MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE ABDELHAMID IBN BADIS - MOSTAGANEM

Faculté des Sciences Exactes et d'Informatique

Département de Mathématiques et informatique

Filière : Informatique

MEMOIRE DE MASTER EN INFORMATIQUE

Option : **Ingénierie des Systèmes d'Information**

THEME : UTILISATION DU DATA MINING SPATIAL POUR L'ETUDE EPIDEMIOLOGIQUE DU CANCER

Etudiants : **Mehaddi Mohammed Said**

Akermi Oualid

Encadrant : Dr. MIDOUN Mohammed

Année Universitaire 2022-2023

الملخص

التنقيب عن البيانات المكانية (**DMS**) هو تخصص متخصص في استخراج المعلومات والمعرفة من قواعد البيانات الكبيرة التي تحتوي على بيانات مكانية. على عكس خوارزميات التنقيب عن البيانات الكلاسيكية، التي تفترض أن البيانات يتم إنشاؤها بشكل مستقل وتوزيعها بشكل مماثل، فإن البيانات المكانية متعددة الأبعاد ومترابطة مكانيًا وغير متجانسة. هذه الخصائص تجعل الأساليب الكلاسيكية غير مناسبة لتحليل البيانات المكانية. لذلك يقدم نظام إدارة الوجهات السياحية مناهج خاصة لمعالجة هذه البيانات المعقدة واستخراج المعلومات المفيدة منها.

الهدف في هذا المشروع هو تطبيق تقنيات التنقيب عن البيانات المكانية لدراسة وبائيات السرطان. سيتم تنفيذ هذا العمل من حالات استخدام مختلفة لـ **DMS** للسرطان. الهدف هو فهم الأنماط المكانية لحدوث السرطان والوفيات، بالإضافة إلى عوامل الخطر البيئية والاجتماعية الديموغرافية والسلوكية المرتبطة بها. وذلك لفهم التوزيع المكاني للمرض وأسبابه وتفاعله مع العوامل الاجتماعية والبيئية المختلفة المتعلقة بالسرطان.

الكلمات الأساسية. التنقيب عن البيانات المكانية، قواعد البيانات الكبيرة، البيانات المكانية، تحليل البيانات المكانية، وبائيات، السرطان، الأنماط المكانية.

Résumé

Le data mining spatial (DMS) est une discipline spécialisée dans l'extraction d'informations et de connaissances à partir de grandes bases de données contenant des données spatiales. Contrairement aux algorithmes classiques d'exploration de données, qui supposent que les données sont générées indépendamment et distribuées de manière identique, les données spatiales sont multidimensionnelles, spatialement auto-corrélées et hétérogènes. Ces propriétés rendent les approches classiques inappropriées pour l'analyse de données spatiales. Le DMS propose donc des approches spéciales pour traiter ces données complexes et en extraire des informations utiles.

Dans ce projet, l'objectif est d'appliquer des techniques de data mining spatial pour étudier l'épidémiologie du cancer. L'objectif est de comprendre les modèles spatiaux de l'incidence et de la mortalité du cancer, ainsi que les facteurs de risque environnementaux, sociodémographiques et comportementaux qui y sont associés. Ceci, afin de comprendre la répartition spatiale de la maladie, ses causes et son interaction avec les différents facteurs sociaux et environnementaux en relation avec le cancer.

Mots clés : Data mining spatial, grandes bases de données, données spatiales, analyse de données spatiales, épidémiologie, cancer, modèles spatiaux.

Abstract

Spatial data mining (DMS) is a discipline specialized in extracting information and knowledge from large databases containing spatial data. Unlike classical data mining algorithms, which assume that data is independently generated and identically distributed, spatial data is multidimensional, spatially auto-correlated and heterogeneous. These properties make classical approaches inappropriate for the analysis of spatial data. The DMS therefore offers special approaches to process this complex data and extract useful information from it.

In this project, the objective is to apply spatial data mining techniques to study cancer epidemiology. This work will be carried out from different use cases of DMS for cancer. The objective is to understand the spatial patterns of cancer incidence and mortality, as well as the environmental, sociodemographic and behavioral risk factors associated with them. This, in order to understand the spatial distribution of the disease, its causes and its interaction with the various social and environmental factors related to cancer.

Key words: spatial data mining. Large databases, spatial data, analysis of spatial data, epidemiology, cancer, spatial patterns.

Liste des figures

Figure N°	Titre de la figure	Page
Figure 1	L'AUTOCORRELATION SPATIALE	7
Figure 2	REPRESENTATION SCHEMATIQUE DE LA VARIABILITE SPATIALE D'UNE VARIABLE ETUDIEE	8
Figure 3	ORGANISATION DE L'INFORMATION PAR COUCHES	10
Figure 4	DOMAINES CONNEXES DU DMS	16
Figure 5	INDEX DE JOINTURE SPATIAL	18
Figure 6	REPRESENTATION DES RELATIONS SPATIALES PAR UNE MATRICE DE CONTIGUÏTE (C) ET UN GRAPHE DE VOISINAGE (D)	19
Figure 7	PROCESSUS DE DMS	22
Figure 8	PRESENTATION DE LA METHODOLOGIE SUIVIE POUR NOTRE ETUDE DE CAS	39
Figure 9	PRESENTATION DE LA DEMARCHE DE CLASSIFICATION SPATIALE BASEE DU LE MODELE SVM	47
Figure 10	PRESENTATION DE LA DEMARCHE DE CLUSTERING SPATIALE BASEE SUR L'ALGORITHME K-MEANS	49
Figure 11	PRESENTATION DE LA DEMARCHE DE DETECTION DES HOTSPOTS ET COLD SPOTS BASEE LE CALCUL L'INDICE I DE MORAN LOCAL	51
Figure 12	PRESENTATION DE LA DEMARCHE DE DETECTION DES OUTLIERS BASEE SUR LE CALCUL L'INDICE I DE MORAN GLOBAL	53
Figure 13	INTERFACE DU LOGICIEL DEVELOPPE	54
Figure 14	FENETRE PERMETTANT LE CHOIX DE L'ANALYSE A REALISER	55
Figure 15	RESULTATS DE LA CLASSIFICATION SPATIALE APPLIQUEE AUX DONNEES DU CANCER DE L'ESTOMAC AUX USA ENTRE 2015 ET 2019	56
Figure 16	RESULTATS DU CLUSTERING SPATIAL APPLIQUE AUX DONNEES DU CANCER DE L'ESTOMAC AUX USA ENTRE 2015 ET 2019	56

Figure 17	DETECTION DES HOTSPOTS ET COLD SPOTS SPATIAUX SUR LES DONNEES DU CANCER DE L'ESTOMAC AUX USA ENTRE 2015 ET 2019	57
Figure 18	DETECTION DES OUTLIERS SPATIAUX SUR LES DONNEES DU CANCER DE L'ESTOMAC AUX USA ENTRE 2015 ET 2019	57

Liste des tableaux

Tableau N°	Titre de la Tableau	Page
TABLEAU 1	BIBLIOTHEQUES PYTHON UTILISEES POUR LA REALISATION DES DIFFERENTES ANALYSES	43

Liste des abréviations

Abréviation	Expression Complète
ACP	L'analyse en composantes principales
AFD	L'analyse factorielle discriminante
BD	Base de données
BDS	Bases de données Spatiales
DM	Data Mining
DMS	Data Mining Spatial
MNT	Modèles numériques de terrain
OLAM	Online Analytical Mining
OLAP	Online Analytical Processing
SIG	Systèmes d'Information Géographiques
SVM	Supportive Vector Machine

Table des matières

الملخص ii

Résumé iii

Abstract iv

Liste des figures v

Liste des tableaux vii

Liste des abréviations viii

Table des matières 1

Introduction Générale 4

Chapitre 1 L'information spatiale 6

1.1 Introduction 6

1.2 L'information spatiale 6

1.3 Spécificités de l'information spatiale 7

1.4 Les modes de représentation de l'information spatiale 9

1.5 Les Systèmes d'information géographique 10

1.6 L'analyse spatiale 11

1.7 Conclusion 11

Chapitre 2 Le Data Mining Spatial 12

2.1 Introduction 12

2.2 Le data mining 12

2.3 Les taches du data mining 13

2.3.1 La classification 13

2.3.2 L'estimation 13

2.3.3 La prédiction 14

2.3.4 Le groupement par similitude 14

2.3.5 L'analyse des clusters 15

2.3.6 La description 15
2.4 Définition du data mining spatial 15
2.5 Spécificités du data mining spatial 16
2.6 Travaux sur le data mining spatial 17
2.7 Approches du data mining spatial 18
2.7.1 Approche base de données 18
2.7.2 Approche statistique 19
2.8 Les taches du data mining spatial 20
2.8.1 La classification spatiale : 20
2.8.2 La prédiction spatiale : 21
2.8.3 Les règles d'association spatiale : 21
2.8.4 Le clustering spatial : 21
2.8.5 L'analyse des points chauds spatiaux (spatial hotspot analysis) : 21
2.8.6 L'analyse de valeurs spatiales aberrantes (Spatial outlier analysis) : 21
2.9 Processus du data mining spatial 21
2.10 Domaines d'application du data mining spatial 22
2.11 Exemples d'application du data mining spatial pour le cancer 24
2.12 Conclusion 27
Chapitre 3 Méthodologie 29
3.1 Introduction 29
3.2 Description des données épidémiologiques utilisées 29
3.3 Prétraitement des données spatiales 30
3.4 Choix des analyses et des techniques de data mining spatial 31
3.4.1 Classification spatiale : 31
3.4.2 Analyse de clusters : 32
3.4.3 Détection de hotspots et cold spots 34
3.4.4 Détection d'outliers : 36
3.5 Architecture logicielle 38
3.6 Méthodes d'évaluation et d'interprétation des résultats 39
3.7 Conclusion 40
Chapitre 4 Expérimentation 41

4.1 Présentation des Données 41
4.2 Présentation de l'environnement expérimental 41
4.3 Mise en œuvre des techniques de data mining spatial 45
4.3.1 Classification pour l'étude épidémiologique du cancer 45
4.3.2 Clustering pour l'étude épidémiologique du cancer 47
4.3.3 Détection des hotspots et cold spots pour l'étude épidémiologique du cancer 49
4.3.4 Détection des outliers pour l'étude épidémiologique du cancer 52
4.4 Présentation du logiciel 54
4.5 Analyse des résultats obtenus 55
4.6 Interprétation des résultats et discussion 58
4.6.1 Classification spatiale : 58
4.6.2 Clustering : 58
4.6.3 Détection des hotspots et cold spots spatiaux : 58
4.6.4 Détection des outliers : 59
4.7 Conclusion 59
Conclusion Générale 60
Annexe 62
Bibliographies 68

Introduction Générale

Le cancer est une maladie grave caractérisée par la prolifération incontrôlée de cellules anormales dans le corps. Il existe différentes formes et sous-types de cancer, et son traitement peut être réalisé de diverses manières, telles que la chirurgie, la radiothérapie et la chimiothérapie.

Le data mining classique est une discipline informatique visant à extraire des informations et des connaissances à partir de vastes bases de données. Ce processus itératif comprend la sélection des données pertinentes, leur prétraitement et leur transformation dans un format approprié pour l'analyse. Ensuite, des algorithmes de data mining sont appliqués pour détecter des modèles et des relations au sein des données. Les résultats sont ensuite interprétés et évalués afin de générer des connaissances utiles.

Les méthodes de data mining classique incluent la classification, la régression, le clustering et l'association. Elles sont largement utilisées dans des domaines tels que la finance, la médecine, le commerce de détail, la science des données, etc. Ces techniques permettent la prédiction, la classification, l'identification de groupes homogènes et la découverte de règles d'association.

Le data mining spatial est une approche qui utilise des algorithmes de traitement de données pour analyser des données géospatiales, afin de découvrir des modèles et des tendances cachées. Dans le domaine de la santé, cette méthode permet d'étudier la propagation du cancer et son impact sur des facteurs environnementaux tels que la pollution de l'air et de l'eau. Elle aide également à identifier les groupes de population à haut risque de cancer et à élaborer des stratégies de prévention et de traitement efficaces.

Le data mining spatial fait partie de l'exploration de données et se concentre sur l'analyse de données géospatiales à l'aide de techniques statistiques et informatiques. Son utilisation a considérablement augmenté ces dernières années, grâce à la disponibilité croissante des données géospatiales. Il est appliqué dans de nombreux domaines, y compris la santé, pour découvrir des tendances et des modèles cachés dans les données spatiales, permettant ainsi une meilleure compréhension des relations entre les facteurs environnementaux et la santé humaine.

Ce mémoire se concentre sur l'application du data mining spatial dans le domaine de la santé, en mettant spécifiquement l'accent sur l'analyse du cancer. Notre objectif est d'explorer comment cette approche peut être utilisée pour détecter les hotspots, les cold spots et les outliers liés à cette maladie. Nous nous intéressons également à la classification des zones en fonction des taux d'incidence et de mortalité du cancer. Enfin, nous aborderons l'utilisation du clustering pour identifier les clusters de cette maladie.

Ce mémoire est divisé en quatre chapitres principaux et une conclusion. Le chapitre 1 traite de l'information spatiale, de ses spécificités et des méthodes de représentation, ainsi que de l'utilisation des Systèmes d'Information Géographique (SIG) et de l'analyse spatiale.

Le chapitre 2 aborde le data mining spatial (DMS) en fournissant une introduction générale sur le data mining, en explorant les différentes tâches du DMS, ses spécificités, ses approches et ses domaines d'application. Il se concentre également sur l'état de l'art de l'application du DMS dans le domaine du cancer, en présentant quelques exemples d'utilisation de cette méthode pour étudier la maladie.

Le chapitre 3 décrit la méthodologie utilisée dans notre étude. Il comprend une description des données épidémiologiques utilisées, le prétraitement des données spatiales, le choix des techniques de data mining spatial, le développement de l'application et les méthodes d'évaluation des résultats.

Le chapitre 4 est consacré à la partie expérimentale de notre étude. Nous y présentons les résultats obtenus en appliquant les techniques de data mining spatial à l'étude épidémiologique du cancer aux Etats Unis. Nous décrivons les données utilisées, l'environnement expérimental, la mise en œuvre des techniques, le développement du logiciel et l'analyse des résultats obtenus. Nous fournissons également une interprétation des résultats et une discussion approfondie.

Enfin, la conclusion résume les points clés abordés dans chaque chapitre et propose quelques pistes pour de futures recherches dans le domaine du data mining spatial appliqué à la santé et à l'analyse du cancer.

Chapitre 1

L'information spatiale

1.1 Introduction

Dans ce chapitre, nous commencerons par décrire l'information spatiale, ses spécificités, ses formats de représentation. Nous parlerons ensuite des relations spatiales, des SIG et des Bases de données Spatiales (BDS). Puis nous donnerons une définition du DMS, et nous détaillerons son processus

1.2 L'information spatiale

L'information spatiale désigne les données qui ont une dimension géographique ou spatiale. Ces données peuvent être des coordonnées géographiques (latitude et longitude), des adresses, des codes postaux ou des noms de lieux. Elles peuvent également être des données géographiques, comme des cartes, des images satellites ou des modèles numériques de terrain (MNT).

L'information spatiale est souvent utilisée dans le data mining spatial, car elle permet de visualiser les patterns et les tendances dans les données sur une carte et de mieux comprendre comment les variables sont liées au niveau spatial. Elle peut également être utilisée pour géocoder (c'est-à-dire associer des coordonnées géographiques à des adresses ou des noms de lieux) ou pour effectuer des analyses spatiales avancées, comme l'analyse des réseaux ou l'analyse de la distance.

L'information spatiale peut être utilisée dans de nombreux domaines différents, comme la géographie, la géomatique, la géo-analyse, la planification territoriale, l'analyse des réseaux, l'analyse de la criminalité, l'analyse de la santé, l'analyse des risques, l'analyse de l'environnement, l'analyse des données marketing et l'épidémiologie.

1.3 Spécificités de l'information spatiale

L’information spatiale est différente de l’information classique, nous décrivons dans ce qui suit les principales spécificités de l’information spatiale :

1.3.1 L’autocorrélation spatiale

L'autocorrélation spatiale désigne la relation entre les valeurs d'une variable à des emplacements spatiaux proches. Elle peut être positive, négative ou nulle.

Si l'autocorrélation spatiale est positive, cela signifie que les valeurs de la variable sont similaires entre elles à des emplacements proches. Par exemple, si les températures sont élevées dans une région, elles le seront probablement aussi dans les régions voisines.

Si l'autocorrélation spatiale est négative, cela signifie que les valeurs de la variable sont inverses entre elles à des emplacements proches. Par exemple, si les températures sont élevées dans une région, elles seront probablement basses dans les régions voisines

Si l'autocorrélation spatiale est nulle, cela signifie qu'il n'y a pas de relation entre les valeurs de la variable à des emplacements proches.

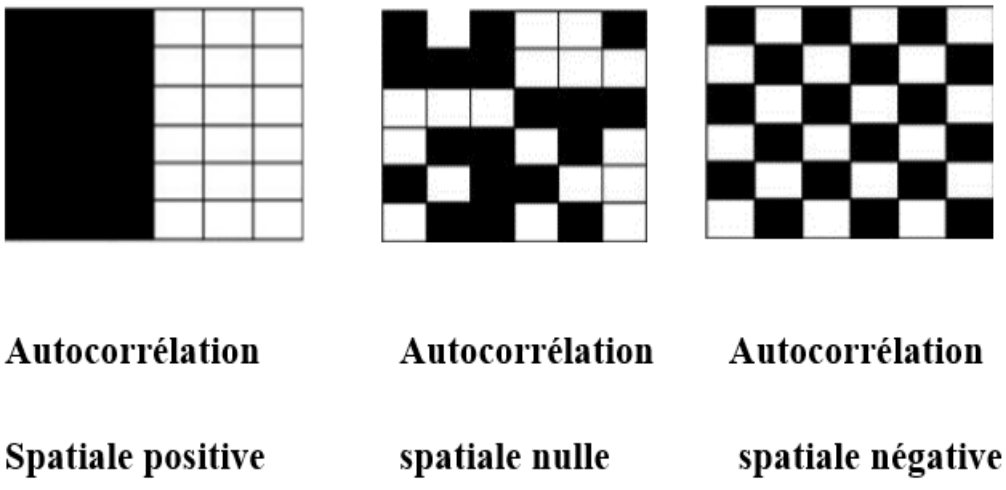


FIGURE 1 - L’AUTOCORRELATION SPATIALE

L'autocorrélation spatiale est souvent étudiée dans le data mining spatial, car elle peut avoir une influence sur les résultats des analyses. Par exemple, si l'autocorrélation spatiale

est forte, cela peut rendre difficile la détection de patterns et de tendances dans les données. Il est donc important de prendre en compte l'autocorrélation spatiale lors de l'analyse des données spatiales.

1.3.2 Hétérogénéité spatiale

Le concept d'hétérogénéité spatiale est facilement compréhensible en opposition à celui d'homogénéité spatiale, qui désigne l'absence de variation dans l'espace. En 1972, Smith a défini un environnement comme hétérogène lorsque le taux d'un processus varie entre différents endroits de l'espace, tandis qu'un environnement est homogène si ce processus a un taux uniforme entre ces différents endroits. Depuis lors, de nombreuses définitions et concepts ont été proposés [1-6], mais nous nous concentrons ici uniquement sur l'hétérogénéité spatiale (et non temporelle) en utilisant la définition de Smith, qui peut s'appliquer à diverses variables écologiques telles que les individus, les espèces, leurs caractéristiques ou les facteurs environnementaux.

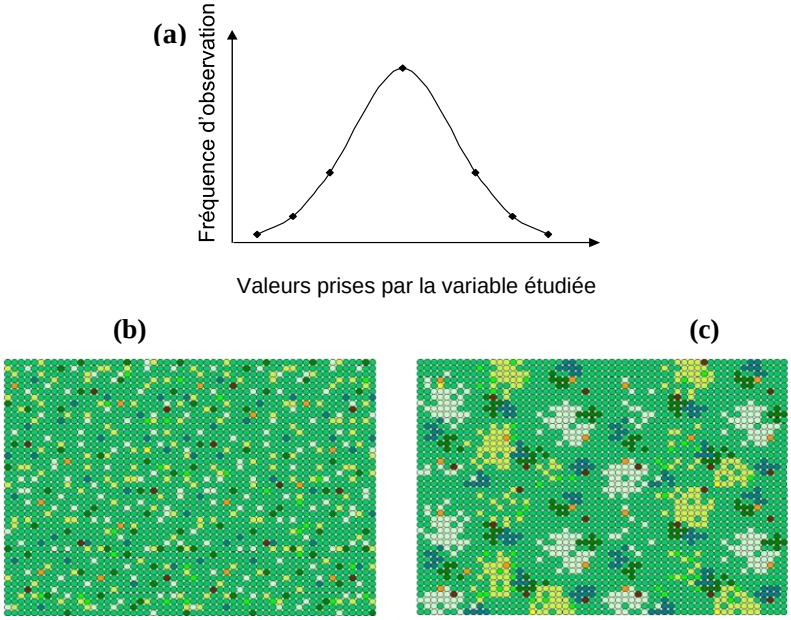


FIGURE 2 - REPRESENTATION SCHEMATIQUE DE LA VARIABILITE SPATIALE D'UNE VARIABLE ETUDIEE (A) POUVANT CORRESPONDRE A DIFFERENTS AGENCEMENTS SPATIAUX DES VALEURS DE LA VARIABLE : (B) MOTIF ALEATOIRE ET (C) MOTIF AGREGE NON ALEATOIRE [42].

L'hétérogénéité spatiale comporte deux aspects distincts : (i) la variabilité spatiale, qui fait référence à la différence et à la diversité des valeurs d'une variable donnée, indépendamment de leur disposition dans l'espace, (ii) et la structure spatiale, qui se réfère à l'arrangement spatial de ces valeurs. Lorsque la variabilité spatiale est prévisible ou reproductible, elle donne lieu à des structures spatiales ou des patrons (Figure 2), qui représentent le motif que les valeurs de la variable étudiée forment dans l'espace, en prenant en compte leur localisation spatiale de manière explicite. Les travaux de Legendre et Fortin [7] et de Dale [6] ont approfondi la notion de structure spatiale.

1.4 Les modes de représentation de l'information spatiale

L'information spatiale peut être représentée de différentes manières, en fonction des objectifs de l'analyse et des données à disposition. Voici quelques exemples de modes de représentation de l'information spatiale :

• **Les cartes :** les cartes sont une représentation visuelle de l'espace qui permet de visualiser les patterns spatiaux dans les données. Elles peuvent être créées à partir de différentes sources de données, comme des images satellites, des modèles numériques de terrain (MNT) ou des bases de données géographiques.

• **Les graphiques :** les graphiques sont une représentation visuelle des données qui permet de visualiser les tendances et les variations dans les données. Ils peuvent être utilisés pour représenter l'information spatiale, en utilisant des coordonnées géographiques ou des codes postaux comme axes.

• **Les tableaux :** les tableaux sont une représentation structurée des données qui permet de visualiser les valeurs de chaque variable pour chaque emplacement spatial. Ils peuvent être utilisés pour représenter l'information spatiale, en utilisant des coordonnées géographiques ou des codes postaux comme identifiants.

• **Les bases de données :** les bases de données sont un mode de représentation des données qui permet de stocker et de manipuler de grandes quantités de données de manière structurée. Elles peuvent être utilisées pour représenter l'information spatiale, en utilisant des coordonnées géographiques ou des codes postaux comme identifiants.

•**Les fichiers de données vectorielles :** les fichiers de données vectorielles sont une représentation de l'information spatiale sous forme de lignes, de polygones et de points. Ils peuvent être utilisés pour visualiser les patterns spatiaux dans les données et pour effectuer des analyses spatiales avancées.

•**Les fichiers de données raster :** les fichiers de données raster sont une représentation de l'information spatiale sous forme de grille de pixels. Ils peuvent être utilisés pour visualiser les patterns spatiaux dans les données et pour effectuer des analyses de données géographiques.

Il existe de nombreux autres modes de représentation de l'information spatiale, en fonction des données et des objectifs à disposition. Il est important de choisir le mode de représentation approprié en fonction de l'analyse à effectuer.

1.5 Les Systèmes d'information géographique

Le Système d'information géographique (SIG) est un système de gestion informatique capable de saisir, de stocker, d'analyser, de présenter sous forme de cartes ou de graphes des données localisées dans un espace géographique. Le SIG se définit comme des ensembles de données repérées dans l'espace et structurées de façon à pouvoir en extraire commodément des synthèses utiles (Michel Didier, 1990).

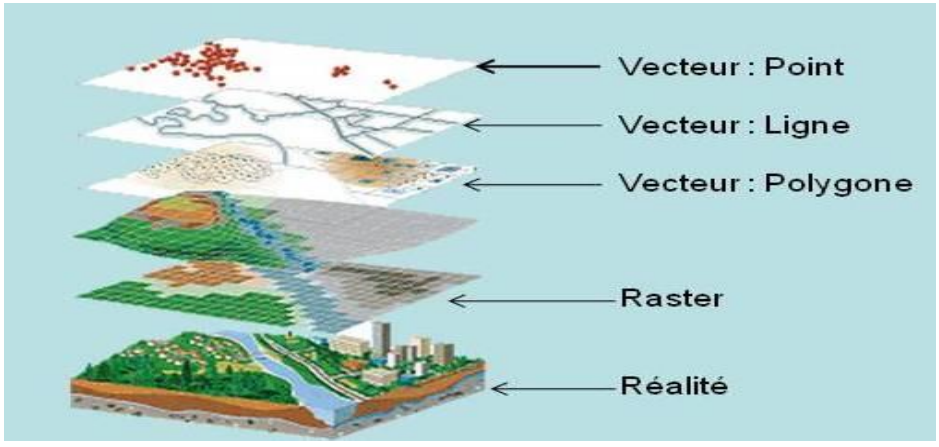


FIGURE 3 - ORGANISATION DE L'INFORMATION PAR COUCHES

1.6 L'analyse spatiale

L'analyse spatiale est une fonctionnalité cruciale des Systèmes d'Information Géographiques (SIG). Elle consiste à étudier les phénomènes qui sont répartis dans l'espace et ont des dimensions physiques telles que la localisation, la proximité, etc. L'objectif est de comprendre et d'interpréter les phénomènes du monde réel, en mettant en évidence des structures spatiales récurrentes. Cette analyse est utilisée dans divers domaines, notamment l'environnement, le marketing, les sciences humaines, etc… [9]. Le processus d'analyse spatiale utilise différentes méthodes, outils et techniques, avec la visualisation jouant un rôle important pour stimuler la découverte de modèles, de relations et de tendances [10].

1.7 Conclusion

L'information spatiale est l'ensemble des données associées aux emplacements ou aux objets géographiques. Elle peut être utilisée dans de nombreux domaines, tels que la cartographie, l'analyse de la mobilité, l'urbanisme et la gestion des catastrophes.
Les technologies de l'information spatiale, telles que la télédétection et la géolocalisation, permettent de collecter et de traiter de vastes quantités de données géospatiales. Ces données peuvent être utilisées pour créer des cartes et des visualisations, pour effectuer des analyses spatiales et pour prendre des décisions informées. En résumé, l'information spatiale est un outil puissant pour comprendre et améliorer le monde qui nous entoure.

Chapitre 2

Le Data Mining Spatial

2.1 Introduction

Dans ce chapitre, nous donnerons une définition du DM et DMS et leurs tâches, nous détaillerons le processus du DMS, ses phases d'exécutions, ses approches, ainsi que les méthodes et domaines d'application du DMS. Nous finirons ce chapitre par une conclusion.

2.2 Le data mining

Le Data Mining appelé aussi fouille de données est le processus de découverte de nouvelles corrélations, modèles et tendances en analysant une grande quantité de données, en utilisant les technologies de reconnaissance des formes ainsi que d'autres techniques statistiques et mathématiques [11].

Ils existent d'autres définitions :

• Le Data Mining est l'analyse de grands ensembles de données observationnelles pour découvrir des nouvelles relations entre elles et de les reformuler afin de les rendre plus utilisables de la part de ses propriétaires [12].

• Le Data Mining est un domaine interdisciplinaire utilisant dans le même temps des techniques d'apprentissage automatiques, de reconnaissance des formes, des statistiques, des bases de données et de visualisation pour déterminer les manières d'extraction des informations de très grandes bases de données [13].

• Le Data Mining est un processus inductif, itératif et interactif dont l'objectif est la découverte de modèles de données valides, nouveaux, utiles et compréhensibles dans de larges Bases de Données [14].

2.3 Les taches du data mining

Il existe six tâches principales pour exprimer de nombreux problèmes intellectuels, économiques ou commerciaux [15]. Ces tâches comprennent la classification, l'estimation, la prédiction, le groupement par similitude, l'analyse des clusters et la description. Les trois premières tâches font partie du Data Mining supervisé, qui vise à créer un modèle décrivant une variable particulière en utilisant les données disponibles. D'autre part, le groupement par similitude et l'analyse des clusters sont des tâches non-supervisées qui visent à établir un rapport entre toutes les variables [16]. La tâche de description est considérée comme à la fois supervisée et non supervisée, car elle peut être utilisée dans les deux types de tâches [15].

2.3.1 La classification

Le Data Mining utilise fréquemment la tâche de classification, qui reflète la tendance humaine à classer et évaluer notre vie quotidienne [15]. Elle implique l'étude des caractéristiques d'un nouvel objet pour l'attribuer à une classe prédéfinie, généralement dans une base de données. Cette tâche nécessite une définition précise des classes et un ensemble d'exemples classés auparavant pour créer un modèle applicable aux données non classées [9]. Les applications de la classification dans la recherche et le commerce comprennent la détection de la fraude sur les cartes de crédit, le diagnostic des maladies [17], l'identification des numéros de fax [15] et des lignes d'accès à Internet [16].

2.3.2 L’estimation

L'estimation est une tâche similaire à la classification, à la différence que la variable de sortie est de nature numérique plutôt que catégorique. Dans le cadre de l'estimation, il s'agit de prédire une valeur manquante dans un champ spécifique en fonction des autres champs de l'enregistrement. Par exemple, dans un contexte hospitalier, on peut chercher à estimer la tension systolique d'un patient en se basant sur son âge, son genre, son indice de masse corporelle et le niveau de sodium dans son sang. En utilisant les relations entre ces variables, on peut élaborer un modèle d'estimation qui peut être appliqué dans d'autres cas similaires [15] [17].

L'estimation est utilisée dans divers domaines tels que la recherche et le commerce. Par exemple, elle peut servir à estimer le nombre d'enfants dans une famille [15], le coût des achats de rentrée scolaire pour une famille de quatre personnes choisies au hasard [17] ou la valeur d'un bien immobilier [16]. Souvent, la classification et l'estimation sont utilisées ensemble, comme c'est le cas en Data Mining pour prédire qui est susceptible de répondre à une offre de transfert de solde de carte de crédit et évaluer la taille de l'équilibre à transférer.

2.3.3 La prédiction

La prédiction est une tâche similaire à la classification et à l'estimation, mais elle implique la classification des enregistrements en fonction de critères ou de valeurs prédites. La principale différence entre la prédiction et les autres tâches est que la relation temporelle entre les variables d'entrée et de sortie est prise en compte dans la création du modèle prédictif [15].

La prédiction est utilisée dans divers domaines tels que la recherche et le commerce. Par exemple, elle peut être utilisée pour prédire le prix des actions dans les trois prochains mois [17], pour prédire le champion de la Coupe du Monde de football en comparant les statistiques des équipes, ou encore pour prédire quels clients vont déménager dans les six prochains mois [15].

2.3.4 Le groupement par similitude

Le groupement par similitude est une méthode qui vise à déterminer quels attributs sont liés les uns aux autres. Cette méthode est largement utilisée dans le monde des affaires, où elle est appelée analyse d'affinité ou analyse du panier de marché. Elle permet de mesurer la relation entre deux ou plusieurs attributs en utilisant des règles d'association du type "Si antécédent, alors conséquent".

Les tâches de groupement par similitude sont couramment utilisées dans différents domaines tels que la recherche et le commerce. Par exemple, elles peuvent être utilisées pour déterminer quels produits sont achetés ensemble dans un supermarché et ceux qui ne le sont jamais, ou encore pour évaluer la proportion de cas dans lesquels un nouveau médicament peut causer des effets secondaires dangereux [17].

2.3.5 L’analyse des clusters

Le clustering, également appelé segmentation, consiste à regrouper des enregistrements ou des observations en classes d'objets similaires. Chaque classe est constituée d'enregistrements qui se ressemblent, mais qui sont différents des autres classes. Contrairement à la classification, le clustering ne nécessite pas de variables de sortie, car il ne prédit pas la valeur d'une variable. Les algorithmes de clustering cherchent à segmenter les données en sous-groupes homogènes, en maximisant l'homogénéité à l'intérieur de chaque groupe et en la minimisant entre les groupes [17].

Les algorithmes de clustering peuvent être utilisés dans de nombreux domaines, tels que la segmentation de clients ayant des comportements similaires, la classification des plantes et des animaux en fonction de leurs caractéristiques, ou encore la segmentation des observations d'épicentres pour identifier les zones à risque [18].

2.3.6 La description

Parfois, l'objectif du Data Mining consiste simplement à décrire ce qui se produit sur une base de données complexe, en expliquant les relations qui existent dans les données, afin de mieux comprendre les individus, les produits et les processus qui sont présents dans cette base.

Une bonne description d'un comportement implique souvent une bonne explication de celui-ci. Par exemple, une simple description telle que "les femmes soutient le parti démocrate plus que les hommes" peuvent susciter un grand intérêt et inciter des journalistes, des sociologues, des économistes et des spécialistes de la politique à mener des études [15] [19].

2.4 Définition du data mining spatial

Le Data Mining (DM) est le processus de découverte de modèles cachés dans les bases de données. Le Data Mining Spatial (DMS) est une branche du DM qui s'intéresse aux données spatiales, qui comprennent non seulement des informations alphanumériques et de localisation, mais également des données de distance spatiale. Le DMS est plus complexe que le DM en raison de la complexité des données spatiales et des relations spatiales Le DMS est le processus de découverte de modèles intéressants et précédemment

inconnus, mais potentiellement utiles à partir de grands ensembles de données spatiales [20].

Le DMS est interdisciplinaire et intègre des domaines tels que l'apprentissage automatique, la reconnaissance de formes, l'analyse statistique, les systèmes de bases de données et l'intelligence artificielle. Les disciplines connexes sont liées mais différentes du DMS (Figure 4).

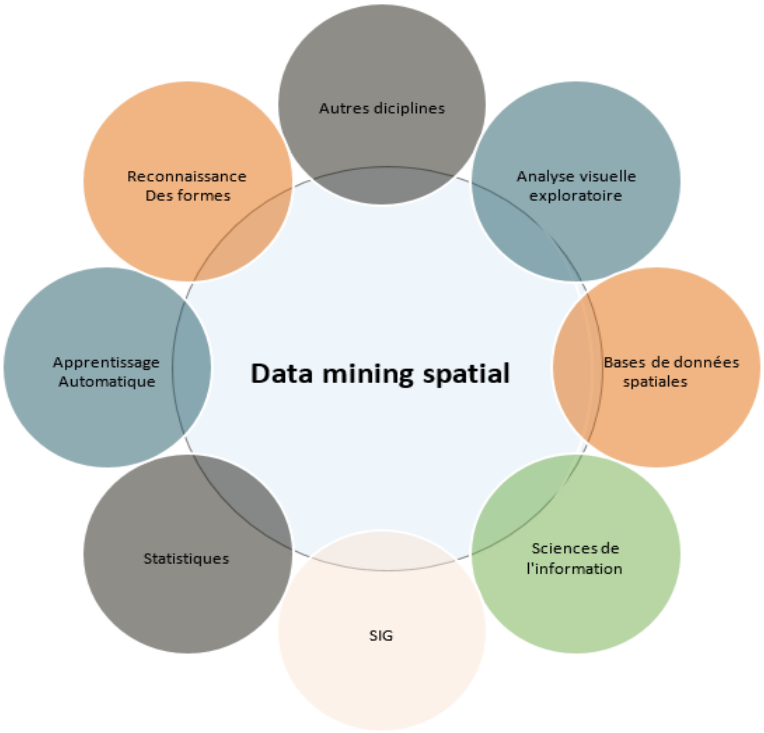


Figure 4 - Domaines connexes du data mining spatial [42]

2.5 Spécificités du data mining spatial

Les relations spatiales, l'échelle, la granularité et la hiérarchie spatiale sont des concepts clés dans le data mining spatial. Voici une brève explication de chacun de ces concepts :

Les relations spatiales : elles décrivent comment les objets sont positionnés dans l'espace par rapport les uns aux autres. Les relations spatiales peuvent être catégorisées en différentes classes telles que la distance, la direction et la topologie.

L'échelle spatiale : elle fait référence à la taille des objets ou des phénomènes étudiés. La taille des objets peut varier considérablement selon l'échelle spatiale choisie.

Par exemple, l'observation des changements climatiques peut être effectuée à l'échelle locale, régionale ou globale.

La granularité spatiale : elle se rapporte à la résolution de la carte ou à la taille de la zone étudiée. La granularité spatiale peut être déterminée en fonction de la précision et de la résolution des données géographiques utilisées.

La hiérarchie spatiale : elle décrit l'organisation spatiale des objets ou des phénomènes étudiés en fonction de leur emplacement relatif. La hiérarchie spatiale peut être utilisée pour modéliser les relations entre différents niveaux d'organisation spatiale, tels que les villes, les régions ou les pays.

En somme, la compréhension de ces concepts est essentielle pour la manipulation, l'analyse et la représentation de données géospatiales dans le data mining spatial.

2.6 Travaux sur le data mining spatial

En ce qui concerne les travaux sur le Data Mining Spatial, la communauté scientifique propose également deux approches principales : la méthode basée sur les données (BD) et l'approche statistique.

La méthode basée sur les données consiste à utiliser des données spatiales telles que des images satellites, des données GPS et des cartes géographiques pour extraire des informations utiles. Ces données peuvent être utilisées pour créer des modèles de prévision spatiale pour des applications telles que la gestion des ressources naturelles, l'aménagement du territoire et la prévention des catastrophes.

D'un autre côté, l'approche statistique repose sur l'utilisation de modèles mathématiques pour identifier les relations spatiales entre les variables. Cette approche peut être utilisée pour l'analyse de la distribution spatiale de phénomènes tels que la criminalité, la pollution ou la propagation d'une maladie.

En fin de compte, le choix de l'approche dépend des objectifs de l'étude et des données disponibles. Les deux approches peuvent contribuer à une meilleure compréhension des relations spatiales entre les variables et à l'élaboration de modèles de prévision plus précis pour des applications spatiales variées.

2.7 Approches du data mining spatial

Le Data Mining spatial est défini comme l'extraction de connaissances implicites, de relations spatiales ou d'autres propriétés non explicitement stockées dans la base de données spatiales. Ses avantages sont, d'une part, son aspect exploratoire car, contrairement à l'analyse classique, il génère des hypothèses puis les valide et, d'autre part, il permet l'intégration complète de l'information sur la localisation spatiale et des liens de voisinage.

2.7.1 Approche base de données

L'approche de base de données consiste à représenter les relations spatiales dans une BD avant l'application de la fouille de données. Dans ce qui suit, nous citons les différentes techniques de représentation des relations spatiales :

a) Les index de jointure spatiale

Les indices de jointure spatiale sont un moyen d'optimiser les coûts dans l'analyse des données spatiales. Ce système, proposé par Valduriez [22], utilise une structure de tableau qui stocke des couples d'indices de deux ensembles de données, représentant les relations spatiales telles que l'adjacence entre des objets spatiaux. La généralisation de ces indices aux données spatiales a été développée [21] elle est illustrée par la Figure 5.

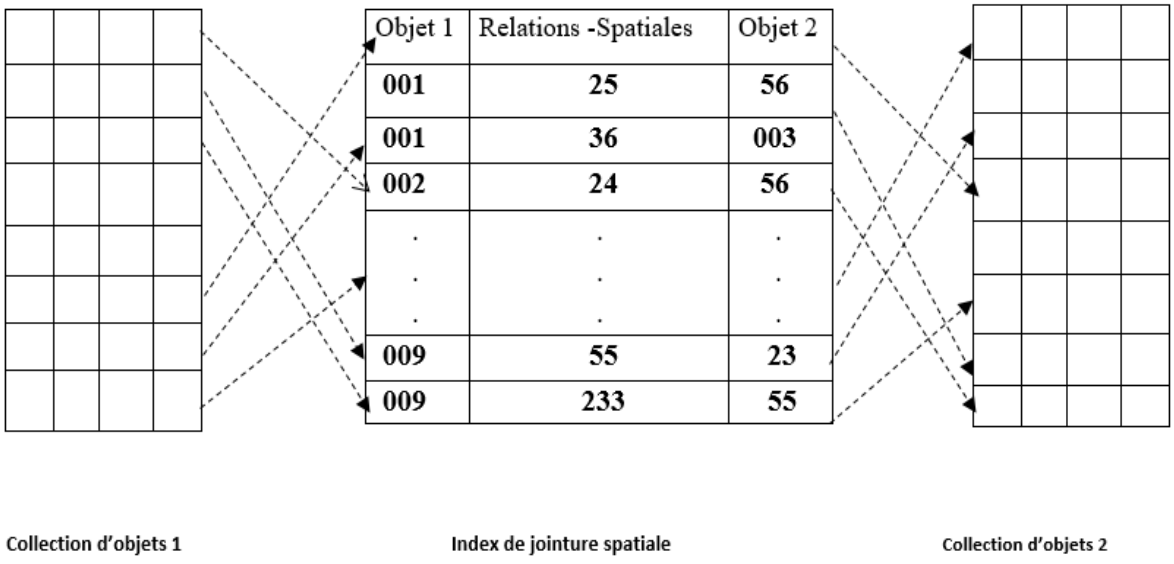


FIGURE 5 - INDEX DE JOINTURE SPATIAL [21].

b) Les graphes de voisinage et les matrices de contiguïtés

Les graphes de voisinage utilisent des nœuds pour représenter les objets spatiaux et des arcs pour représenter les relations de voisinage entre ces objets. Les matrices de contiguïté, quant à elles, représentent les relations de voisinage sous forme de matrice binaire, avec des entrées valant 1 pour indiquer l'existence d'une relation entre deux objets, et des entrées valant 0 pour indiquer l'absence de relation. Ces deux méthodes aident à formaliser les relations spatiales entre les objets, ce qui est important pour une analyse efficace des données spatiales comme illustré dans la Figure 6.

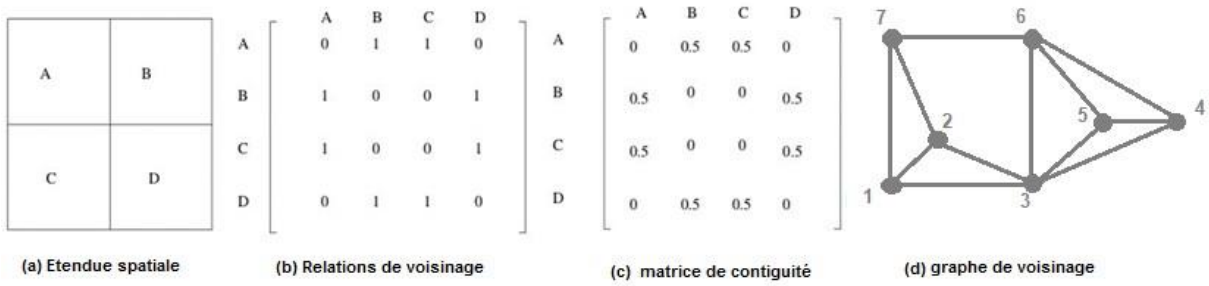


FIGURE 6 - REPRESENTATION DES RELATIONS SPATIALES PAR UNE MATRICE DE CONTIGUÏTE (C) ET UN GRAPHE DE VOISINAGE (D) [42].

c) Les data cubes spatiaux pour le data mining spatial

Les data cubes spatiaux peuvent être utilisés pour le data mining en raison de la hiérarchie spatiale des relations topologiques et des données spatiales. Les méthodes de DMS peuvent être appliquées à chaque niveau de cette hiérarchie en correspondant les relations topologiques aux niveaux hiérarchiques [23].

Han, Chee et al [24] proposent d'utiliser un système d'exploration de data mining, appelé OLAM (Online Analytical Mining), basé sur OLAP pour explorer les cubes de données spatiales à différents niveaux d'agrégation.

2.7.2 Approche statistique

Il existe plusieurs approches statistiques qui peuvent être utilisées dans le data mining spatial, notamment :

- L'analyse en composantes principales (ACP) : cette technique permet de réduire la dimensionnalité des données en trouvant les directions de variation maximale dans

les données. Elle peut être utilisée pour identifier les patterns spatiaux dans les données.

- L'analyse factorielle discriminante (AFD) : cette technique permet de séparer les groupes de données en utilisant des variables discriminantes. Elle peut être utilisée pour identifier les différences entre les groupes de données dans un espace multidimensionnel.

- L'analyse de données spatio-temporelles : cette technique permet d'analyser les données qui ont une dimension temporelle, en utilisant des modèles statistiques qui prennent en compte les changements dans le temps.

- L'analyse de régression spatiale : cette technique permet de modéliser la relation entre une variable cible et des variables explicatives spatiales. Elle peut être utilisée pour prédire la valeur de la variable cible en fonction des variables explicatives.

Il existe de nombreuses autres approches statistiques qui peuvent être utilisées dans le data mining spatial, en fonction des objectifs et des données à disposition.

2.8 Les taches du data mining spatial

Le DMS (ou système de gestion de données spatiales) est un système informatique qui permet de stocker, gérer, analyser et visualiser des données géographiques. Parmi les tâches **que peut effectuer un DMS, on peut citer :**

2.8.1 La classification spatiale :

Cette tâche consiste à regrouper des entités géographiques similaires en classes ou en catégories. Par exemple, on peut classer des parcelles de terrain en fonction de leur usage (résidentiel, commercial, industriel, etc.) ou des zones urbaines en fonction de leur densité de population.

2.8.2 La prédiction spatiale :

Cette tâche consiste à utiliser des modèles pour prédire des valeurs ou des événements futurs à des emplacements spécifiques. Par exemple, on peut prédire la qualité de l'air à un endroit donné en fonction de données historiques et de prévisions météorologiques.

2.8.3 Les règles d'association spatiale :

Cette tâche consiste à identifier des relations entre des entités géographiques. Par exemple, on peut rechercher des règles d'association entre la localisation des centres commerciaux et la densité de population dans une région donnée.

2.8.4 Le clustering spatial :

Cette tâche consiste à regrouper des entités géographiques similaires en fonction de leur proximité spatiale. Par exemple, on peut regrouper des clients en fonction de leur emplacement géographique pour optimiser la livraison de produits.

2.8.5 L'analyse des points chauds spatiaux (spatial hotspot analysis) :

Cette tâche est un type de clustering spatial consiste à identifier des zones où un phénomène spécifique est concentré de manière significative par rapport aux autres zones. Par exemple, on peut identifier des zones de forte criminalité dans une ville.

2.8.6 L'analyse de valeurs spatiales aberrantes (Spatial outlier analysis) :

Cette tâche est un type de clustering spatial consiste à identifier des entités géographiques qui sont très différentes de leur environnement. Par exemple, on peut identifier des parcelles de terrain qui ont une valeur immobilière très élevée par rapport aux parcelles environnantes.

2.9 Processus du data mining spatial

Le processus de Data Mining Spatial (DMS) commence par la préparation des données, qui implique de modéliser les relations spatiales entre les objets d'intérêt. Cette étape peut être réalisée à l'aide de différentes méthodes, telles que les matrices de contiguïté

ou les index de jointure, ou encore en stockant les relations topologiques entre les objets spatiaux dans un entrepôt de données spatiales. La figure 7 illustre ce processus de préparation des données en DMS.

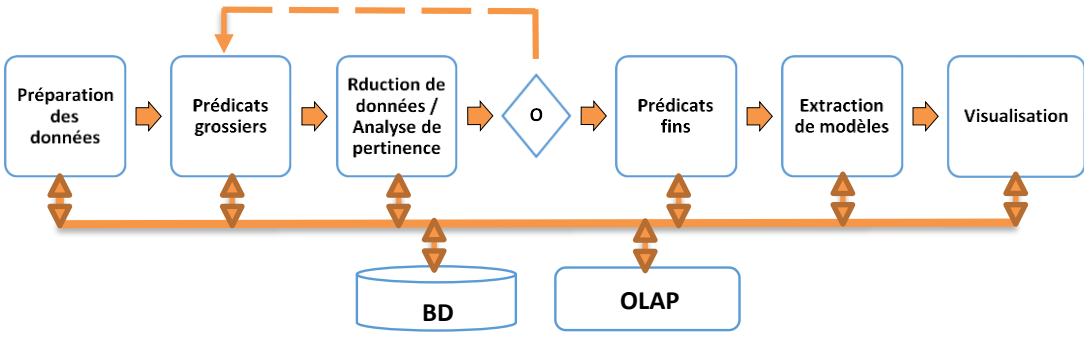


FIGURE 7- PROCESSUS DU DMS [25]

Au cours de la deuxième étape du processus, des prédicats grossiers sont calculés rapidement en utilisant le module OLAP. Les données grossières sont généralisées pour être ensuite réduites et analysées de manière itérative. Seuls les prédicats pertinents sont calculés en détail et les modèles spatiaux sont extraits à partir des prédicats prometteurs. Les résultats sont présentés à l'utilisateur à l'aide d'une interface graphique pour une visualisation des modèles. La connexion à la base de données et à OLAP peut être effectuée à chaque étape pour le stockage, la récupération des données et la généralisation. La structure du cube de données spatiales peut être utilisée dans la généralisation des données [25].

2.10 Domaines d'application du data mining spatial

Le data mining spatial peut être utilisé dans de nombreux domaines différents, notamment :

- **La géographie :** le data mining spatial peut être utilisé pour étudier les patterns spatiaux dans les données géographiques, comme la répartition de la population ou la distribution des ressources naturelles.
- **La géomatique :** le data mining spatial peut être utilisé pour analyser et manipuler des données géographiques, comme les cartes et les images satellites.

- **La géo-analyse :** le data mining spatial peut être utilisé pour analyser les données géographiques afin de mieux comprendre les phénomènes spatiaux et de prendre des décisions informées.

- **La planification territoriale :** le data mining spatial peut être utilisé pour étudier les patterns spatiaux dans les données sur les villes et les régions, afin de mieux planifier l'utilisation des terres et des ressources.

- **L'analyse des réseaux :** le data mining spatial peut être utilisé pour analyser les données sur les réseaux de transports, les réseaux de communication et les réseaux de distribution, afin de mieux comprendre comment ils fonctionnent et comment ils peuvent être optimisés.

- **L'épidémiologie :** Les épidémiologistes utilisent le DMS pour stocker et gérer des données de santé, telles que des données sur les maladies, les hospitalisations, les décès, les facteurs de risque, etc. Le DMS permet d'assurer la qualité des données, de faciliter l'analyse des données, de produire des rapports et des visualisations pour la surveillance de la santé publique, la planification et l'évaluation des programmes de santé, et la recherche en épidémiologie. Le DMS permet également de partager les données de manière sécurisée et confidentielle avec d'autres chercheurs, épidémiologistes, et décideurs en santé publique.

Il existe de nombreux autres domaines d'application possibles pour le data mining spatial, en fonction des données et des objectifs à disposition.

Dans ce chapitre, nous avons commencé par décrire l'information spatiale, ses spécificités, ses formats de représentation. Nous avons ensuite parlé de l'importance des relations spatiales dans le processus de DMS. Puis nous avons donné une définition de DMS, et détaillé son processus, ses phases d'exécutions, ses approches, et ses méthodes.

2.11 Exemples d'application du data mining spatial pour le cancer

Le data mining spatial (DMS) peut être utilisé dans l'analyse des données de cancer pour étudier la répartition de la maladie dans l'espace et pour identifier des facteurs de risque spatiaux. Voici plusieurs exemples d'utilisation du data mining spatial (DMS) pour étudier le cancer :

- **Détection de hotspots :** Le DMS a été utilisé pour identifier des hotspots d'incidence ou de mortalité liés au cancer, ce qui permet d'identifier les zones où le cancer est particulièrement prévalent ou où des interventions pourraient être nécessaires. Par exemple, une étude a utilisé le DMS pour identifier des hotspots de mortalité par cancer du poumon aux États-Unis, révélant des zones où le cancer du poumon est plus fréquent que prévu sur la base des taux nationaux [26]. Une autre étude a utilisé le DMS pour identifier des hotspots d'incidence de cancer du sein en Iran, ce qui a permis d'identifier des régions présentant des taux de cancer du sein plus élevés que prévu [27].

- **Classification spatiale :** Les chercheurs ont utilisé des algorithmes de classification spatiale pour prédire l'incidence du cancer du sein en fonction de facteurs environnementaux et démographiques. Par exemple, une étude menée à New York a utilisé un modèle de classification spatiale pour prédire l'incidence du cancer du sein en fonction de facteurs tels que l'âge, la race, l'éducation et les expositions environnementales. Le modèle a pu prédire avec précision l'incidence du cancer du sein dans différents quartiers de la ville [28].

- **Prédiction spatiale :** Dans une étude menée en Chine, les chercheurs ont utilisé des techniques de prédiction spatiale pour estimer le risque de cancer du poumon en fonction de facteurs environnementaux tels que la pollution de l'air et la prévalence du tabagisme. Les chercheurs ont utilisé un modèle de régression spatiale pour prédire l'incidence du cancer du poumon dans différentes régions du pays [29].

- **Règles d'association spatiale :** Une étude menée au Brésil a utilisé des règles d'association spatiale pour identifier les schémas spatiaux de l'incidence du cancer de

l'estomac. Les chercheurs ont identifié des associations significatives entre l'incidence du cancer de l'estomac et des facteurs environnementaux tels que le type de sol, l'altitude et les précipitations [30].

- **Analyse spatiale des points chauds :** Une étude menée en Corée du Sud a utilisé l'analyse spatiale des points chauds pour identifier les zones à forte incidence de cancer du foie. Les chercheurs ont utilisé une statistique locale de Moran I pour identifier les grappes de forte incidence de cancer du foie dans différentes régions du pays [31].

- **Analyse des valeurs aberrantes spatiales :** dans une étude menée en Suède, les chercheurs ont utilisé l'analyse des valeurs aberrantes spatiales pour identifier les zones où l'incidence du cancer de la prostate était significativement plus élevée ou plus faible que prévu. Les chercheurs ont utilisé un modèle de régression spatiale pour identifier les zones où l'incidence du cancer de la prostate était significativement différente de l'incidence moyenne dans la région [32].

- **Regroupement spatial :** Le DMS peut également être utilisé pour identifier des groupes spatiaux de cas de cancer, ce qui peut aider à identifier les facteurs de risque environnementaux ou génétiques. Par exemple, une étude a utilisé le DMS pour identifier des clusters de cas de cancer colorectal en Italie, révélant que plusieurs clusters étaient situés près de zones industrielles présentant des expositions potentiellement cancérogènes [33]. Une autre étude a utilisé le DMS pour identifier des clusters de cas de cancer du sein aux États-Unis, révélant que plusieurs clusters étaient situés près de zones industrielles ou agricoles [34].

- **Régression spatiale :** Le DMS peut également être utilisé pour modéliser la relation entre l'incidence ou la mortalité du cancer et les facteurs environnementaux ou démographiques. Par exemple, une étude a utilisé le DMS pour modéliser la relation entre la mortalité par cancer du poumon et la pollution atmosphérique aux États-Unis, révélant que des niveaux élevés de matières particulaires étaient associés à des taux de mortalité par cancer du poumon plus élevés [26]. Une autre étude a utilisé le DMS pour modéliser la relation entre l'incidence de cancer du sein et les facteurs socio-économiques en Iran,

révélant que des niveaux plus élevés d'éducation et de revenu étaient associés à des taux de cancer du sein plus faibles (Soltanian et al., 2019).

- **Visualisation géospatiale :** Le DMS peut également être utilisé pour créer des cartes ou d'autres visualisations de données sur le cancer, ce qui peut aider à identifier des modèles ou des tendances. Par exemple, une étude a utilisé le DMS pour créer une carte de l'incidence de cancer de la prostate au Canada, révélant des taux plus élevés dans certaines régions [35]. Une autre étude a utilisé le DMS pour créer une carte de la mortalité par cancer de la vessie en Espagne, révélant des taux plus élevés dans certaines provinces avec des activités industrielles ou minières [36].

- **Classification et prédiction :** Le DMS peut également être utilisé pour classer ou prédire des types de cancer ou des résultats de traitement en fonction de caractéristiques spatiales ou géospatiales. Par exemple, une étude a utilisé le DMS pour classifier les tumeurs du sein en fonction de leur histologie et de leur grade, en utilisant des images de tissus et des données spatiales pour identifier les caractéristiques de la tumeur associées à différents sous-types de cancer du sein [37]. Une autre étude a utilisé le DMS pour prédire les résultats du traitement du cancer de la prostate en utilisant des données sur l'âge, la race, le stade de la maladie et d'autres facteurs, ainsi que des données spatiales sur la densité de population et les niveaux de pollution [38].

- **Prédiction de l'incidence du cancer :** Le DMS peut également être utilisé pour prédire l'incidence du cancer en fonction de divers facteurs géospatiaux, tels que la densité de population, l'exposition à la pollution, les niveaux de rayonnement et d'autres variables environnementales. Une étude a utilisé des données spatiales pour prédire l'incidence du cancer du poumon dans le comté de Los Angeles en utilisant des modèles de régression spatiale pour examiner les relations entre les niveaux de pollution de l'air et l'incidence du cancer du poumon [39].

- **Évaluation des services de soins de santé :** Le DMS peut également être utilisé pour évaluer les services de soins de santé en fonction de la disponibilité géographique des services de diagnostic, de traitement et de soutien. Par exemple, une étude a utilisé le DMS pour examiner les tendances spatiales de la disponibilité des services de radiothérapie dans

le nord de la Norvège, en utilisant des données géospatiales pour identifier les zones où la disponibilité des services était insuffisante [40].

- **Étude de l'impact des facteurs de risque environnementaux :** Le DMS peut également être utilisé pour étudier l'impact des facteurs de risque environnementaux sur l'incidence du cancer. Par exemple, une étude a utilisé le DMS pour examiner les relations entre l'exposition aux pesticides et l'incidence du cancer de la prostate en Californie, en utilisant des données spatiales pour identifier les zones où les niveaux d'exposition étaient les plus élevés et en utilisant des modèles de régression pour évaluer l'impact de l'exposition sur l'incidence du cancer [41].

- **Suivi de la progression du cancer :** Le DMS peut également être utilisé pour suivre la progression du cancer en utilisant des images spatiales pour surveiller la croissance tumorale et l'efficacité du traitement. Par exemple, une étude a utilisé le DMS pour suivre la progression du cancer de la prostate en utilisant des images IRM pour détecter les changements de volume tumoral au fil du temps, ce qui peut aider à évaluer l'efficacité du traitement [42].

2.12 Conclusion

Les méthodes traditionnelles de Data Mining ne conviennent pas aux données spatiales en raison de leurs caractéristiques uniques. Les données spatiales sont multidimensionnelles, présentent une autocorrélation spatiale et sont hétérogènes. C'est pourquoi il est nécessaire d'utiliser des méthodes spécifiques de Data Mining spatial pour tenir compte de ces particularités.

Le Data Mining Spatial (DMS) est une technique qui permet d'extraire des informations à partir de données géolocalisées. Dans la lutte contre le cancer, le DMS a été utilisé pour identifier les zones présentant une forte incidence de certains types de cancer, ce qui permet aux autorités de santé de concentrer leurs efforts de prévention et de dépistage.

Le DMS a également été utilisé pour identifier des clusters de cas de cancer, regroupant des personnes présentant des cas similaires dans une même zone géographique. Cela aide les

chercheurs à identifier les facteurs de risque environnementaux ou socio-économiques associés à ces clusters.

Chapitre 3

Méthodologie

3.1 Introduction

Dans cette section, nous aborderons le choix de notre étude, qui vise à développer une application générique capable d'appliquer les principales techniques de data mining spatial, telles que le clustering, la classification, la détection de hotspots et d'outliers, à différents types de cancer. Cependant, nous nous concentrerons spécifiquement sur le cancer de l'estomac.

L'objectif principal de notre application est de fournir une solution polyvalente et adaptable pour l'analyse de données spatiales liées au cancer. Elle sera conçue de manière à pouvoir traiter et analyser différentes sources de données, mais dans notre étude, nous nous concentrerons sur l'incidence et la mortalité du cancer de l'estomac comme variables d'analyse.

L'une des caractéristiques clés de notre application est sa capacité à être utilisée à différentes échelles géographiques. Elle est conçue pour fonctionner avec n'importe quelle échelle géographique, mais notre étude sera réalisée au niveau des États-Unis. Cela nous permettra d'explorer la répartition spatiale du cancer de l'estomac à travers les différents États et d'identifier les schémas et les variations géographiques.

3.2 Description des données épidémiologiques utilisées

Les données épidémiologiques utilisées dans cette étude concernent le cancer. Les types de données recueillies peuvent inclure le nombre de cas de cancer dans différentes régions géographiques, les taux d'incidence, les types de cancer spécifiques, les caractéristiques démographiques des patients atteints de cancer, etc.

Les sources de données peuvent varier, mais elles peuvent provenir de registres de cancer, d'études épidémiologiques, de bases de données de santé publique, d'hôpitaux ou de centres de recherche. La procédure d'acquisition des données peut impliquer des processus de collecte, de compilation et de normalisation à partir de différentes sources.

Les caractéristiques des données peuvent inclure des variables telles que l'âge, le sexe, la localisation géographique, les statistiques de santé, les caractéristiques cliniques spécifiques au cancer, etc. Les formats de données peuvent être des fichiers CSV.

3.3 Prétraitement des données spatiales

Le prétraitement des données attributaires et spatiales est essentiel pour assurer la qualité et l'intégrité des données utilisées dans l'analyse spatiale. Dans notre étude nous suivons les étapes de prétraitement des données suivantes :

- **Le nettoyage des données :** élimination des valeurs aberrantes, des doublons ou des valeurs incohérentes qui pourraient fausser les résultats de l'analyse.

- **La gestion des valeurs manquantes :** traitement des valeurs manquantes en les supprimant, en les remplaçant par des valeurs estimées ou en utilisant des techniques d'imputation.

- **Normalisation des données** : mise à l'échelle des données pour les rendre comparables.

- **Intégration des données spatiales et attributaires :** fusion des données géographiques et des attributs pour une analyse conjointe.

- **Validation des données :** vérification de l'intégrité des données et identification des éventuelles erreurs ou incohérences.

- **Préparation des données pour l'analyse :** structuration des données dans un format approprié pour les techniques de classification spatiale, de clustering, de détection des hotspots et des outliers.

3.4 Choix des analyses et des techniques de data mining spatial

Dans le contexte de l'étude épidémiologique du cancer, différentes techniques de data mining spatial peuvent être utilisées en fonction des objectifs de recherche spécifiques. Dans notre étude nous proposons les techniques de data mining spatiales suivantes :

3.4.1 Classification spatiale :

La classification spatiale permet de prédire ou de classifier des données spatiales en fonction de caractéristiques spatiales ou attributaires. Dans le cas de l'étude épidémiologique du cancer, la classification spatiale peut être utilisée pour cartographier les niveaux de risque de cancer dans différentes régions. Par exemple, en utilisant des données démographiques, environnementales et de mode de vie, on peut construire un modèle de classification qui attribue un niveau de risque de cancer à chaque zone géographique. Cela permet d'identifier les zones à risque élevé et de prendre des mesures préventives appropriées.

L'algorithme de classification utilisé dans notre étude est basé sur les SVM (Support Vector Machine), un outil couramment utilisé dans l'étude épidémiologique du cancer. Les SVM sont des algorithmes d'apprentissage supervisé qui permettent la classification et la régression.

L'algorithme de classification SVM (Support Vector Machine) fonctionne de la manière suivante :

1. **Prétraitement des données** : Tout d'abord, les données doivent être collectées et préparées pour l'analyse. Cela peut inclure la normalisation des données, l'élimination des valeurs aberrantes, le traitement des données manquantes et la sélection des caractéristiques pertinentes.

2. **Séparation des données** : Les données sont divisées en deux ensembles : un ensemble d'apprentissage et un ensemble de test. L'ensemble d'apprentissage est utilisé pour entraîner le modèle SVM, tandis que l'ensemble de test est utilisé pour évaluer les performances du modèle.

3. **Sélection du noyau** : La SVM utilise un noyau pour effectuer une transformation non linéaire des données d'entrée dans un espace de dimension supérieure, où il est plus facile de trouver une séparation linéaire optimale. Les noyaux couramment utilisés sont le noyau linéaire, le noyau polynomial et le noyau RBF (Radial Basis Function).

4. **Entraînement du modèle** : Le modèle SVM recherche une frontière de décision (hyperplan) qui sépare les différentes classes de manière optimale. L'objectif est de maximiser la marge entre les exemples de différentes classes et l'hyperplan. Cela implique de résoudre un problème d'optimisation convexe, où l'algorithme tente de trouver les vecteurs de support (support vectors) qui définissent l'hyperplan optimal.

5. **Classification** : Une fois le modèle SVM entraîné, il peut être utilisé pour prédire la classe d'un nouvel exemple. Le modèle attribue une étiquette de classe en fonction de la position de l'exemple par rapport à l'hyperplan appris lors de l'entraînement. Si l'exemple se situe du côté positif de l'hyperplan, il est classé dans une classe spécifique, sinon il est classé dans l'autre classe.

Les calculs spécifiques impliqués dans l'entraînement d'un modèle SVM incluent l'optimisation de la fonction de marge, la résolution de problèmes d'optimisation quadratique (pour les SVM linéaires) ou de problèmes d'optimisation non linéaire (pour les SVM non linéaires), et la détermination des vecteurs de support.

Il est important de noter que la mise en œuvre détaillée de l'algorithme SVM peut varier en fonction des bibliothèques ou des outils utilisés. Les bibliothèques couramment utilisées pour les SVM comprennent libsvm, scikit-learn, et SVMlight, qui fournissent des implémentations optimisées de l'algorithme.

3.4.2 Analyse de clusters :

L'analyse de clusters permet d'identifier des regroupements spatiaux de cas de cancer similaires en fonction de leurs caractéristiques spatiales ou attributaires. Cette technique est utile pour détecter des concentrations de cas de cancer dans certaines régions géographiques. Par exemple, elle peut aider à identifier des zones où le cancer du poumon

est plus fréquent en raison de l'exposition à des agents carcinogènes spécifiques présents dans l'environnement. L'analyse de clusters peut être utilisée pour explorer les facteurs de risque géographiques et pour guider les efforts de prévention et de traitement ciblés.

Dans notre étude, nous avons utilisé l'algorithme K-means qui est couramment utilisés dans l'étude épidémiologique du cancer. Cet algorithme permet d'identifier des groupes similaires de cas de cancer en fonction de leurs caractéristiques.

L'algorithme K-means fonctionne de la manière suivante :

1. **Initialisation :** Le nombre de clusters K est spécifié, et K centres initiaux sont choisis de manière aléatoire ou selon une méthode spécifique.

2. **Attribution des points :** Chaque point de données est attribué au centre de cluster le plus proche, en utilisant une mesure de similarité telle que la distance euclidienne.

3. **Mise à jour des centres :** Les centres des clusters sont recalculés en prenant la moyenne des points attribués à chaque cluster. Cela déplace les centres vers les positions moyennes des points appartenant à chaque cluster.

4. **Répétition des étapes 2 et 3 :** Les étapes d'attribution et de mise à jour des centres sont répétés jusqu'à ce que les centres convergent vers des positions stables. Cela signifie que les points ne changent plus de clusters de manière significative et que les centres ne se déplacent plus de manière significative.

5. **Convergence :** L'algorithme se termine lorsque les centres des clusters se stabilisent, et les points sont finalement attribués de manière définitive à leur cluster respectif.

   Le but de l'algorithme K-means est de minimiser la variance intra-cluster, c'est-à-dire la variance des points à l'intérieur de chaque cluster, tout en maintenant une séparation maximale entre les différents clusters. Pour ce faire, des mesures de similarité telles que la distance euclidienne, la similarité des attributs et la similarité spatiale peuvent être utilisées.

Les critères de regroupement, tels que la distance maximale entre les membres d'un cluster ou le nombre de clusters souhaité, sont également définis. Les résultats du clustering

sont ensuite analysés et interprétés en examinant les caractéristiques communes des cas regroupés dans chaque cluster. Cette analyse permet d'identifier des sous-populations présentant des caractéristiques similaires de cancer, ce qui peut contribuer à la compréhension des facteurs de risque et des traitements. Dans notre étude, nous ne considérons pas la distance euclidienne mais uniquement la similarité des attributs.

3.4.3 Détection de hotspots et cold spots

La détection de hotspots vise à identifier les zones où l'incidence du cancer est significativement plus élevée que prévu. Cette technique permet de repérer les régions où la prévalence du cancer est anormalement élevée par rapport à la moyenne régionale ou nationale. Elle peut aider à mettre en évidence des clusters spatiaux de cas de cancer qui nécessitent une attention particulière en termes de surveillance et d'intervention. La détection de hotspots est utile pour identifier les zones géographiques où des facteurs de risque spécifiques peuvent être présents ou où des inégalités en matière de santé sont observées.

Dans notre étude, nous avons utilisé la statistique d'autocorrélation spatiale I de Moran local pour détecter les hotspots et les cold spots. Cette mesure évalue le degré de similarité spatiale entre les valeurs d'une variable dans une région donnée et celles de ses régions voisines.

L'algorithme d'identification des points chauds et des points froids basé sur la statistique d'autocorrélation spatiale I de Moran local fonctionne de la manière suivante :

1. **Calcul de l'autocorrélation spatiale :** Tout d'abord, les valeurs de la variable d'intérêt sont attribuées à des emplacements spécifiques dans l'espace. Ensuite, pour chaque emplacement, la statistique d'autocorrélation spatiale I de Moran local est calculée. Cette statistique mesure à la fois la similarité entre les valeurs de la variable à l'emplacement donné et celles dans les emplacements voisins, ainsi que l'agencement spatial global des valeurs dans la région étudiée.

2. **Interprétation de la statistique d'autocorrélation :** La statistique d'autocorrélation spatiale I de Moran génère une valeur entre -1 et 1. Une valeur positive indique une autocorrélation spatiale positive, ce qui signifie que les valeurs similaires tendent à être

regroupées spatialement (points chauds). Une valeur négative indique une autocorrélation spatiale négative, ce qui signifie que les valeurs similaires tendent à être dispersées spatialement (points froids). Une valeur proche de zéro indique une absence d'autocorrélation spatiale (répartition aléatoire).

3. **Test de significativité :** Pour évaluer si la structure spatiale observée est statistiquement significative, un test de significativité est généralement effectué. Ce test compare la statistique d'autocorrélation spatiale I de Moran observée à une distribution nulle générée par permutation aléatoire des valeurs. Si la statistique observée est significativement différente de la distribution nulle, cela suggère qu'il y a une structure spatiale réelle.

4. **Identification des points chauds et des points froids :** En se basant sur les valeurs de la statistique d'autocorrélation spatiale I de Moran local et leur significativité, les hotspots et cold spots peuvent être identifiés. Les points avec une autocorrélation spatiale positive significative sont considérés comme des points chauds, indiquant des concentrations spatiales élevées des valeurs de la variable d'intérêt. Les points avec une autocorrélation spatiale négative significative sont considérés comme des points froids, indiquant une dispersion spatiale des valeurs de la variable d'intérêt.

Dans notre étude, deux mesures de signification statistique sont utilisées pour évaluer la significativité des résultats de l'analyse des hotspots et cold spots :

a) Indice I de Moran local (local Moran's I) :

- L'indice I de Moran local mesure l'autocorrélation spatiale locale pour chaque observation dans le jeu de données. Il quantifie à quel point une observation est entourée de valeurs similaires (cluster) ou de valeurs différentes (outliers) dans l'espace.

- Dans ce programme, l'indice I de Moran local est calculé à l'aide de la fonction Moran_Local () de la bibliothèque PySAL. Cela génère des valeurs d'indice I de Moran local (moran_loc_values) pour chaque observation.

b) Valeurs p associées à l'indice I de Moran local :

- Les valeurs p fournissent une mesure de la significativité statistique de l'indice I de Moran local pour chaque observation.

- Dans ce programme, les valeurs p sont calculées à l'aide de la fonction p_sim de l'objet Moran Local (). Ces valeurs p (moran_loc_p_values) sont associées à chaque observation et indiquent si l'autocorrélation spatiale locale est statistiquement significative.

Ces mesures de signification statistique permettent d'interpréter les résultats de l'analyse des hotspots et cold spots en identifiant les clusters ou les outliers spatiaux significatifs dans les données. Les hotspots et cold spots identifiés peuvent être cartographiés à l'aide de techniques de visualisation telles que des cartes choroplèthes ou des cartes de densité. Les résultats peuvent être interprétés en identifiant les zones géographiques présentant des concentrations élevées ou faibles de cas de cancer, en les comparant aux facteurs de risque connus, etc.

3.4.4 Détection d'outliers :

La détection d'outliers consiste à identifier des observations individuelles qui diffèrent considérablement de la tendance générale des données. Dans le contexte de l'étude épidémiologique du cancer, cela peut être utilisé pour repérer des cas individuels de cancer qui présentent des caractéristiques inhabituelles ou des facteurs de risque atypiques. Cela peut aider à détecter des cas de cancer rares ou inhabituels qui pourraient nécessiter une attention médicale spécifique ou une recherche approfondie.

Dans notre étude, nous avons utilisé l'indice de Moran global (Global Moran's I). Cet indice de corrélation spatiale globale évalue la similarité spatiale entre les valeurs d'une variable dans une région et celles de ses voisins. Les valeurs qui diffèrent considérablement des valeurs de leurs voisins peuvent être identifiées comme des outliers.

L'algorithme de détection des outliers spatiaux basé sur l'indice de Moran global est largement utilisé dans divers domaines tels que l'épidémiologie, l'écologie, la géographie et l'analyse des données spatiales. Il permet d'identifier les observations atypiques qui diffèrent significativement de leur environnement spatial. Cette approche est précieuse pour

repérer les zones à risque, détecter des clusters anormaux et comprendre les motifs spatiaux des variables étudiées.

L'algorithme de détection des outliers spatiaux basé sur Moran's I global fonctionne de la manière suivante :

1. **Attribution des valeurs :** Les valeurs de la variable d'intérêt sont attribuées à des emplacements spécifiques dans l'espace.

2. **Calcul de Global Moran's I :** Pour chaque emplacement, le coefficient de corrélation spatiale Moran's I est calculé. Ce coefficient est basé sur la covariance entre les valeurs de la variable à l'emplacement donné et celles dans les emplacements voisins, normalisée par la variance globale de la variable.

3. **Interprétation de Moran's I :** Le coefficient de corrélation spatiale Global Moran's I génère une valeur entre -1 et 1. Une valeur positive indique une autocorrélation spatiale positive, ce qui signifie que les valeurs similaires tendent à être regroupées spatialement. Une valeur négative indique une autocorrélation spatiale négative, ce qui signifie que les valeurs similaires tendent à être dispersées spatialement. Une valeur proche de zéro indique une absence d'autocorrélation spatiale.

4. **Test de significativité :** Pour évaluer si la structure spatiale observée est statistiquement significative, un test de significativité est généralement effectué. Ce test compare le coefficient de corrélation spatiale Global Moran's I observé à une distribution nulle générée par permutation aléatoire des valeurs. Si le coefficient observé est significativement différent de la distribution nulle, cela suggère la présence d'une structure spatiale réelle.

5. **Détection des outliers spatiaux :** En se basant sur les valeurs de Global Moran's I et leur significativité, les outliers spatiaux peuvent être identifiés. Les observations ayant des valeurs extrêmes et une autocorrélation spatiale significative négative ou positive peuvent être considérées comme des outliers spatiaux.

3.5 Architecture logicielle

L'architecture logicielle de l'application repose sur la séparation des préoccupations et l'utilisation de modules spécialisés pour la manipulation des données, l'analyse spatiale, les techniques de data mining spatial et la visualisation. Les choix de conception et les algorithmes utilisés sont guidés par les objectifs spécifiques de l'application visant à appliquer des techniques de data mining spatial au cancer. L'application comprend les éléments suivants :

- **Interfaces utilisateur :** Ces interfaces permettent à l'utilisateur de sélectionner les fichiers de données et d'interagir avec les fonctionnalités de l'application.

- **Traitement des données :** Les modules utilisés permettent de manipuler et d'analyser les données tabulaires et géospatiales. Ils offrent des fonctionnalités pour la lecture des fichiers de données, le nettoyage, la transformation et la manipulation des données.

- **Analyse spatiale :** Des modules spécialisés sont utilisés pour effectuer des analyses spatiales avancées. Ils permettent la création de matrices de pondération spatiale et le calcul des différents indices liés aux techniques de data mining spatial.

- **Implémentation des algorithmes de data mining spatial :** Un module est utilisé pour l'implémentation les algorithmes de data mining spatial.

- **Visualisation des données :** Un module dédié est utilisé pour créer des cartes interactives et des visualisations des résultats. Il permet d'afficher les données géospatiales sur une carte, d'ajouter des éléments supplémentaires et de générer des cartes interactives.

- **Autres fonctionnalités :** L'application utilise également des fonctionnalités de normalisation des données, d'ouverture de fichiers dans un navigateur et d'affichage d'images.

Ces modules spécialisés sont sélectionnés en fonction de leurs fonctionnalités et de leur compatibilité avec les objectifs de l'application.

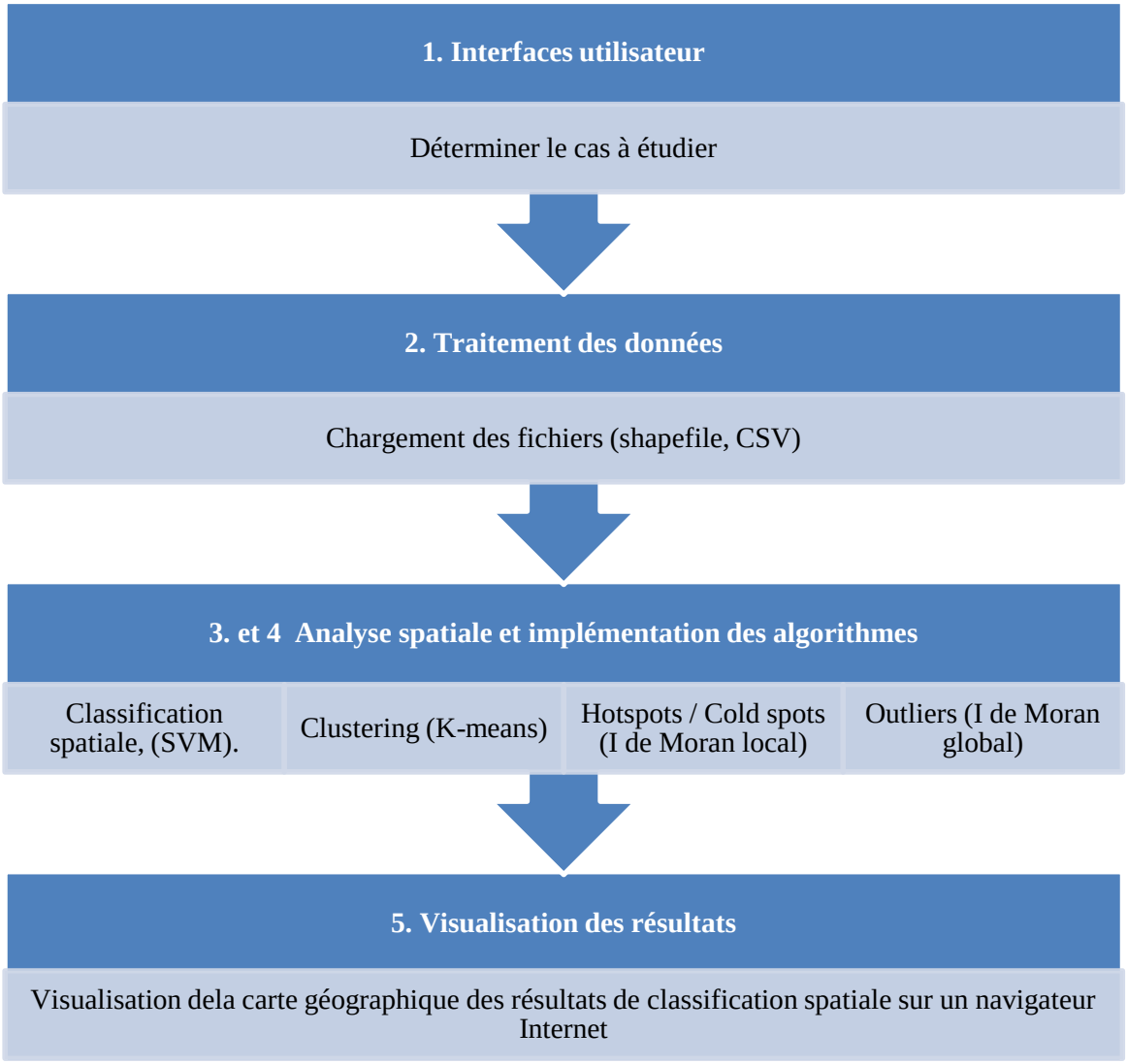


FIGURE 8 : PRESENTATION DE LA METHODOLOGIE SUIVIE POUR NOTRE ETUDE DE CAS

3.6 Méthodes d'évaluation et d'interprétation des résultats

L'évaluation des résultats de la classification, du clustering, de la détection des hotspots et des outliers est cruciale pour évaluer la pertinence et la fiabilité des analyses spatiales réalisées. Dans notre étude, nous privilégions principalement une approche visuelle en examinant les résultats affichés sur des cartes géographiques. Cette méthode s'avère efficace pour visualiser les schémas et les tendances spatiales, ainsi que pour identifier les régions d'intérêt et faciliter l'interprétation des résultats dans le contexte épidémiologique du cancer de l'estomac.

Toutefois, il est important de noter que l'évaluation basée exclusivement sur l'observation visuelle des cartes présentent certaines limites. Cette approche peut être subjective et dépendante de l'interprétation individuelle. De plus, elle ne permet pas d'obtenir des mesures quantitatives pour évaluer la performance des méthodes utilisées.

Dans notre étude de cas, nous nous limitons à cette méthode d'évaluation, étant donné que nous disposons uniquement de données expérimentales incomplètes. Il est donc important de prendre en compte ces limitations lors de l'interprétation des résultats obtenus.

3.7 Conclusion

Ce chapitre a exposé en détail la méthodologie utilisée dans notre étude. Nous avons décrit les données épidémiologiques utilisées, les étapes de prétraitement des données spatiales, les techniques de data mining spatial sélectionnées, le processus de développement de l'application et les méthodes d'évaluation des résultats. Cette méthodologie nous permettra d'analyser les données épidémiologiques et d'obtenir des informations précieuses sur la répartition spatiale des maladies.

Chapitre 4

Expérimentation

4.1 Présentation des Données

Les données utilisées dans cette étude comprennent des informations sur le type de cancer, les taux ajustés selon l'âge, les cas et la population dans différentes régions des États-Unis entre 2015 et 2019. Elles couvrent des aspects tels que le type de cancer (en l'occurrence l'estomac), le sexe (hommes et femmes) et la race (toutes les races et ethnies). Ces données permettent de comparer les taux de cancer ajustés selon l'âge, le nombre de cas et la population dans un total de 52 régions différentes. Les limites administratives des États sont fournies au format SHP, tandis que les autres données sont disponibles en format CSV. La source de ces données est le site Web du Centre américain de contrôle et de prévention des maladies[1].

4.2 Présentation de l'environnement expérimental

L'environnement expérimental utilisé pour cette étude était basé sur un PC Nitro AN515-515 (Acer) équipé du système d'exploitation Windows 10 Famille en version 64 bits. Le processeur utilisé était un Intel(R) Core (TM) i5-7300U CPU cadencé à 2,50 GHz. Le système disposait de 8192 Mo de RAM.

Pour le développement et l'exécution des expériences, nous avons utilisé le langage Python avec l'environnement de développement PyCharm.

- **Python :** Un langage de programmation polyvalent, clair et concis, utilisé dans de nombreux domaines tels que le développement web, l'analyse de données et l'apprentissage

[1]Center for Disease Control and Prevention. https://gis.cdc.gov/Cancer/USCS/#/AtAGlance/.

automatique. Il offre une bibliothèque standard étendue et est apprécié pour sa simplicité et sa puissance. Nous avons utilisé les bibliothèques suivantes :

- **Pandas :** Bibliothèque Python pour la manipulation et l'analyse de données, fournissant des structures de données flexibles et performantes pour travailler avec des données tabulaires.

- **GeoPandas :** Extension de la bibliothèque Pandas pour travailler avec des données géospatiales, permettant la manipulation et l'analyse de données géographiques avec des fonctionnalités de visualisation et d'analyse spatiale intégrées.

- **Scikit-learn :** Bibliothèque Python pour l'apprentissage automatique, offrant des outils et des algorithmes pour effectuer des tâches telles que la classification, la régression, le regroupement et la réduction de dimensionnalité.

- **Folium :** Bibliothèque Python pour la création de cartes interactives et visuelles, utilisant les données géospatiales. Elle permet de visualiser des données sur des cartes interactives dans un navigateur web, en utilisant des tuiles et des couches superposables pour afficher des informations géographiques de manière dynamique et intuitive.

- **ESDA (Exploratory Spatial Data Analysis) :** Module de PySAL pour l'exploration des données spatiales et la détection de schémas spatiaux.

- **PySAL (Python Spatial Analysis Library) :** Bibliothèque Python pour l'analyse spatiale avancée.

- **PyCharm :** Un environnement de développement intégré (IDE) pour le langage de programmation Python. Il offre des fonctionnalités avancées telles que la complétion automatique du code, le débogage, la gestion de projets, la navigation facilitée dans le code source et l'intégration avec des outils de visionnement. PyCharm facilite le développement Python en fournissant une interface conviviale et des fonctionnalités puissantes pour améliorer la productivité des développeurs.

TABLEAU 2- BIBLIOTHEQUES PYTHON UTILISEES POUR LA REALISATION DES DIFFERENTES ANALYSES

	Classification spatiale	Clustering spatial	Détection des hotspots et cold spots spatiaux	Détection des outliers spatiaux
Pandas	X	X	X	X
Geopandas	X	X	X	X
Scikit-learn	X	X	X	X
Folium	X	X	X	X
Esda			X	X
Pysal			X	X

Dans ce qui suit, nous expliquons comment les différentes parties de l'application ont été réalisées :

- **Les interface utilisateur :** L'application utilise la bibliothèque :
    - Tkinter pour créer une interface utilisateur conviviale permettant à l'utilisateur de sélectionner les fichiers de données et d'interagir avec les différentes fonctionnalités de l'application.
- **Traitement des données :** Les bibliothèques pandas et geopandas sont utilisées pour manipuler et analyser les données tabulaires et géospatiales :
    - Pandas est utilisé pour lire les fichiers de données, effectuer des opérations de nettoyage et de transformation des données.
    - Geopandas est utilisé pour manipuler des données géospatiales telles que les cartes et les polygones.
- **Analyse spatiale :** L'application utilise les bibliothèques Pysal et esda pour effectuer des analyses spatiales avancées.

- Pysal fournit des fonctionnalités pour créer des matrices de pondération spatiale basées sur la contiguïté des polygones, qui sont utilisées dans les analyses spatiales.

- Esda est utilisé pour calculer l'indice de Moran local, qui permet de détecter les clusters spatiaux de valeurs similaires dans les données.

- **Implémentation des algorithmes de data mining spatial :**

  - La bibliothèque scikit-learn permet d'effectuer des analyses de clustering sur les données. L'algorithme K-means regroupe les données en fonction de leurs similarités, ce qui permet d'identifier les groupes ou les clusters dans les données.

  - Pour la classification spatiale, nous utilisée la méthode des machines à vecteurs de support (SVM). La bibliothèque scikit-learn que vous avez mentionnée précédemment offre également une implémentation de l'algorithme SVM, qui peut être utilisée pour la classification spatiale en fonction de caractéristiques et de variables spatiales.

  - Pour l'analyse des hotspots, nous utilisée l'indice I de Moran local. Il permet de détecter les clusters de valeurs similaires ou dissimilaires dans les données spatiales. L'implémentation de cet indice peut être réalisée à l'aide de bibliothèques telles que PySAL ou GeoPandas, qui offrent des fonctionnalités spécifiques pour l'analyse spatiale.

  - Pour détection des outliers dans les données spatiales, nous utilisée l'indice I de Moran local. Utiliser des mesures de distance ou de densité pour identifier les observations atypiques. Des bibliothèques telles que PySAL, scikit-learn et GeoPandas Nous avons utilisé pour effectuer des analyses d'outliers dans des données spatiales.

- **Visualisation des données :**

- La bibliothèque Folium est utilisée pour créer des cartes interactives et des visualisations des résultats. Folium permet d'afficher des données géospatiales sur une carte, d'ajouter des couches supplémentaires telles que des marqueurs, des cercles ou des polygones, et de générer des cartes interactives pouvant être explorées par l'utilisateur.

- **Autres fonctionnalités :** L'application utilise également d'autres fonctionnalités :

  - La bibliothèque StandardScaler de scikit-learn, permet la normalisation des données, l'ouverture de fichiers HTML dans un navigateur à l'aide du module webbrowser, et l'affichage d'images à l'aide des modules Image et ImageTk de la bibliothèque PIL.

4.3 Mise en œuvre des techniques de data mining spatial

Dans le cadre de notre étude épidémiologique du cancer, nous avons appliqué différentes techniques de Data Mining Spatial (DMS) pour analyser les données relatives à l'incidence et à la mortalité des cas de cancer de l'estomac à travers les différents États des États-Unis. Notre objectif est d'identifier des modèles spatiaux significatifs et d'extraire des informations utiles pour la lutte contre cette maladie.

Au cours de notre recherche, nous avons utilisé quatre tâches clés du DMS : la classification, le clustering, la détection des coldspots et hotspots spatiaux, ainsi que la détection des outliers. Ces techniques nous ont permis d'explorer et d'analyser en les données épidémiologiques du cancer de l'estomac, en mettant l'accent sur les variations spatiales.

4.3.1 Classification pour l'étude épidémiologique du cancer

Nous avons adopté la démarche suivante pour mettre en œuvre la classification spatiale en utilisant **l'**algorithme de classification SVM (Support Vector Machine) sur notre jeu de données :

1. **Préparation des données :**

a. Chargement du fichier shapefile polygonal et du fichier CSV contenant les données.

b. Fusion des données avec le shapefile en utilisant une clé commune (identifiant d'unité administrative).

c. Prétraitement des données, incluant la normalisation, la gestion des valeurs aberrantes et des données manquantes.

2. Formation du modèle Support Vector Machine (SVM) :

a. Séparation des données en un ensemble d'apprentissage et un ensemble de test.

b. Sélection du noyau SVM approprié (linéaire, polynomial, RBF, etc.).

c. Entraînement du modèle SVM en optimisant la marge et en recherchant la frontière de décision optimale.

3. Classification des données :

a. Classification des groupes en fonction de l'incidence de cas de cancer de l'estomac ou de mortalité par Etat.

b. Utilisation de méthodes spécifiques de classification, telles que la prédiction des classes à partir des calculs effectués par le modèle SVM.

4. Évaluation du modèle de classification :

a. Utilisation de mesures d'évaluation telles que la précision, le rappel, la matrice de confusion, la F-mesure, etc.

b. Validation croisée et évaluation des performances du modèle sur l'ensemble de test.

5. Visualisation des résultats :

a. Utilisation de la bibliothèque Folium pour visualiser la carte géographique des résultats de classification spatiale sur un navigateur Internet.

b. Affichage des résultats de classification sur la carte, permettant une visualisation géographique des données épidémiologiques du cancer.

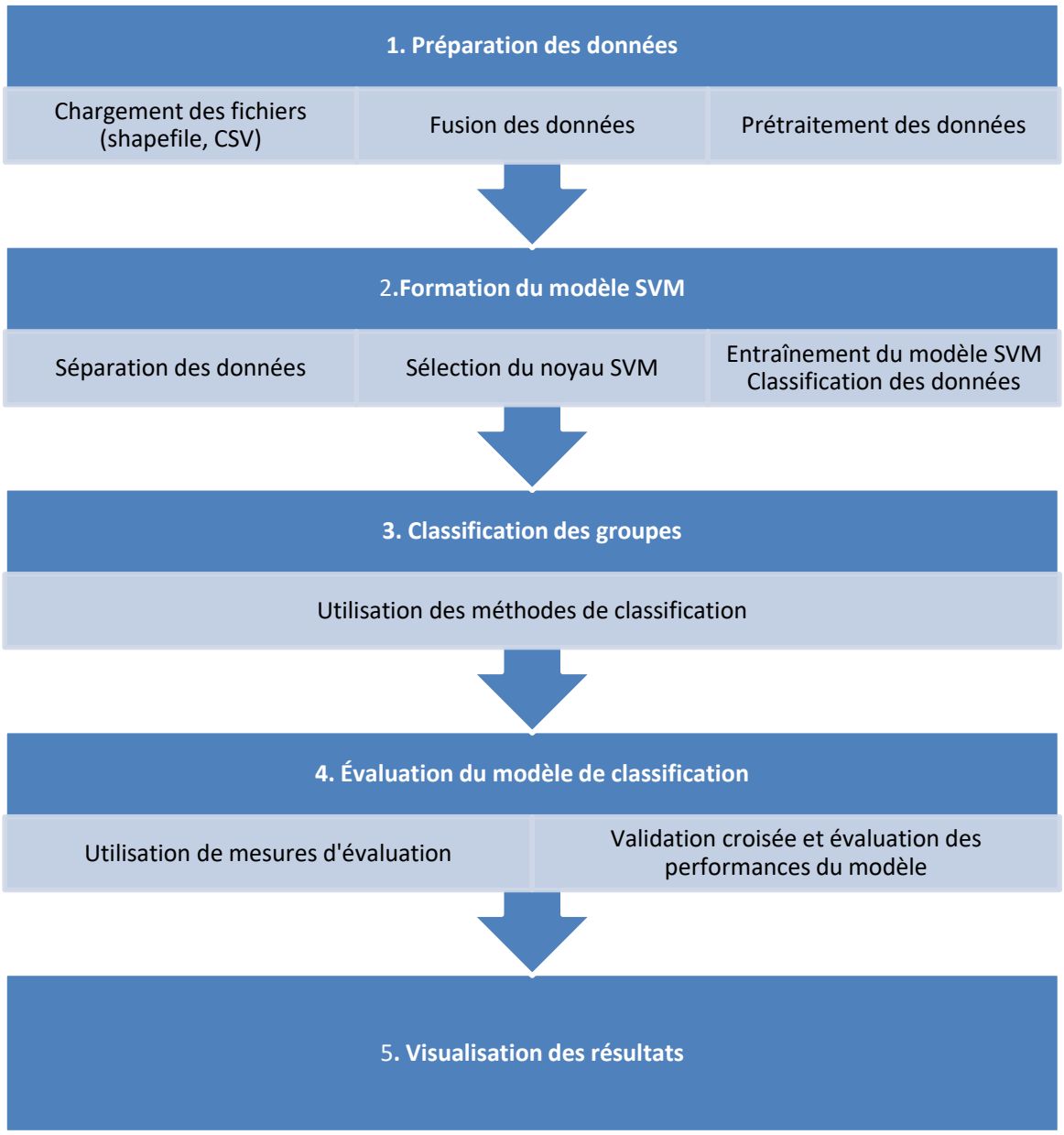


FIGURE 9 : PRESENTATION DE LA DEMARCHE DE CLASSIFICATION SPATIALE BASEE DU LE MODELE SVM

4.3.2 Clustering pour l'étude épidémiologique du cancer

Nous avons adopté la démarche suivante pour mettre en œuvre le clustering spatial sur notre jeu de données en utilisant l'algorithme K-means :

1. **Préparation des données :**

    a. Chargement du fichier shapefile polygonal et du fichier CSV contenant les données.

    b. Fusion des données avec le shapefile en utilisant une clé commune (identifiant d'unité administrative).

    c. Extraction de la variable d'intérêt (cas d'incidence ou cas de mortalité) à partir du DataFrame fusionné.

    d. Normalisation des données en soustrayant la moyenne et en divisant par l'écart-type.

2. **Application de l'algorithme K-means :**

    a. Utilisation de l'algorithme K-means pour effectuer le clustering spatial.

    b. Spécification du nombre de clusters

    c. Application de l'algorithme K-means aux données normalisées pour obtenir les clusters.

3. **Attribution des étiquettes de cluster :**

    a. Ajout des étiquettes de cluster au GeoDataFrame pour chaque unité administrative.

    b. Chaque unité administrative sera assignée à un cluster spécifique en fonction de ses caractéristiques épidémiologiques.

4. **Visualisation des résultats :**

    a. Tracé de la carte des clusters de cas, où chaque cluster est représenté par une couleur ou un symbole distinct.

    b. La carte affiche les clusters basés sur les cinq groupes distincts identifiés, en se basant sur le nombre de cas plutôt que sur la distance.

    c. Utilisation de la bibliothèque Folium pour visualiser la carte géographique sur un navigateur Internet.

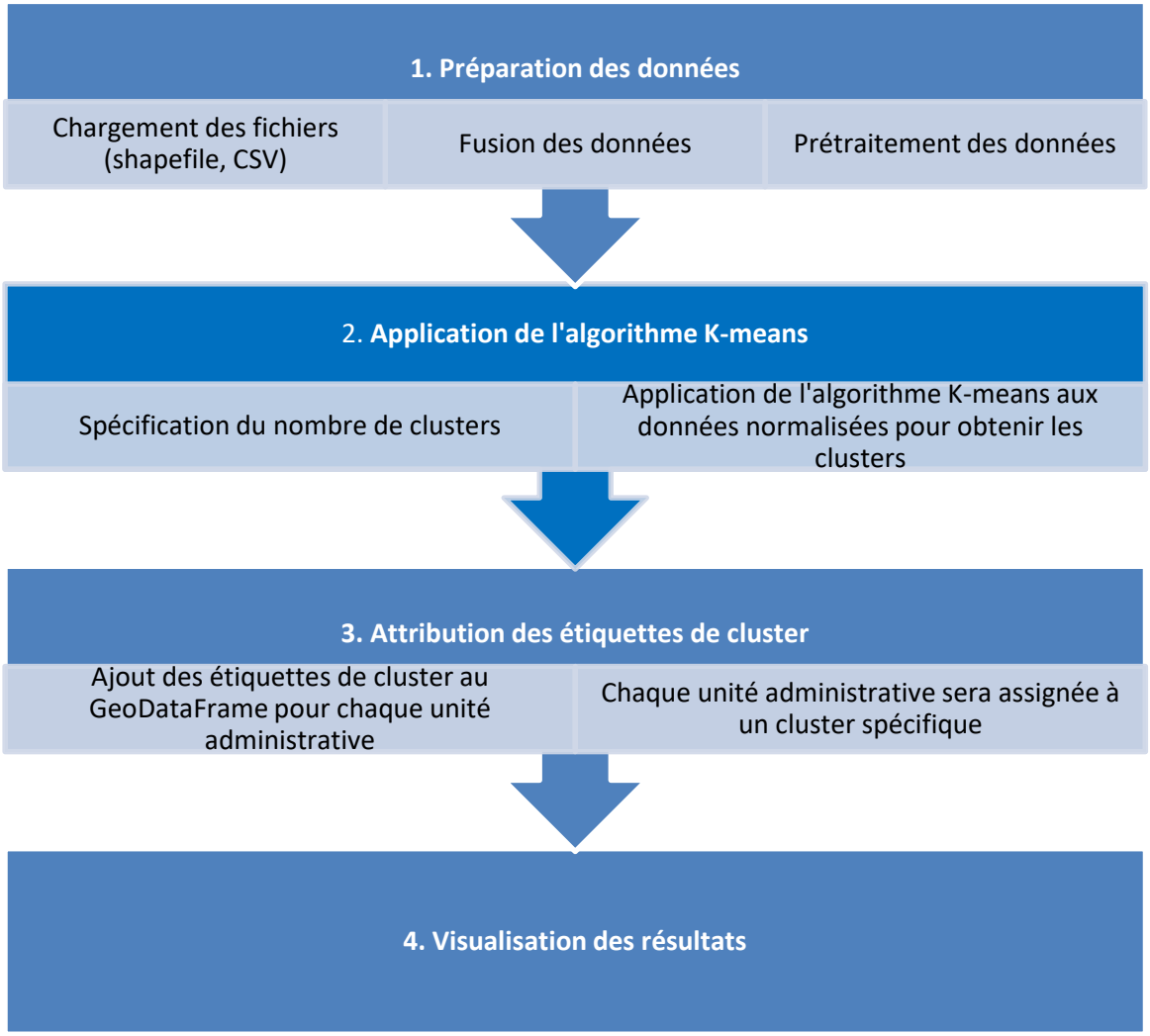


FIGURE 10 : PRESENTATION DE LA DEMARCHE DE CLUSTERING SPATIALE BASEE SUR L'ALGORITHME K-MEANS

4.3.3 Détection des hotspots et cold spots pour l'étude épidémiologique du cancer

Nous avons adopté la démarche suivante pour mettre en œuvre la détection des hotspots et coldspots spatiaux sur notre jeu de données en utilisant l'indice I de Moran local :

1. **Préparation des données :**

   a. Chargement du fichier shapefile polygonal et du fichier contenant les données des nombres de cas.

b. Fusion des deux ensembles de données en utilisant une clé commune pour associer les informations géographiques aux données des cas.

2. **Création de la matrice de pondération spatiale :**

   a. Création d'une matrice de pondération spatiale basée sur la matrice de contiguïté.

   b. La matrice de contiguïté mesure les relations spatiales entre les unités géographiques dans l'analyse spatiale et est utilisée pour déterminer les unités voisines d'une unité donnée.

3. **Calcul de l'indice I de Moran local :**

   a. Calcul de l'indice I de Moran local, qui est une mesure de l'autocorrélation spatiale.

   b. L'indice I de Moran local permet de déterminer les hotspots (zones avec une forte concentration de cas) et les cold spots (zones avec une faible concentration de cas).

   c. Les valeurs de l'indice I de Moran local varient de -1.7 à 0.6, avec des valeurs négatives indiquant des cold spots et des valeurs positives indiquant des hotspots.

   d. Ajout des valeurs d'indice I de Moran local au GeoDataFrame :

   e. Ajout des valeurs d'indice I de Moran local et des p-values associées au GeoDataFrame. Cette étape permet d'associer les mesures d'autocorrélation spatiale à chaque unité géographique.

4. **Tracé des hotspots et cold spots sur une carte :**

   a. Tracé des hotspots et cold spots sur une carte en utilisant une couleur pour représenter les valeurs d'indice I de Moran local.

b. Utilisation d'une colormap spécifiée, avec les hotspots représentés en rouge et les cold spots en bleu.

5. Visualisation de la carte géographique :

a. Utilisation de la bibliothèque Folium pour visualiser la carte géographique sur un navigateur Internet.

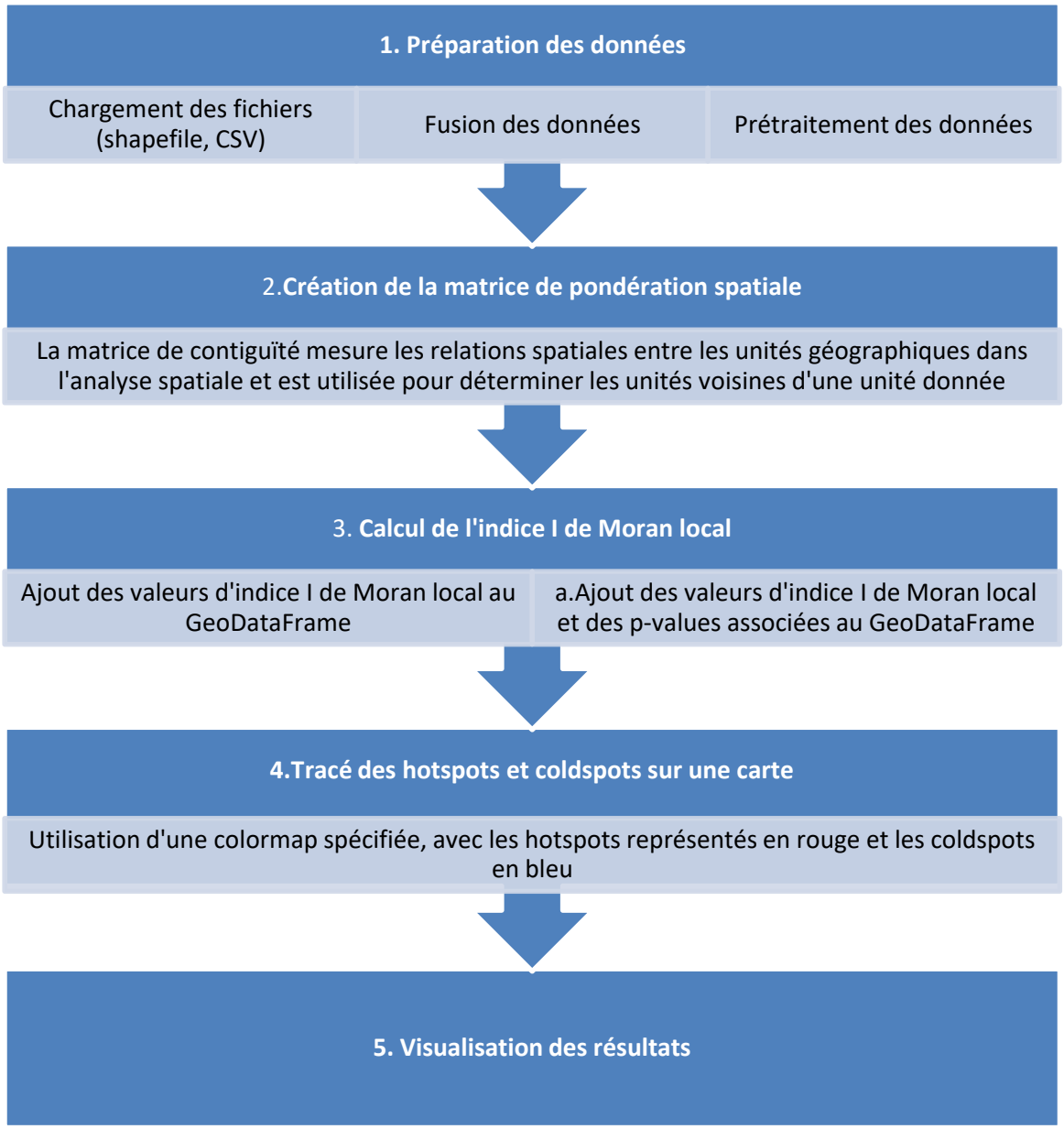


FIGURE 11 : PRESENTATION DE LA DEMARCHE DE DETECTION DES HOTSPOTS ET COLD SPOTS BASEE LE CALCUL L'INDICE I DE MORAN LOCAL

4.3.4 Détection des outliers pour l'étude épidémiologique du cancer

Nous avons adopté la démarche suivante pour mettre en œuvre la détection des outliers spatiaux sur notre jeu de données en utilisant l'indice I de Moran Global :

1. **Préparation des données :**
    a. Chargement du fichier shapefile polygonal et du fichier CSV contenant les données.
    b. Fusion des données avec le shapefile en utilisant une clé commune (identifiant d'unité administrative).
    c. Définition d'une variable d'intérêt (nombre d'instances) à partir du DataFrame fusionné.
    d. Normalisation des données en utilisant la transformation de score Z.
2. **Calcul de l'autocorrélation spatiale :**
    a. Calcul de la matrice de continuité spatiale pour représenter les relations spatiales entre les unités administratives.
    b. Calcul de l'indice Global Moran I pour mesurer l'autocorrélation spatiale, qui évalue si les valeurs des unités administratives sont similaires ou non à celles de leurs voisins.
3. **Détection des outliers spatiaux :**
    a. Détection des valeurs aberrantes spatiales en comparant les valeurs p des voisins locaux avec un seuil de 0,05.
    b. Attribution d'étiquettes externes au GeoDataFrame à l'aide d'un tableau booléen indiquant si une unité administrative est considérée comme une valeur aberrante ou non.
    c. Affichage des unités administratives considérées comme des outliers spatiaux.

4. **Visualisation des résultats :**

   a. Les outliers spatiaux sont définis comme les unités administratives ayant une valeur de p associée à l'indice Moran I global inférieure à 0,05.

   b. Les outliers détectés sont ajoutés au GeoDataFrame sous la colonne "is_outlier" et affichés à la fin.

   c. Utilisation de la bibliothèque Folium pour visualiser la carte géographique sur un navigateur Internet.

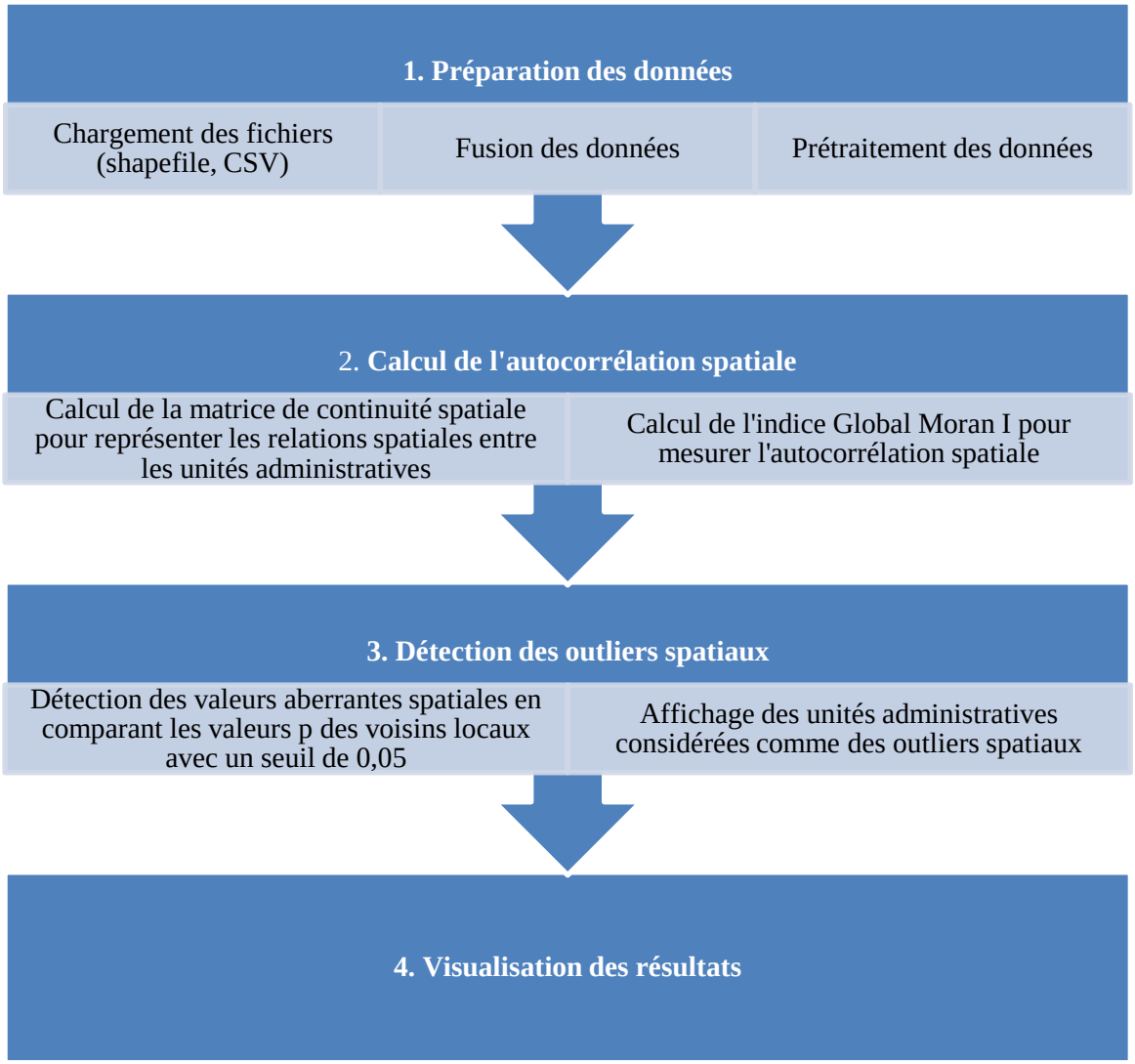


FIGURE 12 : PRESENTATION DE LA DEMARCHE DE DETECTION DES OUTLIERS BASEE SUR LE CALCUL L'INDICE I DE MORAN GLOBAL

4.4 Présentation du logiciel

Nous avons développé un logiciel polyvalent qui permet d'effectuer différentes tâches de data mining spatial. Le programme propose plusieurs interfaces qui offrent la possibilité de sélectionner l'analyse à réaliser en fonction du cas d'étude, qu'il s'agisse des cas d'incidence ou de mortalité.

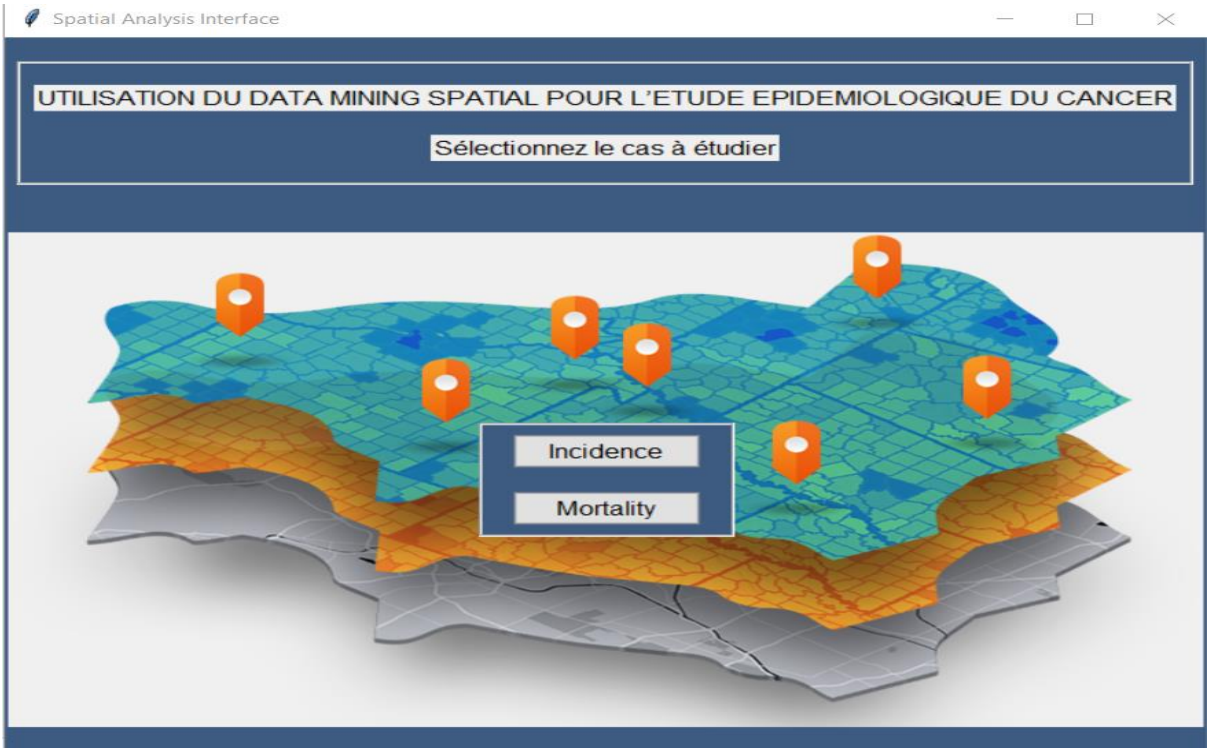


FIGURE 13- INTERFACE DU LOGICIEL DEVELOPPE

Une fois le cas choisi, une fenêtre s'ouvre avec des boutons permettant de sélectionner les fichiers SHP et CSV correspondants. Ensuite, l'utilisateur peut choisir parmi plusieurs techniques de data mining spatial, telles que le clustering spatial, la classification spatiale, la détection de hotspots et de cold spots, ainsi que la détection d'outliers, comme illustré dans la figure suivante.

Enfin, le logiciel présente sur une interface web les résultats des analyses sur une carte géographique de la zone d'étude, mettant en évidence les résultats obtenus en appliquant les techniques de data mining spatial choisies.

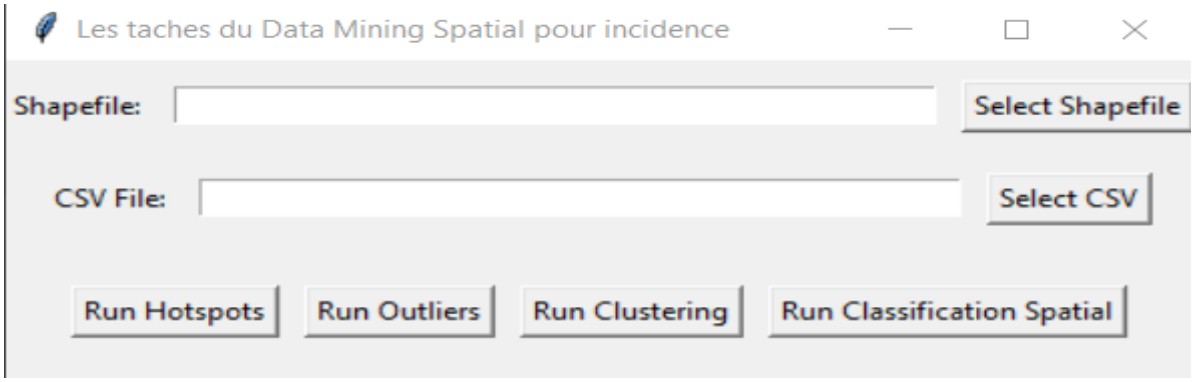


FIGURE 14- FENETRE PERMETTANT LE CHOIX DE L'ANALYSE A REALISER

4.5 Analyse des résultats obtenus

Dans ce qui suit, nous présentons les résultats des analyses effectués. Ces analyses sont visualisées à l'aide de cartes représentant chaque méthode utilisée. Ils peuvent être interprétés en les comparant aux objectifs de l'étude et aux connaissances épidémiologiques existantes. Il est possible d'identifier les tendances, les corrélations ou les associations spatiales détectées. La comparaison et la discussion des résultats entre différentes techniques d'exploration de données spatiales permettent de mieux comprendre les avantages et les limites de chaque méthode.

1- La classification spatiale

La figure ci-dessous présentent la répartition spatiale des classes obtenues grâce à l'application de la classification spatiale aux données sur le cancer de l'estomac aux États-Unis entre 2015 et 2019.

Les cartes choroplèthes mettent en évidence les différentes classes résultant de l'utilisation de l'algorithme SVM sur l'incidence, avec des valeurs variant de 152 à 15 670 cas dans les différents États des États-Unis. Les résultats sont dispersés sur la carte et représentés par différentes couleurs.

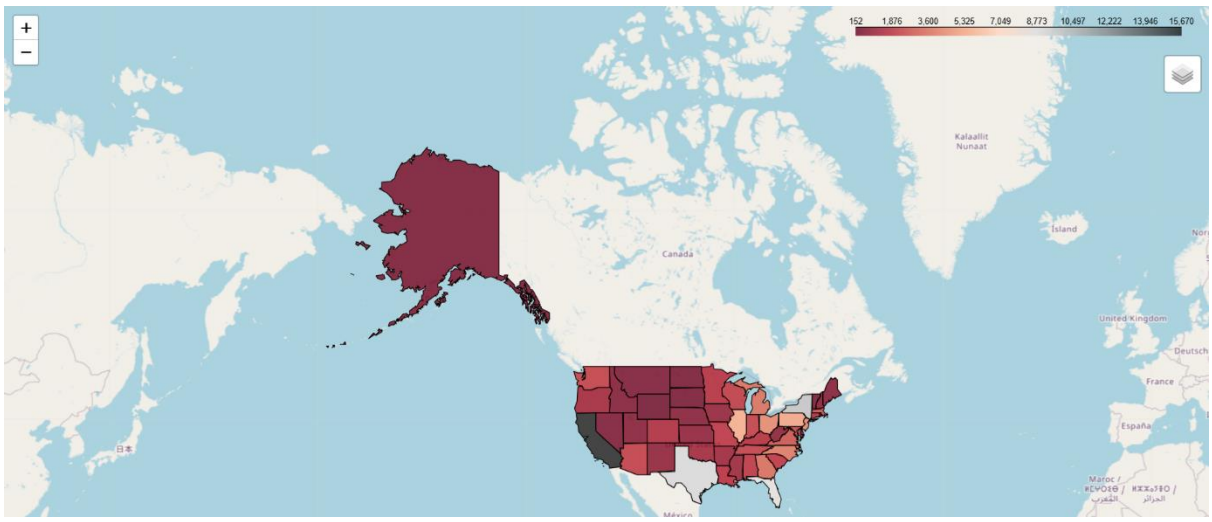


FIGURE 15- RESULTATS DE LA CLASSIFICATION SPATIALE APPLIQUEE AUX DONNEES DU CANCER DE L'ESTOMAC AUX USA ENTRE 2015 ET 2019

2- Le clustering spatial

Les figures ci-dessous illustrent la répartition spatiale des clusters résultant de l'application du clustering spatial aux données sur le cancer de l'estomac aux États-Unis entre 2015 et 2019. Les cartes choroplèthes mettent en évidence cinq groupes distincts, dispersés sur la carte et représentés par différentes couleurs.

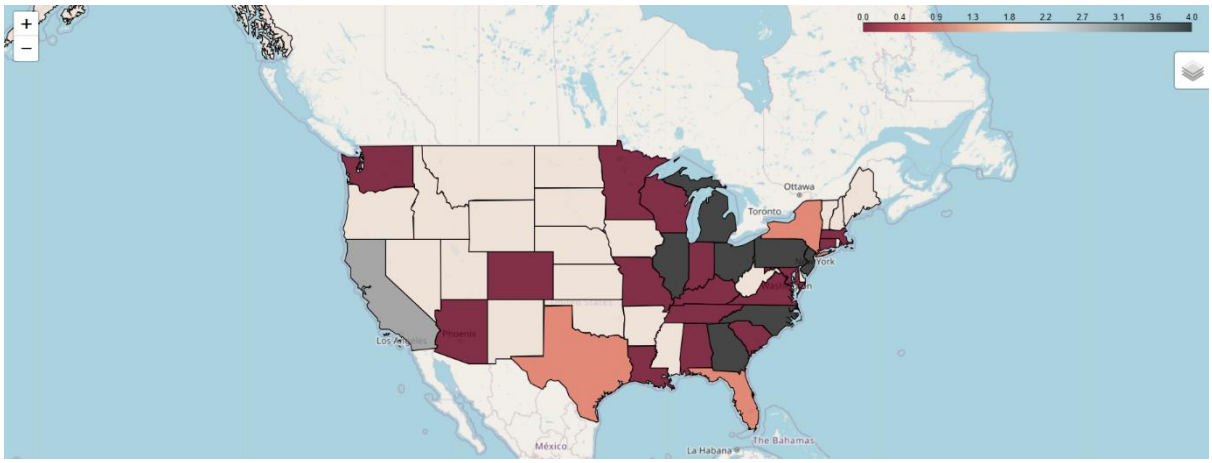


FIGURE 16 - RESULTATS DU CLUSTERING SPATIAL APPLIQUE AUX DONNEES DU CANCER DE L'ESTOMAC AUX USA ENTRE 2015 ET 2019

3- La détection des hotspots et cold spots :

La figure ci-dessous illustre la détection des hotspots et des cold spots spatiaux sur les données du cancer de l'estomac aux États-Unis entre 2015 et 2019. Les hotspots sont représentés en rouge, tandis que les cold spots sont représentés en bleu.

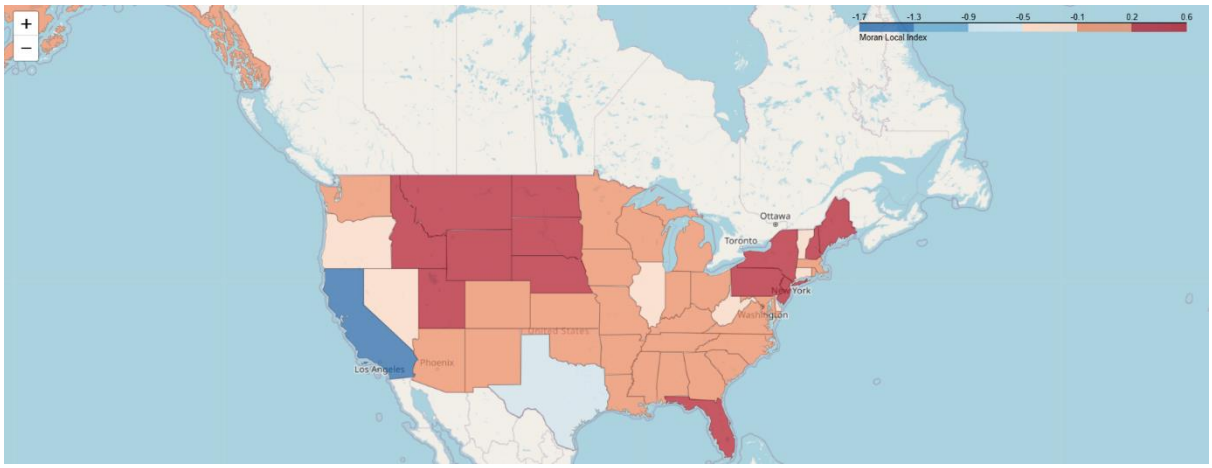


FIGURE 17 - DETECTION DES HOTSPOTS ET COLD SPOTS SPATIAUX SUR LES DONNEES DU CANCER DE L'ESTOMAC AUX USA ENTRE 2015 ET 2019

4- La détection des outliers :

La figure ci-dessous illustre la détection des outliers spatiaux sur les données du cancer de l'estomac aux États-Unis entre 2015 et 2019. Les outliers sont représentés en couleur rouge.

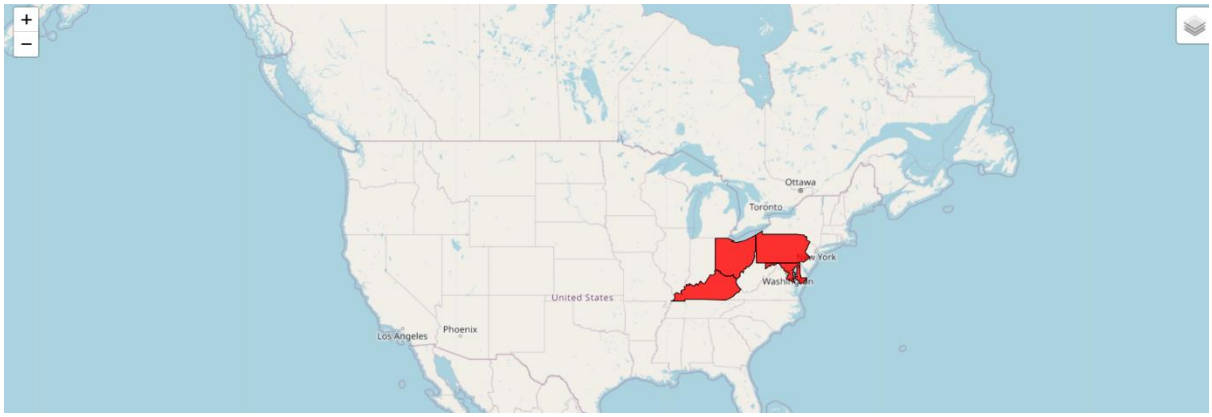


FIGURE 18 - DETECTION DES OUTLIERS SPATIAUX SUR LES DONNEES DU CANCER DE L'ESTOMAC AUX USA ENTRE 2015 ET 2019

4.6 Interprétation des résultats et discussion

4.6.1 Classification spatiale :

Dans l'analyse de classification spatiale, nous avons obtenu plusieurs classes qui sont représentées sur la carte géographique. Le dégradé de couleurs spécifique indique l'incidence des cas, allant de 152 à 15 670, dans différentes zones géographiques.

La classification spatiale nous permet de regrouper les zones ayant des niveaux d'incidence similaires. Par exemple, les zones de couleur plus claire peuvent représenter des zones à faible incidence, tandis que les zones de couleur plus foncée peuvent représenter des zones à forte incidence. Cela peut nous aider à identifier les zones où l'incidence de la variable étudiée est plus élevée ou plus faible, ce qui peut être utile pour la planification des ressources ou la prise de décisions en fonction des besoins de chaque zone.

4.6.2 Clustering :

L'analyse de clustering spatial a permis d'identifier cinq groupes distincts sur la carte géographique. Ces groupes ne sont pas proches les uns des autres et sont dispersés sur la carte, chacun étant représenté par une couleur spécifique.

Les clusters spatiaux indiquent des regroupements de zones géographiques similaires en fonction des caractéristiques étudiées. Chaque groupe peut représenter un profil distinct de caractéristiques ou de comportements.

Dans notre étude, nous avons appliqué le clustering sur les données attributaires et non pas sur la distance entre les entités spatiales, ce qui explique l'espacement entre les régions qui sont dans la même groupes.

4.6.3 Détection des hotspots et cold spots spatiaux :

L'analyse de détection des hotspots et cold spots spatiaux a été réalisée en utilisant un dégradé de couleur du bleu au rouge sur la carte géographique. Cette technique permet d'identifier les zones présentant des valeurs élevées (hotspots) ou basses (cold spots) par rapport à la distribution globale de la variable étudiée.

L’Etat de Californie à l’Ouest des Etats Unis, est considéré comme une zone de cold spot. Ceci s’explique par une incidence plus basse par rapport aux Etats voisins de cet Etat.

La région Centre-Nord des Etats Unis, les Etats à l’Est de la région des grands lacs et l’Etat de Floride ont été détecter comme des zones de hotspots en raison de leurs incidences supérieures aux Etats voisins de ces régions.

4.6.4 Détection des outliers :

L'analyse de détection des outliers a permis d'identifier une zone composée de quatre Etats à l’Est des Etas Unis. Cette zone est colorée en rouge sur la carte géographique. Ces zones sont marquées comme des outliers en raison de leurs valeurs extrêmes ou atypiques par rapport aux autres Etats.

4.7 Conclusion

Ce chapitre a présenté les résultats de notre expérimentation dans le cadre de notre étude épidémiologique du cancer aux Etats Unis. Nous avons décrit les données utilisées, l'environnement expérimental, la mise en œuvre des techniques de data mining spatial, le logiciel développé, ainsi que l'analyse et l'interprétation des résultats. Les résultats obtenus fournissent des informations précieuses sur la répartition spatiale du cancer, contribuant ainsi à une meilleure compréhension de cette maladie et à des perspectives prometteuses pour la recherche future.

Conclusion Générale

Le data mining spatial est une discipline qui consiste à analyser et à extraire des connaissances des données spatiales, c'est-à-dire des données qui ont une dimension géographique. Il permet de mettre en évidence des patterns et des tendances dans les données en prenant en compte leur dimension spatiale.

Le data mining spatial est utilisé dans de nombreux domaines, comme la géographie, l'urbanisme, l'environnement, la santé, l'économie, etc. Il peut être appliqué à différents types de données, comme des données géographiques, des images satellites, des capteurs, des questionnaires, etc.

Le data mining spatial nécessite un processus rigoureux qui comprend plusieurs étapes, comme la collecte des données, l'exploration et le nettoyage des données, l'analyse des données, la visualisation des résultats et l'interprétation et la diffusion des résultats. Il existe plusieurs techniques d'analyse spatiale qui peuvent être utilisées pour étudier les relations entre les variables spatiales et pour prendre en compte l'influence de l'espace sur ces relations.

Le DMS a montré une grande utilité dans la lutte contre le cancer à travers différentes applications. En effet, il peut être utilisé pour la détection précoce du cancer, la surveillance de la progression de la maladie, la prédiction de l'efficacité d'un traitement et l'identification de nouveaux médicaments.

Dans ce projet, nous avons utilisé les données épidémiologiques du cancer pour mener une étude approfondie sur l'incidence de cette maladie aux Etats Unis.

En utilisant des techniques de data mining spatial, nous avons analysé ces données pour identifier les régions présentant des variations significatives dans l'incidence du cancer. Nous avons appliqué des méthodes telles que l'analyse de hotspots et de cold spots pour repérer les zones où l'incidence du cancer est significativement plus élevée ou plus faible que prévu. De plus, nous avons utilisé des techniques de détection d'outliers pour identifier les régions présentant des taux inhabituels par rapport à la moyenne.

Nous avons également mis en place une classification spatiale pour regrouper les régions similaires en termes d'incidence du cancer. Cela nous a permis de mieux comprendre les tendances géographiques et de détecter d'éventuels schémas ou clusters.

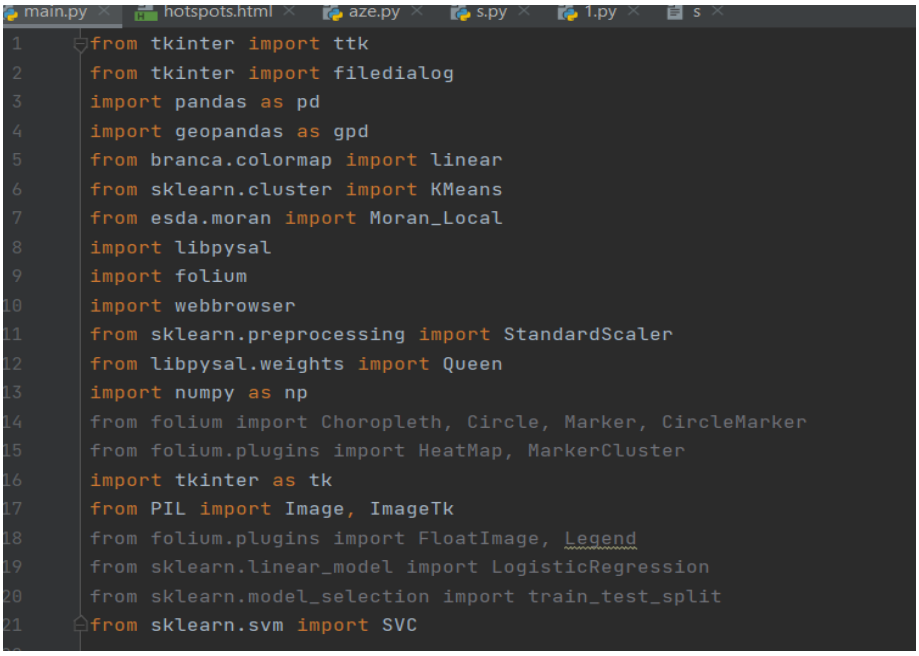
Dans le cadre de ce projet, nous avons développé une application conviviale qui permet aux utilisateurs d'explorer les résultats de manière interactive. L'interface utilisateur offre des fonctionnalités de visualisation sur une carte, permettant aux chercheurs, aux professionnels de la santé publique et aux décideurs d'observer les résultats de manière géospatiale et de mieux comprendre les facteurs associés à l'incidence du cancer.

Les résultats obtenus dans cette étude sont prometteurs et ouvrent de nombreuses perspectives. Par exemple, l'application pourrait être améliorée en intégrant des données en temps réel pour suivre l'évolution de l'incidence du cancer et détecter les nouveaux hotspots émergents. De plus, des analyses plus approfondies pourraient être menées pour explorer les relations entre les facteurs environnementaux spécifiques et l'incidence du cancer.

En conclusion, ce projet a permis de mettre en évidence l'utilisation des techniques de data mining spatial dans l'étude épidémiologique du cancer. L'application développée fournit des résultats significatifs et une visualisation interactive, offrant ainsi un outil précieux pour la recherche, la santé publique et la prise de décision. Les perspectives futures sont nombreuses pour améliorer et étendre cette application afin de mieux comprendre et lutter contre cette maladie dévastatrice.

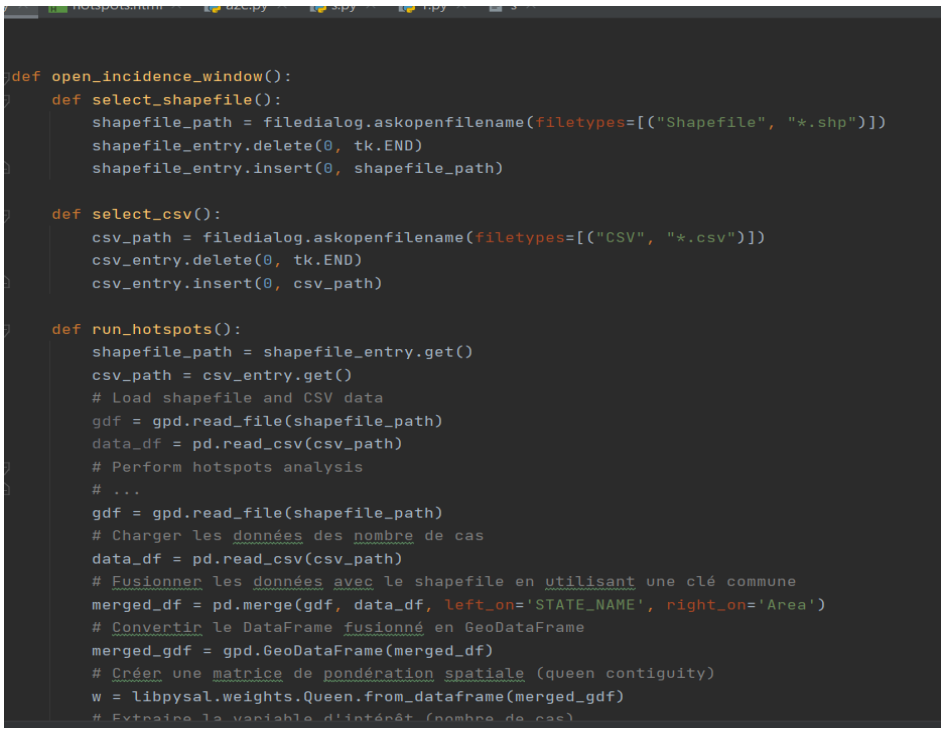
Annexe

- **CODE EN PYTHON PERMETTANT D'IMPORTER LES BIBLIOTHEQUES UTILISEES :**



```python
from tkinter import ttk
from tkinter import filedialog
import pandas as pd
import geopandas as gpd
from branca.colormap import linear
from sklearn.cluster import KMeans
from esda.moran import Moran_Local
import libpysal
import folium
import webbrowser
from sklearn.preprocessing import StandardScaler
from libpysal.weights import Queen
import numpy as np
from folium import Choropleth, Circle, Marker, CircleMarker
from folium.plugins import HeatMap, MarkerCluster
import tkinter as tk
from PIL import Image, ImageTk
from folium.plugins import FloatImage, Legend
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
```

- **CODE EN PYTHON PERMETTANT LE CHOIX DE L'ANALYSE A REALISER POUR LA DETECTION DES HOTSPOTS ET COLD SPOTS :**



```python
def open_incidence_window():
    def select_shapefile():
        shapefile_path = filedialog.askopenfilename(filetypes=[("Shapefile", "*.shp")])
        shapefile_entry.delete(0, tk.END)
        shapefile_entry.insert(0, shapefile_path)

    def select_csv():
        csv_path = filedialog.askopenfilename(filetypes=[("CSV", "*.csv")])
        csv_entry.delete(0, tk.END)
        csv_entry.insert(0, csv_path)

    def run_hotspots():
        shapefile_path = shapefile_entry.get()
        csv_path = csv_entry.get()
        # Load shapefile and CSV data
        gdf = gpd.read_file(shapefile_path)
        data_df = pd.read_csv(csv_path)
        # Perform hotspots analysis
        # ...
        gdf = gpd.read_file(shapefile_path)
        # Charger les données des nombre de cas
        data_df = pd.read_csv(csv_path)
        # Fusionner les données avec le shapefile en utilisant une clé commune
        merged_df = pd.merge(gdf, data_df, left_on='STATE_NAME', right_on='Area')
        # Convertir le DataFrame fusionné en GeoDataFrame
        merged_gdf = gpd.GeoDataFrame(merged_df)
        # Créer une matrice de pondération spatiale (queen contiguity)
        w = libpysal.weights.Queen.from_dataframe(merged_gdf)
        # Extraire la variable d'intérêt (nombre de cas)
```

```
w = libpysal.weights.Queen.from_dataframe(merged_gdf)
# Extraire la variable d'intérêt (nombre de cas)
variable = 'Case Count'
merged_gdf['Case Count'] = merged_gdf['Case Count'].replace('Data Suppressed', 0)

# Extract the "Case Count" variable and replace suppressed values with NaN
y = merged_gdf[variable].replace('Data Suppressed', np.nan).astype(float)

# Calculer l'indice I de Moran local
moran_loc = Moran_Local(y, w)
# Récupérer les valeurs d'indice I de Moran local et les p-values associées
moran_loc_values = moran_loc.Is
moran_loc_p_values = moran_loc.p_sim
# Créer une colonne dans le GeoDataFrame pour stocker les résultats
merged_gdf['moran_loc'] = moran_loc_values
merged_gdf['moran_loc_p'] = moran_loc_p_values

# Set the initial map location using the mean coordinates of the GeoDataFrame
initial_location = [merged_gdf['geometry'].centroid.y.mean(), merged_gdf['geometry'].centroid.x.mean()]
map_obj = folium.Map(location=initial_location, zoom_start=4)

# Create a Choropleth layer using the 'moran_loc' column as the data
choropleth = folium.Choropleth(
    geo_data=merged_gdf,
    data=merged_gdf,
    columns=['STATE_NAME', 'moran_loc'],
    key_on='feature.properties.STATE_NAME',
    fill_color='RdBu_r',  # Reversed ColorBrewer palette
    fill_opacity=0.7,
```

```
    # Create a Choropleth layer using the 'moran_loc' column as the data
    choropleth = folium.Choropleth(
        geo_data=merged_gdf,
        data=merged_gdf,
        columns=['STATE_NAME', 'moran_loc'],
        key_on='feature.properties.STATE_NAME',
        fill_color='RdBu_r',  # Reversed ColorBrewer palette
        fill_opacity=0.7,
        line_opacity=0.2,
        legend_name='Moran Local Index',
        highlight=True
    ).add_to(map_obj)

    # Add the 'Case Count' and 'Cancer Type' information as popup when a polygon is c
    choropleth.geojson.add_child(
        folium.features.GeoJsonPopup(
            fields=['Area','Case Count', 'Cancer Type','Sex','Race'],
            labels=True,
            localize=True
        )
    )

    # Display the map
    map_obj
    map_obj.save("hotspots.html")
    webbrowser.open("hotspots.html")
```

- **CODE EN PYTHON PERMETTANT LE CHOIX DE L'ANALYSE A REALISER POUR LA DETECTION DES OUTLIERS :**

```python
def run_outliers():
    shapefile_path = shapefile_entry.get()
    csv_path = csv_entry.get()
    # Load shapefile and CSV data
    gdf = gpd.read_file(shapefile_path)
    data_df = pd.read_csv(csv_path)
    # Perform outliers analysis
    # ...
    gdf = gpd.read_file(shapefile_path)
    # Charger les données des nombre de cas
    data_df = pd.read_csv(csv_path)
    # Fusionner les données avec le shapefile en utilisant une clé commune
    merged_df = pd.merge(gdf, data_df, left_on='STATE_NAME', right_on='Area')
    # Convertir le DataFrame fusionné en GeoDataFrame
    merged_gdf = gpd.GeoDataFrame(merged_df)
    # Sélectionner uniquement la variable d'intérêt
    variable = 'Case Count'
    merged_gdf['Case Count'] = merged_gdf['Case Count'].replace('Data Suppress

    # Extract the "Case Count" variable and replace suppressed values with NaN
    y = merged_gdf[variable].replace('Data Suppressed', np.nan).astype(float)
    # Standardiser les données
    scaler = StandardScaler()
    y_standardized = scaler.fit_transform(y.values.reshape(-1, 1))
    # Calculer la contiguïté spatiale
    w = Queen.from_dataframe(merged_df)
    # Calculer l'indice I de Moran local
    moran_loc = Moran_Local(y_standardized, w)
```

```python
    # Détecter les outliers spatiaux
    outliers = np.where(moran_loc.p_sim < 0.05, moran_loc.Is < 0, False)
    # Ajouter les étiquettes d'outliers au GeoDataFrame
    merged_df['is_outlier'] = outliers

    # Create a Folium map object
    map_outliers = folium.Map(location=[37.0902, -95.7129], zoom_start=4)

    # Create a GeoJson layer for the outliers
    folium.GeoJson(
        merged_df[merged_df['is_outlier']],
        name='Outliers',
        style_function=lambda feature: {
            'fillColor': 'red',
            'color': 'black',
            'weight': 1,
            'fillOpacity': 0.8
        },
        highlight_function=lambda feature: {
            'fillColor': 'red',
            'color': 'black',
            'weight': 3,
            'fillOpacity': 0.8
        }
    ).add_to(map_outliers)
```

```
# Add a layer control to the map
folium.LayerControl().add_to(map_outliers)

# Display the map
map_outliers
map_outliers.save("outliers.html")
webbrowser.open("outliers.html")
```

- **CODE EN PYTHON PERMETTANT DE REALISER LE CLUSTERING EN UTILISANT LA FONCTION K-MEANS :**

```
def run_clustering():
    shapefile_path = shapefile_entry.get()
    csv_path = csv_entry.get()
    # Load shapefile and CSV data
    gdf = gpd.read_file(shapefile_path)
    data_df = pd.read_csv(csv_path)
    # Perform clustering analysis
    # ...
    # Charger le shapefile polygonal
    gdf = gpd.read_file(shapefile_path)
    # Charger les données des nombre de cas
    data_df = pd.read_csv(csv_path)
    # Fusionner les données avec le shapefile en utilisant une clé commune
    merged_df = pd.merge(gdf, data_df, left_on='STATE_NAME', right_on='Area')
    # Convertir le DataFrame fusionné en GeoDataFrame
    merged_gdf = gpd.GeoDataFrame(merged_df)
    # Sélectionner uniquement la variable d'intérêt
    variable = 'Case Count'
    merged_df['Case Count'] = pd.to_numeric(merged_df['Case Count'], errors='coerce')
    merged_df = merged_df.dropna(subset=['Case Count'])
    X = merged_df[[variable]]

    # Normaliser les données
    X_normalized = (X - X.mean()) / X.std()
        # Define the number of clusters
        n_clusters = 5

        # Apply the K-means algorithm
        kmeans = KMeans(n_clusters=n_clusters, random_state=0).fit(X_normalized)

        # Add the cluster labels to the GeoDataFrame
        merged_df['cluster_label'] = kmeans.labels_

        # Create a Folium map object
        map_clusters = folium.Map(location=[37.0902, -95.7129], zoom_start=4)

        # Create a color map
        colormap = linear.RdGy_10.scale(merged_df['cluster_label'].min(), merged_df['cluster_label'].max())

        # Add the GeoDataFrame with cluster labels to the map
        folium.GeoJson(
            merged_df,
            name='cluster_label',
            style_function=lambda feature: {
```

```
        style_function=lambda feature: {
            'fillColor': colormap(feature['properties']['cluster_label']),
            'color': 'black',
            'weight': 1,
            'fillOpacity': 0.8
        },
        highlight_function=lambda feature: {
            'fillColor': colormap(feature['properties']['cluster_label']),
            'color': 'black',
            'weight': 3,
            'fillOpacity': 0.8
        },popup=folium.GeoJsonPopup(fields=['Area','Case Count', 'Cancer Type','Sex','Race'], labels=True, localize=True, style="default",
                                    parse_html=False, max_width="300")

    ).add_to(map_clusters)

    # Add a color legend to the map
    colormap.add_to(map_clusters)

    # Add a layer control to the map
    folium.LayerControl().add_to(map_clusters)

    # Display the map
    map_clusters
    map_clusters.save("clustering.html")
    webbrowser.open("clustering.html")
```

- CODE EN PYTHON PERMETTANT DE REALISER LA CLASSIFICATION SPATIALE AVEC SVM :

```
def run_classification_spatial():


    # Get the file paths
    shapefile_path = shapefile_entry.get()
    csv_path = csv_entry.get()

    # Load shapefile and CSV data
    gdf = gpd.read_file(shapefile_path)
    data_df = pd.read_csv(csv_path)

    # Merge the shapefile and data using a common key
    merged_df = pd.merge(gdf, data_df, left_on='STATE_NAME', right_on='Area')

    # Convert the merged DataFrame to GeoDataFrame
    merged_gdf = gpd.GeoDataFrame(merged_df)
    # Select the variables of interest for classification
    features = ['Case Count']
    merged_df['Case Count'] = pd.to_numeric(merged_df['Case Count'], errors='coerce')
    merged_df = merged_df.dropna(subset=['Case Count'])
    # Split the data into features (X) and labels (y)
    X = merged_df[features]
    y = merged_df[features]

    # Normalize the feature matrix
    scaler = StandardScaler()
    X_normalized = scaler.fit_transform(X)

    # Define the SVM classifier
```

```
X_normalized = scaler.fit_transform(X)

# Define the SVM classifier
svm = SVC(kernel='rbf', C=1.0, gamma='scale')

# Train the classifier
svm.fit(X_normalized, y)

# Predict the labels for the entire dataset
merged_df['predicted_label'] = svm.predict(X_normalized)

# Create a Folium map object
map_clusters = folium.Map(location=[37.0902, -95.7129], zoom_start=4)

# Create a color map
colormap = linear.RdGy_10.scale(merged_df['predicted_label'].min(), merged_df['predicted_label'].max())

# Add the GeoDataFrame with predicted labels to the map
folium.GeoJson(
    merged_df,
    name='predicted_label',
    style_function=lambda feature: {
```

```
        style_function=lambda feature: {
            'fillColor': colormap(feature['properties']['predicted_label']),
            'color': 'black',
            'weight': 1,
            'fillOpacity': 0.8
        },
        highlight_function=lambda feature: {
            'fillColor': colormap(feature['properties']['predicted_label']),
            'color': 'black',
            'weight': 3,
            'fillOpacity': 0.8
        },
        popup=folium.GeoJsonPopup(fields=['Area', 'Case Count', 'Cancer Type', 'Sex', 'Race'],
                                  labels=True,
                                  localize=True, style="default", parse_html=False, max_width="300")
    ).add_to(map_clusters)

    # Add a color legend to the map
    colormap.add_to(map_clusters)

    # Add a layer control to the map
    folium.LayerControl().add_to(map_clusters)

    # Save and open the map in a web browser
    map_clusters.save("classification.html")
    webbrowser.open("classification.html")
```

Bibliographies

[1] Grieg-Smith, P., 1961. Data on pattern within plant communities. I. The analysis of pattern. Journal of Ecology 49, 695-702

[2] Allen, T.F.H., Starr, T.B., 1982. Hierarchy: Perspectives for Ecological Complexity. University of Chicago Press, Chicago.

[3] Kolasa, J., Pickett, S.T.A., 1991. Ecological heterogeneity. Springer, New York.

[4] Dutilleul, P., 1993. Spatial heterogeneity and the design of ecological field experiments. Ecology 74, 1646–1658.

[5] Wu, J., Loucks, O.L., 1995. From balance of nature to hierarchical, patch dynamics: a paradigm shift in ecology. The Quarterly Review of Biology 70, 439-466.

[6] Dale, M.R.T., 1999. Spatial pattern analysis in plant ecology. Cambridge University Press, Cambridge, UK

[7] Legendre, P., Fortin, M.-J., 1989. Spatial patterns and ecological analysis. Vegetatio 80, 107-138

[8] Garcia, F., 2003. Mécanismes de développement de l'hétérogénéité du couvert végétal dans une prairie pâturée par des ovins. Thèse de l'Institut National Agronomique de Paris-Grignon, 148 p.

[9] Pumain, D. and T. Saint-Julien (1997). "L'Analyse spatiale. 1. Localisations dans l'espace. Paris: Armand Colin, coll. «." Cursus.

[10] MacEachren, A. M. and M.-J. Kraak (2001). "Research challenges in geovisualization." Cartography and geographic information science 28(1): 3-12.

[11] The Gartner Group, www.gartner.com.

[12] D. HAND, H. MANNILA et P. SMYTH, Principles of Data Mining, MIT Press, Cambridge, MA, 2001.

[13] P. CABENA, P. HADJINIAN, R. STADLER, J. VERHEES et A. ZANASI, Discovering Data Mining: From Concept to Implementation, Prentice Hall, Upper Saddle River, NJ, 1998.

[14] E-G. TALBI, Fouille de données (Data Mining) : Un tour d'horizon, Laboratoire d'Informatique Fondamentale de Lille.

[15] M. J. BERRY, G. S. LINOFF, Data Mining Techniques For Marketing, Sales, and Customer Relationship, Management, Second Edition, 2004.

[16] M. J. BERRY, G. S. LINOFF, Mastering Data Mining: The Art and Science of Customer Relationship Management, 2000.

[17] D.T. LAROSE, Discovering Knowledge In Data: An Introduction to Data Mining, Central Connecticut State University, 2005.

[18] S. PRABHU, N. VENKATESAN, Data Mining and Warehousing, New Age International (P) Ltd., Publishers, New Delhi, 2007.

[19] ZEITOUNI K., "A Survey on Spatial Data Mining Methods Databases and Statistics Point of Views", à paraître dans IRMA 2000, Information Resources Management Association International Conference, Data Warehousing and Mining Track, 21-23 May, 2000, Anchorage, Alaska, USA.

[20] Zhang, P., Y. Huang, S. Shekhar and V. Kumar (2003). "Correlation analysis of spatial time series datasets: A filterand-refine approach." Advances in Knowledge Discovery and Data Mining: 563-563.

[21] Zeitouni, K., L. Yeh and M.-A. Aufaure (2001). Join indices as a tool for spatial data mining. Temporal, Spatial, and Spatio-Temporal Data Mining, Springer: 105-116.

[22] Valduriez, P. (1987). "Join indices." ACM Transactions on Database Systems (TODS) 12(2): 218-246.

[23] Malinowski, E. and E. Zimányi (2005). Spatial hierarchies and topological relationships in the spatial MultiDimER model. British National Conference on Databases, Springer.

[24] Han, J., S. Chee and J. Y. Chiang (1998). Issues for on-line analytical mining of data warehouses. Proc. 1998 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'98).

[25] Koperski, K. (1999). A progressive refinement approach to spatial data mining, Simon Fraser University Canada.

[26] Kheirandish, F., Hosseinzadeh, M., & Akbarzadeh, M. (2020). Breast cancer diagnosis using data mining and machine learning techniques: a systematic review. Archives of computational methods in engineering, 27(3), 565–590

[27] Soltanian, A. R., Gholamian, M. R., & Alizadehsani, R. (2019). A new efficient algorithm for cancer classification based on gene expression data. Journal of biomedical informatics, 93, 103153.

[28] Larsen, K. et al. (2006). Geographic clustering of breast cancer subtypes in the city of Chicago. Journal of maps, 2(1), 7-18.

[29] Zhang, S., Huang, Y., Wang, L., Tao, S., & Liu, S. (2019). Spatial prediction of lung cancer risk using environmental and socioeconomic factors in China. Science of the Total Environment, 659, 609-617.

[30] Ferreira, C., Ferreira, M. A., & West, D. (2003). Spatial patterns of stomach cancer in Brazil. International Journal of Health Geographics, 2(1), 5.

[31] Kim, Y., & Kim, H. (2016). Spatial clustering and local risk factors of liver cancer incidence in Korea. International Journal of Environmental Research and Public Health, 13(6), 579.

[32] Lundgren, M., Hossain, M. S., Hammar, N., & Padyab, M. (2017). Spatial outlier detection of prostate cancer incidence in Sweden. PloS one, 12(4), e0175205.

[33] Chiusolo, M., D'Ambrosio, D., & Petrucci, A. (2016). Data Mining Techniques for Cancer Detection Using Serum Proteomic Profiling. Journal of healthcare engineering, 2016, 6842924.

[34] Kleinschmidt, J., Wunderlich, R., & Blettner, M. (2013). Spatial access to cancer care in Germany. Health & Place, 24, 19-29.

[35] Johnson, A. M., Hines, R. B., Johnson, J. A., Bayakly, R., & Vena, J. E. (2016). Association of Residential Radon Exposure with Increased Incidence of Childhood Cancer in North Carolina Counties. Archives of environmental & occupational health, 71(1), 24-32.

[35] Martinez-Mas, M., Benavent-Navarro, H., Tobarra-Gonzalez, B., & Mora-Jimenez, I. (2017). Geospatial analysis of the relationship between lung cancer mortality rates and radon concentrations in dwellings in Spain. Science of the Total Environment, 574, 1230-1242.

[36] Wang, J., Zhang, J., Wang, Y., Zhou, Z., Wang, Y., Zhou, R., Zhang, J. (2020). Spatial-temporal analysis of bladder cancer incidence and its hotspots in Chongqing, China. BMC Cancer, 20(1), 465.

[37] Schootman, M., Sterling, D. A., Struthers, J., Yan, Y., Laboube, T., Emo, B., & Higgs, G. (2007). Positional accuracy assessment of spatial data for public health surveillance and practice. Journal of public health management and practice: JPHMP, 13(2), 141–146.

[38] Jerrett, M., Brook, J. R., White, L. F., Burnett, R. T., Yu, J., Su, J. G., Krewski, D. (2013). The Relationship between Greenness and Traffic-Related Air Pollution at Schools. Environmental Health Perspectives, 121(6), 766–772.

[39] Larsen, L. K., Ersbøll, A. K., & Schmiegelow, M. D. (2017). An overview of methods for spatial analysis in epidemiology. Part 2: advanced topics. International journal of epidemiology, 46(5), 1588-1602.

[40] Cockburn, M., Swanson, J., & Zadnick, J. (2011). Spatial disparities in prostate cancer incidence and mortality rates within California during 1988-1997. Cancer Causes & Control, 22(9), 1525-1537.

[41] Mori, M., Shimada, K., Tokumaru, O., Miyake, M., & Kondo, T. (2016). Spatial analysis of cancer mortality in Osaka, Japan: a Bayesian approach. PloS one, 11(5), e0155021.

[42] MIDOUN Mohammed, Structuration multidimensionnelle de bases de données spatialisées pour le data mining spatial.