la définition suivante reprend les principes sous-jacents et l'esprit d'épidémiologie de la santé publique :

« L'épidémiologie est l'étude de la répartition et des déterminants des états ou événements liés à la santé dans des populations spécifiques et l'application de cette étude au contrôle des problèmes de santé»[20]

Les différentes branches de l'épidémiologie se caractérisent par la nature des questions auxquelles il s'agit d'apporter des réponses, ainsi que par les méthodes utilisées à cet effet. L'épidémiologie a été conçue pour répondre à la question : « Qui a quoi, quand, où et pourquoi? »[21]

Classiquement, on distingue :

8.1 L'épidémiologie descriptive[21]

Il s'agit de la partie de l'épidémiologie qui a pour objet de décrire la fréquence et la répartition de phénomènes de santé ou de déterminants de santé dans les populations, en fonction de caractéristiques humaines, spatiales, temporelles. Il s'agit donc d'apporter des réponses pour les questions suivantes :

Chez qui ? ⇔ Personnes

Où ? ⇔ Lieu

Quand ? ⇔ Temps

Le principe général de l'épidémiologie descriptive est basé sur l'utilisation d'indicateurs simples : par exemple, on calcule des taux de mortalité lorsque l'on s'intéresse aux décès, des taux de prévalence et d'incidence lorsque l'on s'intéresse aux maladies. Ces taux correspondent aux nombres de décès (ou malades) ramenés à la population qu'on étudie.

Les données utilisées par l'épidémiologie descriptive peuvent être issues des statistiques sanitaires (statistiques de mortalité, registres des cancers, ou registres pour d'autres maladies, données issues des déclarations obligatoires pour les maladies transmissibles) ou d'études qui ont été réalisées dans le but de fournir une information spécifique adaptée à un objectif particulier, ou lorsque l'information est défaillante (enquêtes ad hoc).

8.2 L'épidémiologie analytique

Il s'agit de la partie de l'épidémiologie dont l'objet est de mettre en évidence et estimer le lien entre l'exposition à certains facteurs et la survenue ultérieure de maladie (ou événement de santé) au moyen d'enquêtes réalisées chez des individus. La question à laquelle on veut répondre ici est « pourquoi ?»

Les études d'observations mises en place (cas-témoins, ou exposés-non exposés) permettent de comparer les groupes d'individus définis en fonction de la maladie (malade ou non malade) et d'un facteur d'exposition (exposé ou non exposé à ce facteur), en estimant un risque (= probabilité) associé. Lorsque cette relation est établie, on peut ainsi déterminer par combien est multiplié (ou divisé) la probabilité de survenue de la maladie chez les sujets exposés au facteur par rapport aux sujets non exposés à ce facteur.

Selon le champ d'application considéré, on parle d'épidémiologie étiologique (on s'intéresse à l'étude des causes des maladies), d'épidémiologie génétique, d'épidémiologie sociale, d'épidémiologie environnementale.[21]

8.3 L'épidémiologie évaluative

• Démontrer l'efficacité de l'intervention qui est exprimée sous forme d'un état de santé => Évaluation de recherche (ce sont les expériences)

• Vérifier l'efficacité de l'intervention telle qu'elle a été mise en en place dans la pratique habituelle => Évaluation professionnelle (ce sont les études d'observation évaluatives qui utilisent les méthodes de l'épidémiologie descriptive)[21]

9. VSDM en épidémiologie :

Dans le monde du VSDM Data Mining appliqué aux épidémiologies, la première utilisation de ce modèle est sans doute celle de Dr John Snow pour la découvre de la cause probable des cas d'épidémie de choléra qui est survenue en 1854 près de *Broad Street (Londres)*.

Il cartographie alors les cas mortels et de croiser ses données avec les points ou les Londoniens peuvent aller s'approvisionner en eau comme le montre la figure suivant

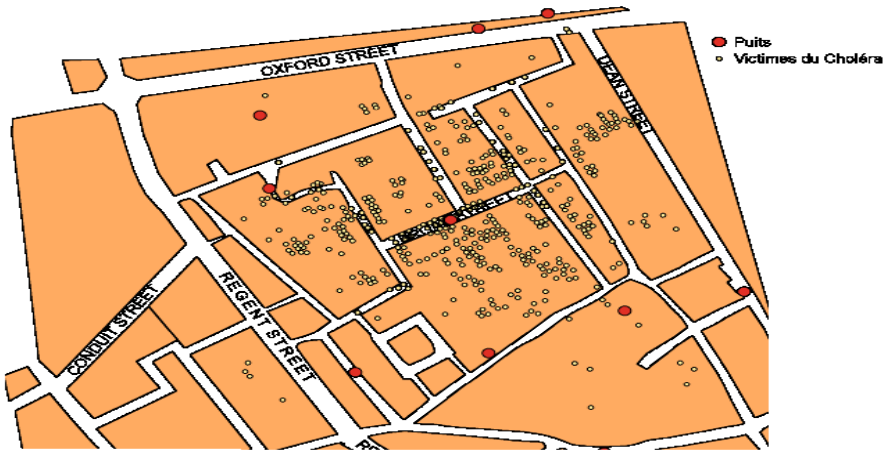


Figure 20: La carte réalisée par le Dr John Snow pour la découvre de la cause probable des cas d'épidémie [22]

La visualisation de cette carte lui permet de trouver les puits infectés, il finit par convaincre les autorités locales de retirer la poignée de la pompe.

Conclusion :

Dans ce chapitre, on as étudier le visuel spatial data mining qui être très performante en matière des précision et rapidité pour l'extraction des connaissances dans les bases de données spatiales.

Le chapitre suivant fera l'objet d'une conception et implémentation d'une application mettant en œuvre les concepts acquis dans les deux premiers chapitres.

Chapitre 3 conception et implémentation

Introduction :

Dans ce chapitre, nous allons aborder les outils et l'algorithme que nous avons choisi pour la modélisation de notre solution. Nous présentons une interface d'interaction intégrée au modèle VSDM.
Le but de notre travail, et d'appliquer les méthodes du visuel spatial data mining sur l'épidémiologie et plus précisément sur le cancer du sein aux Etats Unis d'Amérique.

Comme méthode de VSDM nous avons choisi une approche basée sur les statistiques spatiales. Nous avons d'abord appliqué le clustering pour faire une analyse exploratoire globale et dégager les tendances générales. Puis nous avons effectué une analyse locale basées sur la découverte des Hotspots (points chauds).

Problématique :

C'est un problème qui a été signalé dans un article du mois de décembre 2013 dans le *New York Times*. L'article concerne les taux de mortalité par cancer du sein, il dit que le taux est plus élevé chez les femmes de race noire que chez les femmes de race blanche. Le titre en était "Combattre l'écart racial en matière de survie face au cancer du sein". L'essentiel de l'article a été résumé dans ce paragraphe :

"Les chercheurs... ont découvert que les femmes afro-américaines atteintes du cancer du sein avaient, en moyenne, 40 pour cent de chances en plus de mourir de leur maladie que les femmes de race blanche. Aux Etats-Unis, la disparité des taux de survie chez les patientes souffrant du cancer du sein se traduit par environ 1 700 décès supplémentaires chaque année, ou il y a environ cinq fois plus de femmes de race noire qui meurent chaque jour."

Approche proposée :
Il existe deux approches pour l'analyse et l'extraction de connaissances d'une base de données spatiales. La première est issue des statistiques spatiales et la seconde du domaine des bases de données. On a opté pour l'approche statistique, et on va utiliser les Hotspots (points chauds) et les outlier (valeurs aberrantes) qui sont basés sur les statistiques spatiales « Getis_ord Gi » pour l'analyse globale getis_ord G* pour l'analyse locale.

1. **Statistique Spatial :**

Les statistiques spatiales comportent un ensemble de techniques pour décrire et modeler des données spatiales. De plusieurs manières, elles se prolongent ce que l'esprit et les yeux font, intuitivement, pour évaluer les modèles spatiaux, les distributions, les tendances, les processus et les rapports.
À la différence des techniques statistiques (non-spatial) traditionnelles, les techniques statistiques spatiales emploient réellement l'espace - secteur, longueur, proximité, orientation, ou rapports spatiaux directement dans leurs mathématiques [23].
Comme nous l'avons déjà cité dans le chapitre 1 sur les méthodes de SDM ,les statistiques spatiales concernent :

1.1.Une analyse Globale :

L'analyse global contient des méthodes qui conviennent le mieux à la compréhension des grands schémas et tendances spatiales [24].Avec ces outils, on peut répondre à des questions comme :

Quelles espèces végétales sont les plus concentrées ?

Le modèle spatial de la maladie reflète-t-il le profil spatial de la population ?

Existe-t-il un pic inattendu dans les achats de produits pharmaceutiques ?

.Les outils qu'on retrouve dans l'analyse global sont décrite dans le tableau suivant :

Outils	Description
Distances moyennes du plus proche voisin	Calcule la distance moyenne de chaque caractéristique à son voisin le plus proche en fonction des centres de base
High/low clustering (Getis-Ord general G)	Mesure les concentrations de valeurs élevées ou faibles pour une zone d'étude
Autocorrelation spatial (global Moran's *I*)	Mesure l'autocorrélation spatiale (clustering ou dispersion) en fonction des emplacements et des valeurs des attributs
Analyse spatiale multi-distance (f Ripley's *K* function)	Évalue le Clustering / la dispersion spatiale pour un ensemble de caractéristiques géographiques sur une gamme de distances

Tableau 2 : les outils de l'analyse globale

1.2. L'analyse locale :

Les outils décrits ci-dessus dans le jeu d'outils de l'analyse locale sont des statistiques qui répondent à la question : existe-t-il un regroupement ou une dispersion spatiale statistiquement significatif ? Ces outils d'autre part, identifient où se déroule le regroupement spatial et où se situent les valeurs aberrantes spatiales[23]

Où sont leurs limites nettes entre l'affluence et la pauvreté en Équateur?

Où trouvons-nous des habitudes de dépenses anormales à Los Angeles?

Où voyons-nous des taux de diabète inattendus élevés?

Outils	Description
Cluster et outlier analysis (Anselin's local Moran's *I*)	Compte tenu d'un ensemble de caractéristiques pondérées, identifie des clusters de valeurs élevées ou faibles ainsi que des valeurs aberrantes spatiales
Hot spot analysis (Getis-Ord *i* $G*$)	Compte tenu d'un ensemble de fonctionnalités pondérées, identifie des clusters de fonctionnalités avec des valeurs élevées (points chauds) et des clusters de fonctionnalités à faible valeur (points froids)

Tableau 3 : Les outils de l'analyse locale

- **Hot Spot :**

Dans un ensemble d'entités, cet outil identifie les points chauds et les points froids statistiquement significatifs à l'aide de la statistique Getis-Ord Gi*

L'outil Analyse de points chauds permet de calculer les statistiques Getis-Ord Gi* de chaque entité d'un jeu de données. Les scores z et valeurs p obtenus indiquent l'endroit où les entités de valeurs élevées ou faibles sont agrégées spatialement. Cet outil fonctionne en examinant chaque entité dans le contexte des entités voisines. Une entité dotée d'une valeur élevée est intéressante, mais il ne s'agit pas forcément d'un point chaud statistiquement significatif. Pour être un point chaud, une entité doit avoir une valeur élevée et être entourée d'autres entités également dotées de valeurs élevées. La somme locale d'une entité et de ses voisins est comparée proportionnellement à la somme de toutes les entités. Lorsque la somme locale est très différente de la somme locale attendue, et si la différence est trop importante pour n'être que le fruit du hasard, un score z est généré. Lorsque la correction FDR(Taux de découverte fausse) est appliquée, la signification statistique est ajustée pour prendre en compte les tests multiples et la dépendance spatiale.

✓ **Exemple de script python utilise dans ArcGIS :**

```
# Perform Hot Spot Analysis for assault incidents

# Import system modules

import arcgisscripting

# Create the Geoprocessor object

gp = arcgisscripting.create()

# Local variables...

workspace = "C:/project93/data"

input = "assaults.shp"

collect_output = "collect_output.shp"

collect_count_field = "Count"

hotspot_output = "hotspot_output.shp"

hotspot_output_rendered = "hotspot_output_rendered.lyr"

z_score_field_name = "GiInvDst"

try:

    # Set the current workspace (to avoid having to specify the full path to the feature
classes each time)

    gp.workspace = workspace
```

```
    # Convert assault incidents into weighted point data

    # Process: Collect Events...

    gp.CollectEvents_stats(input, collect_output)

    # Calculate Getis-Ord Gi* statistic

    # Process: Hot Spot Analysis (Getis-Ord Gi*)...

    gp.HotSpots_stats(collect_output, collect_count_field, hotspot_output, "Inverse
Distance", "Euclidean Distance", "None", "#", "#", "#")

    # Render hot spot analysis

    # Process: Z Score Rendering...

    gp.ZRenderer_stats(hotspot_output, z_score_field_name,
hotspot_output_rendered)

except:

    # If an error occurred when running the tool, print out the error message.

    print gp.GetMessages(2)
```

La statistic local Getis-Ord Gi* est donnée comme :

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}}$$

Où x_j est la valeur d'attribut pour la caractéristique j

W est le poids spatial entre la caractéristique i et j

N est égal au nombre total de caractéristiques et :

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\bar{X}\right)^2} \qquad \bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}$$

La statistique G_i^* est un score Z, donc aucun autre calcul n'est nécessaire

- **Hight/Low Clustering**

L'outil High/Low Clustering (Getis-Ord General G) génère une statistique inférentielle, ce qui signifie que les résultats de l'analyse sont interprétés dans le contexte de l'hypothèse nulle. L'hypothèse nulle de cette statistique suppose qu'il n'y a aucune agrégation spatiale de valeurs d'entité. le signe du score z devient important. Si la valeur du score z est positive, l'indice General G observé est plus grand que l'indice General G attendu, indiquant par là que les valeurs élevées pour l'attribut sont agrégées dans la zone d'étude. Si la valeur de score z est négative, l'indice General G observé est plus petit que l'indice attendu, indiquant ainsi que les valeurs faibles sont agrégées dans la zone d'étude. [26]

L'outil High/Low Clustering (Getis-Ord General G) renvoie quatre valeurs : General G observé, General G attendu, score z et valeur de p.

La statistique Getis-Ord General G est calculer comme suit :

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}, \ \forall j \neq i$$

Où x_i et x_j sont des valeurs d'attribut pour les caractéristiques i et j et $w_{i,j}$ est le poids spatial entre la caractéristique i et j

n est le nombre de fonctionnalités dans l'ensemble de données et j = / indique que les fonctionnalités i et j ne peuvent pas être les mêmes caractéristiques quel que soit j =/= i

Le score ZG pour la statistique est calculé comme suit

$$z_G = \frac{G - \mathrm{E}[G]}{\sqrt{\mathrm{V}[G]}}$$

$$\mathrm{E}[G] = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}}{n(n-1)}, \ \forall j \neq i$$

$$\mathrm{V}[G] = \mathrm{E}[G^2] - \mathrm{E}[G]^2$$

✓ **Exemple de script python utilisé dans ArcGIS :[26]**

```
# Analyze the spatial distribution of 911 calls in a metropolitan
area

# using the High/Low Clustering (Getis-Ord General G)

# Import system modules

import arcgisscripting

# Create the Geoprocessor object
```

```
gp = arcgisscripting.create(9.3)

gp.OverwriteOutput = 1

# Local variables...

workspace = "C:\Data\911Calls"

try:

# Set the current workspace (to avoid having to specify the full
path to the feature classes each time)

    gp.workspace = workspace

    # Copy the input feature class and integrate the points to snap

    # together at 500 feet

    # Process: Copy Features and Integrate

    cf = gp.CopyFeatures("911Calls.shp", "911Copied.shp",

                         "#", 0, 0, 0)

    integrate = gp.Integrate("911Copied.shp #", "500 Feet")

    # Use Collect Events to count the number of calls at each
location

    # Process: Collect Events

    ce = gp.CollectEvents("911Copied.shp", "911Count.shp", "Count",
"#")

    # Add a unique ID field to the count feature class

    # Process: Add Field and Calculate Field

    af = gp.AddField("911Count.shp", "MyID", "LONG", "#", "#", "#",
"#",

                     "NON_NULLABLE", "NON_REQUIRED", "#",

                     "911Count.shp")

    cf = gp.CalculateField("911Count.shp", "MyID", "[FID]", "VB")

    # Create Spatial Weights Matrix for Calculations

    # Process: Generate Spatial Weights Matrix...

    swm = gp.GenerateSpatialWeightsMatrix("911Count.shp", "MYID",

                       "euclidean6Neighs.swm",

                       "K_NEAREST_NEIGHBORS",

                       "#", "#", "#", 6,

                       "NO_STANDARDIZATION")
```

```
    # Cluster Analysis of 911 Calls

    # Process: High/Low Clustering (Getis-Ord General G)

    hs = gp.HighLowClustering("911Count.shp", "ICOUNT",

                          "false",

                          "Get Spatial Weights From File",

                          "Euclidean Distance", "None",

                          "#", "euclidean6Neighs.swm")
except:

    # If an error occurred when running the tool, print out the
error message.

print gp.GetMessages()
```

2. Conception :

2.1 Démarche de l'application :

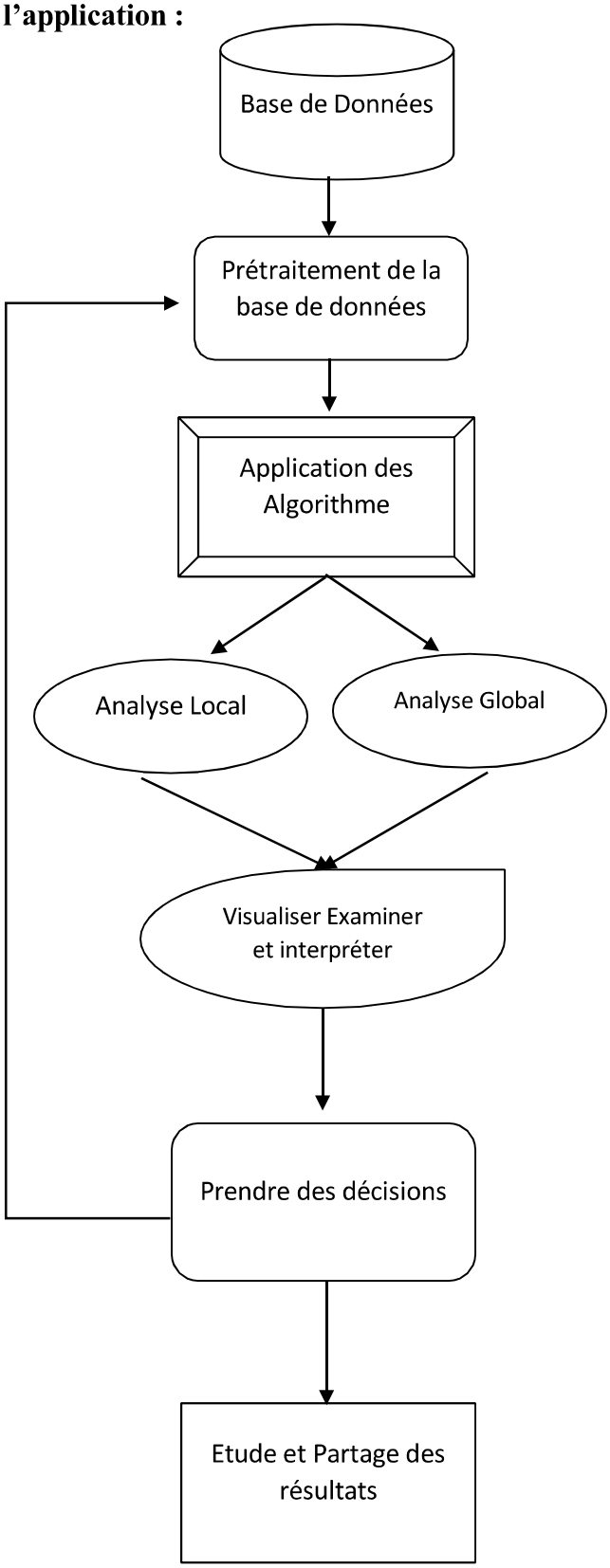


Figure 21 : Architecture de l'application

La figure 21 illustre l'architecture de notre application, détaillée comme suit :

- **Base de données :**

Afin d'atteindre notre objectif, nous nous sommes intéressés à une base de données médicale du Cancer du sein concernant la population des état unis, que nous avons ramené de L'institut national du cancer qui met à disposition des statistiques sur le cancer par l'intermédiaire de son programme « Surveillance, Epidemiology, and End Results (SEER) »

Dans le cadre de notre travaille, les données sur les taux de mortalité par cancer ont été téléchargées afin de nous aider à mieux comprendre les tendances des taux de mortalité par cancer.
Cette base de données contient les taux de mortalité moyens pour les années 2006 à 2010 par cancer du sein chez les femmes de race noire et de race blanche de tous les âges, par comté, et contient aussi les facteur de risque parmi ces facteurs, on peut noter l'obésité, l'inactivité physique et l'abus d'alcool.

- **Prétraitement de la base de données :**

Cette phase consiste à exploiter et nettoyer les données, et les attributs inutiles à l'étude afin exécuter les processus du data mining.

- **Application de l'algorithme :**

On va travailler avec les statistiques spatiales

- Pour l'analyse local on va utiliser : l'analyse des points chaud (Getis-Ord $i\ G*$)
- Pour l'analyse global on va utiliser : Clustering (Getis-Ord i G)

- **Interpréter les résultats :**

Au cours de la tache précèdent, nous avons effectué l'analyse. Cela nous amène à l'interprétation des résultats afin de comprendre les modèles inhabituels ou intéressants. Découvrir si les résultats ont un sens en recherchant une signification sous-jacente, en assemblant les informations cartographiées et en élaborant une vue d'ensemble utile de l'environnement. Évaluez si les résultats offrent une réponse ou une explication satisfaisante.

- **Prendre des décisions :**

Après avoir visualiser, examiner et interpréter les résultats l'utilisateur doit soit valider l'analyse et sauvegarder les résultats soit refait l'analyse avec de nouveaux paramètre**s.**

- **Étude partage des résultats** :

Cette tache se fera dans un autre environnement pour comprendre davantage sur l'étude et avoir un avis extérieur.

3. Implémentation :

3.1 Les outils utilisés :

- **ArcGIS Desktop 10.4.1 :**

ArcGIS est un système complet qui permet de collecter, organiser, gérer, analyser, communiquer et diffuser des informations géographiques. En tant que principale plateforme de développement et d'utilisation des systèmes d'information géographiques (SIG) au monde, ArcGIS est utilisé par des personnes du monde entier pour mettre les connaissances géographiques au service du gouvernement, des entreprises, de la science, de l'éducation et des médias. Ce qui nous intéresse c'est :

- **ArcMap :** c'est l'application principale utilisée dans ArcGIS Desktop pour la cartographie, la mise à jour, l'analyse et la gestion des données.
- **ArcCatalog :** permets d'organiser et gérer les différents types d'informations géographiques.
- **ArcToolbox :** regroupe un ensemble d'outils de conversion de données de gestion des projections d'analyse, de géo traitement, etc.

- **Visual studio 2015 :**

Visual Studio entreprise 2015 est une suite de logiciels de développement pour Windows conçue par Microsoft. Visual Studio est un ensemble complet d'outils de développement permettant de générer des applications web ASP.NET, des services web XML, des applications bureautiques et des applications mobiles. Visual Basic, Visual C++, Visual C# utilisent tout le même environnement de développement intégré (IDE), qui leur permet de partager des outils et facilite la création de solutions faisant appel à plusieurs langages. [25]

- **ArcObjects, 10.4.1** :

Il permet d'intègre les bibliothèques ArcGIS avec des langages de programmation comme Visual Basic, C #, Visual Basic.NET, Java et Python

3.2 Mise en œuvre de l'application :

Cette mise en œuvre reprend les étapes décrites dans l'architecture du système de **la figure22**

3.2.1 Prétraitement de la base de données : dans le but d'exploiter notre base de données nous avions besoin de la nettoyer et de la traiter.

Table

Données_sur_le_cancer_du_sein

OBJECTID *	County	State	Population 2	Taux de mortalité femme noire	geme de mortalité femme noir	Taux de mortalité femme blanche	Period	Game de mortalité femme blanche	Ratio de mortalité	obésité	inactivité physique	% consomm
1961	Pender	North Carolina	52217	55,7	37.4 - 55.7	14,8	2006 - 2010	10.5 - 19.0	3,763514	32,2	24,5	<Null>
1446	Madison	Mississippi	95203	53,2	37.4 - 55.7	38	2006 - 2010	26.9 - 47.8	1,4	28,7	25,9	
2615	Gregg	Texas	121730	51,1	37.4 - 55.7	23,2	2006 - 2010	22.5 - 23.9	2,202586	32,3	30,2	
129	Crittenden	Arkansas	50902	50,2	37.4 - 55.7	20,9	2006 - 2010	19.1 - 21.0	2,401914	37,5	33,3	
2055	Clark	Ohio	138333	50	37.4 - 55.7	25,2	2006 - 2010	24.0 - 26.8	1,984127	32,1	30,3	
28	Etowah	Alabama	104430	49,1	37.4 - 55.7	21,4	2006 - 2010	21.1 - 22.4	2,294393	32,1	34,2	
1988	Wilson	North Carolina	81234	48,2	37.4 - 55.7	25,1	2006 - 2010	24.0 - 26.8	1,920319	32,2	31,6	<Null>
1458	Pike	Mississippi	40404	48	37.4 - 55.7	34	2006 - 2010	26.9 - 47.8	1,411765	39,4	35,9	
2678	McLennan	Texas	234906	46,3	37.4 - 55.7	25,4	2006 - 2010	24.0 - 26.8	1,822835	30,1	25,1	
1203	Frederick	Maryland	233385	46	37.4 - 55.7	21,8	2006 - 2010	21.1 - 22.4	2,110092	26,6	21,9	
2269	Erie	Pennsylvania	280566	45,6	37.4 - 55.7	23,6	2006 - 2010	22.5 - 23.9	1,932203	29,4	29,4	
2905	Spotsylvania	Virginia	122397	44,5	37.4 - 55.7	22,6	2006 - 2010	22.5 - 23.9	1,969027	28,4	19,9	
1968	Robeson	North Carolina	134168	44,2	37.4 - 55.7	23,7	2006 - 2010	22.5 - 23.9	1,864979	41,1	38,7	
2944	Portsmouth	Virginia	95535	43,9	37.4 - 55.7	21,3	2006 - 2010	21.1 - 22.4	2,061033	39,2	29,2	
1116	Ascension	Louisiana	107215	43,7	37.4 - 55.7	18,6	2006 - 2010	10.5 - 19.0	2,349462	32	28,2	
2607	Galveston	Texas	291309	42,9	37.4 - 55.7	22,5	2006 - 2010	22.5 - 23.9	1,906667	29,6	24,9	
2906	Stafford	Virginia	128961	42,9	37.4 - 55.7	26,1	2006 - 2010	24.0 - 26.8	1,643678	29,9	19,8	
41	Lee	Alabama	140247	42,5	37.4 - 55.7	29,5	2006 - 2010	26.9 - 47.8	1,440678	30	27,4	
641	Kankakee	Illinois	113449	42,4	37.4 - 55.7	26,5	2006 - 2010	24.0 - 26.8	1,6	31,1	28,5	
1150	Ouachita	Louisiana	153720	42,1	37.4 - 55.7	19,4	2006 - 2010	19.1 - 21.0	2,170103	31,9	29,7	
1442	Lee	Mississippi	82910	41,9	37.4 - 55.7	32,8	2006 - 2010	26.9 - 47.8	1,277439	33,6	30,9	
1270	Kalamazoo	Michigan	250331	41,7	37.4 - 55.7	26,2	2006 - 2010	24.0 - 26.8	1,591603	28,5	22,3	
2507	Shelby	Tennessee	927644	41,7	37.4 - 55.7	21	2006 - 2010	19.1 - 21.0	1,985714	34,1	28,5	
1843	Erie	New York	919040	41,3	37.4 - 55.7	25,6	2006 - 2010	24.0 - 26.8	1,613281	28,6	26,1	
1775	Atlantic	New Jersey	274549	41	37.4 - 55.7	25,6	2006 - 2010	24.0 - 26.8	1,601563	28,2	25,3	
516	Sumter	Georgia	32819	40,8	37.4 - 55.7	30,1	2006 - 2010	26.9 - 47.8	1,355482	34,7	30,2	
1173	Webster	Louisiana	41207	40,6	37.4 - 55.7	22,2	2006 - 2010	21.1 - 22.4	1,828829	35,6	31,8	
1304	Saginaw	Michigan	200169	40,5	37.4 - 55.7	21,3	2006 - 2010	21.1 - 22.4	1,901408	39,8	31,3	
1932	Halifax	North Carolina	54691	40,3	37.4 - 55.7	33,9	2006 - 2010	26.9 - 47.8	1,188791	38,5	34,3	
1939	Iredell	North Carolina	159437	40,3	37.4 - 55.7	22,6	2006 - 2010	22.5 - 23.9	1,783186	28,4	25,3	
2354	Orangeburg	South Carolina	92501	39,7	37.4 - 55.7	17,7	2006 - 2010	10.5 - 19.0	2,242938	40,4	32,7	
1165	St. Tammany	Louisiana	233740	39,6	37.4 - 55.7	23,1	2006 - 2010	22.5 - 23.9	1,714286	27,5	24,3	
1958	Orange	North Carolina	133801	39,5	37.4 - 55.7	24,2	2006 - 2010	24.0 - 26.8	1,632231	22,7	16,3	
1960	Pasquotank	North Carolina	40661	39,3	37.4 - 55.7	26,2	2006 - 2010	24.0 - 26.8	1,5	33,3	26,2	
435	Douglas	Georgia	132403	39	37.4 - 55.7	22	2006 - 2010	21.1 - 22.4	1,772727	31,2	25,4	
1419	Forrest	Mississippi	74934	39	37.4 - 55.7	14,3	2006 - 2010	10.5 - 19.0	2,727273	36,3	32,4	
1312	Washtenaw	Michigan	344791	38,9	37.4 - 55.7	21,6	2006 - 2010	21.1 - 22.4	1,800926	24,6	19,1	
1027	Fayette	Kentucky	295803	38,8	37.4 - 55.7	20,2	2006 - 2010	19.1 - 21.0	1,920792	30,8	24,4	
2861	Halifax	Virginia	36241	38,7	37.4 - 55.7	21,3	2006 - 2010	21.1 - 22.4	1,816901	33,4	32,9	
1914	Columbus	North Carolina	58098	38,6	37.4 - 55.7	20,9	2006 - 2010	19.1 - 21.0	1,84689	33,8	28,7	

0 (0 out of 290 Selected)

Données_sur_le_cancer_du_sein

Figure 22 : Extrait de la table attributaire « Données_sur_le_cancer_du_sein »

3.2.2 Cartographier certains facteurs de risque : Pour mieux analyser la problématique nous avons fait une analyse spatiale qui montrent la répartition des facteurs de risques les plus influents.

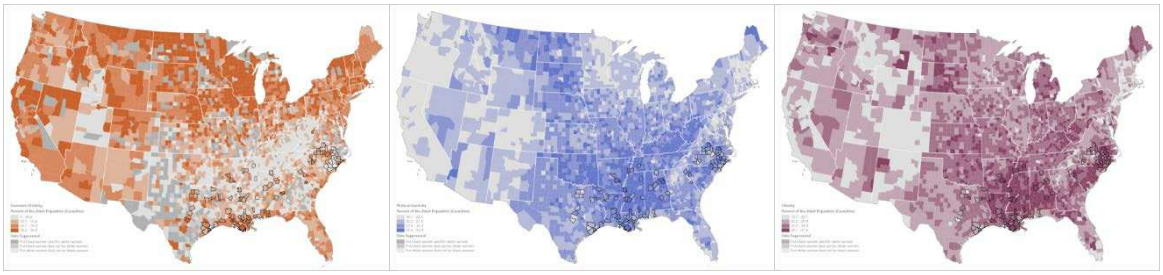
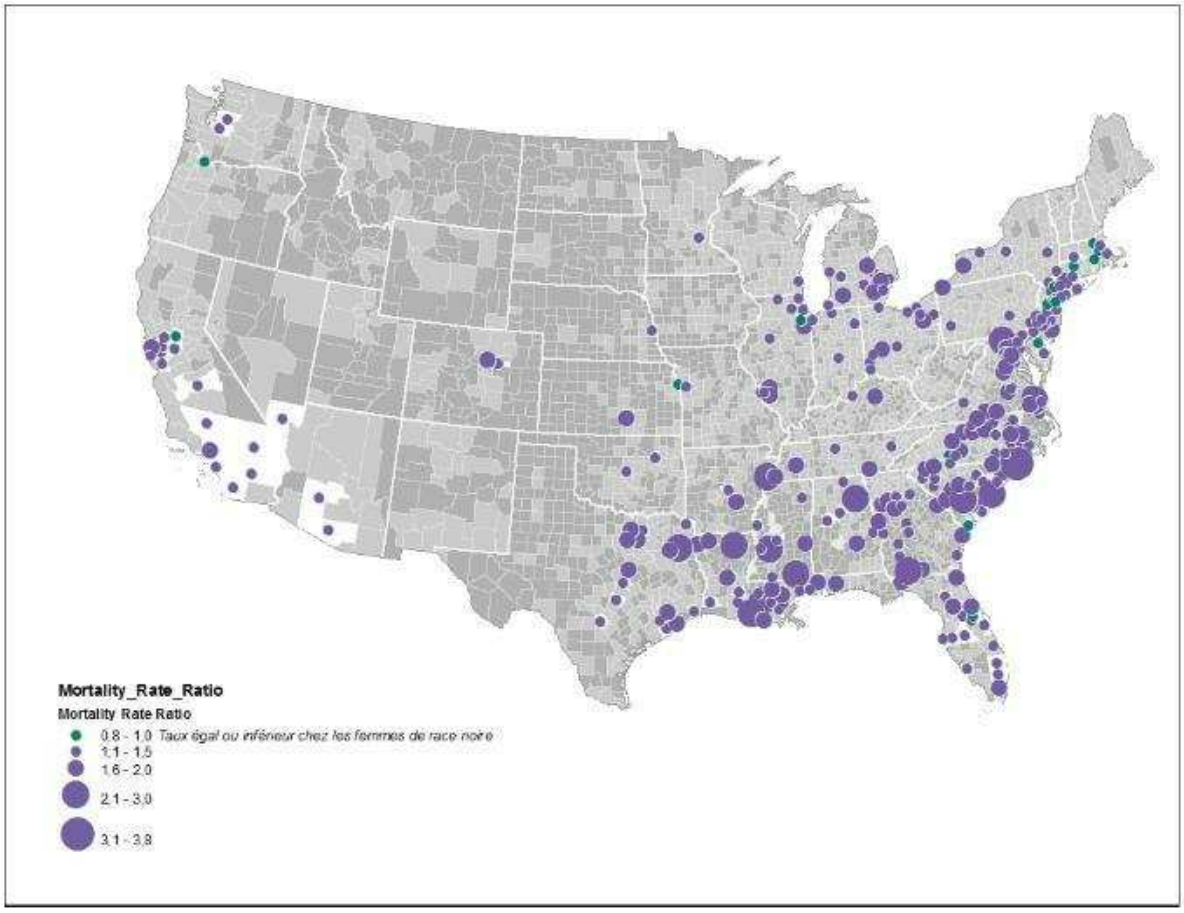


Figure 23 : les facteurs de risque (Abus d'alcool - Inactivité physique – Obésité)

Taux de mortalité :

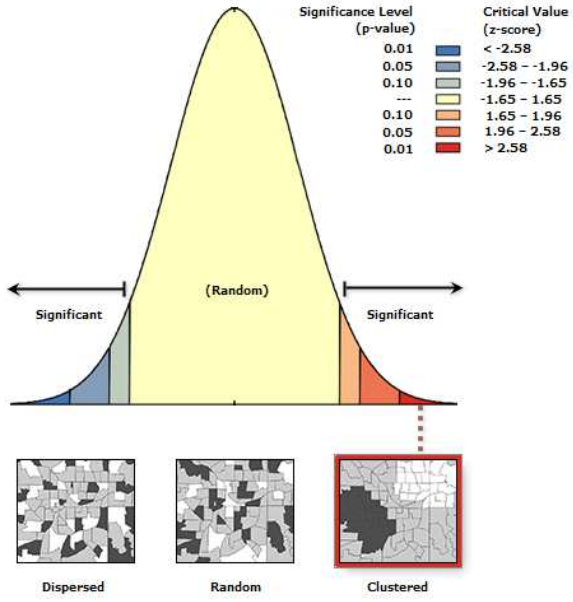


3.2.3 Application des Algorithmes : ArcGIS met à disposition plusieurs outils pour calculer les Statistiques spatiales, on les retrouve dans la fenêtre ArcToolbox dans la droite de l'interface de ArcMap comme le montre la figure suivant :

Figure 24 : ArcToolbox

Analyse global:

Autocorrélation spatial global (Morane I) : l'indice de Moran (ou *I* de Moran) est une mesure de l'autocorrélation spatiale caractérisée par une corrélation entre les mesures géographiquement voisines d'un phénomène mesuré



Global Moran's I Summary	
Moran's Index:	0,150409
Expected Index:	-0,003460
Variance:	0,000301
z-score:	8,862506
p-value:	0,000000

Les données sont autocorrélée est peuvent être claustrer.

Hight/Low Clustering (Getis-Ord i G)

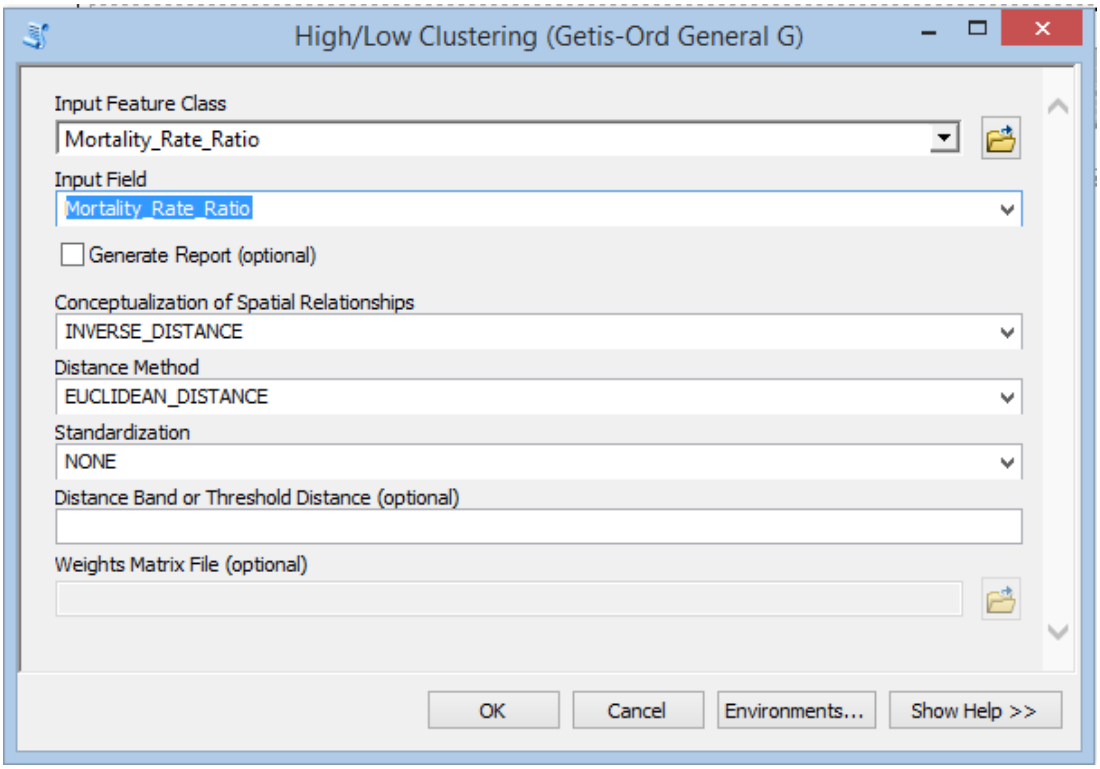


Figure 25 : Réglage des paramètres d'enter et de sortie pour la Hight/Low Clustering

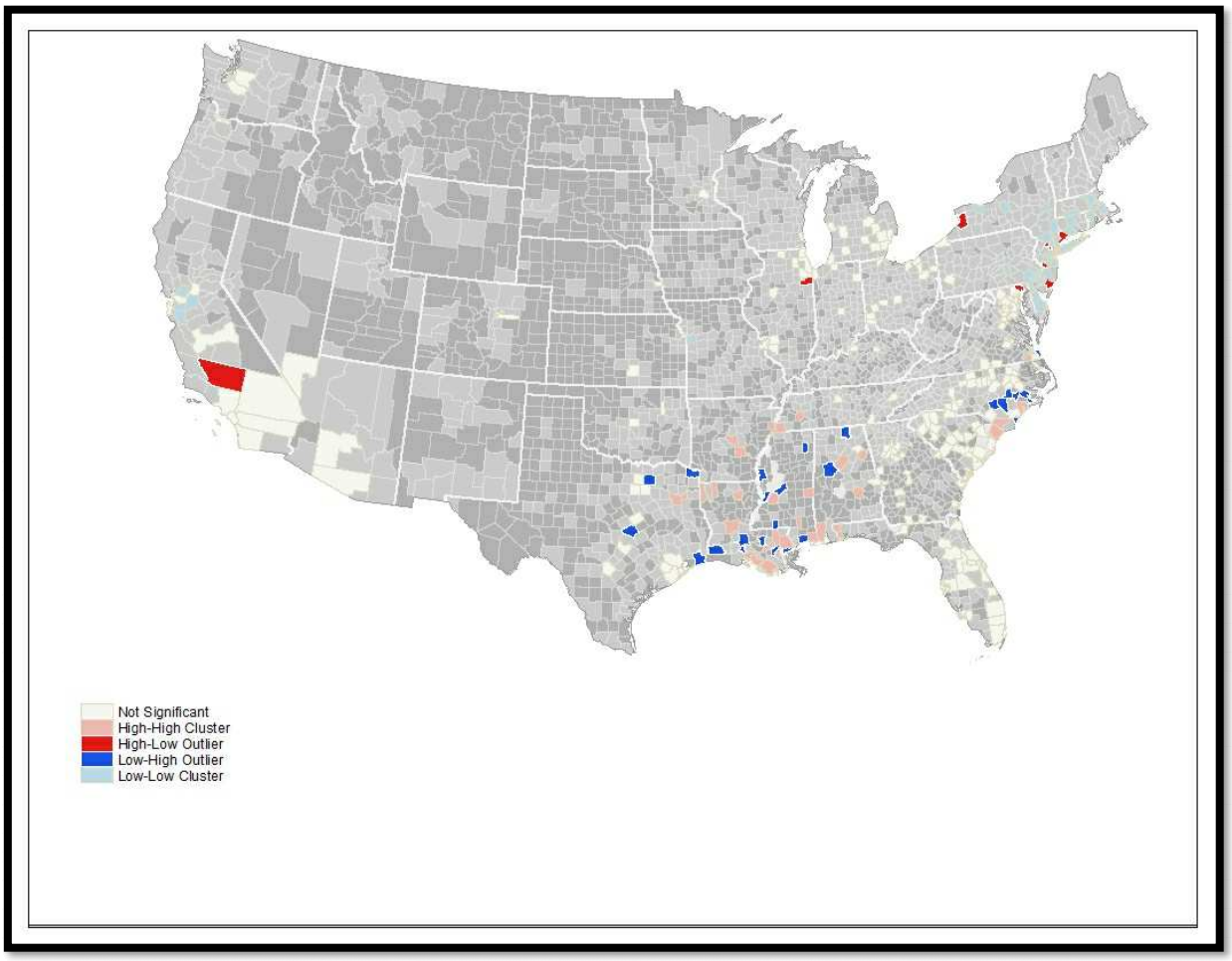


Figure 26 : Visualisation du résultat du clustering Getis-Ord *i G*)

Analyse local: Hot spot (Getis-Ord *i G*∗)

On définit comme entrée le Taux de mortalité et le ratio de mortalité. En sortie, nous aurons le taux des points chaud (Taux_Hot_Spots) que nous enregistrons dans le même répertoire de la Géodatabase déjà créé. Concernant la distance choisie pour effectuer notre analyse on choisit la méthode distance euclidienne.

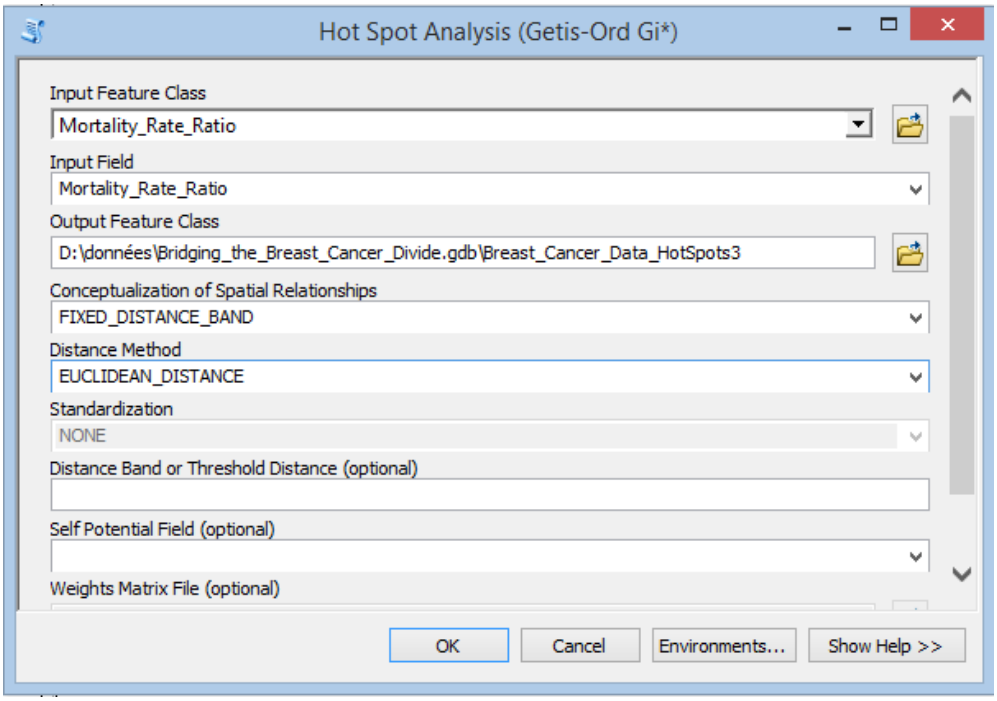


Figure 27 : Réglage des paramètres d'enter et de sortie pour la méthode hot spot

Pour faciliter la lecture de la légende, on a modifié les étiquettes des classes

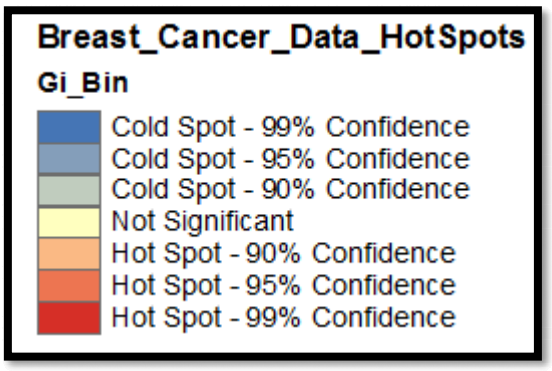


Figure 28: Symbologie des classe

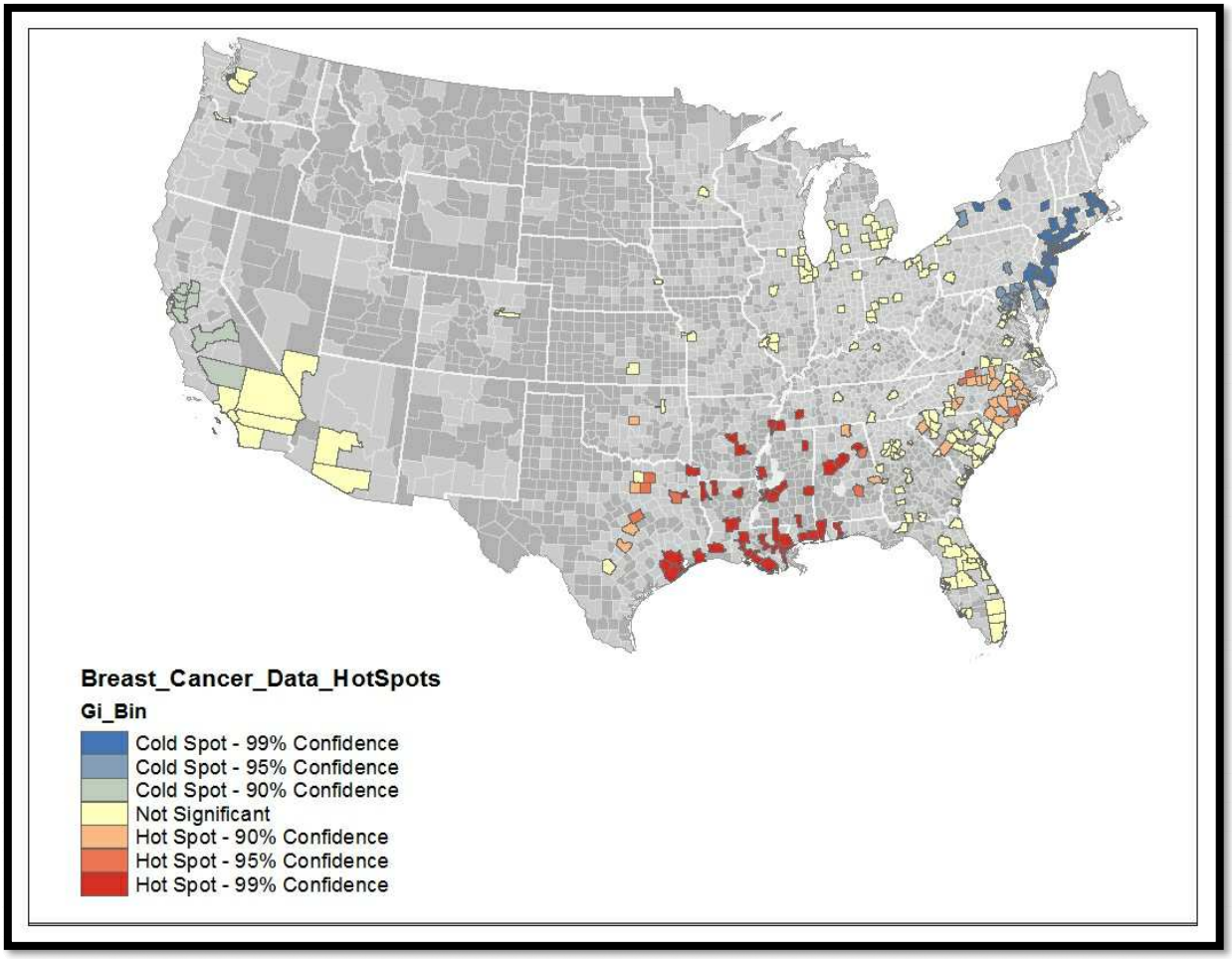


Figure 29 : Visualisation du résultat du hot spot **(**Getis-Ord *i G*∗)

3.5. Étude et partage des résultats :

Cette phase se fera dans un autre environnement crée avec visuel studio une application installable qui contient l'essentiel des fonctionnalités ArcMap (ouvrir ou créé un nouveau Document - ajouter des couches - zoomer – sélectionner et identifier un élément -prendre des mesures)

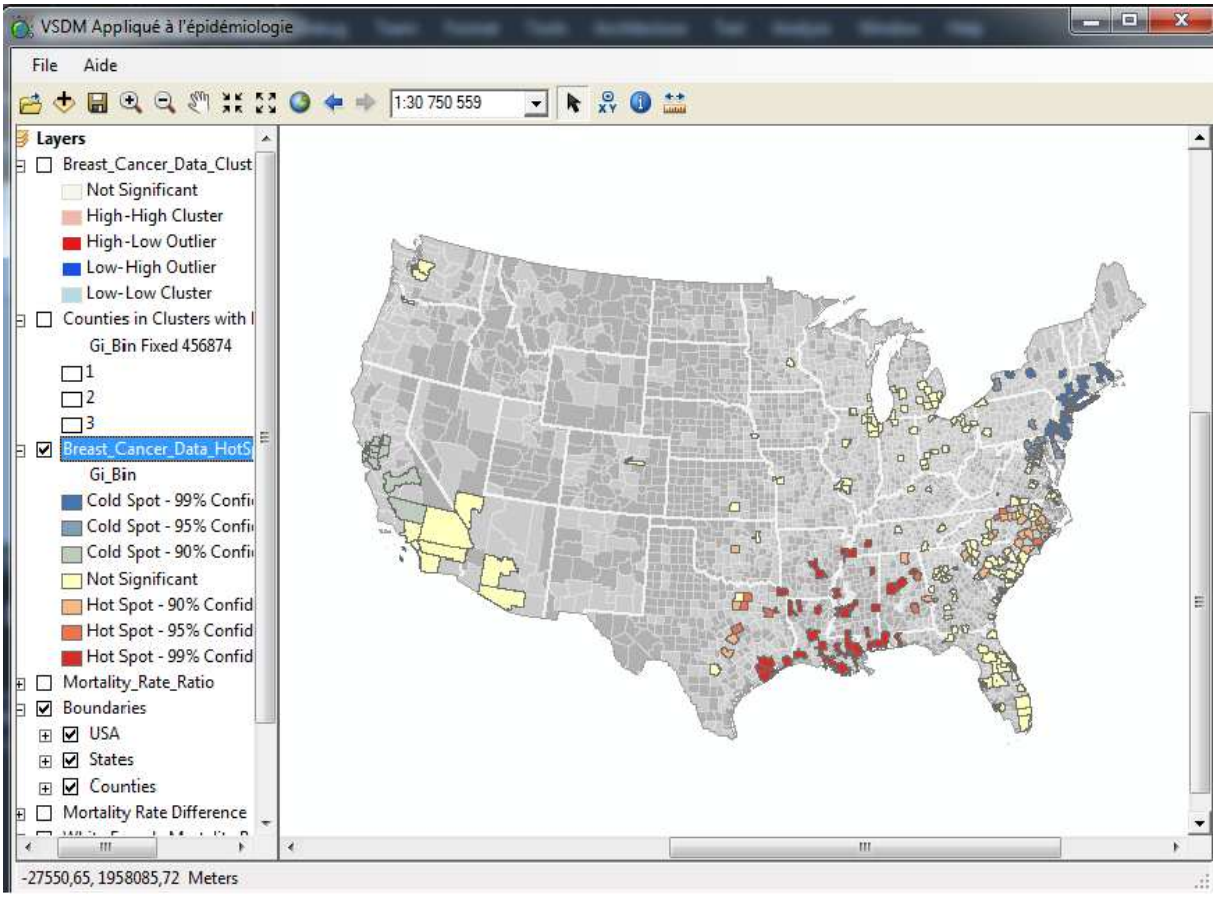


Figure 30 : Affichage des données et résultats sur l'interface créé dans visuel studio.

Etude des résultats du SDM :

- En visualisant des cartes de facteurs de risque, nous pouvons apprendre qu'une relation peut exister entre les taux de mortalité plus élevés chez les femmes de race noire et peut-être l'obésité et l'inactivité physique, mais il ne semble pas y avoir de relation avec l'abus d'alcool. Les données utilisées pour cette partie de l'investigation peuvent être trompeuses, de sorte que les relations entre les taux de mortalité et les facteurs de risque nécessitent un examen plus approfondi.
- On constate en visualisant les points chauds, qu'une grappe significative de taux de mortalité supérieurs dans le centre-sud des États-Unis et en Caroline du Nord. Il existe également une grappe de valeurs faibles dans le nord-ouest, où les femmes de race blanche ont des taux de mortalité plus élevés que chez les femmes de race noire.

Conclusion :

Nous avons présenté dans ce chapitre, en premiers lieux l'approche choisi et la problématique, les **Statistique Spatial** global et local ensuit nous avons présenté l'architecture de notre application basée sur la modélisation du cancer du sein par le VSDM, l'application ainsi réalisée permet d'analyste d'exécuter facilement les algorithmes et interpréter les résultats visuellement.

Conclusion Général

Le fort accroissement de la quantité de données générées quotidiennement, auquel nous assistons depuis plusieurs années, est un phénomène qui semble n'être qu'à ses débuts. Des approches, permettant d'appréhender cet afflux de données les traitements, est le data mining.

Le spatial data mining est plus approprié que le data mining classique en ce qui concerne les bases de données spatiales, sauf qu'il n'est pas très performant si les bases de données sont très volumineuses.

D'autre part le visuel data mining permet une interaction et une analyse rapide des données, car il utilise les capacités d'interprétation visuelle de l'esprit humain. Le seul inconvénient du VDM c'est lorsqu'il est appliqué aux données géospatiales les relations spatiales des données sont difficiles à visualiser efficacement.

Dans ce contexte, ce travail, a eu pour objectif d'étudier le couplage de ces deux approches et de tirer le meilleur profit pour construire des systèmes de découverte de connaissance visuellement activée qui pourraient faciliter le processus automatique de reconnaissances des modèles et des relations dans les données spatiales complexe et permettre à un analyste d'explorer visuellement les données. Et d'appliqué cette technique sur l'épidémiologie. Les résultats ainsi obtenus ont belle et bien confirmé l'utilité de ce couplage.

Comme perspectives nous proposons d'appliquer les autre techniques et Algorithme du visuel spatial data Mining afin d'opter pour celles qui sont les plus adéquates à l'étude épidémiologique.

Bibliographies:

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine,* vol. 17, p. 37, 1996.

[2] F. Gorunescu, *Data Mining: Concepts, models and techniques* vol. 12: Springer Science & Business Media, 2011.

[3] J. Han, K. Koperski, and N. Stefanovic, "GeoMiner: a system prototype for spatial data mining," in *AcM sIGMoD Record*, 1997, pp. 553-556.

[4] R. Srikant and R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements," *Advances in Database Technology—EDBT'96,* pp. 1-17, 1996.

[5] D. Li, S. Wang, and D. Li, *Spatial Data Mining: Theory and Application*: Springer, 2016.

[6] L. Anselin, "Local indicators of spatial association—LISA," *Geographical analysis,* vol. 27, pp. 93-115, 1995.

[7] L. Savary and K. Zeitouni, "Indexed bit map (ibm) for mining frequent sequences," in *European Conference on Principles of Data Mining and Knowledge Discovery*, 2005, pp. 659-666.

[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "Clustering for mining in large spatial databases," *KI,* vol. 12, pp. 18-24, 1998.

[9] M. Ester, A. Frommelt, H.-P. Kriegel, and J. Sander, "Algorithms for Characterization and Trend Detection in Spatial Databases," in *KDD*, 1998, pp. 44-50.

[10] T. Soukup and I. Davidson, *Visual data mining: Techniques and tools for data visualization and mining*: John Wiley & Sons, 2002.

[11] J. Dykes, A. M. MacEachren, and M.-J. Kraak, *Exploring geovisualization*: Elsevier, 2005.

[12] M. Berthold and D. J. Hand, *Intelligent data analysis: an introduction*: Springer Science & Business Media, 2003.

[13] J. Dykes, A. MacEachren, and M. Kraak, "Statistical data exploration and geographical information visualization," 2005.

[14] A. M. MacEachren and M.-J. Kraak, "Research challenges in geovisualization," *Cartography and geographic information science,* vol. 28, pp. 3-12, 2001.

[15] S. J. Simoff, "Visual data mining," in *Encyclopedia of database systems*, ed: Springer, 2009, pp. 3365-3370.

[16] A. E. Akçay, G. Ertek, and G. Büyüközkan, "Analyzing the solutions of DEA through information visualization and data mining techniques: SmartDEA framework," *Expert systems with applications,* vol. 39, pp. 7763-7775, 2012.

[17] D. A. Keim and M. O. Ward, "Visual data mining techniques," ed, 2002.

[18] M. Ankerst, M. Ester, and H.-P. Kriegel, "Towards an effective cooperation of the user and the computer for classification," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 179-188.

[19] L. Yang, "Interactive exploration of very large relational datasets through 3D dynamic projections," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 236-243.

[20] S. T. Sundaram, "Classification Rules by Decision Tree for Disease Prediction," *International Journal of Computer Applications,* vol. 43, pp. 6-12, 2012.

[21] J. Bouyer, *Epidémiologie: principes et méthodes quantitatives*: Lavoisier, 2009.

[22] D. Cameron and I. G. Jones, "John Snow, the Broad Street pump and modern epidemiology," *International journal of epidemiology,* vol. 12, pp. 393-396, 1983.

[23] M. P. Mammen Jr, C. Pimgate, C. J. Koenraadt, A. L. Rothman, J. Aldstadt, A. Nisalak*, et al.*, "Spatial and temporal clustering of dengue virus transmission in Thai villages," *PLoS Med,* vol. 5, p. e205, 2008.

[24] A. Mitchel, "The ESRI Guide to GIS analysis, Volume 2: Spartial measurements and statistics," *ESRI Guide to GIS analysis,* 2005.

Webographie :

[25] https://fr.wikipedia.org/wiki/Microsoft_Visual_Studio
[26] https://desktop.arcgis.com/fr/documentation