République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE ABDELHAMID IBN BADIS MOSTAGANEM

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

MINI PROJET

Présenté par

BOUZID OUAZAA
BOUMEHDI ABDERRAHMANE

Pour obtenir

LE DIPLOME DE MASTER
SPECIALITE : INFORMATIQUE
OPTION : *Système d'Information Géographique*
Intitulé :

Outil d'analyse en ligne pour l'aide à la décision Géospatiale (SOLAP: Spatial On Line Analysis Processing) Application dans la surveillance épidémiologique

Membres de jury :

M^elle^ ZEMRI FARAH AMINA Maître Assistante, Université d'Oran, Algérie
(Co Encadreur)

M^r^ HANNI FOUAD Maitre-Assistant, Université de Mostaganem, Algérie
(Encadreur)

2013/2014

Sommaire

Liste des figures et tableaux

Sommaire

Liste des figures

Chapitre1 : Entrepôt de données, OLAP et SOLAP

Chapitre 2 : Fouille de données spatiales

Liste des tableaux

Chapitre3 : Model Décisionnel proposé

Chapitre4 : Implémentation

Chapitre1 : Entrepôts de données, OLAP & SOLAP

1.1 Introduction

Le challenge des organisations actuelles est de transformer leur système d'information qui avait une vocation de production à un système d'information décisionnel dont la vocation de pilotage devient majeure. Les entreprises ont de nouvelles tendances pour pouvoir accéder à toutes les données de l'entreprise, regrouper les informations dispersées et analyser et prendre des décisions rapidement.

1.2Entrepôt de données

Le concept d'entrepôt de données a été formalisé pour la première fois en 1990 par Bill Inmon[Inm, 96]. Il s'agissait de constituer une base de données *orientée sujet, intégrée et contenant des informations historisées, non volatiles et exclusivement destinées aux processus d'aide à la décision* [Jea, 05].

1.2.1 Généralités sur les entrepôts de données

La définition de l'entrepôt de donnée fait appel à un certain nombre de définitions relatives au domaine du business intelligence qu'on trouve ci-dessous.

- **Entrepôt de données**

Un entrepôt de données (datawarehouse) est définit comme une collection de données organisées par sujet, temporelles et persistantes. Cette collection est destinée à être utilisée dans le processus d'aide à la décision. Les utilisateurs interrogent les données à des fins d'analyse en se basant sur des données historisées, agrégées ou résumées. Ces données peuvent provenir de différentes sources et sont regroupées dans une base unique conçue pour des analystes et des décideurs(Kim, 96 ; Jea, 05).

- **OLAP (On Line Analytical Processing)**

Désigne une catégorie d'applications et de technologies permettant de collecter, stocker, traiter et restituer des données multidimensionnelles, à des fins d'analyse.

- **Datamining**

Désigne une catégorie d'outils d'exploitation d'un entrepôt de données permettant d'effectuer des fouilles " mining " ou d'extraire des connaissances permettant de faire apparaître des corrélations jusqu'alors cachées entre les données[Jea, 05].

- **Entrepôt de données spatial**

Un entrepôt de données spatiales est une collection de données spatiales de qualité, orientée par sujet, non-volatile, variable dans le temps, qui inclut un ensemble d'outils de base permettant d'accéder et d'extraire l'information» [Jea, 05].

- **Datamart**

Les données de l'entrepôt de données étant thématiques, il est donc possible d'extraire des données spécifiques aux besoins d'un secteur ou d'une fonction particulière. Les magasins de données (ou Datamarts) sont des sous-ensemblesd'un entrepôt de données contenant des points de vue spécifiques selon des critères métiers. La *Figure 3* donne un exemple de magasins de données (service Marketing, service Ressources Humaines)[Net I.1]

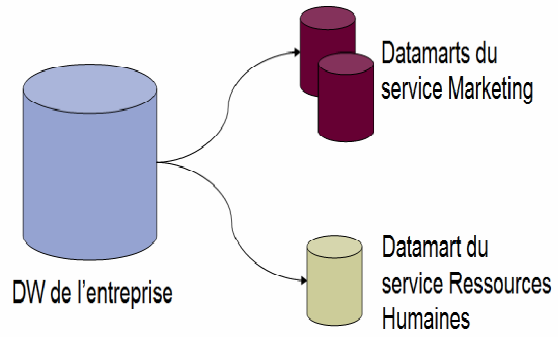


Figure 1.1 Exemples de Magasins de données

1.2.2Architecture générale d'un entrepôt de données

Un entrepôt de données reçoit un flux de données entrant et fournit un flux de données sortant. Le flux entrant passe par un processus d'extraction des données à partir de sources multiples et hétérogènes. Ces données sont ensuite transformées, filtrées, homogénéisées et nettoyées avant d'être stockées dans l'entrepôt de données[Ziy, 10]

La zone de préparation (staging area) est une zone temporaire de stockage qui est souvent détruite après la génération de l'entrepôt de données. La zone de stockage, représentant l'entrepôt de données, est évidemment une zone de stockage permanente des données. La zone de présentation donne accès aux données contenues dans l'entrepôt de données. Elle peut contenir des outils d'analyse programmés (rapports, requêtes, …). Le flux sortant met ces données à la dispositiondes utilisateurs. La *Figure 2* illustre l'architecture générale d'un entrepôt de données [Net I.3].

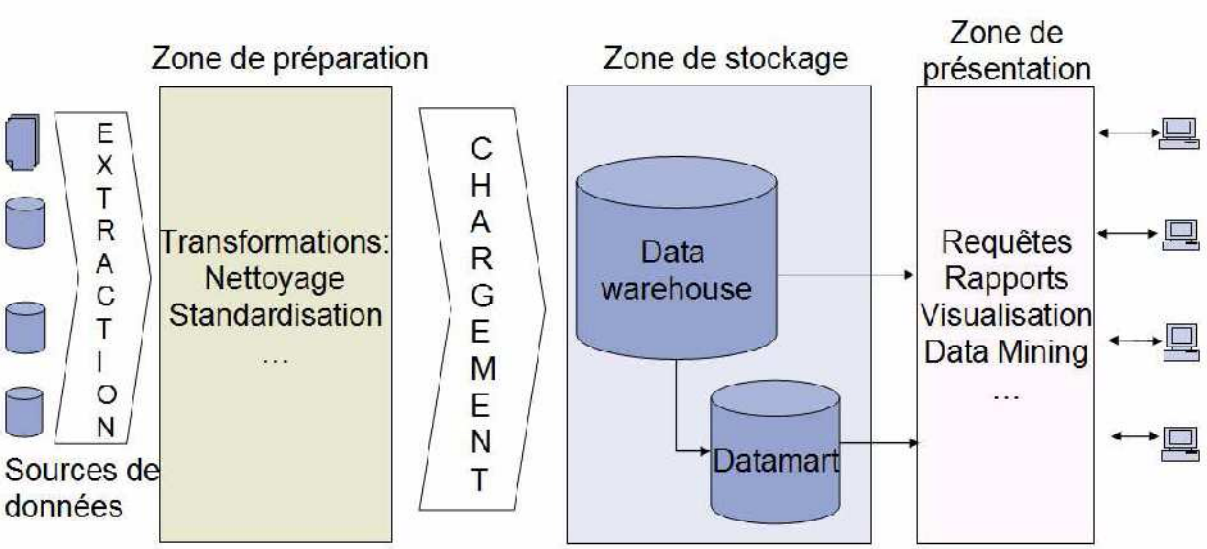


Figure 1.2Architecture générale d'un entrepôt de données

1.2.3 Modélisation des entrepôts de données

Les entrepôts de données ont introduit une nouvelle méthode de conception autour des concepts métiers. On ne parle plus de normalisation au sens relationnel du terme. De nouveaux types de tables ont été également introduits, la table de faits et la table de dimensions. Le modèle en étoile et le modèle en flocon sont de nouveaux modèles de représentation des entrepôts de données[Ben, 12].

1.2.3.1 Les types de modèles

Le modèle en étoile et le modèle en flocon sont les deux modèles les plus utilisés pour modéliser un entrepôt de données.

a. Modèle en étoile

Ce modèle est composé d'une table de faits centrale et de tables de dimensions.

Les dimensions n'ont pas de liens entre elles [Net I.1].

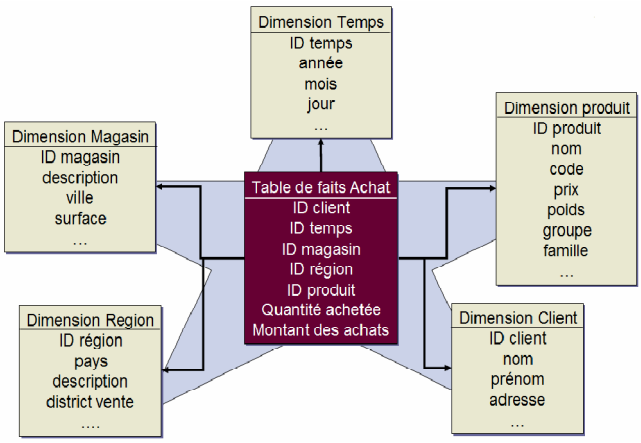


Figure 1.3. Modélisation en étoile

b. Modèle en flocon

Le modèle en flocon est composé quant à lui d'une table de faits et un ensemble de tables de dimension décomposées en sous-hiérarchies. Chaque table de dimension a un seul niveau hiérarchique. La table de niveau hiérarchique le plus bas, ayant donc la granularité la plus fine, est reliée à la table de faits[Net I.1].

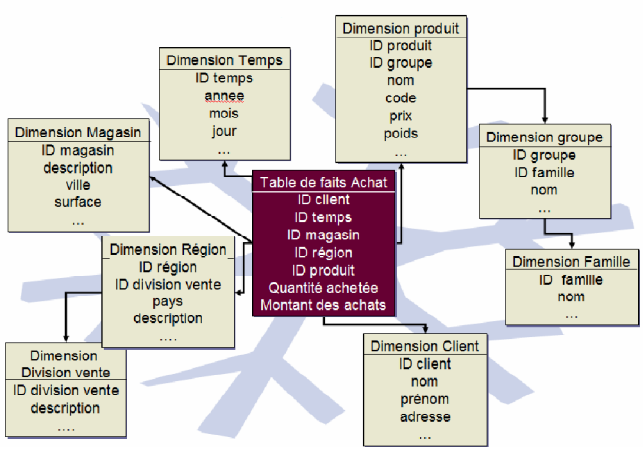


Figure 1.4 Modélisation en flocon

1.2.3.2 Comparaison entre le modèle en étoile et le modèle en flocon

Le tableau suivant montre la différence entre le modèle en étoile et le modèle en flocon en donnant quelques avantages et inconvénients [Zem, 11].

	Avantages	Inconvénients
Modèle en étoile	– Facilité de navigation – Nombre restreint de jointures.	– Redondance des dimensions – toutes les dimensions ne concernent pas les mesures
Modèle en flocon	– Normalisation des dimensions – Economie d'espace disque	– plus complexe – plusieurs jointures dans le traitement des requêtes.

Tableau 1.1 Tableau de comparaison entre Modèle en étoile et le modèle en flocon

1.2.3.3 Modélisation dimensionnelle

L'idée fondamentale de la modélisation dimensionnelle est que presque tous les types de données peuvent être représentés dans une sorte de cube de données, dont les cellules contiennent des valeurs mesurées et les angles sont les dimensions naturelles des données.

- ***Les faits :*** les données observables sur le sujet étudié. Ils représentent ce que l'on veut mesurer. Un fait est formé de **mesures**

- ***Les mesures*** : sont les valeurs numériques correspondant aux informations de l'activité analysée [Tes, 00]**.** Chacune de ces mesures est prise à l'intersection de toutes les dimensions.
- ***La table de faits :*** est la table principale du modèle dimensionnel. Elle modélise le sujet d'analyse et sert à stocker les mesures de l'activité.
- **Les dimensions :** sont les axes d'analyse selon lesquels vont être étudiées **les faits.** Les dimensions sont les points de vue depuis lesquels **les mesures** peuvent être observées. Une dimension se compose de **paramètres.**
- **Les paramètres :** lesinformations faisant varier les mesures de l'activité [Tes, 00]**.**
- **La table de dimension :** c'est la table qui contient le détail sur **les faits.**
- **Granularité :** Nombre de niveaux d'abstraction pour chaque dimension
- **Le cube** : correspond à une vue métier où l'analyste choisit **les mesures** à observer selon certaines **dimensions.** Un cube est une collection de données agrégées et consolidées pour résumer l'information et expliquer la pertinence d'une observation.

La *Table 1* illustre l'exemple d'une entreprise contenant des magasins de vente de produits.la table 2 donne un exemple de la dimension Produit [Net, I.1]

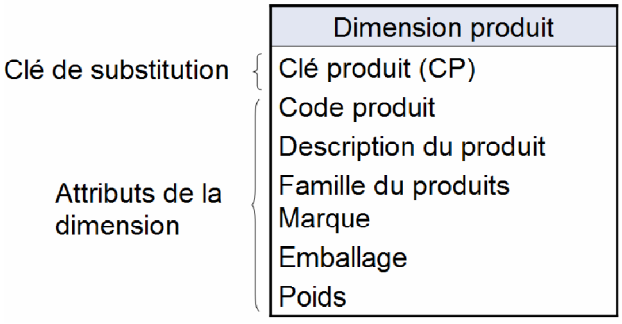


Figure 1.5 Table de dimension Produit

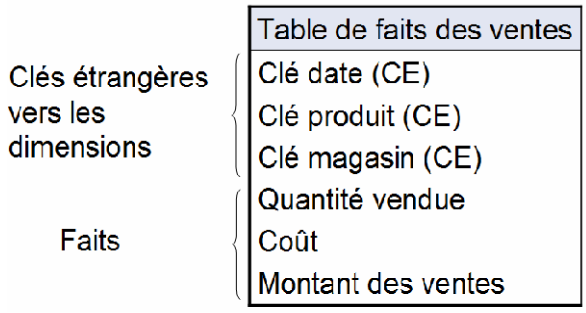


Figure 1.6 Table de faits des ventes

1.2.3.4Le processus décisionnel multidimensionnel

Le modèle multidimensionnel permet d'organiser les données selon des axes représentant des éléments essentiels de l'activité d'une entreprise. Trois niveaux de représentation des données sont définis dans le processus décisionnel :

- **L'entrepôtde données**qui regroupe des données transversales à l'ensemble des métiers de l'entreprise
- **Le magasin de données** qui est une représentation verticale des données portant sur un métier particulier

- **Le cube** de données (ou hypercube).

Le processus décisionnel multidimensionnel consiste en l'exploration de l'hypercube.
L'utilisateur parcourt les données de l'hypercube selon les différents axes d'analyses à la recherche d'informations utiles, dans un processus fortement interactif, itératif et constructif, qui comprend des étapes de formulation des hypothèses, expérimentation etanalyse [Tan, 12].Les utilisateurs interagissent itérativement avec le modèle multidimensionnel pour formuler, modifier et valider leurs hypothèses. Les chemins d'analyse sont imprédictibles, contrairement aux données qui sont définies lors de la conception de l'application. Chaque résultat d'analyse est la conséquence des résultats précédents. Chaque étape du processus d'analyse est représentée par une navigation dansl'hypercube, ou par une requête multidimensionnelle en utilisant les opérateurs OLAP.

En effet dans un processus d'exploration et d'analyse, comparer plusieurs phénomènes est fondamental pour aboutir à une connaissance finale.Corréler plusieurs hypercubes pour avoir une vision unique desdifférentes mesures est donc nécessaire dans le processus d'analyse multidimensionnel.

1.3 OLAP (On Line AnalysisProcessing)

Le but de l'OLAP (*On-Line AnalyticalProcessing*) est de permettre une analyse multidimensionnelle sur des bases de données volumineuses afin de mettre en évidence une analyse particulière des données (il est l'objet d'un questionnement particulier).

1.3.1 Définition de l'OLAP

« Il s'agit d'une catégorie de logiciels axés sur l'exploration et l'analyse rapide des données selon une approche multidimensionnelle à plusieurs niveaux d'agrégation ». Grâce à l'OLAP, les utilisateurs peuvent créer des représentations multidimensionnelles (appelées « *hypercube »s* ou « *cubes OLAP* ») selon les critères qu'ils définissent afin de simuler des situations.

1.3.2Avantages d'OLAP

Les point forts qui font tout le sucées de la technologie OLAP sont principalement les suivants [Pro et al, 02] :

- **La granularité :** Les valeurs sont calculées pour chaque combinaisons de dimensions à l'intérieur du cube, et ceci tant pour les niveaux agrégés que les niveaux détaillés.

- **La navigation et l'exploration** : L'analyse OLAP permet aussi à l'utilisateur d'explorer et de naviguer très facilement à travers les différents niveaux de détail des dimensions à l'aide d'outils de forage.
- **La rapidité d'exécution:** l'approche OLAP est rapide parce que les données sont pré calculées par niveau de détails. Le temps de traitement de l'analyse est alors réduit et répondre très vite aux questions complexes est possible.

1.3.3 Architecture OLAP

Une architecture OLAP générale comprend trois composantes :

- **Base de données :**

– Doit supporter les données agrégées ou résumées

– Peut provenir d'un entrepôt ou d'un marché de données*

– Doit posséder une structure multidimensionnelle (SGDB multidimensionnel ou relationnel)

- **Serveur OLAP :**

– Gère la structure multidimensionnelle dans le SGBD

– Gère l'accès aux données de la part des usagers

- **Module client :**

– Permet aux usagers de manipuler et d'explorer les données

– Affiche les données sous forme de graphiques statistiques et de tableaux

La figure montre les composants de l'architecture OLAP.

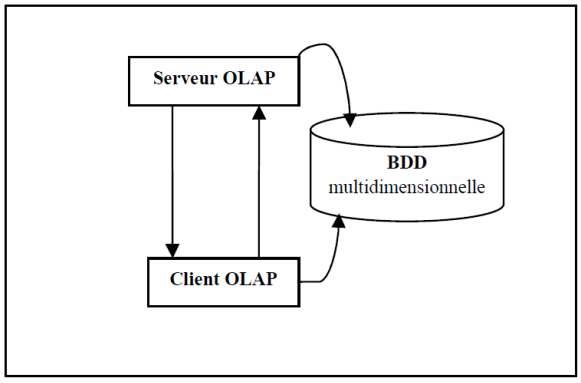


Figure 1.7 les composants de l'architecture OLAP

Les trois composantes peuvent se combiner en plusieurs configurations, selon le type de la base de données accédée : relationnelle, multidimensionnelle, ou hybride (combinée) Bed et al, 05]Ce qui va donner les différents types de l'OLAP.

1.3.4Les Types d'OLAP

OLAP est le fait de faire des analyses sur des bases de données multi dimensionnelles. X-OLAP définit la façon dont seront stockées physiquement les données pour permettre des analyses multi dimensionnelles.

- **R-OLAP**: stocke les données multi dimensionnelles dans un format relationnel (tables, relations)
- **M-OLAP** : les stockent dans un format multi dimensionnel réel
- **H-OLAP** : utilise les deux méthodes pour le stockage
- **D-OLAP** : stocke les données en local pour l'analyse.

1.3.5Les opérateurs OLAP

Les opérateurs OLAP permettent d'explorer les données multidimensionnelles en utilisant les différents concepts de dimensions et hiérarchies.Le cube de données est exploré à l'aide de nombreuses opérations qui permettent sa manipulation. Un panorama des opérateurs OLAP proposés dans la littérature est présenté parRafanelli en (Rafanelli, 2003). Les plus communs sont :

a. Les opérateurs de forage

- ***Roll-up***permet de monter dans les hiérarchies des dimensions, et d'agréger les mesures.
- ***Drill-Down***est l'inverse du Roll-Up et permet de descendre dans une hiérarchie.

b. Les opérateurs de coupe

- ***Slice :*** utilise un prédicat défini sur les membres des dimensions pour couper une partie de l'hypercube limitant le champ d'analyse et permettant à l'utilisateur de se concentrer sur des aspects particuliers du phénomène. En utilisant la terminologie de l'algèbre relationnelle, l'opération de slice est l'équivalent de la sélection.
- ***Dice :*** réduit la dimensionnalité de l'hypercube en éliminant une dimension. Cette opération est équivalente à la projection de l'algèbre relationnelle.
- ***Pivot :***

c. Les opérateursinter-cubent

Drill-Accross :fusionne plusieurs hypercubes en utilisant les axes d'analyse en commun pour comparer leurs mesures. Appelé aussi le forage latérale car il permet de passer d'une mesure à une autre ou bien d'un membre de dimension à un autre.

1.4 SOLAP (Spatial On Line Analysis Processing)

Il est bien connu que les SIG seuls ne présentent pas l'efficacité requise par les applications analytiques (langages d'interrogation, interfaces complexes, temps de traitement longs). L'intérêt d'OLAP pour l'analyse spatiotemporelle a été démontré. Cependant, sans volet cartographique, il est impossible de visualiser la composante géométrique des données. Une solution pourrait être de combiner des technologies spatiales et non-spatiales : SIG et OLAP dans une plate-forme visuelle supportant l'exploration et l'analyse spatio-temporelle faciles et rapides des données selon une approche multidimensionnelle à plusieurs niveaux d'agrégation via un affichage cartographique, tabulaire ou en diagramme statistique.

1.4.1 Définition du SOLAP

La technologie SOLAP est définit par le professeur Yvan Bédard de l'université Laval de Québec comme *"un type de logiciel qui permet la navigation rapide et facile dans les bases de données spatiales et qui offre plusieurs niveaux de granularité d'information, plusieurs thèmes, plusieurs époques et plusieurs modes d'affichage synchronisés ou non : cartes, tableaux et diagrammes "*[Bed et al, 05]

La technologieSOLAP offre de nouvelles fonctions d'aide à la décision non disponibles dans les SIG traditionnels ni dans les outils OLAP.Une technologie SOLAP permet la visualisation cartographique des données, la navigation cartographique dans la carte elle-même ou dans les symboles affichés sur cette carte et ceci selon différents types de forage.

1.4.2 Vocabulaire associé à SOLAP

Nous allons présenter, dans ce qui suit, les concepts fondamentaux liés à la technologie

SOLAP [Bed et al, 05] :

- **Dimensions spatiales**

En plus des démentions descriptives, les outils SOLAP supportent aussi les dimensions (Bédet al. 01 ; Riv et al. 03)

a) **Les dimensions non géométriques :** utilisent une référence spatiale qui est nominale seulement.
b) **Les dimensions spatiales géométriques :** les dimensions géométriques associent une géométrie aux membres de tous les niveaux. (des formes géométriques visualisées et interrogées d'une manière cartographique).
c) **Les dimensions spatiales mixtes :** les dimensions mixtes associent une géométrie aux membres de certains niveaux définis.

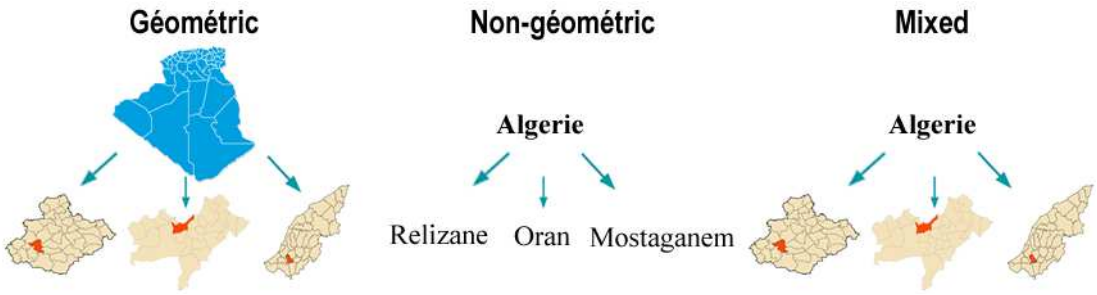


Figure 1.8 Présente un exemple des trois types de dimensions spatiales.

- **Mesures spatiales**

Les outils SOLAP supportent aussi les mesures spatiales [Riv et al, 04].On distingue deux types de mesures spatiales :

a) **Mesures spatiales géométriques :**est le résultat d'un opérateur qui retourne une géométrie. « il s'agit d'un ensemble de coordonnées obtenues à partir des opérateurs d'analyse spatiale d'un SIG » [Bed et al, 05].
b) **Mesures spatiales numériques (non géométriques) :**résulte d'une opération métrique ou des calculs spatiaux **:** surface d'un objet, distance minimale avec l'objet le plus proche, cumul de longueurs sur un réseau).

- **Les Opérateurs spatiaux**

On cite parmi ces opérateurs [Bed et al, 05] :

a) **Les Operateurs spatiaux de navigation :** sont les mêmes operateurs de l'outil OLAP sauf que ses fonctionnalités sont appliquées sur les données géométriques d'une carte géographique on trouve alors le forage spatial, le remontage spatial ; le forage latéral spatial, etc.
b) **Les opérateurs topologiques spatiaux :** sont créés pour faciliter la navigation dans le cube spatial. Ex : adjacent, inclus, intersecte, etc.

c) **Les opérateurs spatiaux temporels :** on peut distinguer quelques exemples d'opérateurs spatio-temporels utilisés dans de différentes applications SOLAP. Ex : précède, en même temps, durant, etc.

1.4.3 Les outils OLAP Spatial

Un outil SOLAP repose sur l'intégration des fonctionnalités SIG et OLAP (Kou, et al. 00 ; Rivest et al. 05).

- **Une composante cartographique**, qui est utilisée pour visualiser les membres de dimensions et/ou les mesures avec une composante spatiale, pour représenter les mesures alphanumériques, grâce à des cartes thématiques, et pour accéder aux opérations de navigation multidimensionnelle.
- **Un entrepôt de données spatiales**, qui doit permettre de modéliser les structures complexes de données associées aux dimensions et aux mesures spatiales.
- **Un client SOLAP** : réétudiées pour permettre en même temps une analyse spatiale et multidimensionnelle
- **un serveur SOLAP** : capable de gérer des requêtes spatio-multidimensionnelles.

1.4.4 Classification des systèmes SOLAP

Différentes systèmes SOLAP, qui peuvent être classifiés en trois différentes typologies, ont été développés. En (Riv, 2000 ; Béd, et al. 05) les solutions SOLAP sont regroupées en trois grandes classes: SIG dominant, OLAP dominant et OLAP-SIG intégrée.

- **SIG dominant**

Dans les solutions SIG dominantes [Hern et al. 05]le serveur OLAP est simulé grâce à une base de données relationnelle modélisée sous forme d'étoile (table de fait et plusieurs tables de dimension). Cette approche offre toutes les fonctionnalités d'un outil SIG : stockage, analyse et visualisation des données spatiales. Par contre, elle doit inclure, dans la base de données, des éléments permettant d'implémenter les opérations OLAP de forage et de coupe, puisqu'il n'existe pas de serveur OLAP pour gérer ces opérations. De plus, toutes les fonctionnalités avancées OLAP comme l'utilisation de mesures dérivées, ne sont pas présentes dans ce type d'outil, ce qui limite ses capacités d'analyse multidimensionnelles.

- **OLAP dominant**

Les outils OLAP dominants [Stolte et al., 2003] utilisent un système OLAP et offrent toutes les fonctionnalités classiques pour l'analyse multidimensionnelle. Par contre, les fonctionnalités SIG sont limitées à une simple représentation cartographique des mesures et

des dimensions spatiales, à la navigation cartographique et à la sélection d'objets géographiques[Bed et al, 05]. Ces solutions ne présentent aucun instrument pour l'analyse spatiale ou d'autres fonctionnalités avancées du SIG nécessaires et complémentaires à l'analyse spatio-multidimensionnelle. De plus, dans ces solutions les opérateurs de forage sur la dimension spatiale sont inexistants ou limités. Par conséquent, dans les solutions OLAP dominant, l'information géographique n'est pas exploitée, ce qui implique que ce type de solution présente quelques limitations pour une analyse spatio-multidimensionnelle concrète[Bed et al, 05].

- **OLAP-SIG intégrée**

Les solutions OLAP-SIG intégrées comme JMap**[Net I.6]**fusionnent toutes les fonctionnalités des deux différents systèmes dans un seul environnent ou les fonctionnalités SIG d'analyse et de visualisation qui sont nécessaires pour l'analyse spatio-multidimensionnelle, complètent les fonctionnalités purement OLAP. Les solutions OLAP-SIG intégrées sont alors les plus adaptées pour une analyse spatio-multidimensionnelle réelle et efficace. Cette intégration peut être vue comme une reformulation des trois niveaux d'une architecture OLAP classique, en utilisant et/ou en ajoutant des fonctionnalités SIG[Hern et al. , 05].

1.4.5 Exemple d'une application SOLAP

Les travaux menés dans le domaine d'aide à la décision spatiale utilisant la technologie SOLAP sont nombreux et dans des domaines d'application différents;on peut citer parmi eux le travail de[Min et al, 12].Cette étude vise à concevoir et mettre en œuvre un prototype « Spatial DataWarhouse » (SDW)etàappliquer des analysesmultidimensionnellesSOLAP pour faire face auxproblèmes liés à lasélection de l'emplacement sur la région du Coré du sud pour faire une mise à jourstratégiquedes données géospatiales tout en minimisant le temps et le cout de la mise à jour. L'implémentation du système a été réalisée en trois phases.

Dans la première phase : les facteurs qui influencent surla politiquede la mise à jourde données géospatialesont été étudiés. Sur la base decette enquête, une table defaitetcinqtables de dimensionont été conçues. Les tables sont structurées dans unschéma en étoile. Les tables du modèle en étoile ainsi configuré, sont implémentées sous l'outil PostGis.

Les tables de dimensions

1. **Table de la dimension spatiale :** Afin de permettre une analyse hiérarchique géospatiale, six niveaux administratifs sont définis où la Corée du Sud est divisée en cinq zones, à savoir Gangwon, Sudo, Yeongnam, Chungcheong et Honamtel (représentent le niveau 5 de la hiérarchie).

2. **Table de la structure de la dimension spatiale :** Les données géospatiales sont disponible dans de nombreux types et formats différents ; les modèles qui étaient considérés dans cette étude sont : le modèle raster, le modèle vecteur, bâtiments 3D et les surfaces.
3. **La table de la dimension Utilisation des bâtiments :** Les bâtiments sont classés dans le résidentiel, le commercial, l'industriel, l'éducatif et d'autres.
4. **La table de la dimension utilisations de la terre.** Les terrains sont classés en 3 classes : zone de développement de restriction, zone de développement prévue, et la zone actuelle d'utilisation de terre.
5. **La dimension temporelle :**Cette étude définit l'intervalle de temps de l'année 2000 ou antérieure à 2009 et fournit les valeurs de données annuelles pour chaque mesure.

La table de fait : contient les clés étrangères de chacune des tables de dimensions plus les mesures qui sont en nombre de 5 et qui sont : le cout, l'âge, le taux de prolongement des bâtiments, le taux de prolongement des terres, le taux de prolongement des routes.

Le modèle en étoile de la table de fait et les cinq tables de dimension est illustrés sur le schéma de la figure ;

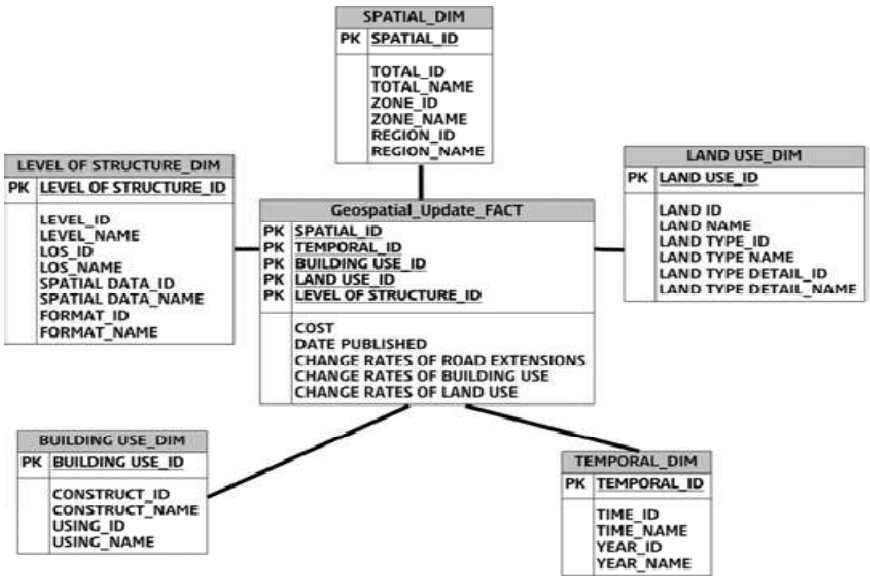


Figure 1.9 Conception du modèle en étoile de l'exemple d'application[Min et al, 12]

Dans la deuxième phase : les différents types etformatsde données provenant desources distribuées sont traitées et transforméesen un formatunifié, puis chargé dans le SDWdéjà conçu. Cette phase, appeléeETLSpatial(Extract, Transform, Load), a été réaliséedans cette étude en utilisant l'open sourceGeoKettle de Pentaho.

La troisième phase consiste enl'implémentation du système SOLAP pour obtenir des informations utiles à partir de l'entrepôt de données spatiales (SDW) modéliséen effectuant une analyse multidimensionnelle. Lesrésultats de l'analysemultidimensionnelle avecSOLAPsont présentéssous forme detableaux de bordspatiaux (des cartes, destableaux et des graphiques) comme le montre la figure 1.10.L'outil qui a été utilisé pour la mise en œuvre du système SOLAP est l'outil JMap SOLAP.

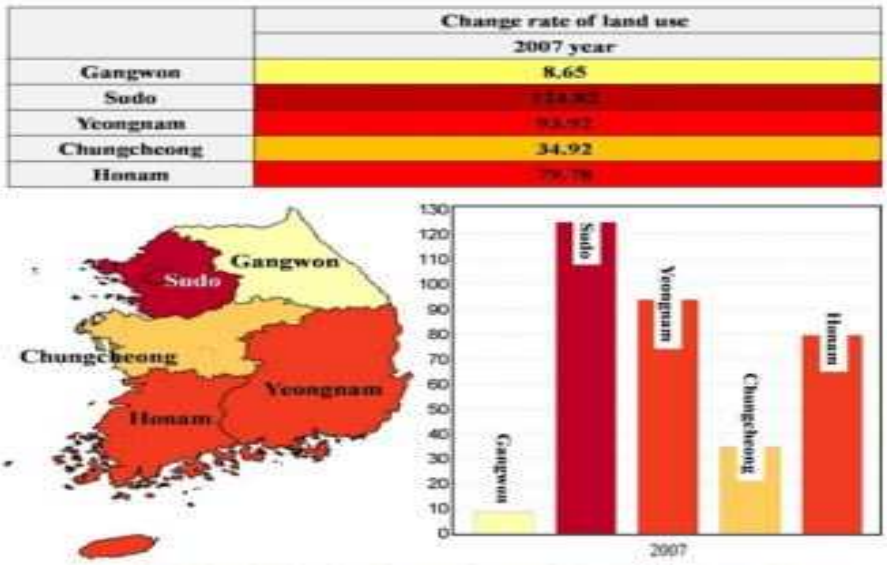


Figure 1.10 configuration du cube de données spatiales et visualisations des résultats[Min et al, 12]

1.5 Conclusion

Dans ce chapitre, nous avons détaillé ces différentes définitions et décrit les modèlesconceptuels pour les bases de données spatio-multidimensionnelles proposés dans la littérature. Ainsi nous avons présenté un panorama des systèmes pour SOLAP, qui révèleque les solutions qui intègrent les fonctionnalités OLAP et SIG sont les plus adéquates à l'analyse spatiale en ligne.Les systèmes SOLAP sont très variés, et peuvent aller de la représentation sur cartes géographiques, au logicielavec carte interactive interrogeant directement la base de données, et ce qu'ellesoit spatiale ou non.

Chpitre2 : Fouille de données spatiales

2.1 Introduction

Jusqu'à 80% des données d'une organisation ont une composante spatiale. Les données spatiales sont de plus en plus nombreuses grâce à l'évolution des outils d'acquisition de données (ex. GPS, images satellite, photo aériennes, etc.) et des méthodes de structuration (ex. raster, vecteur) et de représentation (ex. représentations 2D, 3D). De plus, des outils et des méthodes de représentation des données spatiales (ex. des outils de visualisation) ont été développés pour mettre en évidence les caractéristiques spatiales des données (position, forme, taille, orientation, etc.) et les relations qui existent entre elles (ex. intersection, adjacence, etc.). Afin de faciliter l'interprétation de ces gros volumes de données spatiales, le domaine de la fouille de données(datamining) en général et la fouille de données spatiales en particulier a émergé comme un processus de découverte de règle, relations, corrélations et/ou dépendances à travers une grande quantité de données, grâce à des méthodes statistiques, mathématiques et de reconnaissances de formes.

2.2 Fouille de données

2.2.1 Définition de la fouille de données

La FDD offre des outils et des méthodologies qui peuvent aider à comprendre les données et faire des prédictions.On trouve aussi dans la littérature l'expression Knowledge Discovery in Databases (KDD) ou Extraction de Connaissances à partir de Données (ECD) qui est définit comme le processus global de fouille de données comportant plusieurs étapes dont la fouille de données. Toutefois, par abus de langage, fouille de données et KDD sont souvent confondu. Ainsi, sauf indication contraire, on emploiera le terme fouille de données pour faire référence au processus global dans le KDD[Oua, 10].La fouille de données ne remplace pas les experts, mais les assiste.

2.2.2 Composition de la fouille de données

Le domaine de fouille de données ou data mining converge plusieurs disciplines du domaine informatique, on trouve les bases de données, les entrepôts de donnée…les serveurs de bases de données ou d'entrepôts de données, les bases de connaissances, les engins de fouille de données, les modules d'évaluation du modèle et l'interfaces graphiques pour l'utilisateur [Sch, 04].

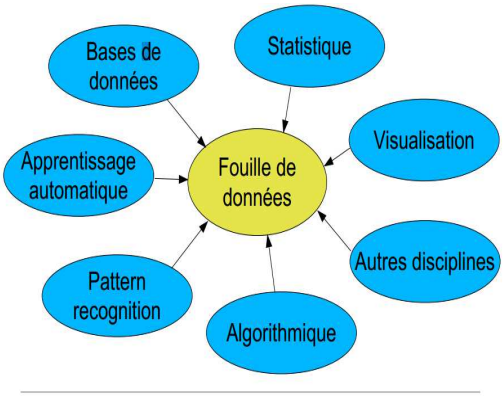


Figure 2.1 disciplines de la fouille de données

C'est un domaine pluridisciplinaire à la confluence de différents domaines: base de données, statistiques, intelligence artificielle, visualisation, parallélisme…[Sch, 04].

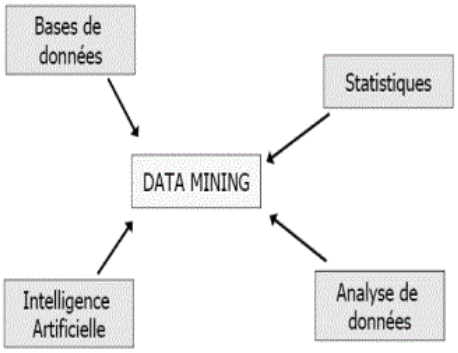


Figure 2.2 domaines de la fouille de données

L'ECD (KDD) est un domaine de recherche à l'intersection de plusieurs disciplines parmi lesquelles les bases de données, l'intelligence artificielle, les statistiques, les reconnaissances de formes, l'apprentissage automatique, la visualisation des données[Oua, 10].

2.2.3 Le processus de la FDD

Le processus de la FDD passe par les étapes suivantes [Abd, 13]:

1. Identification du problème : c'est la phase qui permet de formuler le problème et cerner les objectifs mangeurs.

2. Préparation des données : c'est la phase de la collecte, le nettoyage des données (suppression des doublons, des erreurs de saisie, traitement des informations manquantes, ...), l'enrichissement des données, le codage et la normalisation.

3. Choix du modèle: c'est la phase qui permet de faire le choix d'un type du modèle (classification, …) et une technique (arbres de décision, ...) pour construire ce modèle ainsi que sa validation et évaluation par un expert ou par des statistiques.

4. Utilisation du modèle : c'est la phase de l'application du modèle choisit pour prédire sur de nouvelles données et voir les résultats sur les données.

La figure présente les différentes étapes du processus de la fouille de données.

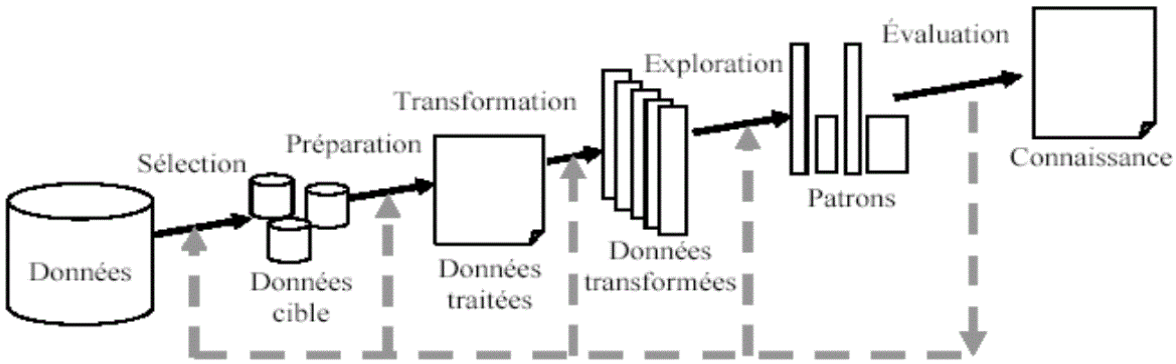


Figure 2.3 Processus du datamining

2.2.4 Typologie des modèles de la FDD

Le type du modèle de la fouille de données peut être choisit selon trois points de vu [Abd, 13]: selon les objectifs, le type d'apprentissage ou le type du modèle obtenu.Les techniques doivent être adaptées au problème considéré[Sch, 04].

2.2.4.1 Selon les objectifs

On distingue, la classification, la prédiction, l'association et la segmentation

- **La classification :** permet d'examiner les caractéristiques d'un objet et lui attribuer une classe
- **La prédiction** : permet de prédire la valeur future d'un attribut en fonction d'autres attributs, par exemple prédire la "qualité" d'un client en fonction de son revenu
- **L'association :** consiste à déterminer les attributs qui sont corrélés : analyse du panier de la ménagère

- **La segmentation :** consiste à former des groupes homogènes à l'intérieur d'une population. Tâche souvent faite avant les précédentes pour trouver les groupes sur lesquels appliquer la classification.

2.2.4.2 Selon le type d'apprentissage

On trouve ici l'apprentissage supervisé et l'apprentissage non supervisé

- **Apprentissage supervisé :** c'est le processus dans lequel l'apprenant reçoit des exemples d'apprentissage comprenant à la fois des données d'entrée et de sortie. Ex : classification, prédiction
- **Apprentissage non supervisé** : c'est le processus dans lequel l'apprenant reçoit des exemples d'apprentissage ne comprenant que des données d'entrée. Ex : Association, segmentation

2.2.4.3 Selon le type de modèles obtenus

Nous avons ici le modèle prédictifs et le modèle descriptif

- **Modèles prédictifs :** utilisent les données avec des résultats connus pour développer des modèles permettant de prédire les valeurs d'autres données. Ex : classification, prédiction
- **Modèles descriptifs :** proposent des descriptions des données pour aider à la prise de décision. Les modèles descriptifs aident à la construction de modèles prédictifs : Association, segmentation

Le tableau 2.1 représente un récapitulatif des modèles de fouille de données

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

Tableau 2.1 Tableau de classification des modèles de fouille de données

2.2.5 Les Méthodes de fouille de données

On cite dans ce titre les méthodes de fouille de données les plus connues[Oua, 10] :

- **K -moyennes (K-means)**

K-Means est une méthode de classification non supervisée et non hiérarchique. C'est une technique de clustering développée par J.MacQueen et mise en œuvre dans sa forme actuelle par E.Forgy. Pour rappel, le clustering est une technique de fouille de données qui permet de classer un ensemble d'objets en groupes de sorte que la similarité intra-groupe soit maximale et celle inter-groupe minimale [Atm, 08].K-Means est un algorithme simple et efficace qui permet de partitionner un ensemble d'objets d'un espace à p dimensions en un nombre fini K de clusters sur la base de la minimisation d'une fonction-objectif.

- **K plus proches voisins (K -ppv ou Knn)**

Fix et Hodges (1951) sont à l'origine de l'approche des k-ppv.Le principe général de la méthode des k-ppv consiste à rechercherparmi l'ensemble d'apprentissage Ω A, contenant l'ensemble des individus et leurs classesd'affectation, un nombre k d'individus parmi les plus proches possibles de l'individu àclasser. Puis, l'individu est affecté à la classe majoritaire parmi ces k individus trouvés. Le nombre k est fixé à priori par l'utilisateur [Atm, 08].

Si k = 1, alors l'individu est affecté à la classe du plus proche voisin de l'ensemble Ω A

- **Naive Bayes**

Les réseaux bayesiens sont des outils de représentationde connaissances en présence d'incertitude. Le succès de ces modèles est fortementlié à leur capacité de représenter et de manipuler des relations de (in)dépendance qui sont importantes pour une gestion efficace des informations incertaines. Les réseauxbayésiens utilisent une représentation basée sur le conditionnement, où lesconnaissances sont structurées sous la forme d'un graphe acyclique orienté. Les nœuds représentent des variables et les arcs codent le lien causal (ou l'influence) entre ces variables.L'incertitude est représentée au niveau de chaque nœud en explicitant toutesles probabilités conditionnelles attachées aux valeurs associées à ce nœud sachantcelles de ses parents. Cette incertitude exprime la force de la relation de causalitéentre les variables [Ben et al, 06]. L'approche bayésienne a pour but de minimiser la probabilité d'erreur de classification.

- **Régression linéaire**

Consiste en l'analyse de la dépendance de plusieurs attributs et de prédire les valeurs de nouveaux enregistrements. La régression consiste en la prédiction des valeurs continues

(numériques) (exemple : quelle sera la valeur d'une maison ou le revenu d'une personne (25milles dollars par an, 10075.99 $, etc.). contrairement à la classification quiprédit plutôt de catégories de valeurs (valeurs discrètes) (exemple : quelle sera la réponse d'un client à une offre ? La réponse pouvant être « J'accepte sans réserve », « j'accepte avec réserve », « je refuse », etc.). La classification et la régression diffèrent également quant à leur mode d'évaluation du résultat de la prédiction[Oua, 10].

- **Réseau de neurones**

Les réseaux de neurones sont des outils très utilisés pour la classification, l'estimation, la prédiction et la segmentation. Ils sont issus de modèles biologiques et sontconstitués d'unités élémentaires : les neurones. Ils sont organisés selon une architecture et ils sont bien adaptés pour les problèmes comprenant des variables continues éventuellement bruitées. Ils obtiennent de bonnes performances, en particulier, pourla reconnaissance de formes[Atm, 08].

- **Arbre de décision**

Le principe des arbres de décision est très utilisé dans l'apprentissage supervisé caril est performant et il permet de générer des procédures de classification sous formede règles.Un arbre de décision est une représentation graphique d'une procédure d'affectationou de classification. Les nœuds internes sont des tests sur les champs ou les attributsX j avec j = 1, .., p. Les feuilles sont les classes Ck avec k = 1, ..., m. Lorsque les testssont binaires, le fils gauche correspond à une réponse positive au test et le fils droit àune réponse négative. Pour classer un nouvel individu, il suffit de descendre dans l'arbre selon les réponses aux différents tests pour l'individu (exemple) considéré [Atm, 08].

- **Règles d'association**

Elles consistent à déterminer les individus qui sont associés. Le principe est la détermination des caractéristiques en fonction des champs, desindividus qu'on trouve réunis dans la même catégorie [Atm, 08].

2.2.6 Classification des techniques de fouille de données

On trouve dans [Chel, 04] une classificationdes différentes techniques de fouille de données selon leurs rôles et leurs type d'analyse:

2.2.6.1 Analyse descriptive

- **Résumer les données :** Moyenne, variante, l'écart-type, analyse factorielle (ACP, AFC), généralisation, caractérisation…
- **Regrouper les données :** K moyenne, les nuées dynamiques, classification croisée, classification hiérarchique
- **Chercher les dépendances entre les données :** corrélation test Khi 2, régression linéaire, régression logistique, analyse factorielle des correspondances, réseaux bayesiens, réseaux de neurones, règles d'associations, graphe d'induction…

2.2.6.2 Analyse prédictive

- **Recherche des règles de classement :** la probabilité conditionnelle, les réseaux de neurones, les réseaux bayesiens, arbre de décision et règles d'association
- **Régression :** régression linéaires, régression logistique, régression multiples

2.3 Fouille de données spatiales

2.3.1 Définition de la fouille de données spatiales

La fouille de données spatiales ou datamining spatial est définie comme l'extraction de connaissances implicites d'une manière automatique ou semi-automatique, de relations spatiales ou d'autres propriétés et connaissances cachées, potentiellement utiles, non explicitement stockées dans des grandes bases de données spatiales [Han et al, 97].

On trouve également le terme GKD ou Geographic Knowledge Discovery définit comme une branche du KDD qui s'intéresse exclusivement à l'information géo-spatiale en tenant compte de sa spécificité et de sa particularité en terme de complexité, hétérogénéité et interdépendance [Oua, 10].

Ses avantages sont d'une part son aspect **exploratoire** dans le sens ou les hypothèses sont générées, contrairement à l'analyse classique. D'autre part, les informations sur la **localisation** spatiale et sur les liens de **voisinage** sont complètement intégrées.

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

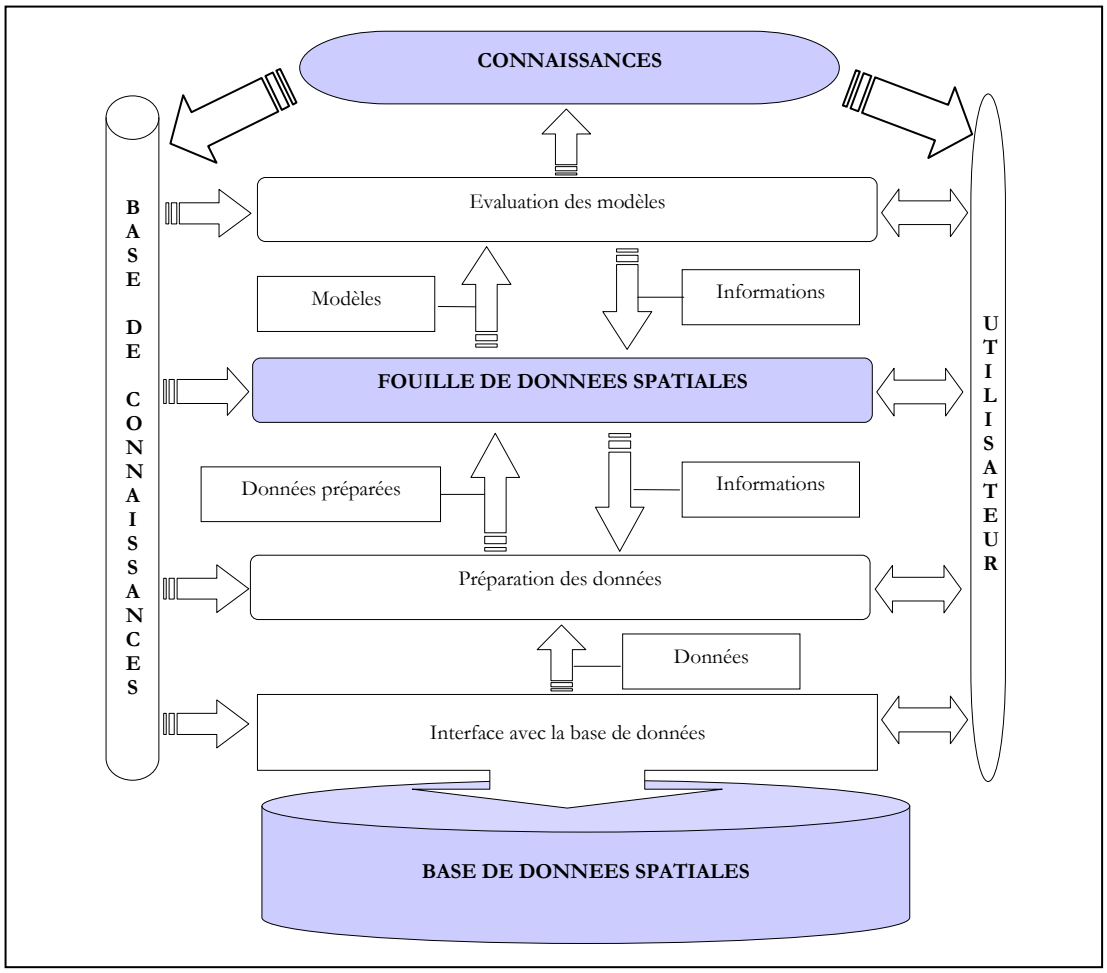


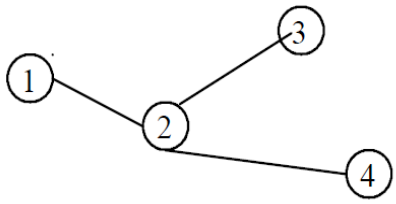
Figure 2.4Place de la FDS dans le processus de découverte des connaissances [Chel, 04]

2.3.2 Caractéristiques des données spatiales

La particularité de la fouille de données géo-spatiales est intrinsèquement liée aux caractéristiques des données traitées[Oua, 10]. Cette différence peut être caractérisée selon cinq points de vue [She et al, 03]: les données, les relations géo-spatiales, l'hétérogénéité, les fondements statistiques et la consommation de ressources.

2.3.3 Matrice et graphe de voisinage

Les relations spatiales sont communément formalisées par la notion de graphe de voisinage, et peuvent être représentées sous forme d'une matrice de voisinage [Lebart, 97]. Celle-ci est une matrice binaire M où M[i,j]=1 si l'objet i est voisin de l'objet j et M[i,j]=0 dans le cas inverse. Ceci est illustré par la figure suivante.



$$M = \begin{vmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{vmatrix}$$

Figure 2.5Graphe de voisinage et matrice de voisinage

La notion de voisinage est générale et peut aussi bien représenter une contiguïté entre formes zonales ou une proximité sur des points. Elle peut être pondérée en qualifiant la proximité par une distance.

2.3.4 Les méthodes de la fouille de données spatiales

Les méthodes types sont un prolongement des tâches de fouille de données intégrant les données et les critères spatiaux.

2.3.4.1 Phase exploratoire

Cette phase permet une description synthétique (**indice d'auto corrélation globale,généralisation, densité, lissage**), de découvrir les écarts donnant les spécificités locales (auto-corrélation locale ou analyse factorielle locale) ou de chercher des regroupements de données (clusters).Cette première phase permet de guider **la phase décisionnelle**.

- **Indice d'auto corrélation globale**

Bien avant l'ère des SIG, des mesures du degré de dépendance aux voisins, dites **d'auto corrélation spatiale globale** ont été étudiées. Elles exploitent, hormis les attributs de l'objet, la matrice de voisinage. Ces mesures sont calculées à l'aide de deux méthodes complémentaires : l'indice de Moran (en 1948) et l'indice de Geary (1954). Dans le cas où les données seraient corrélées, il faut les simplifier afin de faire apparaître une tendance générale.

- **Généralisation spatiale**

Cette méthode est une extension aux données spatiales de la généralisation basée sur l'induction orientée attribut. Elle consiste à substituer les valeurs estimées trop détaillées par des valeurs moins détaillées jusqu'au niveau de détail souhaité, puis à agréger et compter les n-uplets identiques ainsi obtenus. Cette méthode permet de résumer les données et constitue une première étape pour induire **des règles d'associations.** Elle nécessite au préalable de disposer d'une connaissance a priori que l'on trouve dans des « hiérarchies de concepts » définies par des experts. Une hiérarchie de concepts est définie pour un attribut. Elle décrit le passage des concepts les plus spécifiques -correspondant aux valeurs de l'attribut dans la base

de données- aux concepts plus généraux de niveau supérieur. Pour un attribut de type spatial, cette hiérarchie est appelée hiérarchie spatiale et correspond à une relation spatiale d'inclusion entre objets.

Exemple : Le découpage administratif en pays, régions, département, communes, etc. en est un exemple. Pour un attribut non spatial, on parle de hiérarchie thématique.

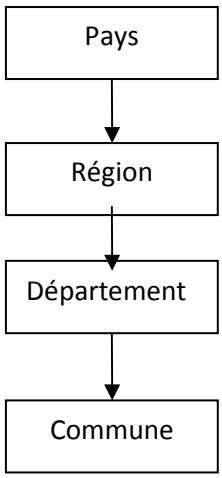


Figure 1.6 Hiérarchie spatiale

Il existe deux types de généralisation :

- **La généralisation à dominante spatiale**

La généralisation à dominante spatiale exploite une hiérarchie spatiale existante en plus des hiérarchies thématiques. Les attributs sont générés par simplification et comptage.

- **La généralisation à dominante non spatiale**

La généralisation à dominante non spatiale n'utilise pas de hiérarchie spatiale mais génère des localisations moins détaillées par fusion d'objets spatiaux. Une induction orientée attribut est faite en utilisant des hiérarchies thématiques, mais en gardant leur description spatiale. Cette induction produit des valeurs d'attributs homogènes (doublons) pour plusieurs objets. Ces objets sont alors fusionnés.

- **Analyse globale par lissage**

L'**analyse multidimensionnelle lissée** est obtenue en remplaçant chaque valeur de la matrice de voisinage par le barycentre de ses voisins. Les méthodes factorielles sont étendues pour prendre en compte la contiguïté puis elles sont appliquées sur le tableau ainsi modifiée.

- **Analyse locale**

A l'inverse de l'analyse globale qui cherche à gommer les particularités, l'analyse locale vise à les faire ressortir pour mettre en évidence les données atypiques. Par exemple, si nous

considérons les variables taux de scolarisation et revenu moyen, il n'existe à priori pas de corrélation au niveau global sur l'ensemble du territoire. Par contre, si l'analyse s'effectue au niveau local sur une région bien précise, une corrélation pourra être trouvée. L'analyse locale peut éventuellement contredire les résultats de l'analyse globale. L'auto corrélation locale consiste à calculer **un indice local d'association spatiale**, dérivé de la formule de l'indice global où l'on remplace la matrice de voisinage par le vecteur ou la ligne de la matrice correspondant à l'objet. L'analyse multidimensionnelle locale est analogue à celle d'analyse lissée présentée ci-dessus. De la même manière, elle procède tout d'abord par transformation du tableau initial, mais en tableau contrasté. Ce dernier correspond à la différence du tableau initial et du tableau lissé présenté dans l'analyse locale. La suite est une analyse factorielle classique sur le tableau ainsi modifié.

- **Clustering**

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

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

2.3.4.2 Phase décisionnelle

Cette phase procède à une analyse plus fine afin d'expliquer les écarts ou de caractériser les groupes (**caractérisation, règles de classement ou d'associations).** Le terme explicatif ici est lié à une intervention de l'analyste qui, à la suite d'une découverte de clusters ou de valeurs atypiques par rapport à une tendance, focalise son analyse sur un sous-ensemble d'objets, sur une partie des variables ou encore sur une zone géographique. Cette partie des données est ensuite analysée dans le but d'expliquer sa particularité par des liens avec certaines valeurs ou

par des règles caractéristiques. Ces méthodes, à l'inverse des méthodes précédentes, opèrent sur plusieurs couches thématiques pour permettre d'expliquer un phénomène suivant les propriétés de son entourage.

- **Caractérisation**

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

Pi = (attribut= valeur) tels que :

(a) sa fréquence relative dans S et dans son voisinage jusqu'à un ordre n'est significativement différente par rapport à sa fréquence relative dans la base (rapport de fréquences supérieur à un seuil donné).

(b) sa fréquence relative est significativement différente dans le voisinage d'une proportion minimum d'objets du sous-ensemble S (proportion supérieure à un seuil de confiance).

La caractérisation découvre en plus le niveau d'extension de ces propriétés aux "voisins". Une propriété caractéristique d'un sous-ensemble S est un prédicat p_i= (attribut= valeur) tels que :

- **Règles d'association**

L'extension de la découverte de règles d'association aux données spatiales permet de générer des règles de type :

X → Y (s, c) avec s comme support et c la confiance telles que X et Y sont des ensembles de prédicats spatiaux et non spatiaux. En d'autres termes, ceci revient à trouver des associations entre des propriétés des objets et celles de leur voisinage.

Exemple : la recherche d'associations impliquant les terrains de golf et les autres entités géographiques (bâtis, infrastructure, etc.) génère un ensemble de règles comme par exemple :

is_a (x, "golf") ->close_to (x, "zone pavillonnaire") (61%, 70%)

- **Classification**

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

parle de segmentation ou de scoring en statistique. La classification s'exprime généralement sous la forme d'un arbre de décision pour lequel l'algorithme de référence est ID3. L'extension au domaine spatial a été définie par l'extension aux propriétés de leurs voisins jusqu'à un ordre N de voisinage

Exemple : il est possible de trouver une règle de type :

Si population élevée et type de voisin = route et voisin de voisin = aéroport **Alors** puissance économique élevée (à 95%).

2.3.5Exemples d'application de fouille de données spatiales

La fouille de données spatiales est un domaine de recherche en pleine expansion. Il offre de nouvelles perspectives pour beaucoup d'applications à caractère décisionnel qui explorent les données de recensements, le trafic routier comme l'analyse du risque d'accident routier, des foyers d'épidémies, de la criminalité (analyse des données cartographiques du crime) et des risques naturels comme les inondations, les feux, les sécheresses et les tremblements de terre, etc. On cite quelques applications dans ce qui suit

2.3.5.1 Application pour la détection des foyers d'épidémies [Net II.1]

C'est en 1855 qu'apparait peut-être la première fouille de données géographique prédictive. Cette année-là, John Snow recherche les causes de l'épidémie de choléra de la fin de l'année 1854 à Londres, et avec une hypothèse novatrice, et une description précise des foyers de la maladie, prouve que l'eau est un vecteur de contamination et trouve la pompe à eau de *Broad Street* qui en est la cause

La figure montre une carte originale de John Snow décrivant le mode de propagation du choléra en 1854 dans le quartier de Soho, Londres, la figure 2 montre la répartition des points d'eau et des victimes de choléra**.**

Figure 2.7Répartition des puits d'eau et des victimes de choléra dans le quartier de Soho a Londre[Net II.1]

2.3.5.2Application dans le domaine de la criminalité[Oua, 10]

Cette application utilise des données concernant la criminalité de la ville de San-Francisco pour diverses périodes notamment 2003-2010. Ces données concernent l'ensemble des appels d'urgence. La méthode utilisée est le Géo clustering qui a permis de catégoriser les crimes en 3 clusters comme le montre la figure

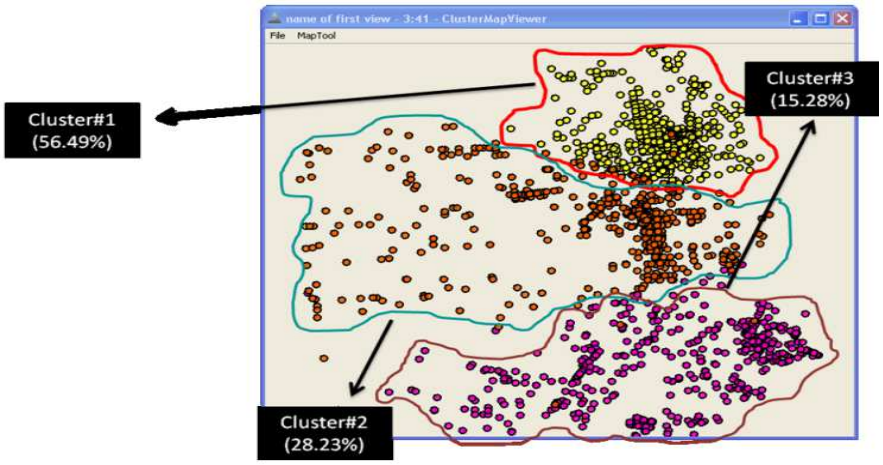


Figure 2.8Répartition des clusters et pourcentage de crime[Oua, 10]

2.3.5.3 Application dans la surveillance de l'accidentologie routière [Chel, 04]

L'analyse du risque routier permet d'identifier les problèmes de sécurité sur le réseau routier en vue de proposer des mesures de sécurité pour y remédier. Le risque routier est estimé à partir du retour d'expérience sur les accidents corporels de la circulation. cette application est basé sur une immense quantité de données numérisées

- **Données thématiques :** sur les accidents, le réseau routier et le flux de véhicules.
- **Données géographiques :** le découpage administratif en communes, quartiers ou îlots, le bâti, la population, etc.

La figure représente la visualisation des différents catégories d'accidents de la route (piétons, 02 roues et autres) ainsi que les alentours des écoles et des administrations(écoles, marchés, gare).

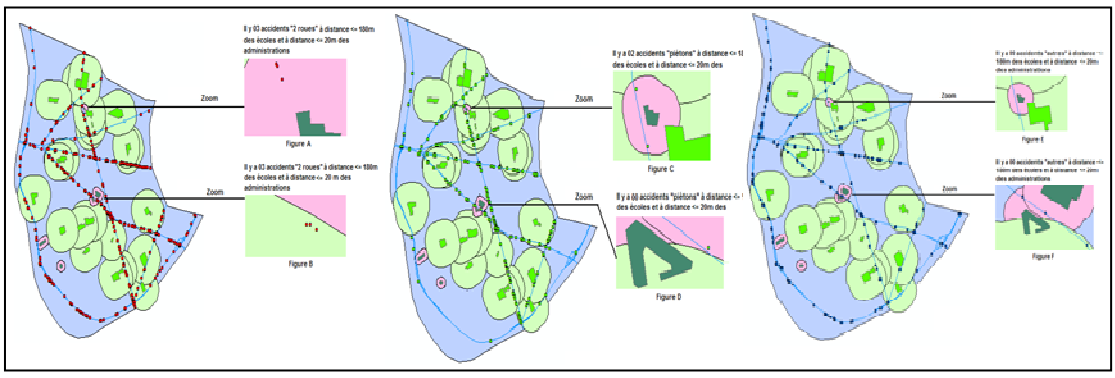


Figure 2.9 La visualisation de différentes catégories d'accidents de la route en fonction des bâtiments autours [Chel, 04]

2.4 Conclusion

Nous avons présenté dans ce chapitre le domaine de la fouille de données et la fouille de données spatiales en spécifiant sa place par rapport au domaine de l'extraction de la connaissance à partir de données dans le processus globale de la prise de décision. Nous avons décrit ensuite les différentes méthodes existantes dans la littérature et nous avons donné plusieurs classifications illustrées par des exemples d'applications réelles issus des travaux de recherche actuels.

Chapitre3 : Model Décisionnel proposé

3.1Introduction

La surveillance épidémiologique a pour but de lutter contre les épidémies. Il est connu qu'on ne peut jamais stopper une épidémie mais par contre on peut toujours limiter les taux de mortalité. Le recours aux technologies de l'information faciliterait grandement la réalisation d'un tel objectif.

La meilleure façon de traiter les questions de la santé publique en générale et la surveillance épidémiologique en particulier est l'élaboration d'un système d'aide à la décision susceptible de fournir les données pertinentes aux auteurs de la santé publique pour la prise de la décision. Le modèle décisionnel proposé servira comme système de supervision permettant la localisation des zones qui présentent un foyer d'épidémie en utilisant les technologies

géo-décisionnelles les plus avancées (Datawarhouse, Datamart, SIG, cubes et hyper cubes, OLAP, SOLAP et tableau de bord spatial).Est l'analyse des observations de larges jeux de données dans le but d'identifier des relations non soupçonnées et de résumer la connaissance incluse au sein de ces données sous de nouvelles formes à la fois compréhensibles et utiles pour l'expert de ces données.

3.2Eléments de la Surveillance Epidémiologique

La surveillance épidémiologique est un processus de prise de décision. La Figure 3.1 illustre la démarche technique adoptée par le processus de la surveillance épidémiologique telle qu'elle se fait sur le terrain [Bou, 00]; ce schéma servira par la suite pour élaborer une démarche décisionnelle dans la surveillance épidémiologique.

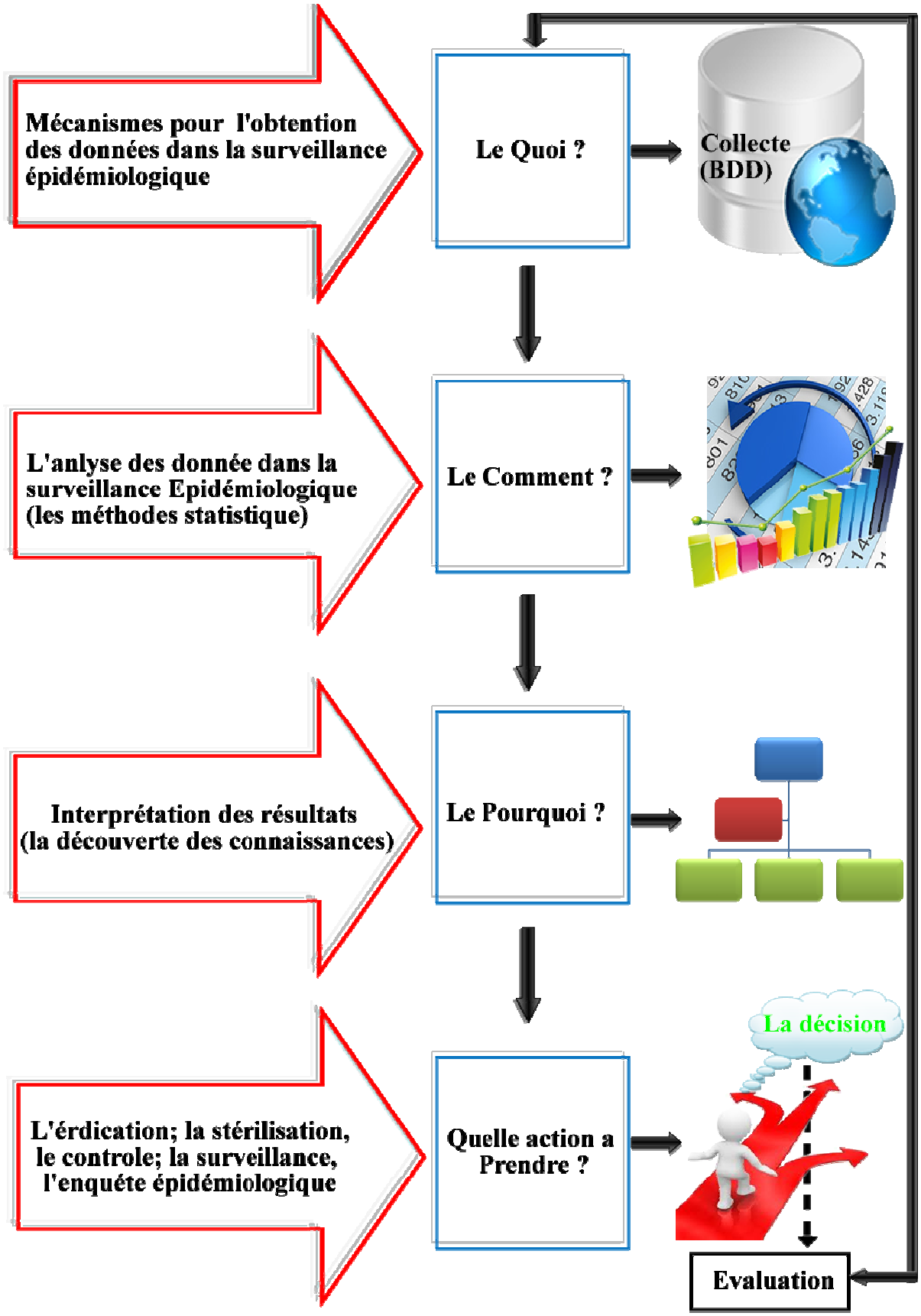


Figure3 1 Processus de la prise de décision dans la surveillance épidémiologique

Le processus de la surveillance épidémiologique passe par plusieurs étapes pour arriver jusqu'à sa finalité qui est la prise de la décision par l'acteur de la santé publique que ça soit un simple épidémiologique ou un décideur de haut niveau [Zem, 11] ;

3.2.1 Etape1 : L'enregistrement des données (la collecte)

L'enregistrement des données sanitaires s'effectue sur la base de notification des maladies prioritaires et à déclaration obligatoire. C'est un processus qui comprend [Bou, 00]:

- La sélection des maladies qui feront l'objet de la surveillance épidémique et le typede données à enregistrer.
- La détermination de la méthode et des instruments de transmission des données.
- L'élaboration des supports d'enregistrement (formulaires, registres)
- La mise en place de circuit de notification avec la fréquence.

L'enregistrement des données devra être complété par l'observation sur le terrain et par des enquêtes épidémiologiques complémentaires

3.2.2 Etape2 : L'analyse des données

L'analyse des données sanitaires est un processus d'exploitation de base qui comprend, d'une part, un traitement des supports de notification et d'autre part, une analyse des données complétée par une interprétation des résultats. L'analyse est uneactivité qui inclut diverses comparaisons des données et qui pour objet [Bou, 00]:

- D'établir les tendances de la maladie afin de détecter des hausses ou des baisseséventuelles d'incidence.
- D'identifier les facteurs associés avec les changements de comportements des processus infectieux.
- De spécifié les localités les plus vulnérables pour appliquer les mesures de lutte.

L'analyse des données peut être achevée par des comparaisons par rapport au temps, aux personnes et aux lieus**.**

3.2.3 Etape3 : Interprétation des résultats d'analyse

Cette étape consiste à trouver les facteurs sociaux et environnementaux favorisant la propagation de l'épidémie. Cet aspect a fait l'objet de toute une étude portant sur le niveau socio-économique du malade et ses conditions de vie et l'influence de ces facteurs sur la maladie de la tuberculose.

3.2.4 Etape4 : La surveillance action et les mesures à prendre (La prise de la décision)

Après l'analyse des données et l'interprétation des résultats, la surveillance épidémiologique doit être complétée par la surveillance action, qui comprend : la retro information, la mise en place des moyens de contrôle, de prévention et de lutte, et les mesures de blocage de tout processus épidémique qui consiste à suivre les étapes du processus de l'éradication de la maladie contagieuse [Bou, 00].

3.3Le modèle décisionnel proposé

Avant de présenter en détail notre model décisionnel proposé pour la surveillance épidémiologique, nous donnons en quelque lignes, un bref aperçu sur le domaine de l'aide à la décision et les systèmes Interactifs d'aide à la décision

3.3.1 L'aide à la décision

L'aide à la décision consiste à assister les décideurs et les aider à mieux exprimer leurs choix et préférence vis-à-vis une situation donnée. On ne cherche donc pas une vérité, mais bien à établir une démarche permettant aux décideurs d'apprendre sur le problème pour mieux choisir.

3.3.2 Les Systèmes Interactifs d'Aide à la Décision (SIAD)

Un SIAD est « un système informatisé qui utilises les connaissances sur un sujet particulier afin d'aider le responsable lors de la prisede décision dans une catégorie de problèmes peu ou pas structurés ». [Bonczek, 1984]

3.3.3 Le modèle de SIAD proposé

En s'appuyant sur les éléments de la surveillance épidémiologique cités ci déçus et les quatre fonctions principales d'un SIAD citées par [Ghy et al, 07] ; nous avons proposé le modèle décisionnel illustré par la figure 3.2. Ce modèle suit les étapes de l'approche décisionnelle spatiale proposées par[Zemri, 11].

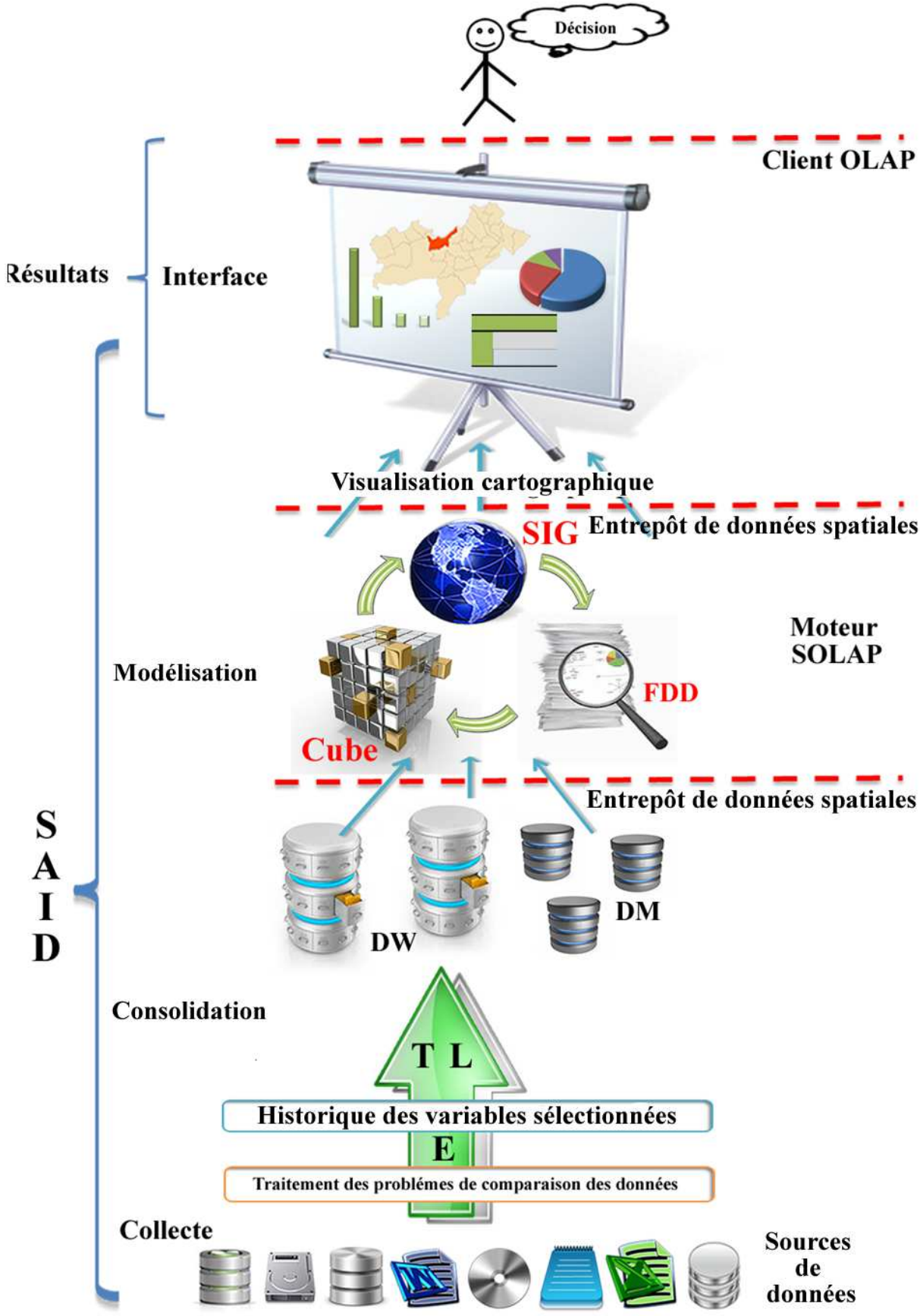


Figure 3 2 Système Interactif d'Aide à la Décision proposé pour la Surveillance Epidémiologique (SIADSE)

Dans le modèle décisionnel proposé pour la surveillance épidémiologique, nous distinguons quatrefonctions fondamentales à savoir : la **Collecte, la Consolidation, la Modélisation et l'Interface.** [Zem, 11] ;

3.3.3.1 Fonction Collecte

Cette première étape se charge de la collection des différentes données de la surveillance prévenant des différentes « **bases de données sources »** ou « **bases de production »** et les emmagasiner dans des bases de données spécialisées « **Entrepôts de données »** et **« Magasins de données ».** On alimentela base de données depuis les données sources par un système de chargement de données**ETL (Extraction, Transformation et Chargement)** destiné à **extraire** des données de diverses sources (bases de données de production, fichiers, Internet, etc.),qui sont souvent hétérogènes en les rendant homogènes, les transformer et les charger dans un entrepôt de données**afin de les analyser [Net 3.1].**

3.3.3.2 Fonction Consolidation (structuration et stockage)

Élément central qui permet aux applications décisionnelles de bénéficier d'une source d'information commune, homogène, normalisée et fiable qui masque la diversité de l'origine des données provenant de différentes sources. La fonction de consolidation est généralement assurée par la gestion des métadonnées, qui assurent l'interopérabilité entre les données. Cette phase comporte tous les outils de création d'un ED en passant par sa Conception, samodélisation et sa structuration.

- **Conception**

Ça concerne la conception multidimensionnelle des données dans des cubes ou hypercubes. Un entrepôt de données peut héberger des milliers de variables mais quelques dizaines seulement sont exploitées pour une activité décisionnelle particulière. Pour cela on utilise des outils décisionnels comme les tableaux, les cubes et les hyper cubes.La conception de Datawarehouse s'articule essentiellement autour de la mémorisation desdonnées dans une base de données unique.L'identification des besoins permettra de sélectionner dans le système d'information opérationnel les données nécessaires pour l'élaboration des informations demandées et les mémoriser dans une base unique.

- **Modélisation dimensionnelle**

L'indicateur de succès d'un projet de Datawarehouse est sa capacité de fournir les informations nécessaires au moment souhaité. Pour aboutir à ce niveau de succès, il faut se baser sur plusieurs niveaux de détails de données, ce qui est assuré par la modélisation multidimensionnelle.La modélisation multidimensionnelle permet la représentation explicite des hiérarchies et même la possibilité de manipuler à la fois le contenu et la structure des données [Kim, 97].

- **Structuration en cubes multidimensionnels**

Les Datawarehouse sont destinés à la mise en place de systèmes décisionnels. Cessystèmes, devant répondre à des objectifs différents des systèmes transactionnels, ont faitressortir très vite la nécessité de recourir à un modèle de données simplifié et aisémentcompréhensible. La modélisation dimensionnelle permet cela. Elle consiste à considérer unsujet d'analyse comme un cube à plusieurs dimensions, offrant des vues en tranches ou des analyses selon différents axes et utilise des opérateurs spécifiques aux cubes pour répondre de manière pertinente aux requêtes des utilisateurs.

3.3.3.3 Fonction Modélisation (Outils d'analyse et d'interprétation)

Ça concerne la phase d'analyse en ligne chargée de la création du cube à partir de l'ED déjà mis en place. Cette étape contient les outils OLAP (On Line AnalyticalProcessing) à savoir les outils d'Analyse, les outils de restitution des données sous différentes formes (graphiques ou tableaux) et les outils d'administration.

- **Analyse des données**

Cette phase est le but du processus d'entreposage des données. Elle doit permettre toutes les analyses nécessaires pour la construction des indicateurs recherchés.

- **Restitution des résultats d'analyse**

L'étape de restitution se fait en utilisant des outils clients d'un entrepôt de données.

Tous les outils pouvant synthétiser, explorer, confirmer, expliquer, prédire les données sont des outils de restitution.

Différents types d'usagers nécessitent différents outils d'exploitation de données. Il en existe pour cela cinq principaux types : les logiciels requêteurs, les logiciels de création de rapports,

les tableaux de bord, les outils OLAP et SOLAP et les outils de fouille de données (data mining).

- **Administration**

La fonction d'administration consiste à assurer la qualité et la pérennité des données aux différents applicatifs ; la maintenance ; la gestion de configuration ; les mises à jour; l'organisation, l'optimisation et la mise en sécurité du système d'information.

La phase d'analyse et de restitution des données conditionne le choix de l'architecture de l'ED et de sa construction.

3.3.3.4 Fonction Interface

Assure le contrôle **d'accès** des utilisateurs, la prise en charge des **requêtes**, la **visualisation** des résultats d'analyse sous différentes formes : tableaux de bord, graphiques, corrélation, simulation, datamining…

3.4 Outils d'investigation

Nous avons utilisé, pour l'élaboration de notre modèle décisionnel proposé, les outils d'investigation suivants :

3.4.1 Outils de stockage : Les systèmes décisionnels actuels comportent deux types d'espaces de stockage à savoir les entrepôts de données et les magasins de données.

- **Datawarhouse (Entrepôt de Données)**

Le rôle de l'entrepôt de données dans une application décisionnelle et de regrouper dans un format homogène des données utiles pour l'aide à la décision provenant soit de sources internes (base de production) ou externes (par exemple provenant d'internet ou dans notre cas provenant des bases des établissements de santé).

L'ED contribue dans l'assemblage et le stockage ainsi que la structuration de nos données de la surveillance et leur préparation pour l'analyse SOLAP.

- **Data Mart (Magasin de données)**

Le magasin de données qui représente un extrait d'informations, orienté sujet, provenant de l'entrepôt est organisé de manière adéquate pour y appliquer des analyses rapides à des fins de prise de décisions ; c'est un entrepôt thématique, principalement dédié à une classe de décideurs. Le magasin de données servira pour la structuration d'un sous ensemble des

données contenant l'entrepôt de données précédemment mis en place qui seront utilisées à un seul type d'analyse.

L'objectif consiste alors à adapter au mieux les structures de données à l'utilisation qui en sera faite.

3.4.2 Outil OLAP (cube ou hyper cube)

« Technologies permettant de collecter, stocker, traiter et restituer des données multidimensionnelles à analyser»

L'interrogation permet de connaitre, mesurer et prévoir (prise de décisions) au travers de la manipulation des données du magasin. On peut considérer plusieurs opérations de manipulation:

- Consultation des données d'un tableau et génération de graphiques;
- Requêtage graphique sur une base de données;
- Application des operateurs multidimensionnels.

3.4.3 Outis de fouille de données (data mining)

> *« Offre des outils et des méthodologies qui peuvent aider à comprendre les données et faire des prédictions »*

On a choisi parmi les algorithmes offerte par l'outil utilisé (SQL server 2012), l'algorithme **« arbre de décision »**ou **« DecisionTree »** qui est une méthodereprésentant une structure de connaissances composée d'une séquence de règles de décision. Il a pour but de trouver les attributs explicatifs et les critères précis sur ces attributs donnant le meilleur classement vis-à-vis d'un attribut à expliquer[Mou, 13].

Notre choix était porté sur cette méthode de datamining « DecisionTree » pour ses avantages multiples tel que la facilité à manipuler des données « symboliques » et l'utilisation des variables d'amplitudes très différentes (données multifactorielles ou multi critères) adaptée à notre étudeen surveillance épidémiologique et qui donne une interprétation et une classification très efficace**.** Néanmoins elle a quelques Inconvénientsa signaler tel que laSensibilité au bruit et points aberrants

3.4.4 Outil d'affichage cartographique (SIG)

Le SIG est définit comme un système informatique de matériel, de logiciel, et de processus conçu pour permettre la collecte, la gestion, la manipulation, l'analyse, la modélisation et l'affichage de données à référence spatiale afin de résoudre des problèmes complexes de

gestion. Les SIG différent selon leurs domaines d'applications et les demandes qu'ils doivent satisfaire. Toutefois, ils ont en commun des fonctionnalités nommées les « 5A » : Abstraction, Acquisition, Archivage, Affichage et Analyse.

3.4Conclusion

Dans ce chapitre, nous avons présenté en détail notre système Interactif d'Aide à la Décision proposé pour la surveillance épidémiologique « SIADSE » en utilisant les deux outils décisionnels SOLAP et le DataMining.

Le model décisionnel suggéré est structuré selon les quatre étapes : Collecte, Consolidation, Modélisation et Interface. En fin de chapitre, nous avons donné les outils d'investigation utilisés pour l'élaboration de notre modèle.

Le prochain chapitre donnera l'aspect technique du Système Interactif d'Aide à la Décision « SIADSE ».

4.1 Introduction

Pour étudier l'épidémie, nous devons prendre en considération: l'environnement, l'espace, le temps et différentes autres fonctions. Pour ce faire nous devons choisir l'outil le plus adéquat. Dans ce chapitre, nous allons détailler notre application qui répond à ces besoins.

Notre travail est constitué de deux parties , d'une part appliquer les principes des systèmes d'information géographiques afin d'extraire les informations utiles à une bonne modélisation, et d'une autre part appliquer les techniques de l'OLAP pour l'analyse et le suivi des données de la surveillance épidémiologique qui faciliterait à l'expert à la fois de prendre des décision pour l'éradication de la maladie et ainsi de valider le modèle.

4.2 Contexte de l'étude

La tuberculose est une maladie infectieuse du poumon et de ses membranes. Contagieuse, elle est la conséquence directe de l'infection par le bacille de Koch. Plus de deux milliards de personnes sont contaminées par la bactérie, ce qui représente près d'un tiers de la population mondiale (chiffres de 2010, OMS [40]), tel que c'est illustré dans la Figure 4.1. Seul un porteur du bacille de Koch sur dix finit par développer la maladie.

Figure 4.1 – Incidences estimées de la tuberculose par pays en 2010.

4.3 Présentation de la zone d'étude

Dans le but de concevoir un système d'information géographique d'aide à la décision pour survie l'épidémiologie , on a choisi comme lieu d'application la wilaya d'Oran plus exactement les commune de la wilaya.

Oran (en arabe : وهران), est la deuxième ville d'Algérie et une des plus importantes du Maghreb. C'est une ville portuaire de la mer Méditerranée, située au nord-ouest de l'Algérie, à 432 km de la capitale Alger, et le chef-lieu de la wilaya du même nom, en bordure du golfe d'Oran.

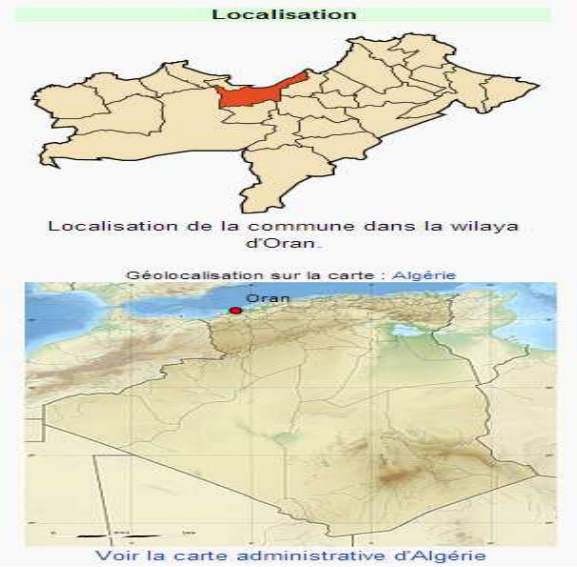


Figure 4.2 Localisation de la zone d'étude (la wilaya d'Oran)

4.4 Définition du projet

Notre projet consiste à aider dans la réalisation de l'objectif majeur de la surveillance épidémiologique dans la wilaya d'Oran: *« Bien cerner le problème de la propagation des épidémies qui est un problème de nature incontrôlable ».*

Pour cela, nous avons élaboré un système d'aide à la décision par l'intégration de l'outil SIG et l'outil base de données déjà existants afin de profiter des avantages de chacun pour garantir de meilleures analyses spatiotemporelles et aider, ainsi, efficacement les acteurs en santé publique.

4.5 Les données de l'étude

Nous avons deux types de données utilisées dans notre étude : les données géographiques (la carte de la wilaya d'Oran) et les données alphanumériques (les données de la surveillance épidémiologique).

4.5.1 la carte d'Oran (donnée géographique)

Les cartes sont les outils utilisés le plus fréquemment pour comprendre les informations spatiales. Qu'il s'agisse d'analyse, de modification, d'illustration de rapports, de concevoir des bases de données ou de les gérer, pour cela nous avons extrait la carte d'Oran à partir de « GoogleMap » pour réussir à faire toute ces opérations en suivant les étapes de création des couches avec la fonctionnalité ArcMap de l'outil Arc Gis d'ESRI.

La figure 4.4 montre la création une classe d'entités de géodatabase en choisissant le nom, le type de couche (polygone) et le système de coordonnées.

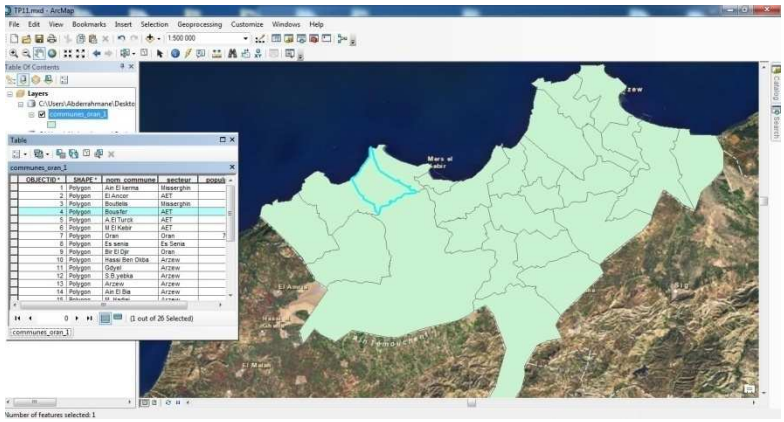


Figure 4.3 – Carte d'Oran (Google Map).

4.5.2 Donnée de la surveillance (données alphanumériques)

Les données de la surveillance de la maladie de la Tuberculoses utilisées dans notre étude, sont extraites d'un fichier Excel qui nous a été parvenu de la direction de la santé de la wilaya d'Oran. En voici un aperçu de ces données sur la figure 4.5.

Commune	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Type P	Type NP	Type IND	Nombre de ..	Class age[..	Class age[..	Class age[..	Class :
A.El Turck	3	2	2	5	6	5	1	12		1	2	6
Ain El Bia	1	2	2	3	7		1	8			1	5
Ain El kerma	1	2	1	2	5		1	6	1		2	1
Arzew	5	17	8	7	25	9	3	37		1	10	13
Benfreha	2	5	4	1	11	1		12		1	3	6
Bethioua	2	2	1	2	6	1	1	8		1	1	2

Figure 4.4 Données de la surveillance épidémiologique.

4.6 Les outils de développements utilisés

Notre choix était porté sur l'outil Microsoft SQL server 2012 pour implémenter notre base de données, l'outil Arc Gis pour le développement de notre système d'information géographique et Eclipse pour le développement de notre interface.

4.6.1 Microsoft SQL Server 2012

Microsoft SQL Server est une application utilisée pour créer des bases de données informatiques pour la famille des systèmes d'exploitation de Microsoft Windows. Il fournit un environnement utilisé pour produire des bases de données accessibles à partir des postes de travail, du web ou d'autres média tels qu'un assistant numérique personnel [Web 4.1].

SQL Server 2012, la plate-forme de gestion et d'analyse des données la plus complète.

4.6.2 ArcGis 10.1

Nous avons choisi ArcGIS 10 Desktop qui est une suite intégrée d'applications SIG professionnelles. Il existe trois niveaux de licence offerts pour ArcGIS : ArcView, ArcEditor et ArcInfo, Les autres composants d'ArcGIS sont ArcCatalog, ArcScene et ArcGlobe. ArcMap, ArcToolbox et ModelBuilder [web, 4.2].

4.6.3 Eclipse

Eclipse est un projet, décliné et organisé en un ensemble de sous-projets de développements logiciels, de la Fondation Eclipse, visant à développer un environnement de production de logiciels libres qui soit extensible, universel et polyvalent, en s'appuyant principalement sur Java.

4.7 Les étapes de création du logiciel

Nous décrivons dans cette section les trois étapes de création de notre projet de réalisation d'un système d'aide à la décision dédié à la surveillance épidémiologique.

4.7.1 Création d'entrepôt de données

Dans cette première étape, nous allons expliquer les phases de création et administration de notre base de données et des différentes tables (dimensions), ainsi que leurs champs dans le moindre détail grâce au SQL Server Management Studio, installé avec SQL Server R2.Il permet de réaliser, avec grande facilité, la plupart des taches communes aux bases de données, de la création de la base jusqu'à la modification de données. Quelques clics suffisent et nul besoin de taper la moindre requête.

4.7.1.1 Création d'entrepôt de données avec « SQL Server management studio »

Avant de pouvoir réaliser une quelconque opération, il nous faut tout d'abord accéder à SQL Server Management Studio. Maintenant commençons par la création de la base de données.

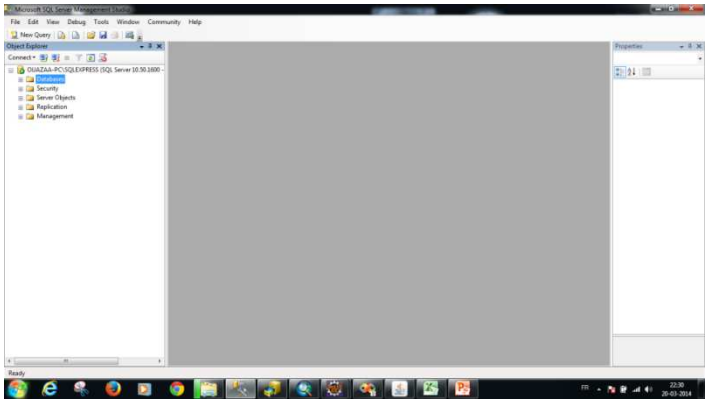


Figure 4.5 Fenêtre de SQL Server management studio

Dans la page d'accueil, un formulaire composé d'un champ de texte, d'une liste et d'un bouton nous permet de spécifier le nom de la base, que nous avons nommé **« BDD Tuberculose»,** ainsi que le type d'interclassement (si aucun n'est spécifié, le type "par défaut" sera utilisé).

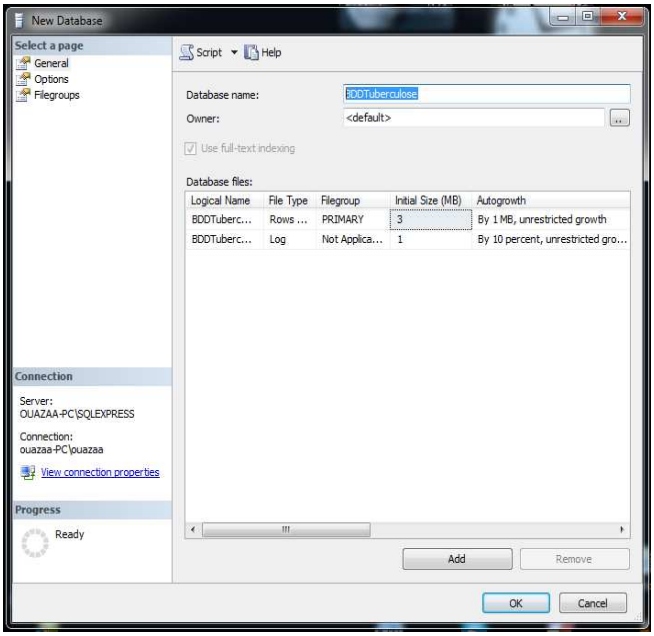


Figure 4.6 Création de la BDD Tuberculose

Après validation, on accède à une autre section ou nous pouvons créer les tables de notre base.

4.7.1.2 Les tables de notre projet

La création de tables est tout aussi simple. Il suffit d'aller dans la base de données **« BDD Tuberculose »** crée auparavant, cliquer sur le sous dossier Table, et choisir New Table. Une nouvelle fenêtre s'affiche dans laquelle il faudra saisir le nom de la table en premier et les champs suivis de leurs types :

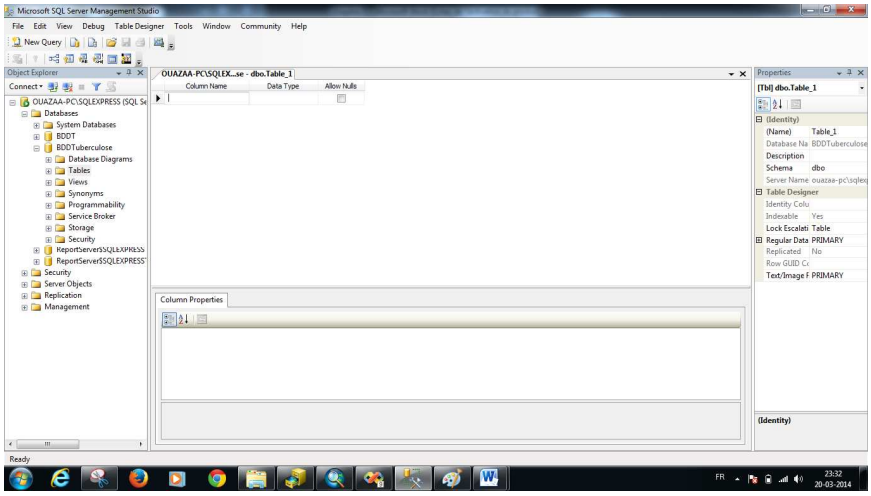


Figure 4.7 Création des tables

Comme exemple, nous allons créer la table « Age » .Elle contient quatre champs.

Une nouvelle page nous permettra d'attribuer à chacun des quatre champs : ID_Age, Age, G_Age, T_Age, ainsi que d'autres éléments, plus spécifiques qui devront être définis dans notre modèle de départ (Primary Key, Unique, Auto_increment, not null, index, etc...).

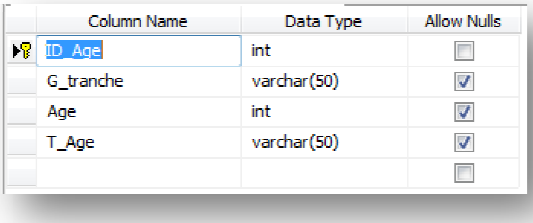


Figure 4.8 Création des champs de la table Age

Ceci fait, un clic sur "sauvegarder" et notre table est prête. Dans le cas où nous nous somme trompé, de type de données ou du nom d'un attribut par exemple, il nous est tout à fait possible de corriger cette erreur en réalisant des modifications sur la table grâce, au lien "design". Par le même procédé, nous avons construit les tables de notre base de données.

4.7.1.3 Insertion des enregistrements

Ainsi, nous avons construit toutes les tables de notre base de données. Il ne reste qu'à les remplir en insérant des enregistrements. Mais la plupart de nos tables se rempliront par l'importation des données à partir des feuilles Excel ou par SQL Server Management Studio.

Expliquons tout de même comment insérer des données avec le module SQL Server Management Studio. Cette insertion est d'une simplicité incroyable. Il suffit de sélectionner table concernée, cliquer sur le bouton droit dessue et choisir EDIT TOP 200 ROWS

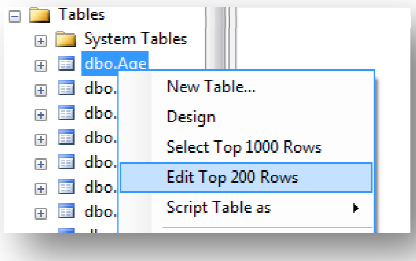


Figure 4.9 Insertion des enregistrements

Remplir les champs par les données comme le montre la figure suivante. Pour le reste, un simple formulaire, avec un champ pour chacun des attributs de la table, et l'enregistrement est terminé.

	ID_Age	G_tranche	Age	T_Age
*	NULL	NULL	NULL	NULL

Figure 4.10 Exemple d'un enregistrement avant insertion

Pour voir l'ensemble des enregistrements réalisés, sélectionner la table concernée, cliquer avec le bouton droit de la sourie dessue et choisir EDIT TOP 200 ROWS : un tableau présente alors tous les enregistrements contenus dans cette table. Il est même possible d'y réaliser des modifications en cliquant sur les champs.

4.7.1.4 Création de jointures entre les tables

Pour solliciter les données de plusieurs tables on doit créer des jointures sur « **SQL Server Visual Studio ».** Pour cela, on doit se connecter à la base de données, en suite cliquer avec le bouton droit de la sourie sur le dossier Database Diagramme et choisir Add New Datagrame. Importer les tables de notre base de données pour établir les jointures entre les tables de dimensions comme le montre la figure 4.11.

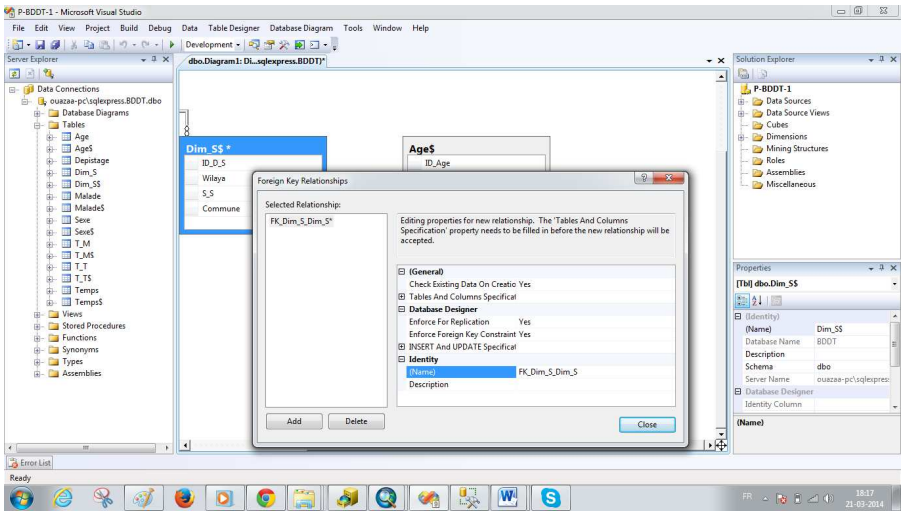


Figure4.11 Création de jointures entres les tables

Puis saisir les clés secondaire dans les champs Foreingn Key pour toutes les tables de dimensions avec la table de fait du milieu pour avoir le modèle en étoile de notre base de données. Ce dernier servira pour la structuration de notre cube de données. Vous trouvez en annexe les requêtes MDX permettant l'interrogation de notre cube de données.

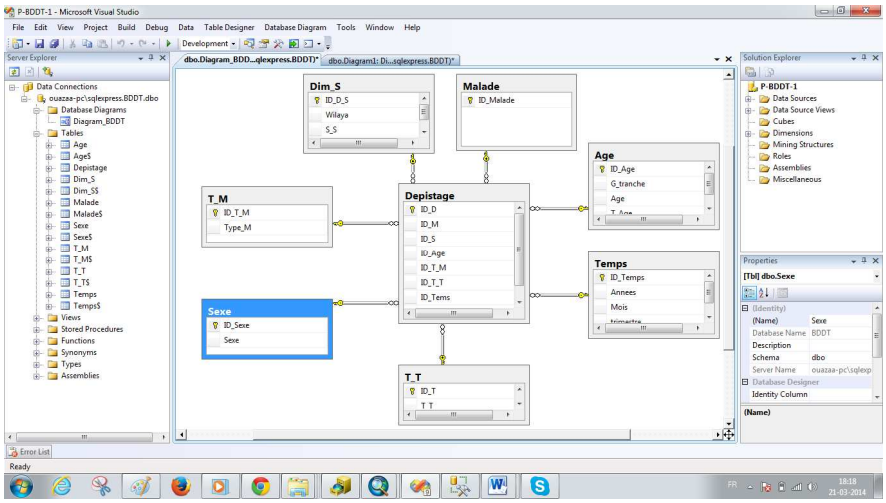


Figure 4.12 Modèle en étoile de notre application

4.7.1.5 Utilisation du Data Mining

L'outil Business Intelligence Development Studio de SQL server 2012 que nous avons utilisé offre la possibilité de rechercher des tendances intéressantes. Etant donné que les résultats des modèles d'exploration de données sont complexes et peuvent être difficiles à comprendre dans un format brut, l'examen visuel des données constitue souvent le moyen le plus simple pour comprendre les règles et les relations découvertes par les algorithmes au sein des données.

- **Méthode1 : Arbre de décision (Decision tree)**

L'algorithme MDT (Microsoft Decision Tree) prédit quelles localisations de la maladie de la tuberculose influencent sur la décision du type de la maladie.

La visionneuse d'arbre de décision fournit le réseau de dépendances de la figure qui affiche les attributs qui déterminent la capacité de prévision du modèle d'exploration de données. La visionneuse du réseau de dépendance renforce nos conclusions selon lesquelles le sexe de l'individu et la localisation de la maladie sont des facteurs importants pour prédire le type de la maladie.

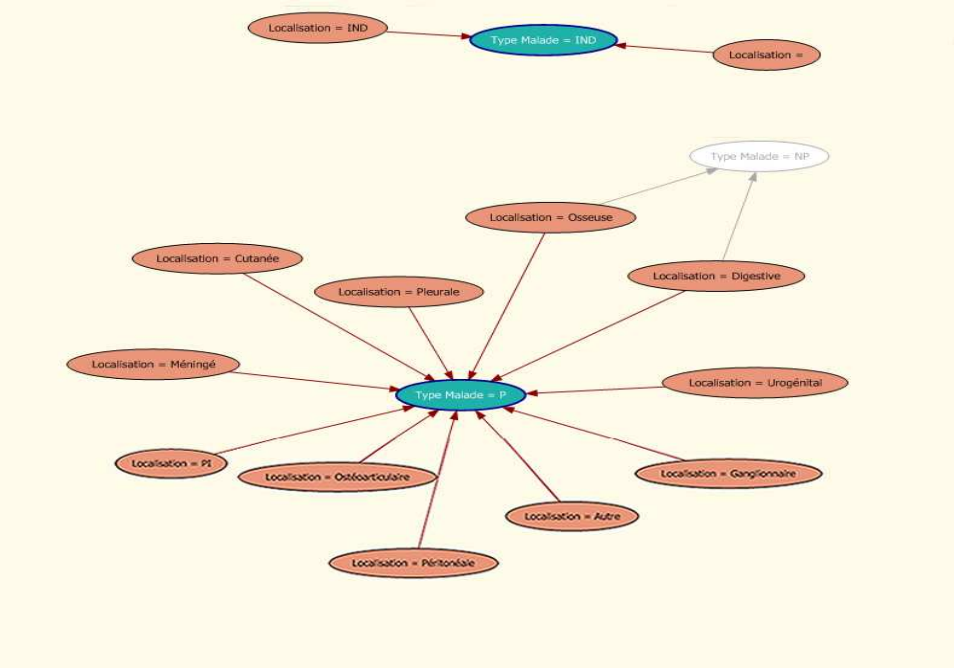


Figure4.13 Réseau de dépendances

- Les deux nœuds bleus représentent le type de maladie P et IND
- Le nœud en blanc représente le type de maladie NP
- Les nœuds orange représentent les différentes localisations de la tuberculose

La figure représente un exemple d'arbre de décision généré après l'exécution le l'algorithme MDT sur notre jeu de données (appliqué sur 2 variables : Type de maladie et localisation).

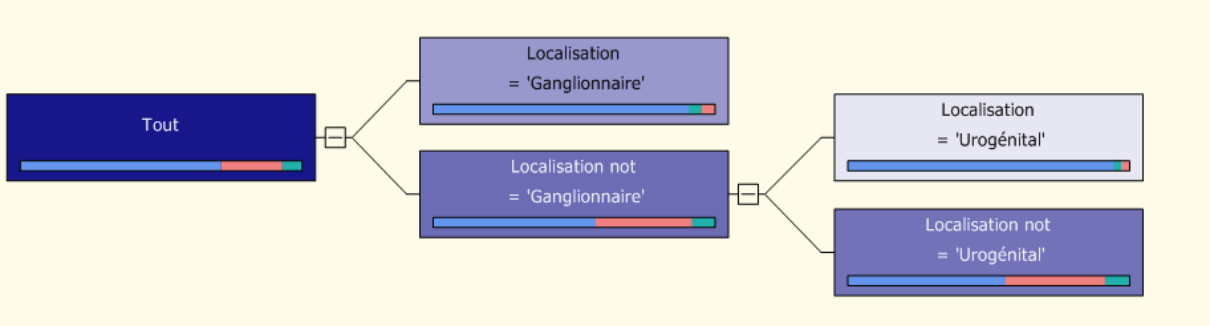


Figure 4.14 Arbre de décision généré

- **Méthode 2 : Réseau Bayesien Naïf (Naive bayes)**

L'algorithme Microsoft Naive Bayes affiche l'interaction entre la localisation de la maladie et le type de la maladie (P, NP ou IND). Le résultat de l'application de TM_Naive Bayes est donné sur la figure

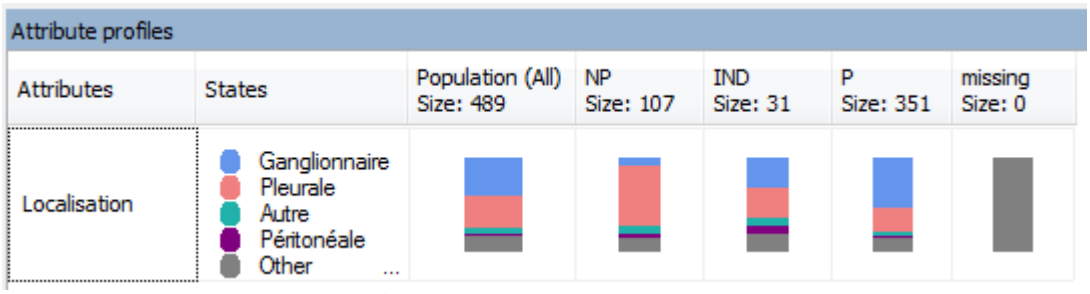


Figure4.15 Résultat de l'application de Naive Bayes

- **Méthode 3 : Classification (Clustering)**

L'algorithme MSC (Microsoft Sequence Clustering) regroupe des cas dans des clusters qui contiennent des caractéristiques semblables. Ces regroupements sont utiles pour l'exploration des données, l'identification d'anomalies dans les données et la création des prédictions.

L'onglet Diagramme de cluster affiche tous les clusters qui sont dans un modèle d'exploration de données. Les lignes entre les clusters représentent le lien logique et sont plus ou moins ombrées selon le degré de similitude entre les clusters. La couleur actuel de chaque cluster représente la fréquence de la variable et l'état dans le cluster. La figure représente un aperçu du résultat du clustering après l'exécution de l'algorithme TM_Clustering.

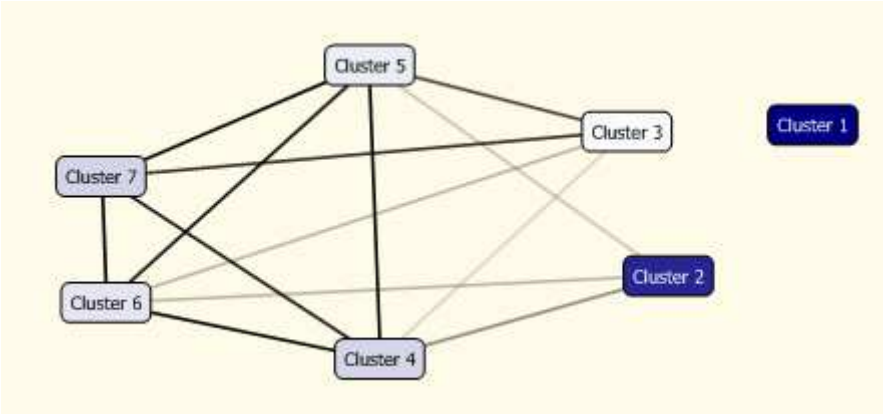


Figure4.16 Diagramme de cluster

L'onglet Profils du cluster présente une vue d'ensemble du modèle TM_Clustering qui contient une colonne pour chaque cluster. La distribution des attributs dans chaque cluster est indiquée sous la forme d'une barre de couleur comme le montre la figure

Figure 1.17 Profils des clusters générés

Nous avons constaté une forte connexion entre les deux clusters 6 et 7. Ceci se voit bien sur l'affichage des caractéristiques de ces deux clusters comme le montre la figure

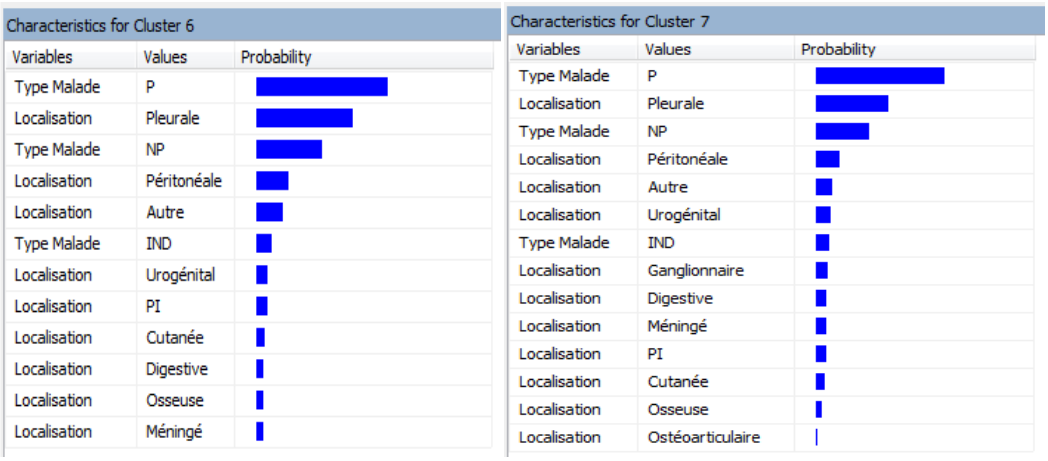


Figure 4.18 Similitudes entre le cluster 6 et le cluster 7

Le cluster 1 est un cluster orphelin qui n'a aucune similitude avec les autres clusters la figure montre ses caractéristiques

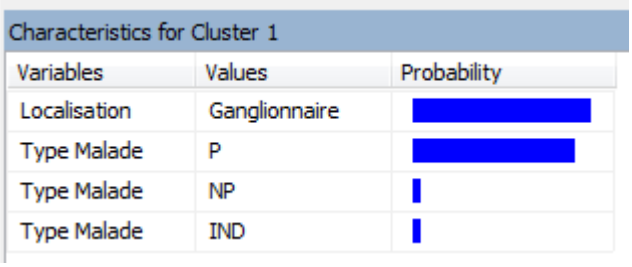


Figure4.19 Caractéristiques du cluster 1

On peut lire à partir de ce cluster que la probabilité que la maladie qui a la localisation « Ganglionnaire » est forte d'être de type Prouvé.

Méthode 4 : Les règles d'association

Nous voulons maintenant ajouter un nouvel attribut qui influence lui aussi sur le type de la maladie en plus de la localisation. Il s'agit du sexe de l'individu (Féminin ou Masculin).

Pour découvrir l'influence des deux variables : localisation de la maladie et sexe de l'individu, nous avons appliqué l'algorithme « Association Rules » qui a généré le graphe de dépendances de la figure. Les règles d'association générées sont listées en annexe.

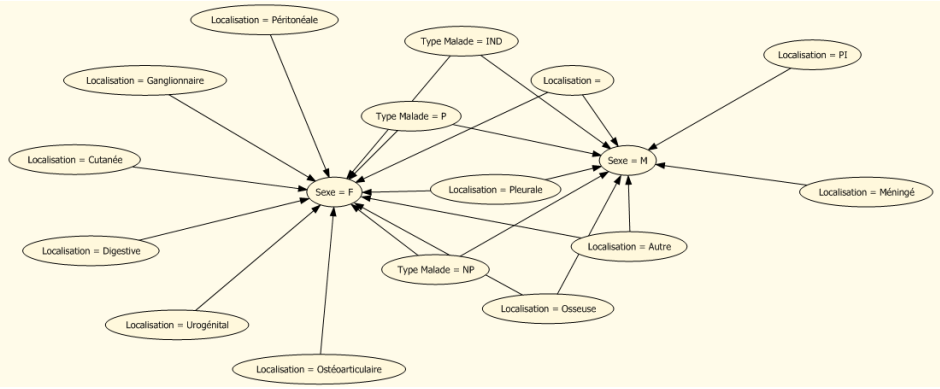


Figure 4.20 Graphe de dépendances pour l'algorithme Association Rules

4.7.2 Création de la carte pour notre projet

Pour pouvoir visualiser les données stockées sur notre base de données (SQL server) on aura besoin d'une carte numérique du site concerné par notre étude. Nous disposons d'une carte de format JPEG qu'il faudra numériser par la suite avec ARCMAP 10.1.

Pour cela, On doit se connecter à notre répertoire (dossier) dans le menu **catalog** et choisir **folder connections** puis ouvrir le répertoire pour appeler la carte d'Oran décrite précédemment. En fin sélectionner la carte 800px-DZ-3101-Oran.svg.png et la glisser dans la fenêtre d'ArcMap.

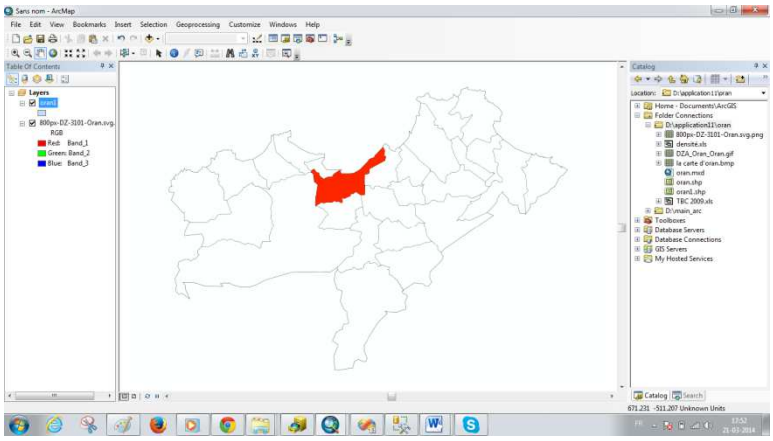


Figure 4.21 Sélection de la carte 800px-DZ-3101-Oran.svg.png sur Arc Gis

Par la suite on aura besoin de créer une nouvelle couche pour faire la numérisation de la carte, pour cela on sélectionne et on clique avec le bouton droit sur le répertoire et on choisit New puis ShapeFile. Une nouvelle fenêtre s'ouvrira dans laquelle il faudra saisir le nom de la couche du Design et choisir le type polygone par ce qu'on est en train de créer les communes et les secteurs socio sanitaires de la wilaya d'Oran. La figure suivante monte une numérisation complète de la carte

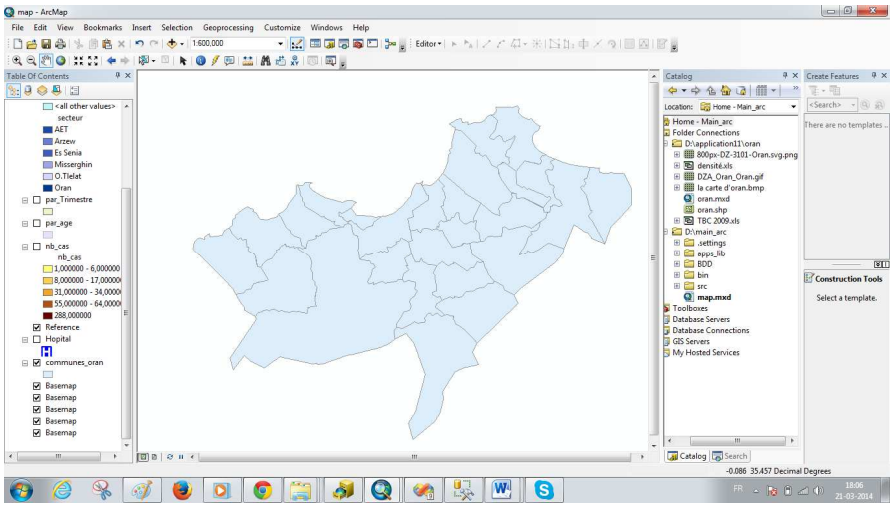


Figure4.22 Numérisation de la carte d'Oran sous Arc Gis

4.7.3 Création de l'interface de l'application

L'interface de notre application contient deux onglets : Le premier onglet **« insertion »** sert à insérer des données dans la base de données. Le deuxième onglet **« Map »** sert à afficher les résultats des requêtes sur la carte.

- **Onglet Insertion**

Contient les champs de saisie de l'âge ; le sexe ; la date de déclaration de la maladie, le type de la maladie, la localisation et la commune de résidence du malade.

Le bouton « valider » sert à enregistrer les données saisies dans notre base BDD Tuberculose crée sous SQL server et connectée avec éclipse.

Figure 4.23Onglet Insertion de l'application

- **Onglet Map**

Sert à afficher la carte de la wilaya d'Oran à partir de Arc Gis et effectuer les différentes analyses sur les donnés puis afficher les résultats sur la carte. La figure 4.24 montre les résultats d'analyse des taux d'incidences de la tuberculose dans toutes les communes de la wilaya d'Oran en fonction du type de la maladie sous forme de secteurs avec une légende interactive à gauche de l'interface.

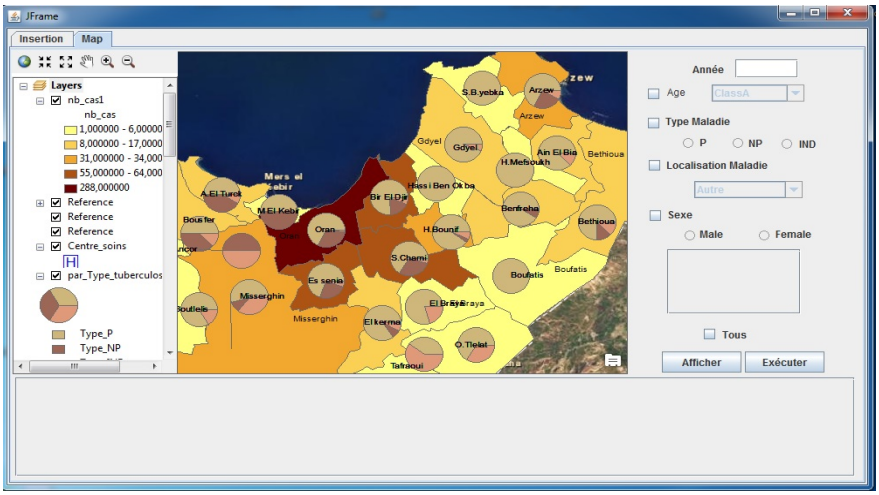


Figure 4.24 Onglet Map de l'application

Pou ce faire, il faudra établir une connexion entre la carte d'Oran importée et numérisée sur ArcGis et l'interface développée sur éclipse. Nous donnons ici une partie du code source nous permettant de connecter éclipse à ArcGis.

```
private MapBean getMapBean() {
    if (mapBean == null) {
        try {
            //this.mapBean.clearLayers();
            mapBean = new MapBean();
            mapBean.setBounds(new Rectangle(190, 2, 516, 398));
            mapBean.setDocumentFilename("C:\\Users\\Abderrahmane\\Desktop\\sig_app_tp Boumehdi & bouzid\\TP11.mxd");
            mapBean.addLayerFromFile("C:\\Users\\Abderrahmane\\Desktop\\sig_app_tp Boumehdi & bouzid\\nb_cas1.lyr", 0);
            mapBean.refresh(esriViewDrawPhase.esriViewGeography, null, null);

        } catch (java.lang.Throwable e) {
            // TODO: Something
        }
    }
    return mapBean;
}
```

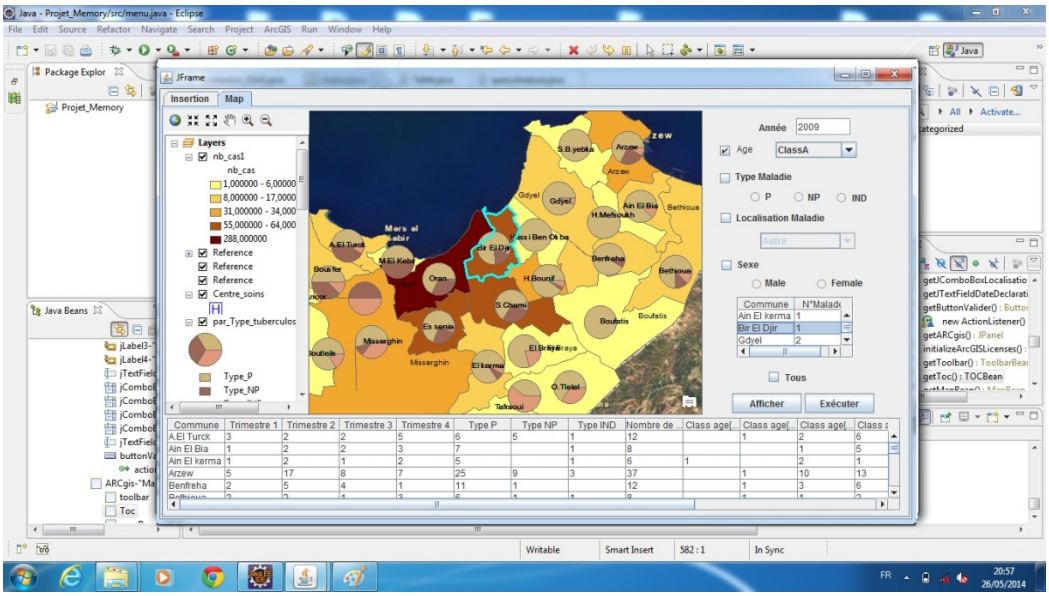


Figure 2.25Résultat finale de la connexion de Arc Gis avec l'interface

4.8 Conclusion PAIRESPICTIFE

Ce chapitre était consacré essentiellement au développement d'une application java crée sous éclipse et basée sur le concept SOLAP fondé sur l'intégration de l'outil OLAP réalisé sous SQL srever 2012 et l'outil SIG développé sous Arc Gis.

Précisément, dans ce chapitre, Nous avons découvert les capacités de SQL server 2012 en terme de création d'entrepôt de données, de déploiement de cube et exploration des données par les outils de data mining.

Par faute du nombre de pages limité, on ne peut aller au détail près de cette implémentation sinon l'application a atteint son objectif majeur qui n'est autre que l'expérimentation des techniques de fouille de données sur notre cube de données dédié à la surveillance épidémiologique. Ainsi nous pouvons conclure que les deux parties SOLAP et datamining sont entièrement opérationnelles.

Il manque cependant, une réelle interprétation des résultats que nous avons obtenus par un expert du domaine de la surveillance épidémiologique afin de valider les règles d'association générés ainsi que les modèles suggérées.

Annexe

Règle d'association

Nous donnons dans cette annexe, les règles générées de l'application de l'algorithme **« Association Rules »** sur notre jeu de données en spécifiant l'importance et la probabilité pour chaque règle d'association.

Probability	importance	rule
1	0,160866668	Localisation = Digestive, Type Malade = P -> Sexe = F
1	0,179320229	Localisation = Digestive -> Sexe = F
1	0,229027334	Localisation = PI, Type Malade = IND -> Sexe = M
1	0,225151385	Localisation = Péritonéale, Type Malade = P -> Sexe = F
1	0,160866668	Localisation = Ostéoarticulaire -> Sexe = F
1	0,160866668	Localisation = Ostéoarticulaire, Type Malade = P -> Sexe = F
1	0,080248309	Localisation = Osseuse, Type Malade = IND -> Sexe = F
1	0,080248309	Localisation = Digestive, Type Malade = NP -> Sexe = F
1	0,080248309	Localisation = PI, Type Malade = NP -> Sexe = F
1	0,080248309	Localisation = Urogénital, Type Malade = NP -> Sexe = F
1	0,080248309	Localisation = Cutanée, Type Malade = NP -> Sexe = F
1	0,176757866	Localisation = Méningé, Type Malade = NP -> Sexe = M
0,8	0,111478909	Localisation = Cutanée -> Sexe = F
0,783	0,145296756	Localisation = Péritonéale -> Sexe = F
0,75	0,17810655	Localisation = Autre, Type Malade = IND -> Sexe = M
0,75	0,08079273	Localisation = Péritonéale, Type Malade = IND -> Sexe = F
0,75	0,08079273	Localisation = Cutanée, Type Malade = P -> Sexe = F
0,714	0,179476268	Localisation = Méningé -> Sexe = M
0,667	0,150311606	Localisation = Méningé, Type Malade = P -> Sexe = M
0,667	0,150311606	Localisation = Osseuse, Type Malade = P -> Sexe = M
0,659	0,126272085	Localisation = Ganglionnaire, Type Malade = P -> Sexe = F
0,639	0,110229333	Localisation = Ganglionnaire -> Sexe = F
0,636	0,145199693	Type Malade = IND, Localisation = Pleurale -> Sexe = M
0,632	0,049771617	Localisation = Autre, Type Malade = P -> Sexe = F
0,625	0,034687027	Localisation = Autre, Type Malade = NP -> Sexe = F
0,625	0,132818685	Type Malade = IND, Localisation = Ganglionnaire -> Sexe = M
0,625	0,132818685	Localisation = PI -> Sexe = M
0,615	0,035065703	Localisation = Urogénital -> Sexe = F
0,6	0,11026523	Localisation = PI, Type Malade = P -> Sexe = M
0,594	0,13156348	Type Malade = IND -> Sexe = M
0,591	0,100303028	Type Malade = P -> Sexe = F
0,585	0,151424375	Localisation = Pleurale, Type Malade = P -> Sexe = M
0,583	0,013143974	Localisation = Urogénital, Type Malade = P -> Sexe = F
0,581	0,01721863	Localisation = Autre -> Sexe = F
0,58	0,194588725	Localisation = Pleurale -> Sexe = M
0,565	0,122230707	Type Malade = NP, Localisation = Pleurale -> Sexe = M

0,556	-0,00760762	Localisation = Péritonéale, Type Malade = NP -> Sexe = F
0,538	0,081234652	Localisation = Osseuse -> Sexe = M
0,514	0,081961838	Type Malade = NP -> Sexe = M
0,5	-0,045936914	Localisation = Osseuse, Type Malade = NP -> Sexe = F
0,5	-0,045579465	Localisation = -> Sexe = F
0,5	0,051376906	Localisation = Osseuse, Type Malade = NP -> Sexe = M
0,5	0,050930092	Localisation = -> Sexe = M
0,5	-0,045579465	Localisation = , Type Malade = IND -> Sexe = F
0,5	0,050930092	Localisation = , Type Malade = IND -> Sexe = M
0,5	0,05206204	Type Malade = NP, Localisation = Ganglionnaire -> Sexe = M
0,5	-0,046483736	Type Malade = NP, Localisation = Ganglionnaire -> Sexe = F
0,486	-0,072448358	Type Malade = NP -> Sexe = F
0,462	-0,077356477	Localisation = Osseuse -> Sexe = F
0,444	0,009308622	Localisation = Péritonéale, Type Malade = NP -> Sexe = M
0,435	-0,119429925	Type Malade = NP, Localisation = Pleurale -> Sexe = F
0,42	-0,175446942	Localisation = Pleurale -> Sexe = F
0,419	-0,02232708	Localisation = Autre -> Sexe = M
0,417	-0,016928492	Localisation = Urogénital, Type Malade = P -> Sexe = M
0,415	-0,149871977	Localisation = Pleurale, Type Malade = P -> Sexe = F
0,409	-0,112966944	Type Malade = P -> Sexe = M
0,406	-0,137785167	Type Malade = IND -> Sexe = F
0,4	-0,113598811	Localisation = PI, Type Malade = P -> Sexe = F

Les requêtes MDX

MDX est l'acronyme de Multi Dimensional eXpression. C'est un langage de requêtes OLAP pour les bases de données multidimensionnelles. Ce langage est fait pour naviguer dans les bases multidimensionnelles et pour définir des requêtes sur tous leurs objets (dimensions, hiérarchies, niveaux, membres et cellules) afin d'obtenir (simplement) une représentation sous forme de tableaux croisés. MDX ressemble à SQL par ses mots clé SELECT, FROM, WHERE, mais : SQL construit des vues relationnelles et MDX construits des vues multidimensionnelles des données.

Comparaison entre le langage MDX et le langage SQL

Ce tableau illustre une analogie entre termes multidimensionnels (MDX) et relationnels (SQL) :

Multidimensionnel (MDX)	Relationnel (SQL)
Cube	**Table**
Niveau (Level)	**Colonne** (chaine de caractère ou valeur numérique)
Dimension	**plusieurs colonnes liées** ou une table de dimension
Mesure (Measure)	**Colonne** (discrète ou numérique)
Membre de dimension (Dimension member)	**Valeur** dans une colonne et une ligne particulière de la table

Exemples d'une requête MDX pour afficher les mesures de notre projet

```
select [Measures].[Tuberculose Count] on Columns,
non empty[Espace].[Nom Commune].[Nom Commune]  on rows
 from  [Memory]
 where ({[Age].[Id Age].&[1]:[Age].[Id Age].&[4]},[Maladie].[Type Maladie].&[P]);
```

Des requêtes MDX sur java de notre projet

```
cellSetA1 = statement.executeOlapQuery("select [Measures].[Tuberculose Count] on columns,[Espace].[Nom
Commune].[Nom Commune] on rows from [Memory] where ({[Age].[Id Age].&[0]:[Age].[Id
Age].&[4]},[Temps].[Annee].&["+annee+"]);");
            cellSetA2 = statement.executeOlapQuery("select [Measures].[Tuberculose Count] on
columns,[Espace].[Nom Commune].[Nom Commune] on rows from [Memory] where ({[Age].[Id Age].&[5]:[Age].[Id
Age].&[14]},[Temps].[Annee].&["+annee+"]);");
            cellSetA3 = statement.executeOlapQuery("select [Measures].[Tuberculose Count] on
columns,[Espace].[Nom Commune].[Nom Commune] on rows from [Memory] where ({[Age].[Id Age].&[15]:[Age].[Id
Age].&[24]},[Temps].[Annee].&["+annee+"]);");
            cellSetA4 = statement.executeOlapQuery("select [Measures].[Tuberculose Count] on
columns,[Espace].[Nom Commune].[Nom Commune] on rows from [Memory] where ({[Age].[Id Age].&[25]:[Age].[Id
Age].&[34]},[Temps].[Annee].&["+annee+"]);");
            cellSetA5 = statement.executeOlapQuery("select [Measures].[Tuberculose Count] on
columns,[Espace].[Nom Commune].[Nom Commune] on rows from [Memory] where ({[Age].[Id Age].&[35]:[Age].[Id
Age].&[44]},[Temps].[Annee].&["+annee+"]);");
            cellSetA6 = statement.executeOlapQuery("select [Measures].[Tuberculose Count] on
columns,[Espace].[Nom Commune].[Nom Commune] on rows from [Memory] where ({[Age].[Id Age].&[45]:[Age].[Id
Age].&[54]},[Temps].[Annee].&["+annee+"]);");
            cellSetA7 = statement.executeOlapQuery("select [Measures].[Tuberculose Count] on
columns,[Espace].[Nom Commune].[Nom Commune] on rows from [Memory] where ({[Age].[Id Age].&[55]:[Age].[Id
Age].&[64]},[Temps].[Annee].&["+annee+"]);");
            cellSetA8 = statement.executeOlapQuery("select [Measures].[Tuberculose Count] on
columns,[Espace].[Nom Commune].[Nom Commune] on rows from [Memory] where ({[Age].[Id Age].&[65]:[Age].[Id
Age].&[72]},[Temps].[Annee].&["+annee+"]);");
```