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 :

Détection des hotspots et cold spots spatiaux :

Application à l'épidémie de la covid-19

Etudiantes : **Benkerdagh Hadjer**

Ghalmi Meriem Zahia

Encadrant(e) : **Dr. MIDOUN Mohammed**

Année Universitaire 2022-2023

الملخص

تحتوي معظم قواعد البيانات الكبيرة المتاحة حاليًا على مكون مكاني قوي. النظام المسؤول عن استخراج هذه المعلومات والمعرفة هو التنقيب في البيانات. يتم تنفيذ اكتشاف المعرفة من خلال تطبيق خوارزميات تلقائية تتعرف على الأنماط في البيانات. تفترض خوارزميات التنقيب عن البيانات الكلاسيكية أن البيانات يتم إنشاؤها بشكل مستقل وتوزيعها بشكل مماثل. البيانات المكانية متعددة الأبعاد ومترابطة مكانيًا وغير متجانسة. تجعل هذه الخصائص خوارزميات التنقيب عن البيانات الكلاسيكية غير مناسبة للبيانات المكانية، وتنتهي صحة افتراضاتها الأساسية. في هذا المشروع، ستكون مسألة إجراء تحليلات وبائية للإحصاءات المتعلقة بـ **Covid-19** ستستند هذه التحليلات إلى اكتشاف البقع الباردة والنقاط الساخنة المكانية خلال تطور وباء كوفيد -19، منذ بداياته وحتى اليوم.

الكلمات الأساسية: التنقيب عن البيانات المكانية، اكتشاف البقع الباردة والنقاط الساخنة المكانية, Covid-19

Résumé

La plupart des grandes bases de données actuellement disponibles ont une forte composante spatiale. La discipline chargée de l'extraction de ces informations et connaissances est le data mining. La découverte de connaissances est effectuée en appliquant des algorithmes automatiques qui reconnaissent des modèles dans les données. Les algorithmes d'exploration de données classiques supposent que les données sont générées de façon indépendante et identiquement distribuées. Les données spatiales sont multidimensionnelles, spatialement autocorrélées et hétérogènes. Ces propriétés font en sorte que les algorithmes de data mining classique sont inappropriés pour les données spatiales, et que leurs hypothèses de base cessent d'être valables. Dans ce projet, il s'agira d'effectuer des analyses épidémiologiques sur les statistiques relatives à la Covid-19. Ces analyses seront basées sur la découverte des cold spots et hotspots spatiaux tout au long de l'évolution de l'épidémie de la Covid-19.

Mots clés : Data mining spatial, Hotspots, Cold spot, Covid 19

Abstract

Most of the large databases currently available have a strong spatial component. The discipline responsible for extracting this information and knowledge is data mining. Knowledge discovery is performed by applying automatic algorithms that recognize patterns in data. Classical data mining algorithms assume that data is independently generated and identically distributed. Spatial data is multidimensional, spatially autocorrelated and heterogeneous. These properties make classical data mining algorithms inappropriate for spatial data, and their basic assumptions cease to be valid. In this project, it will be a question of carrying out epidemiological analyzes on the statistics relating to Covid-19. These analyzes will be based on the discovery of spatial cold spots and hotspots throughout the evolution of the Covid-19 epidemic.

Key words: Spatial data mining, Hotspots, Cold spot, Covid 19

Liste des figures

Liste des tableaux

Liste des abréviations

Abréviation	Expression Complète	Page
SIG	Systèmes d'Information Géographiques	16
DM	Data Mining	17
BD	Base de données	20
DMS	Data Mining Spatial	21
GRW	Geographically Weighted Régression	24
OMS	Organisation mondiale santé	28
CNN	Convolutional neural network	28
UCLA	Université de Californie à Los Angeles	28
Getis-Ord G	The High/Low Clustering (Getis-Ord General G)	35
CSV	Comma-separated values	39
SHP	Shapefile	39

Table des matières

Introduction Générale

Le coronavirus, appartenant à la famille des Coronaviridae, a été identifié pour la première fois dans la province chinoise de Wuhan le 8 décembre 2019. Les symptômes de la maladie durent généralement environ cinq jours, avec une période d'incubation pouvant varier de deux à quatorze jours. La maladie, désignée par l'Organisation mondiale de la santé, s'est rapidement propagée dans le monde entier, devenant une pandémie. Elle a touché plus de 674 millions de personnes et entraîné plus de 6,86 millions de décès dans plus de 196 pays. Les pays les plus touchés sont les États-Unis, le Brésil, la Russie, l'Inde, l'Espagne et l'Italie.

L'objectif de ce travail est de réaliser des analyses épidémiologiques basées sur les statistiques relatives à la Covid-19 dans le monde. Ces analyses se concentreront sur l'identification des hotspots et cold spots spatiotemporels tout au long de l'évolution de l'épidémie, depuis son apparition jusqu'à aujourd'hui.

Pour atteindre cet objectif, ce mémoire commence par un chapitre qui introduit la notion d'information spatiale en fournissant des définitions et en expliquant les spécificités et les formats de représentation associés. Nous aborderons également le concept d'analyse spatiale.

Le deuxième chapitre est consacré au data mining spatial. Nous présenterons les définitions du data mining spatial, ainsi que les différentes tâches, approches et domaines d'application associés à cette discipline.

Le troisième chapitre est consacré à méthodologie, nous présenterons le contexte et l'objectif de notre étude, qui consiste à développer une solution pour la détection des hotspots et cold spots dans le cadre de la pandémie de la Covid-19.

Dans le quatrième chapitre, nous détaillerons notre expérimentation, en commençant par présenter les modèles de données utilisés et les outils que nous avons employés dans notre projet. Nous exposerons également l'algorithme sélectionné pour la réalisation de notre solution. De plus, nous décrirons l'interface utilisateur que nous avons développé, et nous présenterons notre étude de cas spécifique dédiée à la détection des hotspots et cold spots, dans le but d'analyser les taux d'incidence de la Covid-19 en Italie.

Enfin, nous clôturerons ce mémoire par une conclusion générale, qui sera présentée dans la dernière section.

Chapitre 1

L'information spatiale

1.1 Introduction

L'information spatiale est une représentation d'objets ou de phénomènes réels situés dans l'espace à un instant donné. Les données spatiales sont généralement stockées sous forme de coordonnées et de topologie, qui peuvent être représentées spatialement sur une carte, généralement traité et analysé par SIG. Dans ce chapitre, nous commencerons par décrire l'information spatiale, ses spécificités, ses formats de représentation. Nous parlerons ensuite des SIG, de l'analyse spatiales. Nous finirons ce chapitre par une conclusion.

1.2 L'information spatiale

Également connue sous le nom de donnée géospatiale ou d'information géographique, ontrouve plusieurs définitions de cette notion :

« Représentation d'un objet ou d'un phénomène réel, localisé dans l'espace à un moment donné [1].

« L'information géographique peut-être définie comme une information relative à un objet géographique ou à un phénomène du monde terrestre, décrit plus ou moins complètement : par sa nature, son aspect, ses caractéristiques diverses, et par son positionnement sur la terre » [2].

Plus simplement, on dira que l'information géographique est la combinaison de deux informations qui définissent un objet :

- Géométrique : Localisation, forme et dimension de l'objet
- Sémantique : Attributs décrivant l'objet

Les données spatiales sont généralement stockées sous forme de coordonnées et de topologie, elles peuvent être représentées spatialement sur une carte et sont souvent traitées etanalysées par les systèmes d'information géographique. Dans les ensembles de données spatiales, il existe à la fois des données spatiales et des données non spatiales. Les données nonspatiales sont des nombres, des caractères ou des types logiques. Les données spatiales peuventêtre représentées par des points, des lignes ou des polygones, elles possèdent des coordonnées géographiques référencées sur la surface terrestre.

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

Anselin et Bera (1998) définissent l'autocorrélation spatiale comme la coïncidence de la similarité de valeur avec la similarité de localisation. L'autocorrélation spatiale positive se traduit par une tendance à la concentration dans l'espace de valeurs faibles ou élevées d'une variable aléatoire [3]. Il est généralement confronté à trois types de localisation : points représentant par exemple des localisations d'unités de production ou de distribution, lignes, connectées entre elles ou non, comme un réseau routier ou fluvial. Les données sont parfois fournies pour des aires géographiques comme des régions ou des pays, et le nombre de ces points, de ces lignes ou de ces zones est supposé fini.

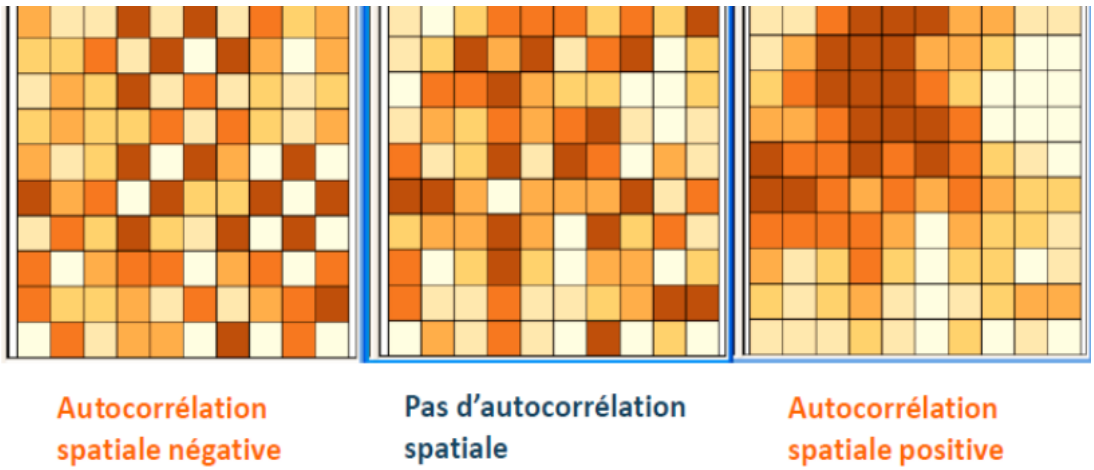


Figure 1: Représentation les différents types de corrélations spatiales

L'autocorrélation spatiale a deux sources principales : les processus d'interactions et les localisations proches. La diffusion d'un phénomène à partir d'un ou de plusieurs lieux d'origineimplique que l'intensité de la mesure dépend de la distance à l'origine

Il existe différents types de corrélations spatiales. Elle peut être positive, négative ou neutre.

- **Corrélation spatiale positive :** Une autocorrélation spatiale positive se produit lorsque des valeurs similaires se regroupent sur une carte.

Corrélation spatiale négative : L'autocorrélation spatiale négative se produit lorsque des valeurs dissemblables sont côte à côte sur une carte.

- **Pas de corrélation spatiale :** L'absence de corrélation spatiale signifie qu'il n'y a pas de relation entre les valeurs d'une variable dans des emplacements géographiques

1.3.2 Hétérogénéité spatiale

L'hétérogénéité spatiale est une propriété d'un paysage ou d'une population dans laquelle différentes concentrations d'individus sont inégalement réparties dans une zone.

L'hétérogénéité spatiale peut être locale ou stratifiée. La première est appelée hétérogénéité spatiale locale, qui fait référence au phénomène selon lequel la valeur d'attribut d'un lieu est différente de son environnement environnant, comme les points chauds ou les points froids, la seconde est appelée hétérogénéité spatiale stratifiée, qui fait référence à des phénomènes où lavariance intra-strate est inférieure à la variance inter-strates, comme les écorégions et les classes d'utilisation des terres [4].

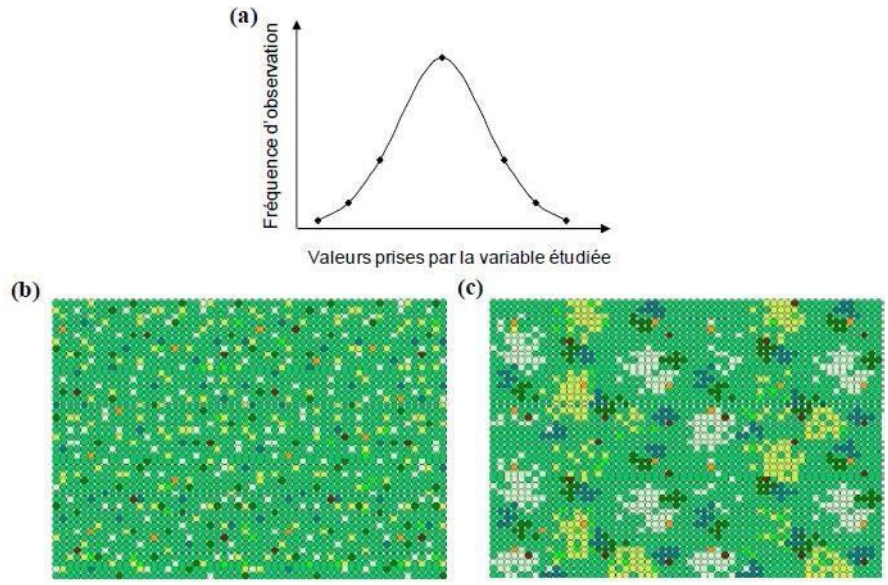


Figure 2: Représentation l'Hétérogénéité spatiale

Cette figure représente les motifs spatiaux :

a) Donne une représentation schématique de la variabilité pouvant correspondre à différents agencements spatiaux des valeurs de la variable.
b) Motif aléatoire : peuvent être décrites comme un modèle dans lequel les emplacements des individus sont indépendants les uns des autres.
c) Motif agrégé non aléatoire : est un modèle dans lequel les localisations des individus ne sont pas indépendantes les unes des autres et sont regroupées.

1.4 Les modes de représentation de l'information spatiale

Les SIG exploitent deux modes de représentation numérique de cette information :

Le mode vecteur : les entités sont représentées au moyen de formes géométriques. Les objets ponctuels, linéaires ou zonaux sont décrits par un ensemble de points déterminant leur contour.

Le mode raster ou matriciel : Les données raster sont représentées sous la forme d'une matrice de points. Les données de type raster sont principalement des photographies numériques, des images satellites ou des plans scannés, L'unité de base des données rasterest le pixel.

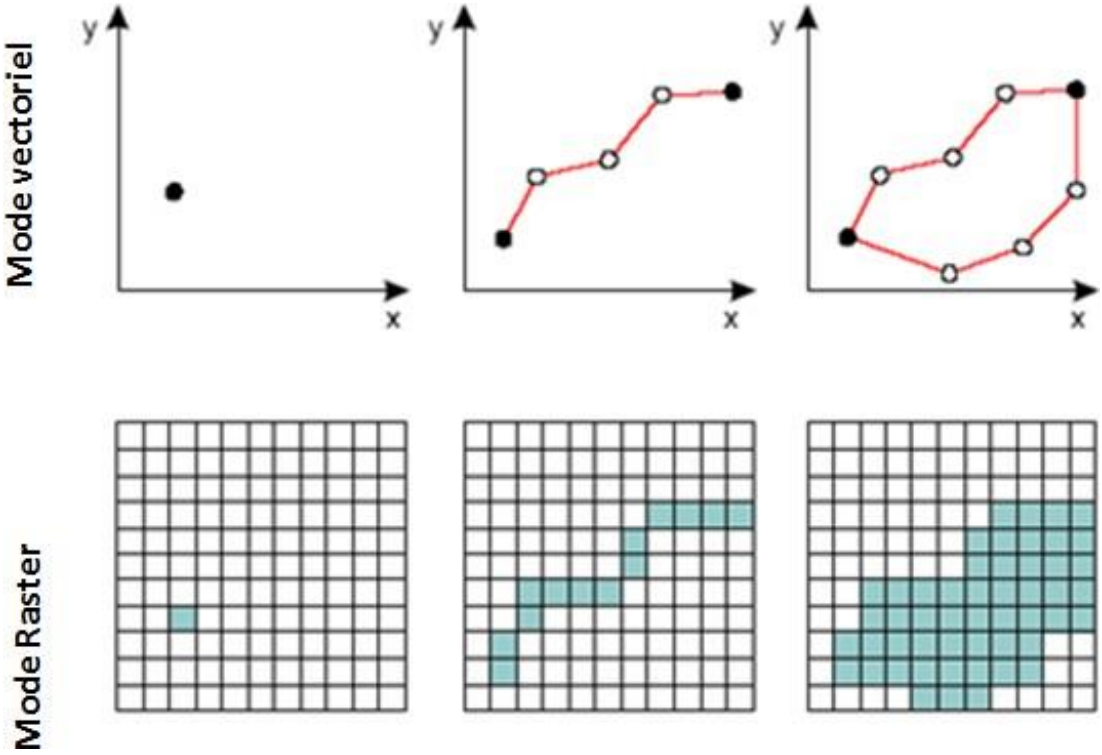


Figure 3: Représentation des formats vectoriels et raster

1.5 Systèmes d'information géographique

Un système d'information géographique ou SIG (en anglais, Geographic Information System ou GIS) est un système d'information conçu pour recueillir, stocker, traiter, analyser, gérer et présenter tous les types de données spatiales et géographiques [5].

Le terme de SIG décrit un système d'information qui intègre, stocke, analyse et affiche l'information géographique. Les applications liées aux SIG sont des outils qui permettent aux utilisateurs de créer des requêtes interactives, d'analyser l'information spatiale, de modifier et d'éditer des données par l'entremise de cartes et d'y répondre cartographiquement

1.6 L'analyse spatiale

L'analyse spatiale est définie comme le processus d'étude des entités en examinant, évaluant et modélisant les caractéristiques des données spatiales telles que les emplacements, les attributs et leurs relations qui révèlent les propriétés géométriques ou géogra-

phiques des données. Les méthodes d'analyse spatiale ont pour objectif de déterminer les caractéristiques de la distribution spatiale des individus géographiques (les ménages par exemple) ou de leurs valeurs (par exemple : le nombre de cas de COVID dans le ménage) [6].

En plus simple : l'analyse spatiale est un type d'analyse géographique qui cherche à expliquer les modèles de comportement humain et son expression spatiale en termes de mathématiques et de géométrie, c'est-à-dire l'analyse de localisation

1.7 Conclusion

Les données spatiales sont multidimensionnelles, spatialement auto corrélées et hétérogènes. Dans ce chapitre, nous avons commencé par décrire l'information spatiale, ses spécificités, ses formats de représentation. Puis nous avons parlé de l'analyse spatiale. Le chapitre suivant sera dédié au data mining spatial.

Chapitre 2

Le Data Mining Spatial

2.1 Introduction

Le DM est généralement défini comme l'extraction de connaissances intéressantes dans le but de révéler des modèles difficiles à mettre en évidence. Le DMS est une branche du DM, qui est un domaine interdisciplinaire, qui est en relation avec d'autres domaines tels que les statistiques spatiales, l'analyse spatiale et les bases de données spatiales.

Dans ce chapitre, nous commencerons par décrire Le DM, Puis nous donnerons une définition du DMS, et nous détaillerons son processus, ses phases d'exécutions, ses approches, ses tâches et des exemples d'utilisation du DMS pour l'analyse de la pandémie de la Covid 19. Nous finirons ce chapitre par une conclusion.

2.2 Le data mining

Le data mining est une technique dont le but est d'évaluer l'information et extraire des connaissances à partir de grandes quantités de données. La technologie est devenue un outil important pour augmenter les revenus des entreprises et répondre aux problèmes dus à l'énormequantité de données à exploiter [7].

2.3 Les taches du data mining

L'analyse de beaucoup de problèmes intellectuels, économiques ou même commerciauxpeuvent être exprimés en termes des six tâches suivantes :

- La classification
- L'estimation

- La prédiction
- Le groupement par similitude
- L’analyse des clusters
- La description

Les trois premières tâches sont des exemples de Data Mining supervisé, qui consistent à créer des modèles prédictifs en utilisant des données étiquetées pour prédire des variables spécifiques. Cela implique de cibler des données spécifiques, les regrouper en fonction de leur similarité et les analyser. Le clustering, en revanche, est une tâche non supervisée dont l'objectif est de regrouper les données sans utiliser d'étiquettes préalables, afin d'identifier des structures ou des relations intrinsèques dans les données. La description mentionnée peut être associée à ces deux types de tâches, car elle implique à la fois une analyse ciblée de variables spécifiques et une analyse non supervisée de la relation entre toutes les variables.

2.3.1 La classification

La classification est un processus essentiel pour comprendre notre vie quotidienne, impliquant la catégorisation d'objets, de données ou de phénomènes en classes bien définies à l'aide d'exemples classés. Elle trouve de nombreuses applications dans différents domaines de recherche et de commerce. Quelques exemples d'utilisation de la classification incluent :

- Détection de fraudes dans l'utilisation de cartes de crédit.
- Diagnostic de maladies en identifiant les symptômes et les signaux précurseurs.
- Identification des numéros de téléphone destinés à la transmission de fax.
- Sélection de la ligne téléphonique appropriée pour accéder à Internet.

2.3.2 L’estimation

L'estimation est un processus qui vise à prédire ou à compléter une valeur manquante dans un champ spécifique. Elle est similaire à la classification mais se concentre sur l'estimation plutôt que sur la catégorisation.

Voici quelques exemples d'utilisation de la tâche d'estimation dans divers domaines de recherche et de commerce :

- Estimer le nombre d'enfants dans une famille en se basant sur des informations démographiques.
- Estimer le montant qu'une famille de quatre personnes sélectionnées au hasard dépenserait pour la rentrée scolaire en utilisant des données statistiques.
- Estimer la valeur d'un bien immobilier en se basant sur des facteurs tels que la taille, l'emplacement et les caractéristiques.

2.3.3 La prédiction

La prédiction se concentre sur la relation entre les variables d'entrée et les variables de sortie, permettant ainsi d'estimer ou de prévoir une valeur future. Elle est largement utilisée dans divers domaines de recherche et de commerce. Voici quelques exemples d'utilisation de la tâche de prédiction :

- Prédire les cours des actions pour les trois prochains mois en analysant les tendances historiques, les facteurs économiques et les nouvelles du marché.
- Prédire le vainqueur de la Coupe du monde de football en se basant sur l'analyse des performances des équipes, des statistiques des joueurs et d'autres facteurs pertinents.
- Prédire quels clients déménageront dans les 6 prochains mois en analysant des données telles que les changements d'adresse précédents, les modèles de comportement et les caractéristiques démographiques.

2.3.4 Le groupement par similitude

La tâche la plus couramment utilisée dans le domaine du commerce est l'analyse d'association, qui permet de mesurer les relations entre deux ou plusieurs attributs. Elle vise à identifier les liens et les associations entre des variables dans un ensemble de données. Voici quelques exemples d'utilisation de la tâche d'analyse d'association dans les domaines de la recherche et du commerce :

- Découvrir quels produits sont achetés ensemble et ceux qui ne sont jamais achetés ensemble dans un supermarché. Cela permet de comprendre les habitudes d'achat des clients, de recommander des produits complémentaires et d'optimiser le placement des produits dans les rayons.

- Déterminer la proportion de cas dans lesquels un nouveau médicament peut avoir des effets dangereux. Cela permet d'identifier les combinaisons de médicaments ou de conditions médicales qui pourraient entraîner des interactions indésirables ou des effets secondaires, aidant ainsi à prendre des décisions éclairées en matière de sécurité et de traitement.

2.3.5 L’analyse des clusters

Le clustering est une méthode permettant de regrouper des enregistrements ou des observations en classes d'objets similaires. Les algorithmes de clustering visent à maximiser l'homogénéité au sein des classes et à minimiser la variance entre les classes. Cette approche est largement utilisée dans divers domaines, notamment :

- Découvrir des groupes de clients ayant des comportements similaires, ce qui permet de mieux comprendre les segments de marché et d'adapter les stratégies de marketing en conséquence.
- Classer les plantes et les animaux en fonction de leurs caractéristiques communes, facilitant ainsi l'identification et la classification des espèces.
- Segmenter les observations des épicentres sismiques pour identifier les zones à risque élevé, ce qui peut contribuer à la prévention des catastrophes naturelles et à la planification d'urgence.

2.3.6 La description

Parfois, l'objectif du Data Mining est simplement de décrire les relations existantes dans les données afin de mieux comprendre les individus, les produits et les processus. Par exemple, la constatation selon laquelle les femmes soutiennent davantage le parti démocrate peut susciter un grand intérêt et promouvoir des études menées par des journalistes, des sociologues et des spécialistes en politique. Ces analyses des données permettent de mettre en évidence des tendances et des modèles qui peuvent aider à comprendre les comportements, les préférences et les opinions des individus, ainsi que leurs relations avec des facteurs socio-économiques, politiques ou culturels.

2.4 Définition du data mining spatial

Le data mining spatial, est l'exploration de données ou encore l'extraction de connaissances implicite de relations spatiales ou autre propriétés non explicitement stockés dans les bases de données spatiale à partir d'une grande masse de données géographiques [8].

2.5 Approches du data mining spatial

Il existe deux approches pour l'analyse et l'extraction de connaissances d'une base de données spatiales. La première est issue des statistiques spatiales et la seconde du domaine desbases de données. Très peu de liens existent aujourd'hui entre ces deux types de recherches. Malgré cela, elles permettent parfois de résoudre les mêmes tâches d'analyse et ont certains points en commun [9].

2.5.1 Approche statistique

Les méthodes statistiques sont les méthodes les plus courantes pour analyser et modéliserdes données spatiales. Les statistiques spatiales ne satisfont pas l'hypothèse de distributions identiques et indépendantes. Les méthodes statistiques utilisées par le DMS peuvent s'appliquer sur les données décrites ci-dessous :

2.5.1.1 Les données géostatistiques

La géostatistique traite de l'analyse de la continuité spatiale, de la faible stationnarité (contraire de l'hétérogénéité spatiale) des données ponctuelles spatiales dans des sous-espaces continus. Il fournit une suite d'outils statistiques tels que le Krigeage et les problèmes d'agrégation spatiale.

2.5.1.2 Les données laticielles

Le champ de recherche est discret et fixe, et les sites de recherche sont agencés selon le réseau formé par le graphe de voisinage. Des statistiques d'autocorrélation spatiale peuvent êtredéfinies pour mesurer la corrélation des attributs non spatiaux dans les régions voisines.

2.5.1.3 Les données ponctuelles

Les processus ponctuels spatiaux sont des modèles de distribution spatiale de points dansl'espace, et contrairement aux données géostatistiques, la variable aléatoire est l'emplacement.Les statistiques d'analyse spatiale peuvent être utilisées pour détecter ces points chauds.

2.5.1.4 Statistiques des réseaux spatiaux

La plupart des recherches sur les statistiques spatiales se sont concentrées sur l'espace euclidien, mais les réseaux sont importants et posent des défis inhérents aux applications aux sciences de l'environnement et à l'analyse de la sécurité publique.

2.5.2 Approche de base de données

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

2.5.2.1 Les index de jointure spatiale

Une façon de réduire ce coût consiste à utiliser des indexes de jointure. Structure Les index de jointure ont été proposés par Valduriez en 1987 [10]. Le principe de la technologie comprend l'utilisation d'une structure de type tableau pour stocker deux paires d'index dans une base de données. La technique convertit la présence de relations spatiales, telles que la relation de contiguïté entre une paire d'objets spatiaux. La généralisation des index de jointure aux données spatiales a été proposée par Valduriez, elle est illustrée dans la Figure 4 [10].

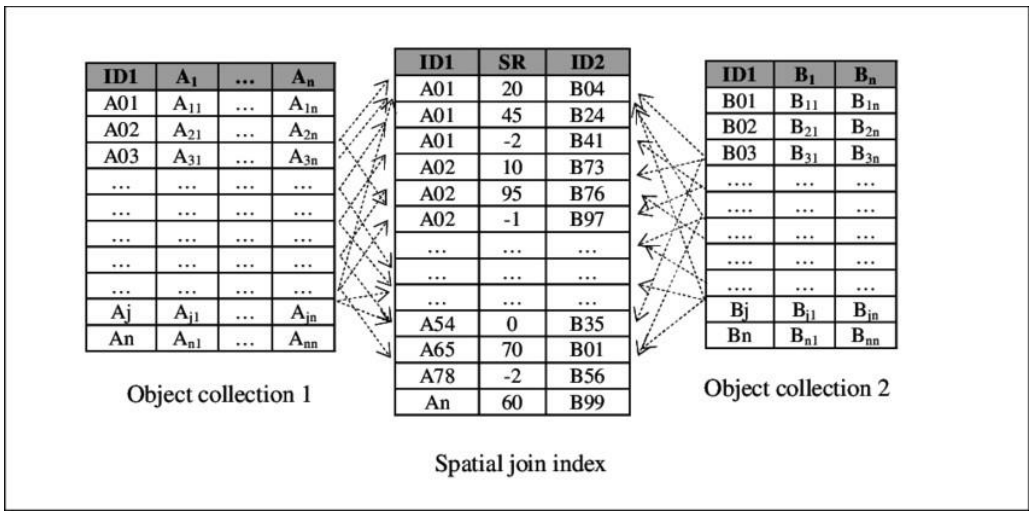


Figure 4 : Index de jointure spatial

2.5.2.2 Les matrices de contiguïtés

Une matrice de contiguïté a une structure très similaire. Elle encode les relations spatialesPar paire d'indices. C'est la structure utilisée pour analyser le problème espace de données. Lescritères peuvent être des critères de distance ou de relation spatiale.

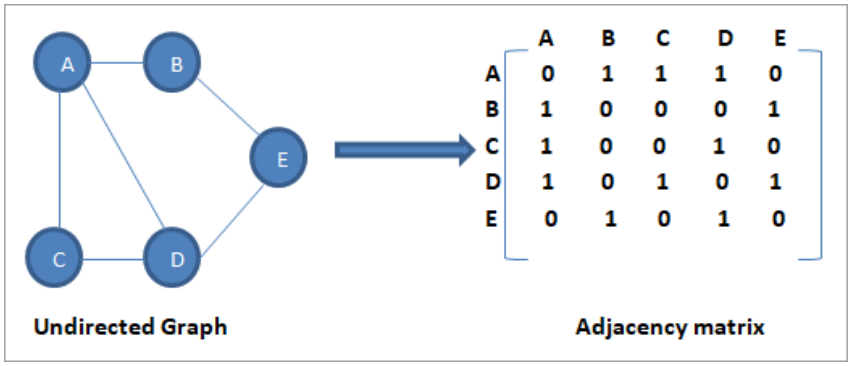


Figure 5: Illustration d'une matrice de contiguïté

2.5.2.3 Les data cubes spatiaux pour le data mining spatial

La motivation pour l'exploration de données à l'aide de cubes de données spatiales est liée au concept de hiérarchie spatiale. En fait, la relation topologique représente la structure des relations hiérarchiques qui correspondent à des relations sémantiques hiérarchiques entre des agrégations spatiales.

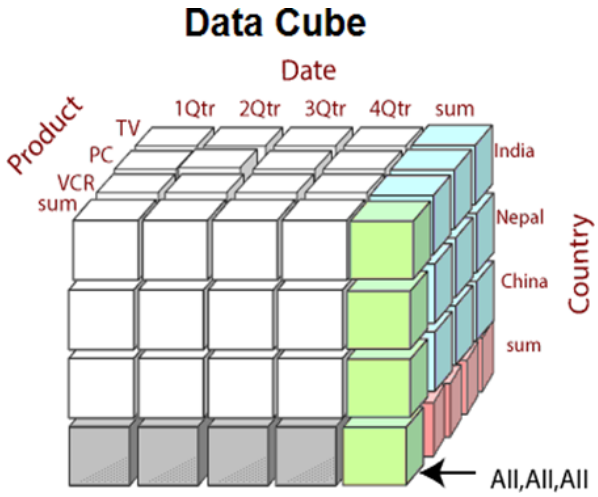


Figure 6: Les data cubes spatiaux

D'autre part, les données à dimension spatiale sont structurées selon une hiérarchie spatiale. Ces données peuvent être agrégées à chaque niveau de la hiérarchie en réalisant une correspondance entre les relations topologiques et les hiérarchies. La hiérarchie spatiale peut alors être réalisée et les méthodes appliquées sur chaque niveau de la hiérarchie spatiale

2.6 Les taches du data mining spatial

2.6.1 La classification spatiale

Les règles de classification spatiale regroupent un ensemble de données de manière à maximiser la similarité des fonctionnalités dans les grappes et minimiser la similitude entre deux grappes. Les objets spatiaux sont répartis en différents groupes en prenant en compte la similarité de leurs caractéristiques et en faisant en sorte que la différence entre les objets d'un même groupe soit aussi grande que possible. Les algorithmes de classification couramment utilisés peuvent être basés sur la partition, la hiérarchie, la densité et les grilles. Ces méthodes sont en mesure de tenir compte de plusieurs couches thématiques et en même temps d'étendrecritères discriminants pour régler les effets de voisinage [11].

2.6.2 La prédiction spatiale

La prédiction spatiale est importante pour les événements survenus à des emplacements géographiques particuliers, et la modélisation des dépendances spatiales à l'aide

des modèles classiques de régression améliore la précision globale. Le modèle Geographically Weighted Regression (GWR) est utilisé pour les prédictions locales par contre le modèle des moindres carrés ordinaire. La régression pondérée géographiquement est un outil puissant pour explorerl'hétérogénéité spatiale [12].

2.6.3 Les règles d'association spatiale

Koperski définit la règle d'association spatiale comme une règle de la forme :

P1 Λ P2…Λ Q1 Λ Q2 ... Λ Qn (I.1)

Pi ; Qj : sont des prédicats où au moins un des prédicats est un prédicat spatial.

La co-location est un type spécial d'association spatiale et l'autocorrélation spatiale est utiliséepour mesurer la force des relations entre les objets spatiaux du même type [13].

2.6.4 Le clustering spatial

L'objectif du clustering ou du regroupement est de trouver le nombre optimal de clusters dans un ensemble de données, fournissant des connaissances sur les modes de partition spatiale globale des objets dans le jeu. Il existe deux approches de regroupement : spatial et non-spatial[14].

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

Les hotspots se réfèrent à la zone où certains événements se produisent fréquemment, et il existe deux principaux types de méthodes de détection des points chauds spatiaux : l'analyse spatiale de points et les statistiques d'autocorrélation spatiale locale [15].

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

La détection des valeurs aberrantes est utilisée pour extraire des exceptions intéressantes dans les ensembles de données par le DMS. Il s'agit d'un exercice subjectif pour déterminer si une observation est ou n'est pas un résultat aberrant [16].

2.7 Processus du data mining spatial

Le processus détaillé du DMS peut être divisé en cinq phases décrites comme suit : enquête surla demande, sélection des données, prétraitement des données, conversion des données,exploration des données, représentation et évaluation des connaissances. Dans la Figure 7, le flux de SDM est décrit [17].

2.7.1 Investigation

Il est nécessaire de comprendre les données existantes et les informations avant l'exploration de données. Comprendre pleinement les problèmes à résoudre et donner une définition claire sur l'objectif de l'exploration de données. Par conséquent, l'enquête sur la demande est la première étape nécessaire du DMS en fonction de la tâche orientée vers l'application réelle.

2.7.2 La sélection de données

La sélection des données est effectuée après la fin de l'enquête sur la demande et la connaissance de la demande sans ambiguïté. La sélection des données consiste à déterminer lasource de données à utiliser dans l'exploration de données et à collecter les enregistrements dedonnées stockés dans la base de données conformément aux normes en cours d'établissement.Certaines règles ou méthodes de sélection des données doivent être établies ou adoptées pour sélectionner les données nécessaires au cours du processus.

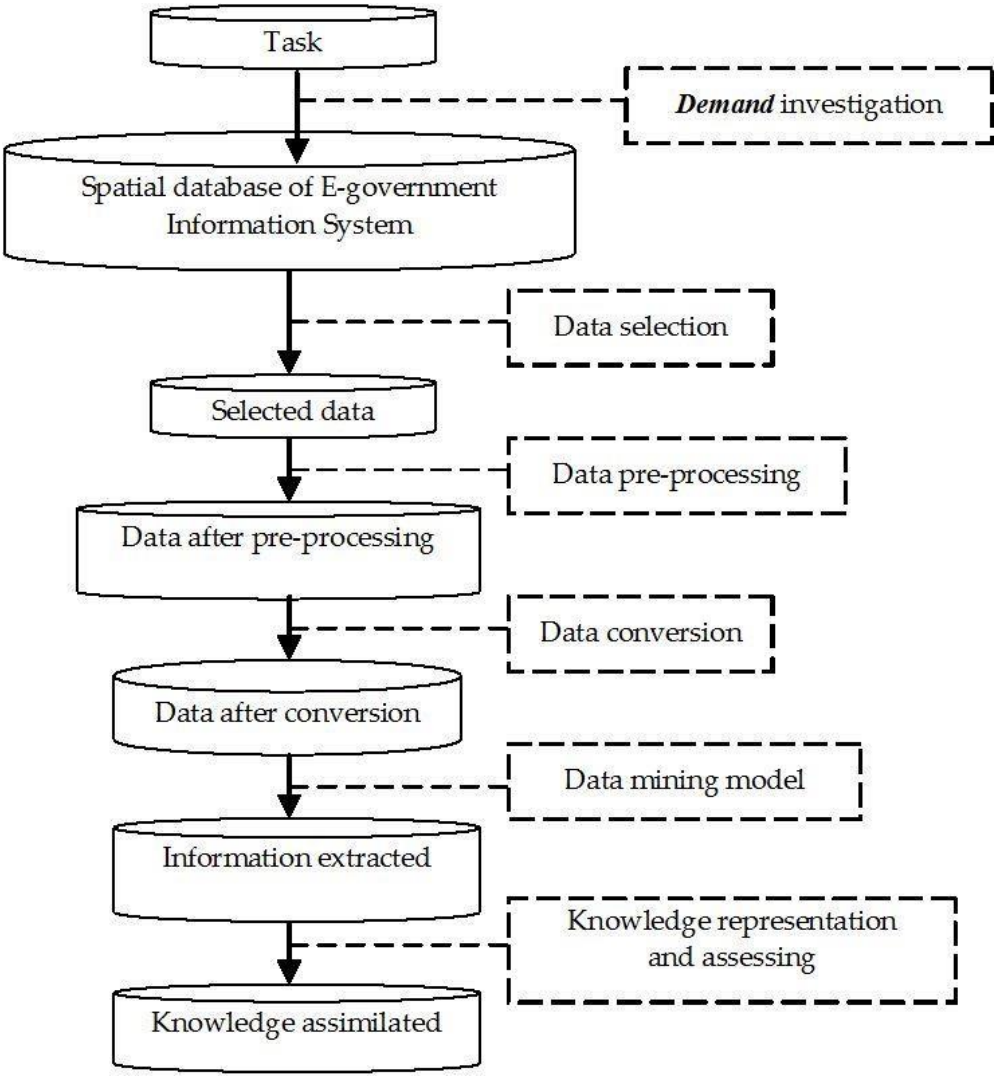


Figure 7 : Processus du DMS

2.7.3 Traitement des données

Une fois la sélection des données terminée, il est nécessaire de procéder au prétraitement des données. Le travail principal de cette étape consiste à effectuer un nettoyage des données visant les données stockées dans la base de données du système d'information du gouvernementélectronique et à supprimer les informations inutiles, et à transformer les données requises en un format de données unifié. En outre, la conversion des données se fera par le processus de fusion et d'intégration des données pour les convertir en données avec identification après prétraitement des données.

2.7.4 Construction du modèle de data mining spatial

C'est l'étape importante pour construire le modèle d'exploration de données. Des modèles d'exploration de données correspondants doivent être construits en visant différentes exigences,telles que l'arbre de décision, le clustering, etc. Il est décisif pour le prétraitement des données de sélectionner un certain type de modèle. L'exploration de données peut être effectuée en visant les données stockées dans la base de données du système d'information du gouvernementélectronique et l'extraction de modèles peut également être effectuée après avoir déterminé le modèle d'exploration de données.

2.7.5 Représentation et évaluation des connaissances

Les résultats doivent être interprétés lorsque l'exploration des données est terminée. Au cours du cours, une partie du processus peut être renvoyée aux étapes de traitement avant afin d'obtenir une connaissance plus efficace. Et l'extraction des connaissances peut être effectuée à plusieurs reprises afin d'obtenir des informations plus efficaces à interpréter comme les connaissances nécessaires à la prise de décision à l'avenir. Enfin, les performances du modèle appliqué nécessitent d'évaluer et d'améliorer constamment l'algorithme pour répondre aux besoins des différentes applications réelles.

2.8 Domaines d'application du data mining spatial

Le data mining spatial peut être utilisée pour la compréhension des données spatiales, la découverte des relations entre les données spatiales et non spatiales, la construction de bases de connaissances spatiales, l'optimisation des requêtes, la réorganisation des données dans desbases de données spatiales, la saisie des caractéristiques générales de manière simple et concise [18]. Ainsi, la technique peut être appliquée pour la détection, la cartographie et la prédiction detout phénomène qui manifeste une composante spatiale.

Le DMS trouve son application aussi bien dans le domaine public, dans le domaine scientifique que dans le secteur privé. Mais les objectifs ne sont pas identiques. En exploitant les données géographiques, Le DMS sert à la classification automatique d'objets spatiaux, ou bien à découvrir des régions dignes d'intérêt, ou des objets rares dans l'immensité de notre univers. En archéologie, les données géographiques et la fouille de données spatiales sont exploitées pour trouver de nouveaux sites. Le data mining spatial est utilisée

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

2.9 Exemples d'application du data mining spatial pour l'analyse de la Covid-19

Dans ce qui suit, nous citons des exemples d'application du data mining spatial pour l'analyse de la Covid-19 :

- **Détection de hotspots :** Les hotspots de COVID-19 sont généralement associés à une augmentation rapide du nombre de cas dans une région spécifique. Pour identifier ces hotspots, différentes approches ont été utilisées, notamment l'analyse de l'imagerie par le Data Mining Spatial (DMS). Cette méthode permet d'analyser les images afin de détecter les zones présentant des caractéristiques spécifiques et une concentration inhabituelle de l'épidémie. Par exemple, l'Organisation mondiale de la santé (OMS) a conclu que cette augmentation pourrait être attribuée à divers facteurs [19], tels que la densité de population, la mobilité des individus, le manque de distanciation sociale et les déplacements.
- **Classification spatiale :** Plusieurs équipes de recherche ont utilisé l'imagerie satellite pour cartographier la répartition géographique de l'épidémie de la Covid-19. Par exemple, l'analyse spatiale joue un rôle crucial dans la surveillance de l'épidémie au niveau des villes [20], Les autorités peuvent utiliser ces analyses pour mieux comprendre les dynamiques temporelles et spatiales de l'épidémie, ainsi que pour répondre aux besoins urgents des populations vulnérables, notamment les personnes âgées, les enfants et les travailleurs exposés. De plus, des modèles basés sur des réseaux de profondeur convolutionnelle (CNN) peuvent être utilisés pour détecter le virus en classifiant les images radiographiques [21].
- **Prédiction spatiale** : La prédiction spatiale pour la COVID-19 est une étude qui a été notamment menée par des chercheurs de l'UCLA et de Microsoft Research. Cette étude a utilisé des données de géolocalisation pour prédire les tendances des maladies dans la ville de Los Angeles, Les modèles de prédiction spatiale peuvent être utilisés pour

prévoir l'épidémie de COVID-19 à l'aide de la méthode d'analyse de la moyenne mobile intégrée autorégressive (ARIMA) [22] ainsi que du système d'information géographique (SIG) et des algorithmes d'apprentissage automatique [23]. Les méthodes existantes de prédiction des risques de COVID-19 se concentrent principalement sur les cas confirmés.

- **Règles d'association spatiale :** Une étude menée par des chercheurs de l'Université de Hong Kong a utilisé des règles d'association spatiale pour déterminer la relation entre la mortalité par COVID-19 et des facteurs géographiques tels que la densité de population, le pourcentage d'adultes âgés et la présence de comorbidités. Les chercheurs ont découvert que les zones à forte densité de population et les populations vieillissantes avaient des taux demortalité plus élevés.
- **Analyse des valeurs aberrantes spatiales :** Une autre étude à procéder à la détection des valeurs aberrantes, elle est basé sur l'outil ArcGIS Geoprocessing pour identifier les points aberrants spatiaux et temporels, ainsi que les agrégats de points chauds et froids statistiquement significatifs [24].
- **Data mining spatiotemporel :** Une autre étude a évalué la dynamique spatiotemporelle de l'évolution de la Covid-19 en utilisant des systèmes d'information géographique et des méthodes d'analyse spatiale pour détecter les valeurs aberrantes spatiales [25].
- **Regroupement spatial :** Une étude a analysé la transmission spatiale du Covid-19 par les transports publics et privés en Chine [26]. Une autre étude s'est penchée sur les déterminants socio-économiques de l'hospitalisation et de la mortalité, tous deux associés au Covid-19 d'unepart et à la surmortalité d'autre part [27].
- **Régression spatiale :** Une étude a identifié un sous-ensemble de déterminants spatiaux de la Covid 19 aux États-Unis à l'aide d'une analyse de cluster et d'une régression [28], Les résultats ont révélé que les zones urbaines et les lieux comptant une plus grande proportion d'individus étaient associés à un nombre plus élevé de cas et de décès.
- **Visualisation géospatiale** : L'analyse géospatiale et le SIG ont été utilisés pour étudier la dimension géographique de la pandémie de COVID-19. Cela comprend la visualisation et l'analyse des virus dans le contexte de cartes Comme la carte de l'Université Johns Hopkins et la carte MSD Manuals [29]. Ces cartes permettent la combinaison

de visualisations géographiques avec d'autres facteurs d'influence tels que l'âge, le sexe, la profession et la création de visualisationsde données.

- **Classification et prédiction :** Plusieurs études ont proposé des modèles de prédiction et de classification de la COVID-19 basés sur des variables dérivées des patients [26]. Par exemple, des recherches ont mis en évidence l'impact de la structure par âge de la population et des liens intergénérationnels sur les différences dans les décès liés au COVID-19 entre les pays. Une étude de cohorte descriptive longitudinale a analysé les paramètres de numération globulaire complète, l'âge du patient, le sexe et les comorbidités, suggérant ainsi l'utilisation de la démographie des patients et des résultats de la prise en charge comme biomarqueurs cliniques prédictifs. Dans le cadre de cette étude, la tomodensitométrie thoracique a été utilisée comme outil pour évaluer la gravité de la COVID-19 chez les personnes symptomatiques présentant un risque élevé [30].

Le DMS est donc une méthode utile pour explorer les données spatiales et géospatiales liées au COVID-19. Il peut être utilisé pour explorer les modèles spatio-temporels et les modèles aberrants de COVID-19 au niveau de la ville, ainsi que pour analyser les effets spatio-temporels des facteurs moteurs sur les incidences de COVID-19. Il peut également être utilisé pour analyser la dynamique de la transmission du COVID-19 et les facteurs de risque associés et visualiser des informations géographiques.

2.10 Approches de détection des hotspots et cold spots spatiaux

La détection des hotspots et cold spots spatiaux est une tâche importante du DMS qui est largement utilisée pour l'analyse des pandémies et notamment la pandémie de la Covid 19. Pour la mettre en œuvre, plusieurs approches peuvent être utilisées :

2.10.1 Approches basées sur la localisation spatiale

Les approches basées sur la localisation spatiale sont une approche géographique qui étudie les localisations et les interactions spatiales en tant que composantes actives des phénomènes étudié.

2.10.2 Approches basées sur l'analyse des données épidémiologiques

Les approches basées sur l'analyse des données épidémiologiques sont des méthodes d'analyse qui se concentrent sur l'utilisation de données de santé pour comprendre les tendances et les modèles de maladies dans une population donnée. Ces approches nécessitent souvent l'utilisation de nouvelles approches statistiques pour l'analyse des grands jeux de données en épidémiologie.

2.10.3 Approches basées sur les réseaux sociaux et les médias numériques

Les approches basées sur les réseaux sociaux et les médias numériques sont des méthodes d'analyse qui se concentrent sur l'utilisation des réseaux sociaux numériques et des médias sociaux pour comprendre les comportements et les interactions des individus en ligne.

2.10.4 Approches basées sur l'apprentissage automatique

Les approches basées sur l'apprentissage automatique sont des méthodes d'analyse qui se concentrent sur l'utilisation de l'intelligence artificielle pour apprendre à partir de données et améliorer les performances de prédiction. Ces approches sont utilisées dans divers domaines, tels que la qualité d'usinage de pièces métalliques, la gestion des données de géolocalisation, la traduction automatique adaptative en profondeur, etc.

2.10.5 Approches combinées

Les approches combinées peuvent faire référence à une variété de sujets dans différents domaines, mais généralement, elles font référence à l'utilisation de plusieurs méthodes ou techniques pour atteindre un objectif commun. Par exemple, dans la recherche scientifique, la combinaison de différentes approches peut conduire à des connaissances plus approfondies et générer de nouvelles connaissances qui ne peuvent être obtenues par une seule approche.

2.11 Conclusion

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

Dans ce chapitre, nous avons d'abord décrit le data mining et ses taches. Nous avons ensuite parler du DMS et de l'importance des relations spatiales dans le processus de DMS. Puis, nous avons présenté les taches du DMS et des exemples d'application du DMS pour l'analyse d'une pandémie comme la Covid 19. Nous avons enfin cité les méthodes de détection des hotspots et cold spots spatiaux.

Dans le chapitre suivant, nous décrirons la méthodologie que nous avons suivi pour réaliser notre projet.

Chapitre 3

Méthodologie

3.1 Introduction

Dans ce chapitre, nous présenterons le contexte et l'objectif de notre étude, qui vise à développer une solution pour la détection des hotspots et cold spots dans le contexte de la pandémie de la Covid-19.

L'objectif principal de notre application est de fournir une solution adaptable pour l'analyse de données spatiales liées à la Covid-19. Nous avons conçu notre application de manière à pouvoir traiter et analyser différentes sources de données. Cependant, dans le cadre de notre étude, nous nous concentrerons spécifiquement sur l'incidence du Covid-19 dans les différentes régions d'Italie en tant que variables d'analyse.

Une caractéristique essentielle de notre application est sa capacité à être utilisée à différentes échelles géographiques. Cela nous permettra d'explorer la répartition spatiale de la Covid-19 à travers les différentes régions et d'identifier les schémas et variations géographiques associés.

3.2 Description de la méthode proposée

La méthode proposée dans notre étude repose sur le processus de data mining spatial et implique plusieurs étapes clés. Tout d'abord, nous effectuons une étape de prétraitement des données, qui comprend le nettoyage des données, la gestion des valeurs manquantes et l'intégration des données spatiales et attributaires. Cette étape garantit la qualité et l'intégrité des données utilisées dans notre analyse.

Ensuite, nous procédons à l'implémentation de techniques de détection des hotspots et cold spots. Nous choisissons des indicateurs appropriés, tels que le taux d'incidence ou la mortalité, qui mesurent l'activité virale de la COVID-19. Ces indicateurs sont utilisés pour cartographier la distribution spatiale de la maladie et identifier les zones à risque élevé ou faible de propagation.

Une fois que les hotspots et cold spots sont identifiés, nous passons à l'étape d'analyse des résultats. Nous examinons les caractéristiques spatiales des clusters détectés pour obtenir une compréhension approfondie de la répartition spatiale de la COVID-19.

Enfin, nous procédons à l'étape d'interprétation des résultats. Nous analysons les résultats obtenus en fonction du contexte épidémiologique. Cela nous permet de tirer des conclusions pertinentes sur les zones à risque élevé ou faible de propagation de la maladie, ainsi que de formuler des recommandations pour la prise de décision et la planification des mesures de lutte contre la COVID-19.

3.3 Prétraitement des données spatiales

Le prétraitement des données attributaires et spatiales est important afin d'assurer la qualité et l'intégrité des données utilisées dans notre analyse spatiale. Dans notre étude, nous avons suivi les étapes de prétraitement des données suivantes :

- **Nettoyage des données :** Cette étape consiste à éliminer les valeurs aberrantes, les doublons ou les valeurs incohérentes qui pourraient altérer les résultats de notre analyse.
- **Gestion des valeurs manquantes :** Dans notre cas, il n’y avait pas de valeurs manquantes.
- **Intégration des données spatiales et attributaires :** Nous avons fusionné les données géographiques et les attributs associés pour permettre une analyse conjointe et une meilleure compréhension des relations spatiales.
- **Validation des données :** Une vérification rigoureuse de l'intégrité des données a été réalisée pour détecter d'éventuelles erreurs ou incohérences. Cela nous permet de travailler avec des données fiables et de garantir la précision de nos résultats.

- **Préparation des données pour l'analyse :** Les données ont été structurées dans un format adapté aux techniques de détection des hotspots.

En effectuant ces étapes de prétraitement, nous nous assurons que les données utilisées dans notre analyse sont fiables, cohérentes et prêtes à être soumises à des techniques d'analyse spatiale avancées.

3.4 Choix des indicateurs pour la détection des hotspots et cold spots

Pour l'identification des hotspots et des cold spots spatiaux liés à la Covid-19, plusieurs indicateurs peuvent être utilisés. Parmi eux, on peut mentionner :

- **Taux d'incidence**

Le taux d'incidence est un indicateur épidémiologique qui mesure le nombre de nouveaux cas d'une maladie apparus dans une population au cours d'une période de temps déterminée. En épidémiologie, le taux d'incidence rapporte le nombre de nouveaux cas d'une pathologie observés pendant une période donnée – population

- **Taux de mortalité**

Le taux de mortalité est un indicateur démographique qui mesure le nombre de décès dans une population donnée sur une période de temps donnée. Il peut être utilisé pour mesurer la mortalité dans une région ou un pays.

- **Taux de croissance**

Le taux de croissance est un indicateur économique qui mesure l'évolution d'une grandeur (PIB, chiffre d'affaires, salaire, etc.) d'une période à l'autre (mois, trimestre, année) et est exprimé en pourcentage

- **Taux de positivité des tests**

Le taux de positivité, qui mesure le nombre de personnes testées positives pour la première fois depuis plus de 60 jours rapporté au nombre total de personnes testées positives ou négatives sur une période donnée

- **Taux de reproduction**

Le taux de reproduction de l'épidémie, est un indicateur qui mesure la capacité d'un virus à se transmettre d'une personne contaminée à une personne non malade.

Dans le cadre de notre étude, nous avons choisi d'utiliser le taux d'incidence comme indicateur principal dans notre étude en raison de sa pertinence pour évaluer la propagation du Covid-19. Il permet de mesurer le nombre de nouveaux cas confirmés sur une période spécifique, ce qui reflète la dynamique de transmission du virus. Le taux d'incidence nous permet également de comparer la situation épidémiologique entre différentes régions et de suivre les variations temporelles de la maladie. C'est un indicateur largement utilisé et valide dans la recherche épidémiologique.

3.5 Choix des algorithmes de détection des hotspots et cold spots

Il existe plusieurs algorithmes de détection des hotspots et cold spots spatiaux. Nous pouvons citer les algorithmes suivants :

- **Indice de Moran :** L'indice de Moran est une mesure de l'autocorrélation spatiale, qui caractérise la corrélation d'un signal entre des emplacements proches dans l'espace. Il est utilisé pour analyser les différences géographiques dans divers domaines, tels que la santé, l'économie et la géographie.
- **Indice de Getis-Ord :** L'indice Getis-Ord est une mesure de l'autocorrélation spatiale qui relie l'autocorrélation spatiale aux interactions spatiales. Il est utilisé pour mesurer des concentrations de valeurs élevées ou faibles pour une zone d'étude donnée.
- **Analyse de quadrants :** L'analyse par quadrant a été utilisée pour classer les associations spatiales en quatre types. Le nuage de points est divisé en quatre quadrants, les cas des quadrants supérieur droit et inférieur gauche indiquent une autocorrélation spatiale positive, et les cas des quadrants supérieurs gauche et inférieur droit indiquent une autocorrélation spatiale négative.
- **Analyse de densité de kernel :** L'estimation de densité de noyau est utilisée pour estimer la fonction de densité de probabilité d'une variable aléatoire. Il s'agit d'estimer la densité à chaque point d'une grille en additionnant les contributions des points de don-

nées proches. La densité à chaque point est calculée en additionnant les valeurs de toutes les surfaces du noyau où elles se chevauchent au centre de la cellule raster.

- **L'analyse spatiale du cluster scan de Kulldorff :** La statistique de balayage spatial de Kulldorff est une méthode de détection de grappes spatiales de maladies. Il est largement suggéré pour détecter les clusters locaux appropriés par rapport à d'autres méthodes spatiales
- **L'analyse spatiale de Ripley K :** L'indice K de Ripley (Ripley 1977) est une méthode d'analyse spatiale pour Décrire comment les modèles de points sont répartis sur une région d'intérêt donnée. Ce Le K de Ripley permet aux chercheurs de déterminer si un phénomène d'intérêt (par exemple, les arbres) semble être dispersés, regroupés ou distribués au hasard dans tout le zone d'étude. Cette méthode est similaire à la fonction I de Moran, mais elle peut Décrire des modèles de points sur plusieurs échelles définies par l'utilisateur.

Tableau 1 : Tableau de comparaison des paramètres et des Seuils de détection pour chaque algorithme.

Algorithme	Paramètres	Seuils
Indice de Moran	Attribut d'intérêt, voisins spatiaux	Valeurs z-score
Indice de Getis-Ord Analyse de quadrants	Attribut d'intérêt, voisins spatiaux	Valeurs z-score
Analyse de densité de kernel	Taille de la fenêtre, poids spatial, attribut d'intérêt	Valeurs de densité
L'analyse spatiale du cluster scan de Kulldorff	Taille de la fenêtre, poids spatial, attribut d'intérêt	Statistiques du ratio d'incidence
L'analyse spatiale de Ripley K	Taille de la fenêtre, attribut d'intérêt	Statistiques de la fonction de K
La méthode de l'indice de surprenance de Gini	Attribut d'intérêt, voisins spatiaux	Valeurs de l'indice de surprenance de Gini
L'analyse spatiale de Join Count	Attribut d'intérêt, voisins spatiaux	Valeurs de l'indice de Join Count

- **La méthode de l'indice de surprenance de Gini :** L'indice de Gini Surprise est une mesure de l'inégalité de la population qui peut décomposer et analyser la contribution des facteurs à l'inégalité.

- **L'analyse spatiale de Join Count :** Join Count est une méthode d'analyse spatiale utilisée pour évaluer l'association spatiale de données nominales et pour tester l'auto-corrélation spatiale. Il peut également être utilisé pour examiner la dépendance spatiale entre les unités territoriales.

Le tableau ci-dessus présente une comparaison des paramètres et des seuils utilisés pour chaque algorithme :

Chaque algorithme utilise des paramètres spécifiques pour l'analyse spatiale et des seuils pour déterminer les résultats significatifs. L'indice de Moran et l'indice de Getis-Ord utilisent des attributs d'intérêt et des voisins spatiaux, avec des valeurs z-score pour évaluer l'autocorrélation spatiale. L'analyse de densité de kernel et l'analyse du cluster scan de Kulldorff utilisent des tailles de fenêtre, des poids spatiaux et des attributs d'intérêt pour identifier les clusters significatifs. L'analyse spatiale de Ripley K utilise une taille de fenêtre et des attributs d'intérêt pour mesurer la fonction de K. La méthode de l'indice de surprenance de Gini utilise des attributs d'intérêt et des voisins spatiaux pour calculer l'indice de surprenance de Gini. L'analyse spatiale de Join Count utilise des attributs d'intérêt et des voisins spatiaux pour calculer l'indice de Join Count. Le choix de l'algorithme dépendra des objectifs spécifiques de l'étude et des données disponibles.

Dans le cadre de notre étude, nous avons opté pour l'algorithme Getis-Ord G afin de détecter les hotspots et les cold spots spatiaux liés au COVID-19. Nous avons fait ce choix en raison de la large utilisation et de l'adéquation de cet algorithme pour identifier les zones où les taux d'incidence ou de mortalité sont statistiquement significatifs et où la maladie est concentrée. En utilisant l'algorithme Getis-Ord G, nous pouvons obtenir une compréhension approfondie des zones où la prévalence du COVID-19 est plus élevée et où des mesures ciblées peuvent être prises pour la prévention et le contrôle de la maladie.
Les étapes de fonctionnement de cet algorithme sont les suivantes :

1. **Collecte des données :** Tout d'abord, les données géospatiales appropriées sont collectées, telles que les cas de Covid-19 par région ou la mortalité.
2. **Calcul des indicateurs locaux :** L'algorithme calcule l'indicateur local Getis-Ord G pour chaque unité spatiale de la région d'étude. Cet indicateur mesure la concentration spatiale d'un attribut particulier (par exemple, le nombre de cas de Covid-19) dans un voisinage donné.

3. **Définition du voisinage** : Un voisinage est défini pour chaque unité spatiale en spécifiant une distance de recherche ou un nombre de voisins à considérer. Cela permet de déterminer la proximité spatiale des unités voisines.
4. **Calcul du score Z :** En utilisant l'indicateur local Getis-Ord G, un score Z est calculé pour chaque unité spatiale. Le score Z est une mesure statistique standardisée qui permet de quantifier à quel point l'observation diffère de la moyenne dans des unités d'écart-type.
5. **Interprétation des scores Z :** Les scores Z obtenus sont interprétés pour identifier les hotspots (valeurs élevées et statistiquement significatives) et les cold spots (valeurs faibles et statistiquement significatives). Les hotspots indiquent des concentrations spatiales élevées de l'attribut étudié, tandis que les cold spots indiquent des concentrations spatiales faibles.
6. **Cartographie des résultats :** Les résultats sont généralement représentés sur une carte, où les hotspots sont affichés en utilisant une couleur distincte (par exemple, rouge), tandis que les cold spots sont représentés avec une autre couleur (par exemple, bleu).

En utilisant ces étapes, l'algorithme Getis-Ord G permet d'identifier et de cartographier les hotspots et cold spots spatiaux, ce qui facilite la compréhension des modèles de concentration spatiale des attributs étudiés. Ces informations sont précieuses pour la prise de décisions stratégiques et l'élaboration de politiques appropriées pour la lutte contre la pandémie.

3.6 Conclusion

Ce chapitre a introduit notre approche pour la détection des hotspots et cold spots dans le contexte de la pandémie de COVID-19. Nous avons décrit la méthode proposée, en mettant l'accent sur le prétraitement des données spatiales, le choix des indicateurs appropriés, ainsi que la sélection des algorithmes de détection des hotspots et cold spots. Dans le chapitre suivant, nous appliquerons cette méthodologie à notre étude de cas, ce qui nous permettra d'obtenir des informations pour la prise de décision et l'élaboration de stratégies de lutte contre la COVID-19.

Chapitre 4

Expérimentations

4.1 Introduction

Dans ce chapitre, nous allons aborder des modèles de données ainsi que les outils que nous avons utilisés dans notre projet, et ainsi que l'algorithme que nous avons choisi pour la modélisation de notre solution, puis nous allons créer une interface utilisateur conviviale pour permettre aux utilisateurs d'explorer les résultats de manière interactive et de visualiser les résultats obtenus sur une carte.

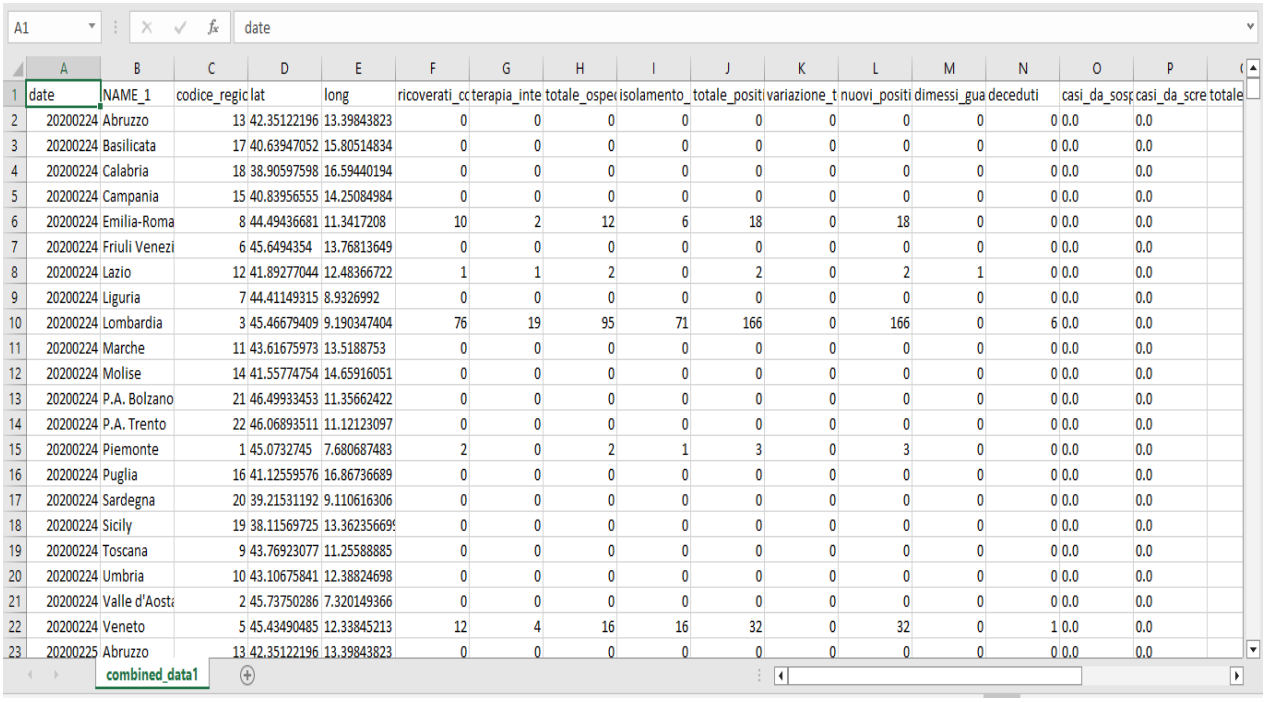
4.2 Description des données utilisées

Les données utilisées dans cette étude proviennent d'un fichier CSV fourni par l'API Italia coronavirus tracker, qui recense les données relatives au COVID-19 en Italie. Ces données comprennent des informations essentielles telles que la date, le nombre de cas confirmés, le nombre de décès et le nombre de guérisons, ventilés par région.

Avant d'entreprendre notre analyse spatiale, nous avons effectué un prétraitement minutieux des données afin de garantir leur qualité et leur cohérence. Les étapes de prétraitement ont inclus la vérification des valeurs manquantes, la gestion d'éventuelles duplications et la normalisation des variables si nécessaire.

La Figure 8 illustre de manière synthétique les données relatives au COVID-19 en Italie, en mettant en évidence la répartition spatiale des cas confirmés dans les différentes régions du pays. Des codes ou des noms uniques sont utilisés pour identifier chaque région, facilitant ainsi l'analyse spatiale.

En complément, les limites administratives des régions italiennes sont également fournies dans au format SHP (Shapefile). Cela permet de délimiter précisément chaque région et d'effectuer une analyse spatiale plus précise.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	date	NAME_1	codice_regic	lat	long	ricoverati_cc	terapia_inte	totale_ospe	isolamento_	totale_positi	variazione_t	nuovi_positi	dimessi_gua	deceduti	casi_da_sosp	casi_da_scre
2	20200224	Abruzzo	13	42.35122196	13.39843823	0	0	0	0	0	0	0	0	0	0.0	0.0
3	20200224	Basilicata	17	40.63947052	15.80514834	0	0	0	0	0	0	0	0	0	0.0	0.0
4	20200224	Calabria	18	38.90597598	16.59440194	0	0	0	0	0	0	0	0	0	0.0	0.0
5	20200224	Campania	15	40.83956555	14.25084984	0	0	0	0	0	0	0	0	0	0.0	0.0
6	20200224	Emilia-Roma	8	44.49436681	11.3417208	10	2	12	6	18	0	18	0	0	0.0	0.0
7	20200224	Friuli Venezi	6	45.6494354	13.76813649	0	0	0	0	0	0	0	0	0	0.0	0.0
8	20200224	Lazio	12	41.89277044	12.48366722	1	1	2	0	2	0	2	1	0	0.0	0.0
9	20200224	Liguria	7	44.41149315	8.9326992	0	0	0	0	0	0	0	0	0	0.0	0.0
10	20200224	Lombardia	3	45.46679409	9.190347404	76	19	95	71	166	0	166	0	6	0.0	0.0
11	20200224	Marche	11	43.61675973	13.5188753	0	0	0	0	0	0	0	0	0	0.0	0.0
12	20200224	Molise	14	41.55774754	14.65916051	0	0	0	0	0	0	0	0	0	0.0	0.0
13	20200224	P.A. Bolzano	21	46.49933453	11.35662422	0	0	0	0	0	0	0	0	0	0.0	0.0
14	20200224	P.A. Trento	22	46.06893511	11.12123097	0	0	0	0	0	0	0	0	0	0.0	0.0
15	20200224	Piemonte	1	45.0732745	7.680687483	2	0	2	1	3	0	3	0	0	0.0	0.0
16	20200224	Puglia	16	41.12559576	16.86736689	0	0	0	0	0	0	0	0	0	0.0	0.0
17	20200224	Sardegna	20	39.21531192	9.110616306	0	0	0	0	0	0	0	0	0	0.0	0.0
18	20200224	Sicily	19	38.11569725	13.36235669	0	0	0	0	0	0	0	0	0	0.0	0.0
19	20200224	Toscana	9	43.76923077	11.25588885	0	0	0	0	0	0	0	0	0	0.0	0.0
20	20200224	Umbria	10	43.10675841	12.38824698	0	0	0	0	0	0	0	0	0	0.0	0.0
21	20200224	Valle d'Aosta	2	45.73750286	7.320149366	0	0	0	0	0	0	0	0	0	0.0	0.0
22	20200224	Veneto	5	45.43490485	12.33845213	12	4	16	16	32	0	32	0	1	0.0	0.0
23	20200225	Abruzzo	13	42.35122196	13.39843823	0	0	0	0	0	0	0	0	0	0.0	0.0

Figure 8 : Représentation tabulaire des dataset de la COVID-19 en Italie

4.3 Environnement de développement

Dans ce projet, nous passons en revue les différents langages et outils utilisés lors du développement de notre application.

4.3.1 Visual studio code

Visual Studio Code est un éditeur de code source léger mais puissant. On possède souvent une solution à laquelle nous sommes habitués, qui nous convient la majorité du temps et on a peur de se retrouver perdu et de perdre par la même occasion en productivité. Pourtant, Visual Studio Code rassure la majorité des nouveaux utilisateurs dès les premières heures d'utilisation [32]. Nous utilisons Visual studio code pour faciliter notre travail.

4.3.2 PYTHON

Le langage de programmation Python a été créé en 1989 par Guido van Rossum, aux Pays-Bas. Le nom Python vient d'un hommage à la série télévisée Monty Python's Flying

Circus dont G. van Rossum est fan. La première version publique de ce langage a été publiée en 1991

Ce langage de programmation présente de nombreuses caractéristiques intéressantes, il est : multiplateforme, gratuit, un langage de haut niveau, orienté objet, simple et très utilisé en bio-informatique et plus généralement en analyse de données. Il demande relativement peu de connaissance sur le fonctionnement d'un ordinateur pour être utilisé, c'est un langage interprété. Nous avons utilisé les bibliothèques suivantes :

- **Pandas**

C'est une bibliothèque dédiée à la manipulation et l'analyse des données et sa structure de données clé est appelée DataFrame. Les DataFrames vous permettent de stocker et de manipuler des données tabulaires.

- **Geopandas**

Une bibliothèque pour la manipulation de données géographiques en Python qui propose des outils pour l'analyse spatiale, y compris la détection de hotspots et cold spots.

- **Folium**

Une bibliothèque pour la visualisation de données géographiques en Python qui permet de créer des cartes interactives et D'afficher les résultats de l'analyse spatiale.

- **PySide**

C'est une bibliothèque open-source pour la création d'interfaces graphiques utilisateur (GUI) en utilisant Python. Il est basé sur la bibliothèque Qt, une bibliothèque multiplateforme pour la création d'interfaces graphiques.

4.4 Implémentation de l'algorithme de détection de hotspots et cold spots

Dans notre projet, nous avons utilisé la statistique Getis-Ord Gi* pour analyser les hotspots et cold spots. Cette méthode nous permet de détecter les zones présentant une concentration spatiale élevée (hotspots) ou faible (cold spots) des données étudiées. Elle attribue à chaque zone un score z, qui quantifie à quel point cette zone diffère de la moyenne dans des unités d'écart-type.

Le score z, également appelé z-score, est une mesure statistique standardisée. Il permet de comparer une observation à la distribution des valeurs et d'évaluer si elle diffère

significativement de la moyenne. Ainsi, l'utilisation du score z nous permet de détecter différents types de hotspots et cold spots, tels que les nouveaux hotspots ou cold spots, les hotspots et cold spots consécutifs, intensifs, persistants, sporadiques, oscillants ou historiques.

Il est important de souligner que l'indice de Moran et le score z diffèrent dans leur objectif et leur utilisation. L'indice de Moran est spécifiquement utilisé pour évaluer l'autocorrélation spatiale, c'est-à-dire la similarité entre les valeurs dans les zones voisines. En revanche, le score z est une mesure statistique plus générale qui permet d'évaluer la différence entre une valeur observée et une distribution de référence.

```
# Calcul de la moyenne et de l'écart-type des nouveaux cas quotidiens
  mean = data["nouveaux_cas"].mean()
  std = data["nouveaux_cas"].std()
```

Faire une colonne dans data qui contient la valeur de score z pour chaque région

Code

```
# Calcul du score z pour chaque région
  data["score_z"] = (data["nouveaux_cas"] - mean) / std
```

Après détection des hotspots et coldspots selon la valeur de score_z si <-1.96 hotspots et coldspots >1.96

Et mets dans une DataFrame hotspots seuls w coldspots seuls

Code

```
# Identification des coldspots et des hotspots
    coldspots = data[data["score_z"] < -1.96]
    hotspots = data[data["score_z"] > 1.96]
```

Téléchargez la carte OpenStreetMaps (utilisation bibliothèque foluim) et on utilise la fonction GeoJson pour téléchargez les fichier shapfils sur la carte OpenStreetMaps

Code

```
# Création d'une carte OpenStreetMaps
    m = folium.Map(location=[45.5, 9],zoom_start=5)

    folium.GeoJson(italy, style_function=lambda x: {'weight': 2, 'color': 'black',
'fillOpacity': 0}).add_to(m)
```

Apres la création d'une carte thématique pour les hotspots et coldspots on utilisent la fonction choroplèthe et aussi l ajoute "info-bulle pour afficher la propriété NAME_1 "

Code

```
# Création d'une carte thématique pour les coldspots
    folium.Choropleth(
        geo_data=italy,
        name='Coldspots',
```

Figure 9 : Code en Python pour le calcul de la statistique Gestis Ord Gi* et le calcul des zones e hotspots et cold spots.

Le score z peut être calculé à l'aide de l'algorithme suivant :

1. Calculer la moyenne (m) et l'écart-type (s) de l'ensemble de données.
2. Soustraire la moyenne (m) de la valeur individuelle (x) que vous voulez convertir en score z.
3. Diviser la différence obtenue dans l'étape 2 par l'écart-type (s) de l'ensemble de données
4. Le résultat obtenu est le score z de la valeur individuelle.

La formule mathématique pour calculer le score z est la suivante : $z = (x - m) / s$ où :

Z est le score z de la valeur individuelle.

X est la valeur individuelle qu'on veut convertir en score z.

M est la moyenne de l'ensemble de données.

S est l'écart-type de l'ensemble de données.

La figure 9 montre le code en Python basé sur le calcul de la statistique Getis Ord Gi*.

4.5 Réalisation du logiciel de détection de hotspots et cold spots

Notre projet a abouti au développement d'un logiciel spécifiquement conçu pour l'analyse des hotspots et cold spots spatiaux. Ce programme offre plusieurs interfaces conviviales qui permettent aux utilisateurs de sélectionner l'analyse souhaitée, que ce soit pour les cas d'incidence ou de mortalité liés à une problématique donnée.

L'interface principale du logiciel est illustrée dans la figure ci-dessous. Elle présente un espace de visualisation spatiale de la région d'étude, offrant une représentation cartographique claire et précise. Cette interface propose également six boutons fonctionnels essentiels pour l'utilisateur :

- Données : Ce bouton permet d'accéder à un tableau récapitulatif affichant les chiffres totaux des cas étudiés, fournissant ainsi une vue d'ensemble statistique.

- Diagramme : En cliquant sur ce bouton, l'utilisateur peut générer un diagramme à barres représentant graphiquement les données des cas étudiés, facilitant ainsi leur interprétation et leur comparaison.

- Carte : Ce bouton offre la possibilité de visualiser les données des cas étudiés sous forme de carte, permettant une analyse spatiale plus approfondie et une meilleure compréhension de leur répartition géographique.

- Limites administratives : permet à l'utilisateur de visualiser les limites administratives de la région d'étude. En cliquant sur ce bouton, une fenêtre s'ouvre affichant les limites géographiques de la région sous forme de polygones.

- Cold spots / Hotspots : Ce bouton permet de générer des résultats spécifiques à la détection des cold spots et hotspots, c'est-à-dire aux zones présentant une faible concentration des cas étudiés.

De plus, un bouton "Quitter" est également disponible pour permettre à l'utilisateur de quitter le logiciel lorsque nécessaire.

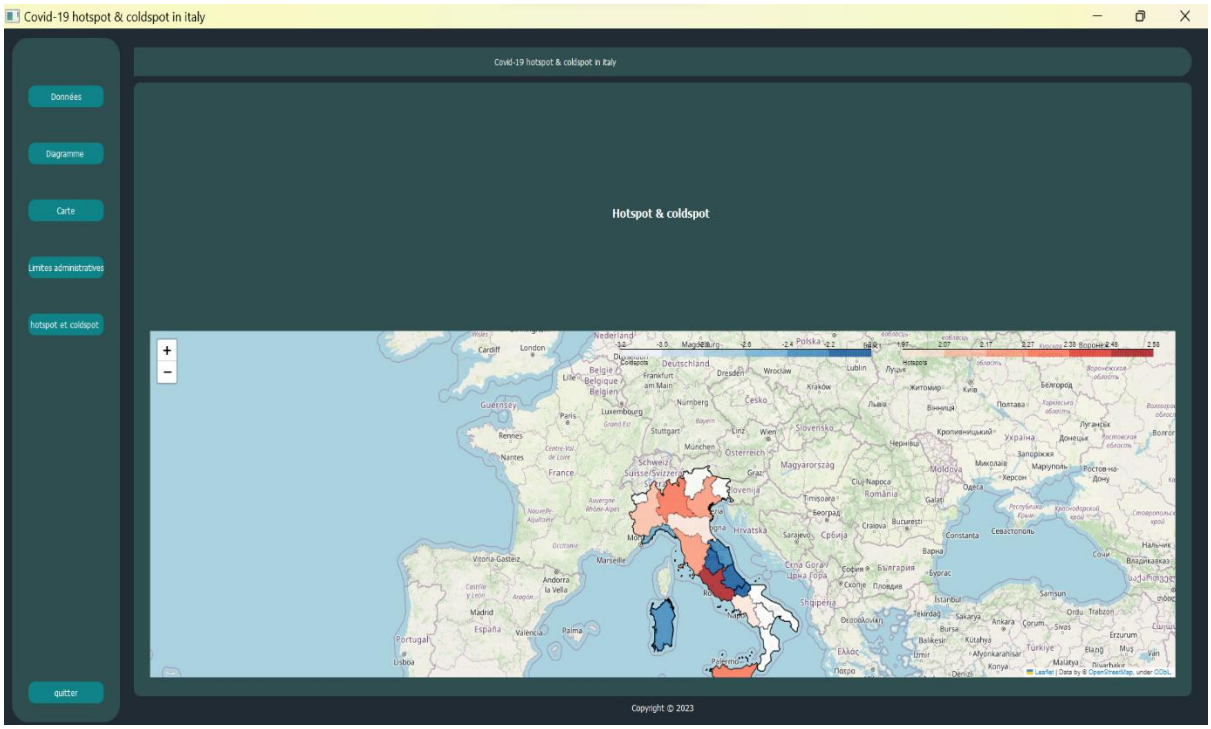
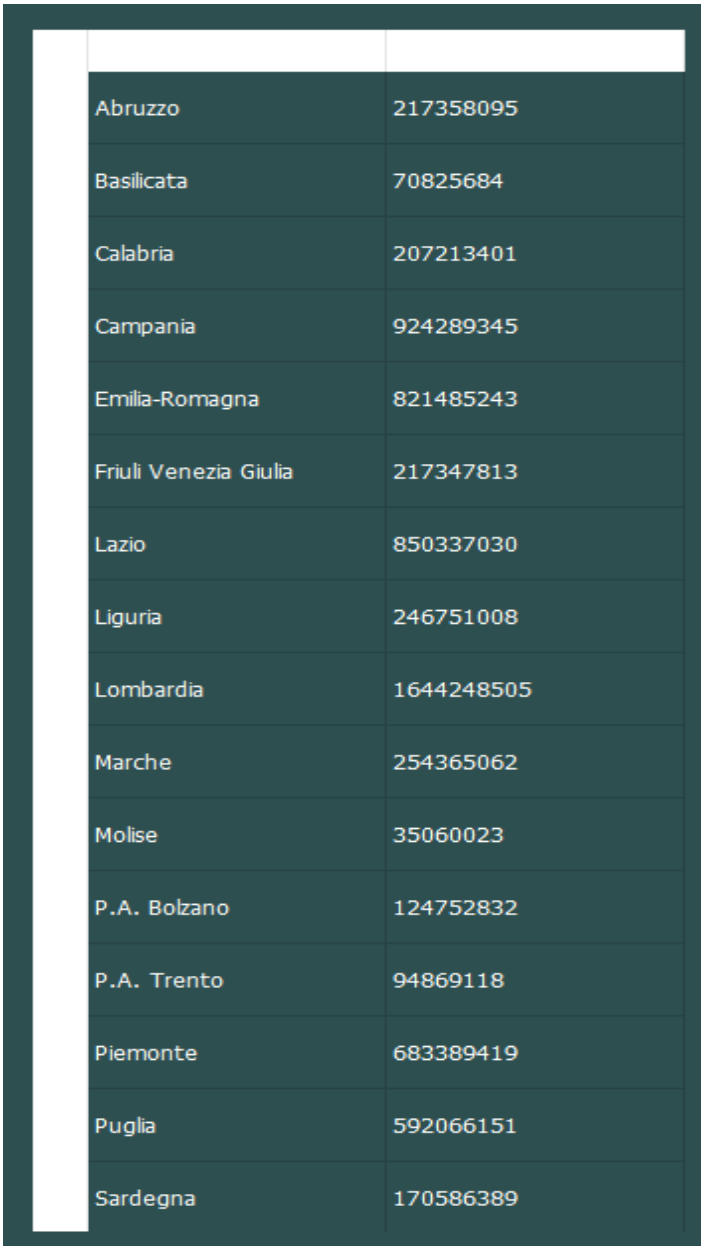


Figure 10 : Aperçu de l’interface principale

Lorsque l'utilisateur clique sur le bouton "Données", une deuxième interface s'affiche, présentant un tableau contenant les noms des régions administratives ainsi que les valeurs d'incidence ou de mortalité des cas de COVID-19 pour chaque région.

Cette interface permet à l'utilisateur d'accéder rapidement et facilement aux informations spécifiques sur les régions étudiées. En affichant les données dans un tableau structuré, il devient aisé de comparer les valeurs d'incidence ou de mortalité entre les différentes régions et d'identifier les variations.



Abruzzo	217358095
Basilicata	70825684
Calabria	207213401
Campania	924289345
Emilia-Romagna	821485243
Friuli Venezia Giulia	217347813
Lazio	850337030
Liguria	246751008
Lombardia	1644248505
Marche	254365062
Molise	35060023
P.A. Bolzano	124752832
P.A. Trento	94869118
Piemonte	683389419
Puglia	592066151
Sardegna	170586389

Figure 11 : Interface de visualisation des données d'incidence ou de mortalité du COVID-19 par région

Lorsque l'utilisateur clique sur le bouton "Carte", une interface dédiée s'ouvre, affichant une carte interactive de la région d'étude.

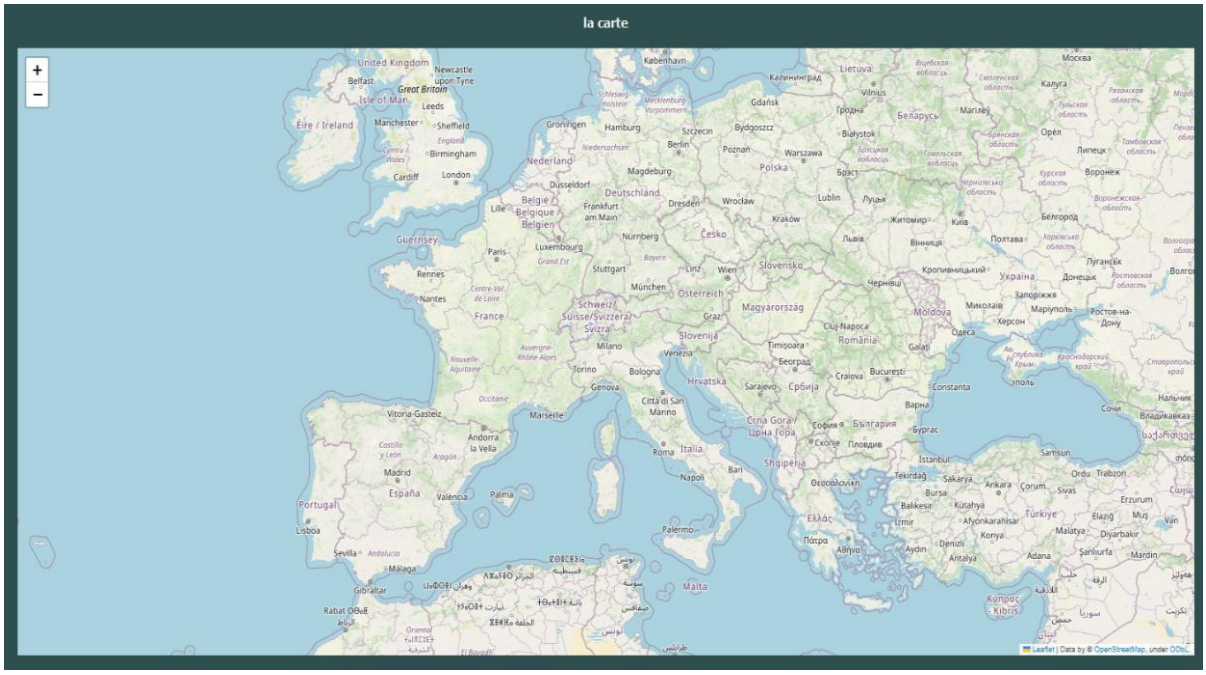


Figure 12 : : Interface de visualisation de la région d'étude

En cliquant sur le bouton "Limites Administratives", une interface dédiée s'affiche, présentant les limites administratives de la région d'étude.

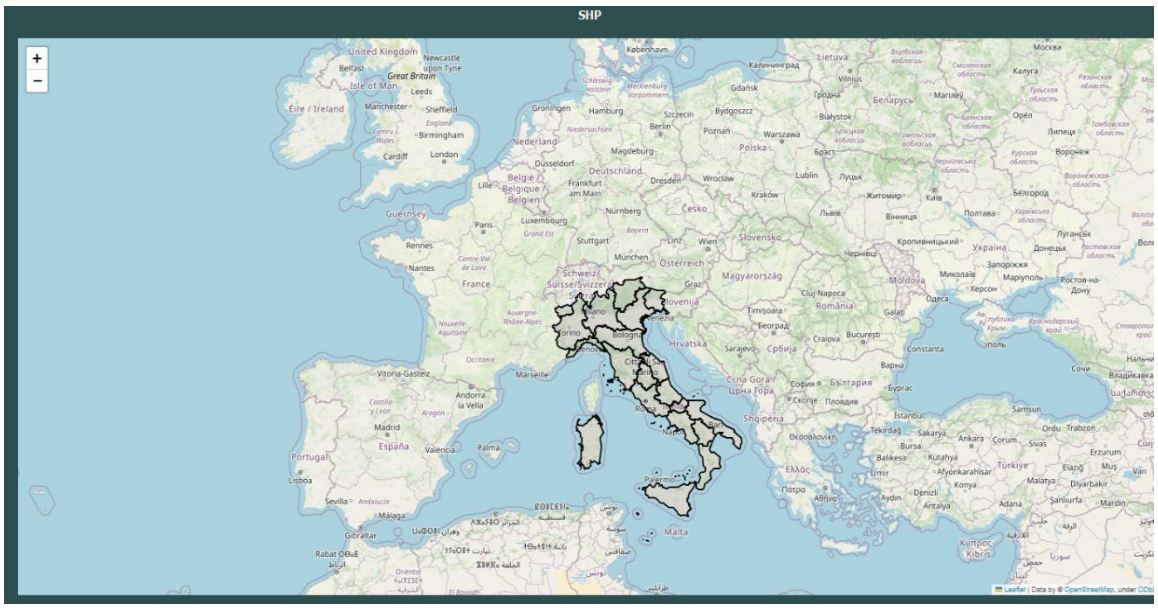


Figure 13 : Interface de visualisation des limites administratives de la région d'étude

En utilisant le bouton "Diagramme", une interface s'affiche, présentant l'évolution des cas quotidiens de COVID-19 sur une période donnée. Dans notre cas d'étude, l'analyse se concentre sur la période allant de janvier 2020 à janvier 2023.

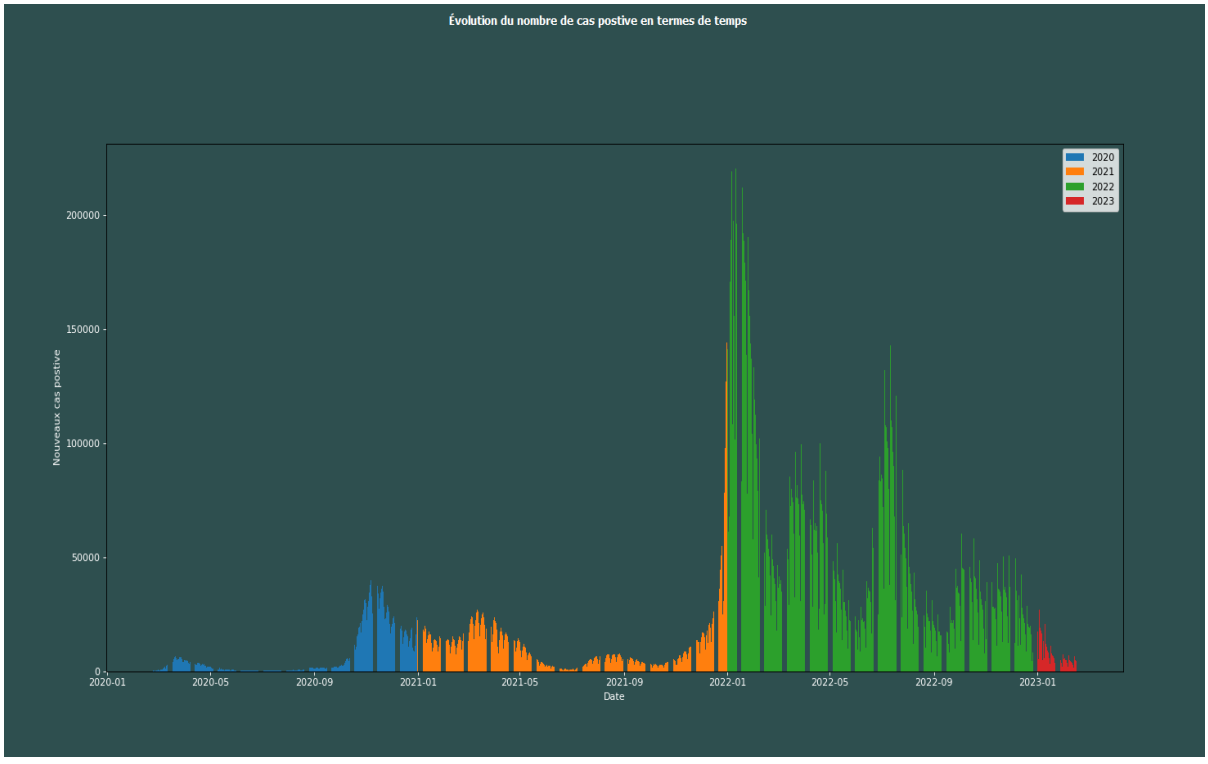


Figure 14 : L'interface « Diagramme » montrant l'évolution des cas d'incidence du Covid-19 de janvier 2020 à janvier 2023

4.6 Résultats d'analyse de la détection des cold spots et hotspots

La figure ci-dessous présente les résultats de la détection des hotspots et des cold spots spatiaux basée sur les données d'incidence du COVID-19 en Italie. La légende e la carte indique que les hotspots sont représentés en rouge, tandis que les cold spots sont représentés en bleu. Cette codification couleur permet d'identifier visuellement les zones avec des taux d'incidence élevés (hotspots) et les zones avec des taux d'incidence plus faibles (cold spots).

Ces informations spatiales peuvent aider à identifier les zones nécessitant une attention particulière en termes de ressources médicales, de mesures de prévention et de contrôle de la maladie.

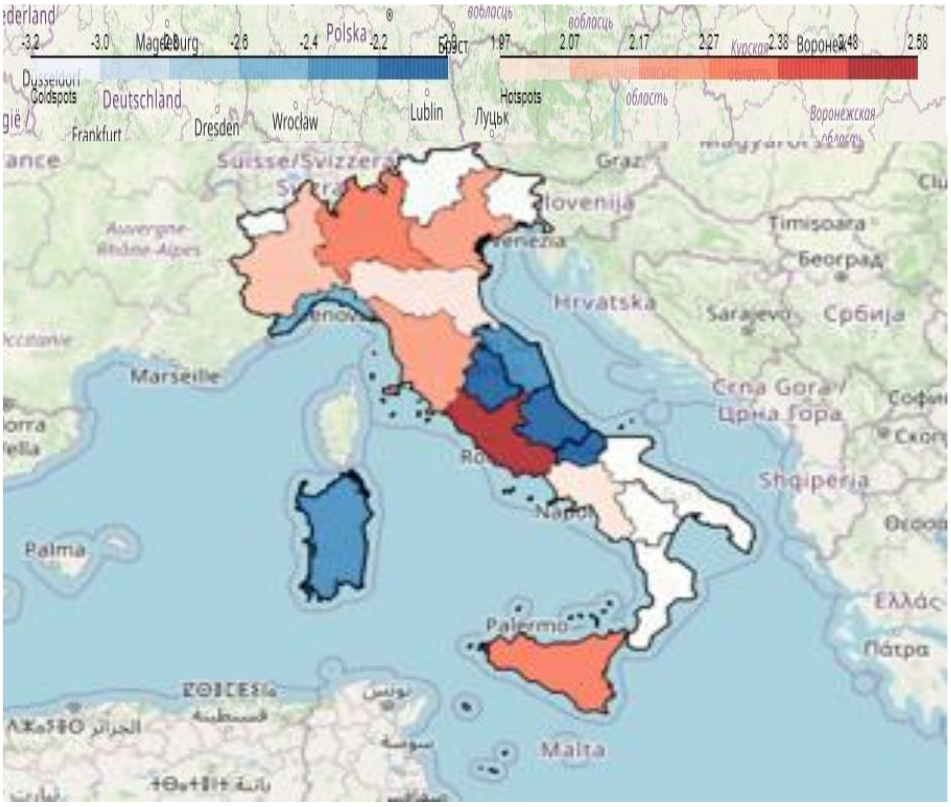


Figure 15 : Visualisation de la détection des Cold spots et hotspots

On constate en visualisant les points chauds et froids, qu'une grappe significative de taux d'incidence supérieurs dans le nord (couleur rouge) En d'autres termes, Les régions du nord ont été plus durement touchées que les régions du sud du pays (couleur

bleu), Et il y a certaines régions du sud-ouest qui n'ont pas été touchées par la maladie (couleur blanc).

4.7 Interprétation et discussion des résultats

L'analyse des résultats met en évidence plusieurs tendances intéressantes. Tout d'abord, les régions du nord de l'Italie, telles que la Lombardie, le Piémont, la Vénétie et la Toscane, apparaissent comme des hotspots majeurs avec des taux d'incidence élevés. Cela peut être attribué à divers facteurs, tels que la densité de population, les liens économiques et les voyages internationaux dans ces régions. La concentration de cas dans ces zones peut être préoccupante car elle peut entraîner une pression accrue sur les systèmes de santé locaux et nécessiter des mesures de contrôle et de prévention plus strictes.

D'un autre côté, les régions du Molise, des Abruzzes et des Marches dans l'Est du pays, se démarquent comme des cold spots avec des taux d'incidence relativement faibles. Cela peut indiquer que ces régions ont réussi à contenir la propagation du virus grâce à des mesures efficaces de prévention et de contrôle. Il serait intéressant d'étudier les stratégies spécifiques mises en place dans ces régions pour comprendre les bonnes pratiques et les leçons apprises qui pourraient être appliquées ailleurs.

La Sicile, en tant qu'exception avec une incidence élevée malgré sa localisation géographique dans le sud du pays, pourrait nécessiter une attention particulière. Des facteurs tels que les mouvements de population, les infrastructures de santé et les comportements individuels pourraient contribuer à cette situation. Des mesures ciblées et une meilleure compréhension des dynamiques locales seraient nécessaires pour réduire l'incidence dans cette région.

En général, l'identification de hotspots et de cold spots permet de mieux comprendre la répartition spatiale des cas de COVID-19 en Italie. Cela peut aider à cibler les ressources et les interventions dans les régions les plus touchées, tout en identifiant les zones où les mesures de contrôle peuvent être assouplies. Il est crucial de poursuivre la surveillance et l'analyse spatiale pour une gestion efficace de la pandémie et une prise de décision éclairée.

D’un autre côté, l’analyse du diagramme d’incidences montre comment le nombre de cas quotidiens de COVID-19 a évolué au fil du temps. En se basant sur l'identification du premier cas, il a été supposé que le virus fût déjà présent en Italie à partir de janvier 2020, mais à un taux relativement faible. Au cours de la période entre janvier 2020 et 2021, on observe une augmentation progressive du nombre de cas de COVID-19. Cependant, à partir de janvier 2022 jusqu'en 2023, on constate une augmentation significative, correspondant à ce qui est communément appelé la deuxième vague de la pandémie. Cette période se caractérise par une augmentation très importante des cas de COVID-19.

Cependant, vers janvier 2023, nous observons une diminution des cas de COVID-19. Cela peut être le résultat de divers facteurs tels que la mise en œuvre de mesures de santé publique plus strictes, la vaccination généralisée, l'immunité collective croissante ou d'autres interventions efficaces. Cette diminution des cas indique une amélioration de la situation de la pandémie à ce stade.

L'analyse de l'évolution temporelle des cas de COVID-19 et la détection des cold spots et hotspots permet de mieux comprendre les différentes phases de la pandémie en Italie et d’identifier les régions les plus touchées par la pandémie. Cela fournit des indications sur les périodes de propagation accrue de la maladie ainsi que les périodes de maîtrise ou d'amélioration. Ces informations sont essentielles pour orienter les décisions de santé publique, mettre en place des stratégies de prévention et d'intervention, et évaluer l'efficacité des mesures prises pour lutter contre la pandémie.

4.8 Conclusion

Dans ce chapitre, nous avons procédé à une expérimentation approfondie, en commençant par une présentation détaillée des modèles de données utilisés ainsi que des outils employés dans notre projet. Nous avons également exposé l'algorithme que nous avons sélectionné pour mettre en œuvre notre solution. De plus, nous avons décrit en détail l'interface utilisateur que nous avons développé, et nous avons présenté une étude de cas spécifique dédiée à la détection des hotspots et cold spots, dans le but d'analyser les taux d'incidence de la Covid-19 en Italie.

Conclusion Générale

Les méthodes conventionnelles de Data Mining ne conviennent pas aux données spatiales en raison de leurs caractéristiques spécifiques. Les algorithmes traditionnels supposent une indépendance et une distribution uniforme des données, ce qui ne correspond pas à la réalité des données spatiales qui sont multidimensionnelles, auto-corrélées spatialement et hétérogènes. C'est pourquoi le Data Mining Spatial joue un rôle essentiel pour tenir compte de ces particularités.

Dans ce mémoire, nous avons commencé par fournir une description détaillée de l'information spatiale, en mettant en évidence ses propriétés et ses formats de représentation. Ensuite, nous avons souligné l'importance des relations spatiales dans le processus de Data Mining Spatial (DMS) et fourni une définition complète du DMS, en détaillant ses différentes phases d'exécution, ses approches et ses méthodes.

Pour notre projet, nous avons utilisé les données de la Covid-19, en nous concentrant spécifiquement sur l'incidence de la maladie dans les différentes régions d'Italie. En utilisant des techniques de Data Mining Spatial telles que l'analyse des hotspots et des cold spots, notre objectif était d'identifier les régions présentant une incidence significativement plus élevée ou plus faible que prévu de la Covid-19. Pour cela, nous avons proposé un algorithme basé sur la statistique Getis-Ord G. Cette statistique est largement utilisée dans l'analyse spatiale pour détecter les hotspots et les cold spots.

Enfin, nous avons développé une interface utilisateur conviviale permettant aux utilisateurs d'explorer de manière interactive les résultats de notre analyse et de visualiser les résultats sur une carte. Cette application sera précieuse pour les chercheurs, les professionnels de la santé publique et les décideurs, leur permettant de mieux comprendre les tendances de la Covid-19 et les facteurs qui y sont associés.

En termes de perspectives, nous proposons d'explorer davantage d'autres techniques et algorithmes du Data Mining Spatial afin de sélectionner ceux qui conviennent le mieux à l'étude épidémiologique. En continuant à améliorer notre compréhension de l'analyse spatiale des données de la Covid-19, nous pourrons fournir des informations précieuses pour lutter contre la pandémie.

Annexe A

Description détaillée du contenu du fichier CSV

Tableau 2 **:** Explication des colonnes du fichier CSV

Nom de colonne	**Le sens**
Date	La date de l'enregistrement des données
NAME_1	Le nom de la région
codice_regione	Le code de la région
Lat	La latitude de la région
Long	La longitude de la région
ricoverati_con_sintomi	Le nombre de personnes hospitalisées avec des symptômes
terapia_intensiva	Le nombre de personnes en soins intensifs
totale_ospedalizzati	Le nombre total de personnes hospitalisées
isolamento_domiciliare	Le nombre de personnes en isolement à domicile
totale_positivi	Le nombre total de personnes positives à la COVID-19 (hospitalisées et en isolement à domicile)
variazione_totale_positivi	La variation du nombre total de personnes positives par rapport au jour précédent
nuovi_positivi	Le nombre de nouveaux cas positifs enregistrés
dimessi_guariti	Le nombre de personnes guéries et sorties de l'hôpital
Deceduti	Le nombre de décès enregistrés
casi_da_sospetto_diagnostico	Le nombre de cas positifs détectés par test diagnostique

Annexe 2

Code en python permettant la détection des hotspots et cold spots avec la Statistique Getis-Ord Gi*

```
# Calcul de la moyenne et de l'écart-type des nouveaux cas quotidiens
   mean = data["nouveaux_cas"].mean()
   std = data["nouveaux_cas"].std()
```

Faire une colonne dans data qui contient la valeur de score z pour chaque région

Code

```
 # Calcul du score z pour chaque région
   data["score_z"] = (data["nouveaux_cas"] - mean) / std
```

Après détection des hotspots et cold spots selon la valeur de score_z si <-1.96 hotspots et cold spots >1.96

Et mets dans une DataFrame hotspots seuls w cold spots seuls

Code

```
# Identification des cold spots  et des hotspots
    cold spots  = data[data["score_z"] < -1.96]
    hotspots = data[data["score_z"] > 1.96]
```

Téléchargez la carte OpenStreetMaps (utilisation bibliothèque foluim) et on utilise la fonction GeoJson pour téléchargez les fichier shapfils sur la carte OpenStreetMaps

Code

```
# Création d'une carte OpenStreetMaps
    m = folium.Map(location=[45.5, 9],zoom_start=5)

     folium.GeoJson(italy, style_function=lambda x: {'weight': 2, 'color':
'black', 'fillOpacity': 0}).add_to(m)
```

Apres la création d'une carte thématique pour les hotspots et cold spots on utilisent la fonction choroplèthe et aussi l ajoute "info-bulle pour afficher la propriété NAME_1 "

Code

```
# Création d'une carte thématique pour les cold spots
    folium.Choropleth(
        geo_data=italy,
        name='Cold spots',
        data=cold spots ,
        columns=['NAME_1', 'score_z'],
        key_on='feature.properties.NAME_1',
        fill_color='Blues',
        fill_opacity=0.8,
        line_opacity=0.2,
        legend_name='Cold spots',
        nan_fill_color='white',
        nan_fill_opacity=0.8
    ).add_to(m)

    # Création d'une carte thématique pour les hotspots
    hotspots_layer = folium.Choropleth(
        geo_data=italy,
        name='Hotspots',
        data=hotspots,
        columns=['NAME_1', 'score_z'],
        key_on='feature.properties.NAME_1',
        fill_color='Reds',
        fill_opacity=0.8,
        line_opacity=0.2,
        legend_name='Hotspots',
        nan_fill_color='transparent'  # Set the non-hotspot areas to trans-
parent
    )
    hotspots_layer.geojson.add_child(
        folium.features.GeoJsonTooltip(['NAME_1'])  # Add a tooltip to dis-
play the NAME_1 property
    )
    hotspots_layer.add_to(m)
```

Bibliographie

[1] Olivier Balay. La représentation de l'environnement sonore urbain à l'aide d'un Système d'Information Géographique. Lyon : s.n., Octobre 1999.

[2] Jean Denègre, François Salgé. Les systèmes d'information géographique. Presses Universitaires de France. 2004. p. 128.

[3] Julie Le Gallo. Hétérogénéité spatiale Principes et méthodes. Dans Économie & prévision 2004/1. p. 151 à 172.

[4] Julie Le Gallo. Économétrie spatiale : l'autocorrélation spatiale dans les modèles de régression li-néaire. Dans Économie & prévision 2002/4. p. 139 à 157.

[5] RAPPORT DU SIG ET DE LA TELEDETECTION DANS LA MODELISATION SPATIALE DU RISQUE NATUREL. Espace-UMBM. [En ligne] 2021. http://dspace.univ-msila.dz:8080/xmlui/handle/123456789/25342

[6] L'analyse spatiale. Édition science et bien commun. [En ligne] https://scienceetbiencommun.pressbooks.pub/evalsantemondiale/chapter/spatiale

[7] Mémoire. [En ligne] http://eprints.univ-batna2.dz/193/1/Djazia%20CHAMI.pdf

[8] Le Data Mining Spatial et les bases de données spatiales. ResearchGate. [En ligne] https://www.researchgate.net/publication/228445173_Le_Data_Mining_Spatial_et_les_bases_de_donnees_spatiales

[9] Espace numerique Universite Abd El Hamid Ibn Badis Mostaganem. [En ligne] http://e-biblio.univ-mosta.dz/bitstream/handle/123456789/9622/MINF199.pdf
[10]OneDrive.[Enligne]

https://onedrive.live.com/?authkey=%21AKhxjLzJPd%2DEZ9I&id=2E1019D5E7DC378%2%201142&cid=02E1019D5E7DC378&parId=root&parQt=sharedby&o=OneUp

[11] Research on Spatial Data Mining in E-Government Information System. CHAPTER METRICS OVERVIEW. [En ligne] 2012. https://www.intechopen.com/chapters/38581

[12] OneDrive.[Enligne] https://onedrive.live.com/?authkey=%21AKhxjLzJPd%2DEZ9I&id=2E1019D5E7DC378%2%201142&cid=02E1019D5E7DC378&parId=root&parQt=sharedby&o=OneUp

[13] Process of spatial data mining. youtube. [En ligne] https://www.youtube.com/watch?v=97DMx7TFV5U

[14] Détection de hotspots. World Health Organization. [En ligne] https://www.who.int/emergencies/disease-outbreak%20-news/item/2020-DON264

[15] Régis Darques, Julie Trottier. Ville et épidémiologie spatiale. 2021. p. 15 à 39

[16] An Efficient CNN Model for COVID-19 Disease Detection Based on X-Ray Image Classification. Hindawi. [En ligne] 2021. https://doi.org/10.1155/2021/6621607

[17] An Efficient CNN Model for COVID-19 Disease Detection Based on X-Ray Image Classification. Hindawi. [En ligne] 2021. https://doi.org/10.1155/2021/6621607

[18] Spatial prediction of COVID-19 epidemic using ARIMA techniques in India. National Library of Medicine. [En ligne] 16 Jul 2020. https://pubmed.ncbi.nlm.nih.gov/32838022/

[19] Spatial Prediction of COVID-19 in China Based on Machine Learning Algorithms and Geographical-ly Weighted Regression. Hindawi. [En ligne] 2021.

https://www.hindawi.com/journals/cmmm/2021/7196492/ .

[20] Emerging Hot Spot Analysis. sfdep.[En ligne] https://sfdep.josiahparry.com/articles/understanding-emerging-hotspots.html

[21] Mapping and Spatial Pattern Analysis of COVID-19 in Central Iran Using the Local Indicators of Spatial Association (LISA). BMC Public Health. [En ligne] 08 December 2021. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-021-12267-6

[22] Analyse des valeurs aberrantes locales (Exploration des modèles spatio-temporels). esri. [En ligne] https://pro.arcgis.com/fr/pro-app/latest/tool-reference/space-time-pattern- mining/localoutlieranalysis.html

[23] Bakari RAMADANE. Evaluation de la dynamique spatio-temporelle de l'évolution de la covid a Li-breville par une approche machine learning. memoireonline. [En ligne] https://www.memoireonline.com/11/22/13465/Evaluation-de-la-dynamique-spatio- temporelle-de-levolution-de-la-covid-a-Libreville-par-une-appr.html

[24] Spatial transmission of COVID-19 via public and private transportation in China. National Library of Medi-cine. [En ligne] 14 Mar 2020. https://pubmed.ncbi.nlm.nih.gov/32184132/

[25] Covid-19 analyse spatiale de l'influence des facteurs socio-économiques sur la prévalence et les consé-quences de l'épidémie dans les départements français. ResearchGate. [En ligne] April2020. https://www.researchgate.net/publication/340808369_Covid-19_analyse_spatiale_de_l%27influence_des_facteurs_socio-economiques_sur_la_prevalence_et_les_consequences_de_l%27epidemie_dans_les_departements_francais

[26] Analyzing the spatial determinants of local Covid-19 transmission in the United States.National Library of Medicine.[En ligne] 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7498441/

[27] CORONAVIRUS RESOURCE CENTER. JOHNS HOPKINS UNIVERSITY of MEDICINE. [En ligne] https://coronavirus.jhu.edu/

[28] Visual studio. Documentation. [En ligne] https://visualstudio.microsoft.com/

[29] "CORONAVIRUS RESOURCE CENTER," [Online]. Available: https://coronavirus.jhu.edu/ .

[30] "Comparative Study Based on Analysis of Coronavirus Disease (COVID-19) Detection and Prediction Using Machine Learning Models," 20 Nov 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8605773/ .

[31] "Scribbr," [Online]. Available: https://www.scribbr.fr/methodologie/collecte-de-donnees .

[32] "Visual studio," [Online]. Available: https://visualstudio.microsoft.com/ .