Titre: Title:	Perspectives d'intégration entre le data mining et les systèmes d'information géographique (S.I.G.) : étude de cas en analyse du marché des meubles aux États-Unis
Auteur: Author:	Thi Thu Hoa Le
Date:	2009
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Le, T. T. H. (2009). Perspectives d'intégration entre le data mining et les systèmes d'information géographique (S.I.G.) : étude de cas en analyse du marché des meubles aux États-Unis [Master's thesis, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/8474/

UNIVERSITÉ DE MONTRÉAL

PERSPECTIVES D'INTÉGRATION ENTRE LE DATA MINING ET LES SYSTÈMES D'INFORMATION GÉOGRAPHIQUE (S.I.G): ÉTUDE DE CAS EN ANALYSE DU MARCHÉ DES MEUBLES AUX ÉTATS-UNIS

THI THU HOA LE
DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE INDUSTRIEL)
AOÛT 2009

Your file Votre référence
ISBN: 978-0-494-57255-9
Our file Notre référence
ISBN: 978-0-494-57255-9

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé:

PERSPECTIVES D'INTÉGRATION ENTRE LE DATA MINING ET LES SYSTÈMES D'INFORMATION GÉOGRAPHIQUE (S.I.G): ÉTUDE DE CAS EN ANALYSE DU MARCHÉ DES MEUBLES AUX ÉTATS-UNIS

présenté par: LE Thi Thu Hoa

en vue de l'obtention du diplôme de: Maîtrise ès sciences appliquées

a été dûment accepté par le jury d'examen constitué de:

M. TRÉPANIER Martin, ing., Ph.D., président

M. AGARD Bruno, Doct., membre et directeur de recherche

M. FRAYRET Jean-Marc, Ph.D., membre

REMERCIEMENTS

Je tiens tout d'abord à remercier Monsieur Bruno Agard, mon directeur de recherche, qui a su m'accompagner, me guider et m'encourager dès mon arrivée à l'École Polytechnique de Montréal.

Mes remerciements s'adressent également à Messieurs Stéphane Deveault et Stéphane Deschesnes de l'entreprise Canadel pour leurs conseils et leurs contributions dans le traitement et l'interprétation des données empiriques de cette étude.

Mes remerciements vont aussi à Monsieur Trépanier Martin pour sa gentillesse de me permettre d'utiliser le serveur.

Un grand merci à ma famille au Viet Nam et mes amis au Canada qui m'ont encouragée et ont toujours été avec moi dans les périodes difficiles de ma vie à l'étranger.

Je remercie enfin le programme de bourses canadien pour la francophonie (PCBF) non seulement pour leur support financier mais ainsi pour leur soutien et leur aide durant mes études au Canada.

RÉSUMÉ

Le data mining et les SIG (Systèmes d'Information Géographique) existent tous deux depuis plusieurs années. Chacun a ses propres méthodes, techniques, approches d'analyse et de visualisation des données. Les SIG et le data mining sont maintenant utilisés largement pour appuyer la prise de décision dans des secteurs aussi variés que les affaires, l'environnement, la santé, la logistique ou encore la sécurité.

Au cours de la dernière décennie, plusieurs entreprises de SIG commencent à réaliser l'importance d'intégrer les techniques de data mining dans leurs produits. Le data mining peut contribuer à trouver des corrélations significatives, des règles et des tendances cachées dans les grandes bases de données de SIG. D'un autre côté, les SIG par leurs capacités de visualisation, montrent aussi des potentiels pour la pratique du data mining.

L'intégration actuelle entre data mining et les SIG peuvent être considérées selon les deux approches ci-dessous :

- Les techniques du data mining s'appliquent aux données spatiales. Cette approche est aussi connue sous le nom de data mining spatial.
- Les SIG sont utilisés comme outil de vérification, d'analyse et de visualisation des résultats du data mining.

Par une revue de littérature, l'objectif principal de ce mémoire est de présenter le potentiel d'intégration entre le data mining et les SIG. En collaboration avec une entreprise de meuble canadienne dans un cas d'étude, nous montrerons comment l'intégration entre le data mining et les SIG peut améliorer mutuellement leurs résultats. Nous développerons un système basé web qui intègre les techniques de segmentation du data mining et les services d'un SIG (Google Map) afin d'analyser

le marché des meubles aux États-Unis. Les résultats obtenus nous confirment les avantages de cette intégration.

ABSTRACT

The theories and applications in the field of data mining and Geographic Information Systems (GIS) have been developed several years. They both have been shown in many studies in methodologies and techniques in data analysis and visualization. GIS and data mining are now widely used to support decision making in various sectors as in business, environment, public health, logistics and security.
During the last decade, several GIS companies realize the importance and start to integrate data mining techniques in their products. Data mining contributes facilities to find significant correlations, rules and trends hidden in large databases like GIS. On the other hand, the GIS with its capabilities of its display and virtualization, is the potential implementation for the data mining methodologies.
The current researches of the integration between data mining and GIS can be considered in 2 approaches:

- The implementation of data mining techniques to spatial data. This approach is also known as spatial data mining.
- GIS is used as a tool to verify, visualize and analyze the results of data mining.

The potential of integration data mining and GIS will be studied more in the literature review section. In collaboration with a Canadian furniture company in a case study, we present an integration of data mining and GIS and its benefits to improve the outcome, regarding to a separate techniques. We develop a web based system that integrates the segmentation techniques of data mining and Googleő Map services to analyze the US furniture market. The results confirm the benefits of this integration.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

SIG	Systèmes d'information géographique
DM	Data Mining
CRISP – DM	Cross-Industry Standard Process for Data Mining
AFC	Analyse factorielle des correspondances
ACM	Analyse des correspondances multiples
DIME – GBF	Dual Independent Map Encoding- Geographic Database Files
GPS	Global Positioning System
TIGER	Topologically Integrated Encoding and Referencing
WWF	World Wildlife Fund for Nature
SGBD	Système de Gestion de Bases de Données
SGBDR	Système de Gestion de Bases de Données Relationnel
SGBDOO	Système de Gestion de Bases de Données Orienté-Objet
SGBDRO	Système de Gestion de Bases de Données Relationnel-Objet
COD	Clustering with Obstructed Distance
CLARANS	Clustering Large Applications based on RANdomized Search
COE – CLARANS	Clustering with Obstacle Distance based on CLARANS
GDBSCAN	Generalized Density-Based Spatial Clustering of Applications with Noise
SCART	Spatial Classification and Regression Trees
OLAP	Online Analyticial Processing
API	Application Programming Interface
BSS	Between Sum of Square
ISODATA	Iterative Self-Organizing Data Analysis Techniques
ALENAl	Accord de Libre-Échange Nord-Américain

CHAPITRE 1

INTRODUCTION

1.0.1 Contexte

Le data mining et les SIG sont des domaines relativement récents, qui sont apparus vers le début des années 1960. Néanmoins, ils se sont développés indépendamment depuis plusieurs années. Chacun ayant des méthodes, des techniques et des approches spécifiques.

Les SIG sont été à l'origine développés pour la gestion de l'environnement et de la terre. Au cours de la dernière décennie, la portée, la couverture et le volume de la géographie numérique se développent rapidement grâce au progrès en sciences géomatiques. Les domaines d'applications des SIG ne se limitent plus à la géographie. Les SIG ont des capacités de stockage, gestion, manipulation et visualisation des données spatiales. Ils sont donc actuellement utilisés comme outil de planification et d'aide à la décision dans beaucoup de domaines (Longley *et al.*, 2005). Ils couvrent dors et déjà des champs relatifs à l'agriculture, l'environnement, le géomarketing, la santé, la pollution, le transport, etc. Pourtant, avec l'augmentation rapide du volume des données géographiques et la complexité de leurs relations, les méthodes traditionnelles d'analyse spatiale ne peuvent pas aisément découvrir des modèles, ni des règles ou de nouvelles connaissances cachées dans les bases de données spatiale.

D'un autre côté, le data mining est un processus qui réunit des théories et techniques d'analyse, d'extraction et de représentation pour découvrir des connaissances dans les grandes bases de données. De nos jours, tous les domaines produisent des

quantités énormes de données. Le data mining est donc la clé de succès pour toutes les entreprises. Le magazine en ligne ZDNET News a considéré le data mining comme "un des développements technologies les plus révolutionnaires des dix prochaines années" (Larose, 2005). Cependant, le data mining a aussi rencontré beaucoup de défis au cours de la dernière décennie. Un de ses défis est la demande des nouvelles techniques de visualisation.

À la fin du $20^{è}me$ siècle, le potentiel de l'intégration entre le data mining et les SIG a commencé à attirer l'attention des chercheurs. Dans cette intégration, les SIG peuvent bénéficier de techniques efficaces du data mining pour la manipulation, l'exploration et l'analyse des données spatiales. Le data mining offre également de nombreux outils de modélisation qui ne sont pas dans la fonctionnalité d'analyse des SIG tels que les arbres de décision et les règles d'association.

Par ailleurs, les SIG peuvent s'intégrer aux nombreuses étapes d'un processus de data mining en particulier lors de la phase de prétraitement, d'évaluation et de déploiement. Avec leurs capacités d'analyse, d'interprétation et de visualisation, les SIG peuvent améliorer les résultats finaux du data mining.

Pourtant, les recherches de l'intégration entre le data mining et les SIG sont encore récentes. Les premiers travaux sur le data mining spatial étaient proposés par (Koperski et Han, 1995) en 1995. Depuis, de nombreuses recherches ont concerné ce domaine et des implémentations commencent à apparaître dans des produits. Les intégrations actuelles du data mining et du SIG peuvent être considérées selon deux approches ci-dessous :

- Approche 1: les techniques du data mining sont appliquées aux données spatiales. Cette approche est aussi connue sous le nom de data mining spatial.
- Approche 2: les SIG sont utilisés comme outil d'analyse, de vérification et de visualisation des résultats du data mining.

1.0.2 Objectif

Le data mining et les SIG sont des sujets très larges. Par conséquent, nous ne pouvons pas aborder ces deux sujets de façon exhaustive dans le cadre de ce mémoire de maîtrise. A travers une analyse de la littérature, notre premier objectif est de montrer le potentiel et les techniques d'intégration entre le data mining et les SIG. Basé sur une étude de cas, nous démontrerons comment l'intégration des techniques de segmentation du data mining au SIG peut apporter des avantages dans le domaine de l'analyse du marché.

1.0.3 Organisation du mémoire

Ce mémoire est structuré en cinq chapitres.
Le chapitre 1 présente le contexte actuel et l'intérêt de l'intégration entre le data mining et les SIG.

Le chapitre 2 a pour objectif de se familiariser avec les concepts de base du data mining et du SIG et de faire ressortir le potentiel de l'intégration de ces deux domaines. La première partie du chapitre résumera les étapes d'un processus du data mining, les techniques de prétraitement des données et ses défis actuels. La seconde partie synthétisera l'évolution des SIG, ses composants, ses fonctionnalités et aussi ses défis face au progrès de la technologie.

Les deux approches de l'intégration entre le data mining et les SIG seront présentées dans le chapitre 3. Nous discuterons des caractéristiques des données spatiales, des méthodes du data mining appliquées aux données spatiales et de leur applications dans la première approche. Dans la deuxième approche, nous discuterons des techniques de géo-visualisation au data mining etde leur applications empiriques.

Ensuite, nous présenterons deux systèmes principaux d'intégration entre le data mining et les SIG. Finalement, quelques remarques sur l'état de recherche actuelle de ces deux approches viendront clore ce chapitre.

Le chapitre 4 présentera notre étude de cas sur l'intégration entre les techniques de segmentation du data mining et les SIG en analyse du marché des meubles américain. Ce chapitre contient quatre parties principales. Dans l'introduction, nous décrirons d'abord le contexte actuel du marché des meubles américain et de l'entreprise. Puis, nous proposerons une approche qui vise à déterminer les États potentiels et la concurrence sur le marché américain à l'aide de techniques de segmentation du marché et de SIG. Ensuite, nous ferons une revue de littérature de la segmentation du marché. Dans la partie suivante, nous expliquerons en détail les 3 étapes de la méthodologie proposée: l'étape du data mining, l'étape du SIG et l'étape de l'intégration. Les résultats des 3 étapes seront exposés dans la troisième partie. Nous finirons ce chapitre avec quelques conclusions et perspectives pour cette étude de cas.

Dans le dernier chapitre, nous conclurons en résumant le potentiel, les défis de cette intégration et exposerons quelques perspectives de recherches.

CHAPITRE 2

CONCEPTS DE BASE

Ce chapitre concerne les concepts de base des deux sujets abordés: le data mining et les systèmes d'information géographique (SIG). Dans la partie consacrée aux concepts de base du data mining, nous examinerons d'abord les six étapes du processus de data mining. Puis, nous porterons notre attention sur les techniques de prétraitement des données. Les tendances futures du data mining seront discutées par la suite. La deuxième partie sera consacrée aux concepts de base des SIG. Nous aborderons en premier lieu les cinq composants et les fonctionnalités principales d'un SIG. Nous soulignerons ensuite quelques tendances futures des SIG.

2.1 Concepts de base du data mining

Depuis des années 1980, la quantité de données augmente de manière exponentielle grâce aux progrès de l'informatique et de la capacité de stockage. Le data mining est né du besoin de l'exploitation des connaissances cachées dans les grandes bases de données. Lorsqu'on traduit littéralement le terme data mining en français, on obtient "fouille des données". Pourtant, en pratique on fait souvent référence au terme "d'extraction de connaissance". (Fayyad *et al.*, 1996a) définit le data mining comme : "un processus non-trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données". C'est un processus itératif qui met en œuvre un ensemble de techniques d'analyse des données, de statistique, d'intelligence artificielle et d'interface homme-machine pour découvrir des connaissances dans les données. Aujourd'hui, le data mining est largement

appliqué dans plusieurs secteurs, surtout dans les affaires. Le sondage effectué sur le portail web www.kdnuggets en juillet 2007 montrait que parmi les secteurs utilisant le data mining, celui des affaires représentait 80%.

2.1.1 Le processus du data mining

Le processus du data mining le plus utilisé empiriquement est proposé en 1996 par CRISP-DM (le Cross-Industry Standard Process for Data Mining), voir (Larose, 2005). Le processus CRISP-DM comprend six étapes itératives et adaptatives comme illustré dans la figure 2.1. La séquence de l'étape suivante dépend le plus souvent des résultats de l'étape précédente.

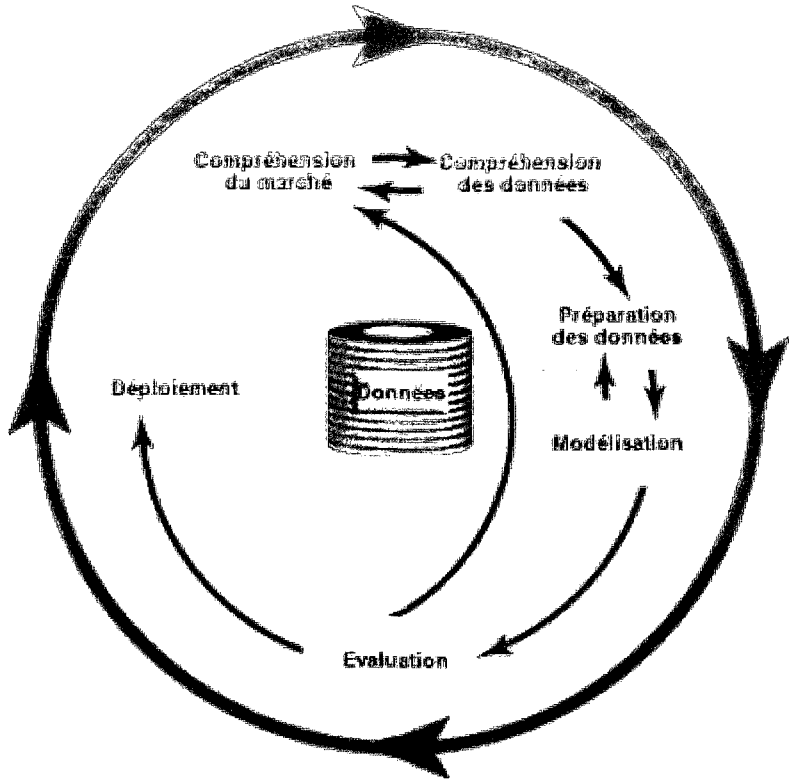


Figure 2.1 Processus CRISP-DM du data mining (Larose, 2005)

2.1.1.1 Compréhension de la problématique

La réalisation d'un projet de data mining est un travail de collaboration. Elle demande la participation d'experts du domaine, d'experts en base de données et d'experts du data mining. A la première étape, les experts doivent se réunir pour identifier le problème et cerner les objectifs. Le problème est ensuite décomposé en sous-problèmes plus faciles à résoudre. A la fin de cette étape, le problème devrait être transformé en une ou plusieurs tâches du data mining (prédiction, classification, etc.).

2.1.1.2 Compréhension des données

La plupart des projets de data mining ont besoin de données de plusieurs sources. Le but de cette étape est de disposer des données nécessaires pour permettre d'extraire de l'information pertinente et fiable. Elle est entre autre composée des phases ci-dessous:

- Identifier les sources de données disponibles pour le projet.
- Recueillir les données.
- Décrire et analyser les données collectées pour découvrir les perspectives initiales.
- Évaluer le niveau de qualité des données.
- Déterminer et sélectionner les données cibles selon les objectifs.

2.1.1.3 Préparation des données

La qualité du résultat du data mining est basée sur la qualité des données à exploiter. Les données collectées sont souvent incomplètes et bruitées. Elles doivent subir une préparation dans le but d'être traitées par le data mining. (Pyle, 1999) a estimé que cette étape compte pour environ 60% de temps et de l'effort dans un projet de data mining. Les opérations ci-dessous sont nécessaires pour la préparation des données:

- Sélectionner les données : choisir les données qui vont être utilisées pour chaque objectif.
- Réduire l'espace de recherche en choisissant les cas et les variables appropriés à l'analyse. Le choix des variables devrait être guidé par l'expert du domaine.
- Intégrer les données de sources variées dans une base de données identique si nécessaire.
- Nettoyer les données et compléter les données manquantes.
- Transformer les données dans le format de données pertinent pour les algorithmes de data mining.
- Dans certains cas, il faut préparer un échantillon test et un échantillon vérifié.

Dans la section 2.1.2, nous expliquerons plus en détail les techniques de prétraitement des données qui vont être utilisées dans notre étude de cas.

2.1.1.4 Modélisation

Cette étape constitue le cœur du processus de data mining. La modélisation consiste à appliquer les algorithmes de data mining aux données afin de trouver des patrons intéressants dans la base de données.

- Sélectionner des algorithmes de data mining appropriés afin d'atteindre les objectifs du projet. Il faut noter que souvent, plusieurs algorithmes peuvent être utilisés pour le même problème de data mining.
- Déterminer les paramètres des algorithmes pour optimiser les résultats. Par exemple, dans le cas des réseaux de neurones, le nombre de couches et le nombre de neurones dans chaque couche affectent fortement le résultat.
- Dans certain cas, il faut revenir à l'étape de préparation des données pour transformer à nouveau des données aux besoins spécifique d'un algorithme de data mining particulier.

(Tufféry, 2005) et (Han et Kamber, 2006) divisent les techniques de data mining en deux catégories: les techniques descriptives et les techniques prédictives.
Les techniques descriptives, qui n'ont pas de variables cibles, ont pour but de décrire les patrons, les relations existantes dans les données. Par opposition, les techniques prédictives ont pour but d'établir une fonction entre les entrées et les sorties. Le tableau 2.1, adapté de (Tufféry, 2005), représente les algorithmes communs à ces deux catégories.

Tableau 2.1 Classification des techniques du data mining

Classifications des techniques	
Catégorie	Algorithmes
Techniques descriptives	Analyse factorielle des correspondances AFC
	Analyse des correspondances multiples ACM
	Méthodes de partitionnement: k-moyenne, k-medoids, etc.
	Méthodes de classification hiérarchiques
	Méthodes de classification par réseaux de Kohonen
	Recherche des règles d'associations
	Recherche de séquences similaires
Techniques prédictives	Arbres de décision
	Réseaux de neurones
	Régression linéaire
	Modèle log-linéaire
	K-plus proches voisins

2.1.1.5 Évaluation

Une fois le modèle de connaissance construit, l'étape d'évaluation a pour objectif de vérifier et d'évaluer le résultat du modèle. Dans le cas où plusieurs modèles ont été développés par le biais de plusieurs algorithmes ou méthodes, ils doivent être comparées pour optimiser les performances ou réduire le taux d'erreur. Cette étape contient les phases suivantes:

- Évaluer les résultats des modèles fournis par l'étape de modélisation selon les objectifs définis dans la première étape.

- Dans le cas où les résultats ne répondent pas aux objectifs, il faut revenir aux étapes précédentes.
- Décider des modèles pour le déploiement.

2.1.1.6 Déploiement

L'étape finale consiste à déployer les modèles choisis et à transférer les résultats de data mining aux utilisateurs.

- Planifier le déploiement.
- Présenter les résultats du data mining aux utilisateurs finaux.
- Le déploiement ne signifie pas la fin du projet. Les données changeant en temps réels, le suivi et la maintenance sont indispensables pour assurer la qualité du résultat.

Les utilisateurs finaux ne sont pas des experts du data mining. Il est donc nécessaire de trouver une façon de visualiser les résultats du data mining afin de faciliter la prise de décision par l'utilisateur final.

2.1.2 Les techniques de prétraitement des données

Comme nous l'avons mentionné dans la section précédente, le prétraitement des données est une étape qui influence hautement la qualité de résultat final du data mining. D'après (Han et Kamber, 2006), l'étape de prétraitement des données requiert les opérations suivantes : l'intégration, la réduction de l'espace, le nettoyage et la transformation des données. Les sections suivantes expliquent les techniques relatives à ces opérations.

2.1.2.1 Intégration

Les données utilisées pour un projet de data mining sont généralement collectées de plusieurs sources, avec des formats différents (par exemple : les données XML, les données CSV, etc.). Une fois que le choix des variables et des sources de données est réalisé, les données devront être acquises et intégrées pour être utilisées lors des prochaines étapes. L'intégration des sources de données est effectuée soit par approche virtuelle soit par approche matérialisée (Hacid et Reynaud, 2004). Dans l'approche virtuelle, l'intégration est réalisée sans toucher aux sources de données d'origine. Elle est basée sur une interface (souvent appelé médiateur) qui envoie les requêtes vers les sources originales (voir figure 2.2). Cette approche permet une intégration en temps réel. Cependant, elle pose des difficultés au niveau de la traduction des requêtes pour être interprétées par les différentes sources.

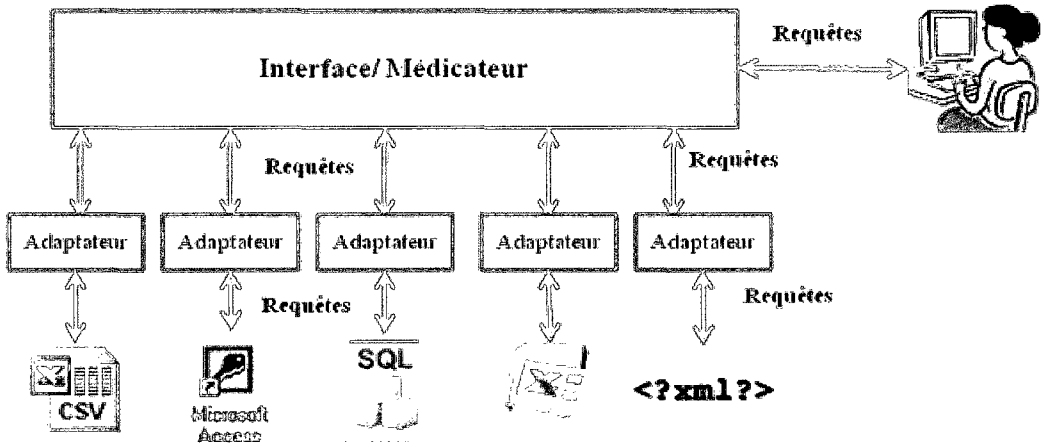


Figure 2.2 Intégration virtuelle des sources de données

Dans l'approche matérialisée, les données sont extraites de différentes sources et combinées pour être stockées de manière centralisée dans une base cible (voir figure 2.3). Ce stockage permet à l'utilisateur d'avoir un accès unique à toutes les sources originales. L'approche matérialisée est très performante car les requêtes sont di-

rectement exécutées dans le référentiel. Le problème principal de cette approche est la mise à jour entre les données référentiel et les sources de données originales.

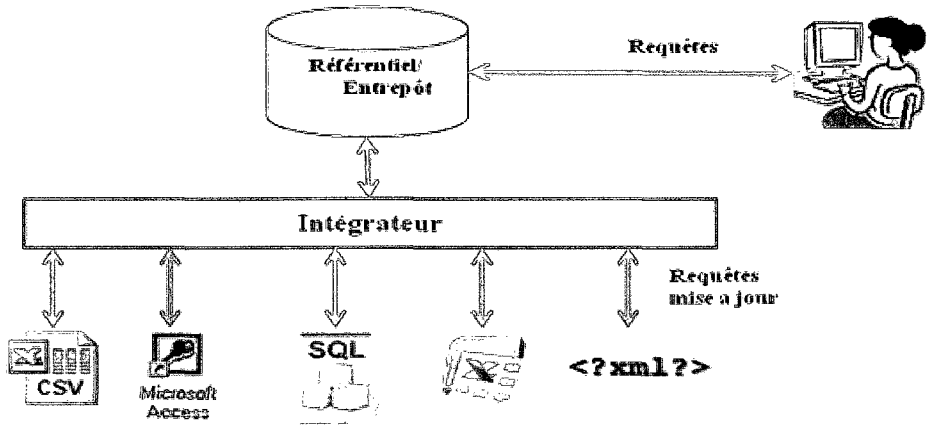


Figure 2.3 Intégration matérialisée des sources de données

2.1.2.2 Réduction de l'espace

Les données collectées contiennent souvent des variables inutiles. La réduction de l'espace du problème consiste à réduire le nombre de variables et le nombre de modalités des variables (Tufféry, 2005). Dans la segmentation du marché, cette étape est souvent faite en collaboration avec un expert du domaine. Elle sert non seulement à réduire l'espace des données mais aussi à augmenter la qualité et la performance de l'étape de fouille des données. Les principales tâches de cette étape sont:

- D'éliminer les variables non pertinentes par rapport à l'objectif à atteindre.
- Si possible, de grouper plusieurs variables en une seule : par exemple, le nombre de personnes de 1 à 8 ans peut être additionné à celui des personnes de 9 à 15 ans pour fournir le nombre de personnes de 1 à 15 ans.

- De regrouper les modalités des variables discrètes et qualitatives dans le cas ou elles sont trop nombreuses ou elles ont la même signification fonctionnelle.

2.1.2.3 Nettoyage des données

Des données manquantes, des valeurs aberrantes ou encore des données redondantes peuvent délivrer des résultats instables. Il est donc nécessaire de procéder au nettoyage des données brutes avant de les exploiter. Cette opération comprend les tâches ci-dessous (Tufféry, 2005):

- **Traiter les valeurs rares ou manquantes**: La technique la plus simple est de remplacer ces valeurs par la valeur la plus fréquente, la moyenne ou la médiane. Pourtant, l'inconvénient de cette technique est qu'elle déforme la distribution de la variable imputée (Tufféry, 2005). (Rubin, 1976) propose la technique d'imputation multiple qui remplace chaque valeur manquante par une valeur présumée. Les techniques de régression, de classification ou d'arbre de décision examinent les individus de même profil que ceux ayant une valeur manquante, puis remplacent chaque valeur manquante par la moyenne de la classe de l'individu.

- **Traiter les valeurs aberrantes ou extrêmes**: Une erreur de saisie ou de calcul peut engendrer des valeurs aberrantes, donc nécessaires à corriger. Quant aux valeurs extrêmes, ce sont des données valides et intéressantes à détecter. Les techniques ci-dessous sont souvent utilisées (Tufféry, 2005):
  - Examiner l'histogramme graphique des variables
  - Imputer par régression.
  - Imputer par le plus proche voisin.

- **Traiter les données redondantes**: Les techniques communes pour identi-

fier et éliminer les données redondantes sont :

- Utiliser des requêtes pour supprimer les doublons de la base de données.
- Analyser la corrélation pour identifier les données redondantes.

2.1.2.4 Transformation des données

Chaque algorithme du data mining requiert des formats de variables différents. Par exemple, les méthodes de classification ont souvent besoin de variables continues, tandis qu'on a besoin de variables discrètes pour les méthodes d'associations. Les techniques les plus utilisées pour cette opération sont : la normalisation, la discrétisation et la création des nouvelles variables.

- **Normalisation**: La normalisation met les données à l'échelle pour arriver à les faire entrer dans l'intervalle spécifié. Il existe plusieurs méthodes de normalisation. La normalisation min-max, la normalisation par le test Z en sont des exemples.

- **Discrétisation**: La discrétisation est utilisée dans le cas d'un besoin en données discrètes. Les variables continues seront découpées en classes basées sur des échelles choisies.

- **Création des nouvelles variables**: Dans certains cas, il s'avère nécessaire de créer des variables nouvelles à partir de variables initiales. Par exemple, pour calculer l'âge du client, on peut se baser sur sa date de naissance et la date actuelle.

2.1.3 Les défis du data mining

De nos jours, le développement de la technologie, les nouveaux domaines, les nouveaux besoins posent au data mining plusieurs défis. Nous discuterons ici des trois défis principaux du data ming qui ont été proposés récemment par (Han et Kamber, 2006).

2.1.3.1 Méthodologie et interactions avec l'utilisateur

Les utilisateurs actuels du data mining sont variés et ont des intérêts hétérogènes. Le data mining doit répondre à cette diversité en intégrant les différences techniques telles que la classification, la prédiction, l'association, la discrimination, etc. Ces techniques peuvent être appliquées aux mêmes bases de données afin de fournir des vues de connaissance différentes.

De plus, avec la quantité de données actuellement traitée, il est difficile de déterminer le niveau d'extraction de connaissance. Cela pose également des défis pour le prétraitement des données. Le data mining devrait être interactif pour les utilisateurs et leur permettre de raffiner le résultat de différentes manières. Les techniques de data mining devraient être en mesure de traiter des données brutes et incomplètes.

(Han et Kamber, 2006) ont aussi souligné l'importance de la présentation et de la visualisation des résultats du data mining. Les connaissances exploitées sont utiles tant qu'elles sont intéressantes, et surtout compréhensibles par l'utilisateur. La visualisation des différents types de données, l'interaction avec l'utilisateur et l'usabilité de l'interface sont des problèmes majeurs dans l'étape de déploiement du processus de data mining.

2.1.3.2 Performance

La plupart des algorithmes actuels du data mining posent des problèmes de performance en travaillant avec une grande base de données tant en termes d'occupation mémoire qu'en termes de temps de traitement. Les problèmes de performances comprennent: l'efficacité, la parallélisation et la distributivité des algorithmes du data mining.

- **Efficacité et mise à l'échelle des algorithmes de data mining**:
  L'efficacité est un critère important pour la mise en œuvre d'un projet de data mining. En effet, les algorithmes ont besoin d'améliorations pour pouvoir travailler avec une grande base de données dans un temps acceptable.

- **Parallélisation, distributivité et possibilités incrémentales des méthodes de data mining**:
  Aujourd'hui, il est commun de traiter des données se mesurant en terabytes. Cette quantité de données motive le développement d'algorithmes distribués. Ces algorithmes traitent en parallèle des parties de données et les résultats sont fusionnés par la suite.

2.1.3.3 Diversité des données

- **Données relationnelles et types complexes**:
  Les données actuelles sont variées : les données relationnelles, les données multimédias, les données spatiales, les données textuelles etc. Cela pose la question d'améliorer les algorithmes de data mining pour les adapter aux nouveaux types de données. On note au cours de ces dernières années que les recherches sur la fouille de données spatiales, le text mining, etc. ont bien été développées.

- **Bases de données hétérogènes et systèmes global d'information**: L'internet présente une source de données importante pour le data mining. Pourtant, les données de l'internet sont souvent hétérogènes, semi-structurées ou non structurées. Il devient très difficile de réaliser du Web mining dans ce domaine de l'exploration de données hétérogènes. La découverte de connaissances à partir de ces données pose donc un grand défi au data mining.

2.2 Concepts de base des Systèmes d'Informatique Géographique(SIG)

2.2.1 L'évolution du SIG

Un SIG signifie un système d'information géographique. Dans la littérature sur les SIG, plusieurs définitions sont proposées. Nous citons ici la définition la plus utilisée, qui est proposé par le comité scientifique du colloque intégration de la photogrammétrie et de la télédétection dans les SIG, Strasbourg 1990 : "Un système d'information géographique est un système informatique permettant, à partir de diverses sources, de rassembler et organiser, de gérer, d'analyser et de combiner, d'élaborer et de présenter des informations localisées géographiquement contribuant notamment à la gestion de l'espace". Nous marquons ici quelques événements importants dans trois grandes périodes de l'évolution des SIG (Longley *et al.*, 2005):

- L'ère de l'évolution de la fin des années 1950 au milieu des années 1970 : Le premier système d'information géographique est développé en 1963 par l'Inventaire des terres du Canada pour identifier les ressources territoriales et ses potentiels. Ce projet a introduit la notion de SIG. En 1967, le US Census Bureau a développé DIME-GBF (Dual Independent Map Encoding-Geographic Database Files), qui gère les données numériques de toutes les rues des États-Unis pour supporter l'agrégation des données du recensement.

- L'ère de la commercialisation dans les décennies 1980 et 1990 avec l'apparition de plusieurs produits commerciaux de SIG tels que ArcGIS, MapInfo, MapQuest, etc. Des nouvelles technologies telles que le GPS (Global Positioning System), TIGER (Topologically Integrated Encoding and Referencing), etc. ont également apporté de grandes avancées pour les SIG. Le marché des SIG a aussi grandement bénéficié de l'adoption du paradigme de l'Internet depuis 1993. Les SIG sur l'internet facilitent l'utilisation et le partage des données. Les domaines d'application des SIG sont ainsi élargis.

- L'ère de l'exploitation depuis le milieu des années 1990 a vu l'augmentation des utilisateurs de SIG dans plusieurs domaines, le croisement des SIG avec d'autres technologies telles que le système d'aide à la décision, le data mining, etc.

2.2.2 Les composants d'un SIG

Un SIG est constitué de cinq composants indispensables comme présenté dans la figure 2.4. Nous expliquons brièvement ces composants dans les paragraphes suivants (Longley *et al.*, 2005).

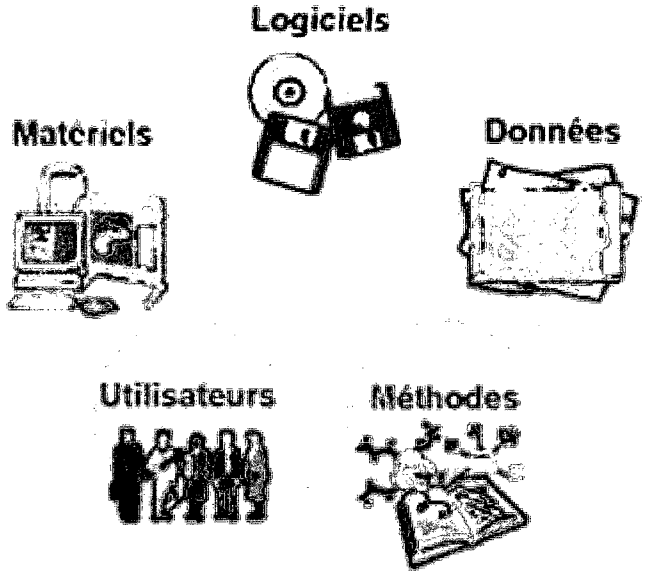


Figure 2.4 Les compostants d'un SIG

2.2.2.1 Matériels

Les composantes matérielles dans les SIG sont des serveurs, des équipements des utilisateurs, des matériels d'acquisition des données, etc. Des équipements des utilisateurs contemporains ne sont plus limités aux ordinateurs de bureau. En effet, l'accès à internet via les PDA ou les téléphones cellulaires permet aujourd'hui aux utilisateurs d'accéder aux SIG.

2.2.2.2 Données

La base de données est le cœur de tous les SIG. Les données des SIG peuvent être constituées en interne ou acquises auprès de producteurs de données. Il y a deux types de données géospatiales: les données raster et les données vectorielles. Les

données raster sont constituées d'une matrice ou d'une grille régulière de pixels, chaque pixel affichant un attribut unique. Elles sont principalement des photographies numériques, des photographies scannées, des images satellites ou des plans scannés. Les données vectorielles quant à elles sont constituées d'objets tels que des points, des lignes, des polygones, des surfaces ou des volumes représentant les objets visibles sur le terrain. La figure 2.5 représenté par l'image de ces deux types de données.

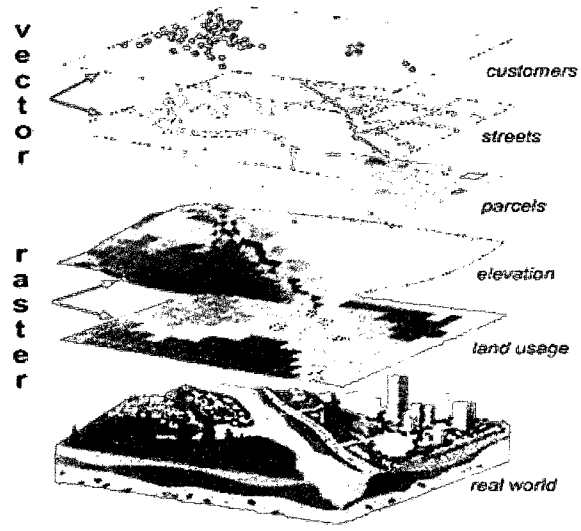


Figure 2.5 Les données raster par rapport aux données vectorielles (ESRI, 2008)

2.2.2.3 Méthodes

Les SIG sont utilisés dans des domaines très divers tels que la planification territoriale, l'économie, le transport ou la gestion des risques naturels. Chaque domaine a des exigences différentes. En effet, la mise en œuvre et l'utilisation d'un SIG ne sauraient être réalisées sans la prise en compte des méthodes et des procédures du domaine appliqué. Ces méthodes permettent d'intégrer les matériels, les logiciels et

les données du SIG par l'ensemble des utilisateurs afin de répondre aux demandes spécifiques d'un projet SIG.

2.2.2.4 Logiciels

Si la base de données peut être comprise comme le cœur d'un SIG, alors les logiciels peuvent servir de pont relient les quatre autres composants. Ils sont utilisés par les utilisateurs pour appliquer les méthodes aux données spatiales et afficher les résultats sur les matériels.

Les logiciels de SIG actuels peuvent être catégorisés ainsi : les logiciels commerciaux et les logiciels libres. Le tableau 2.2 liste quelques logiciels actuels de SIG couramment utilisés.

Tableau 2.2 Liste des logiciels SIG

Les logiciels de SIG	
Catégorie	Logiciels
Commerciaux	ArcGIS (ArcInfo, ArcView, etc) d'ESRI; AutoCAD Map 3D d'Autodesk ; Bentley Map de Bentley Systems; GeoMap GIS de GEOMAP ; MapInfo de Pitney Bowes Software - MAPINF
Libres	MapWindow GIS; g3DGMV (3D Graphical Map Viewer); Openmap; OrbisGIS; Google Map MashUp; SAGA GIS;

Qu'il soit commercial ou libre, un logiciel SIG doit assurer les quatre fonctions fondamentales suivantes qui permettent d'organiser, de représenter et de gérer les informations géographiques : l'acquisition, l'archivage, l'affichage et l'analyse. Nous en discuterons dans la section 2.2.3

2.2.2.5 Utilisateurs

Il y a peu de temps, utiliser un SIG demandait aux utilisateurs certaines connaissances en géographie et en informatique. Les SIG en général, s'adressent à des domaines variés. Les utilisateurs ont besoin de connaissances spécifiques du domaine pour exploiter pleinement les capacités des SIG. Aujourd'hui, l'évolution des SIG facilite leur utilisation et élargit la communauté des utilisateurs de SIG. Cependant, les utilisateurs jouent toujours un rôle de premier ordre dans le succès de la mise en œuvre d'un SIG.

2.2.3 Les fonctionnalités d'un SIG

Dans cette section, nous résumerons les quatre fonctionnalités principales que tous les SIG doivent assurer. Les détails de ces fonctionnalités peuvent se trouver dans le livre "Geographic Information Systems and Science" de (Longley *et al.*, 2005). Un SIG commence par acquerir des données cartographiques, les stocker et les gérer dans une base de données relationnelles pour pouvoir effectuer des analyses, puis produire des cartes géographiques (voir figure 2.6).

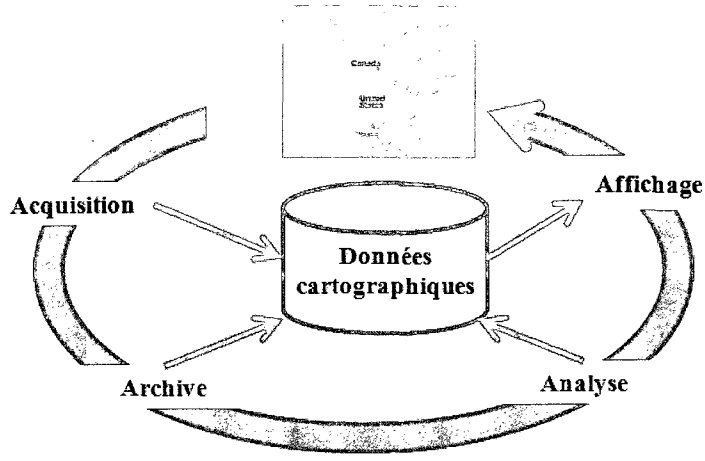


Figure 2.6 Les fonctionnalités d'un SIG (Longley *et al.*, 2005)

2.2.3.1 L'acquisition

La collecte des données est la tâche la plus importante d'un SIG. En effet, les coûts de saisie de données peuvent représenter jusqu'à 85% du coût d'un SIG. Le processus de collecte des données cartographiques est presenté dans la figure 2.7.

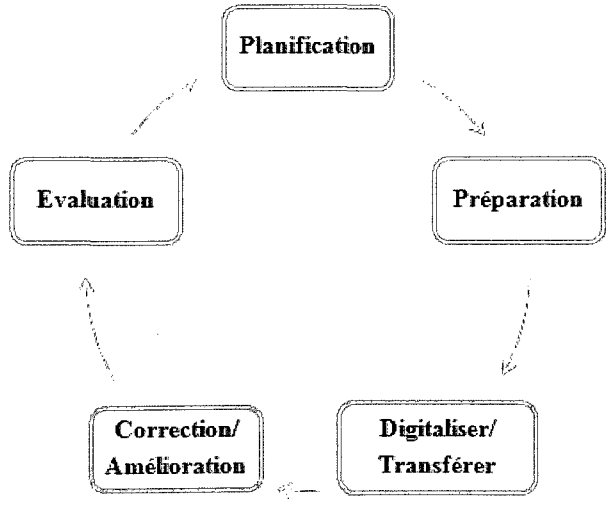


Figure 2.7 Processus d'acquise les données pour les SIG

L'acquisition des données cartographiques peut être effectuée soit par la saisie, soit par la transformation des données.

- La saisie des données:

La saisie des données fournit la source des données de base pour le SIG. En effet, les données collectées par cette méthode peuvent être utilisées directement par le SIG. Les images numériques de la télédétection satellite, les photographies aériennes numériques, les cartes scannées, les cartes topographiques, les mesures de GPS comptent parmi les sources principales pour la saisie des données cartographiques.

- La transformation des données:

Au delà des sources de données directes, le SIG peut aussi bénéficier des données géographiques de plusieurs sources externes telles que les données environnementales du WWF(World Wildlife Fund for Nature), les données socio-démographiques des gouvernements, etc. Pourtant, les données collectées par des sources externes sont souvent encodées dans des formats différents (shapefile, DXF, AutoCAD ou même en Excel). Il est donc nécessaire de les transformer dans un format qu'un SIG peut traiter.

2.2.3.2 L'archivage

Au cours des dernières années, le volume et la complexité des données géographiques augmente très vite, il est donc essentiel d'utiliser un SGBD (Système de Gestion de Bases de Données) pour faciliter le stockage, l'organisation, la gestion, la mise à jour et les extractions de données. Il existe plusieurs types de SGBD mais en général, le type de SGDB le plus utilisé actuel en système d'information géographique est un SGBDR (Système de Gestion de Bases de Données Relationnel). La mise en place d'une base de données spatiale doit prendre en compte les méthodes de gestion au niveau logique mais aussi au niveau physique.

Au niveau logique, les données géométriques et les données attributaires peuvent

être gérées simultanément ou séparément par les systèmes de gestion de bases de données. Dans le premier cas, les données spatiales sont ramassées dans un même enregistrement contenant la description de la géométrie et la valeur de chacun de ses attributs. Dans le deuxième cas, il y a au moins deux fichiers. L'un contenant les données géométriques, l'autre contenant les attributs. Cette approche est la plus utilisée dans les systèmes de gestion des bases de données spatiales actuelles car elle permet de gérer indépendamment les différentes informations et donc de réduire les risques dans le cas d'un accident(un serveur en panne ou une destruction des données).

Au niveau physique, les bases de données peuvent être construites par l'architecture centralisée ou l'architecture distribuée comme illustré à la figure 2.8.

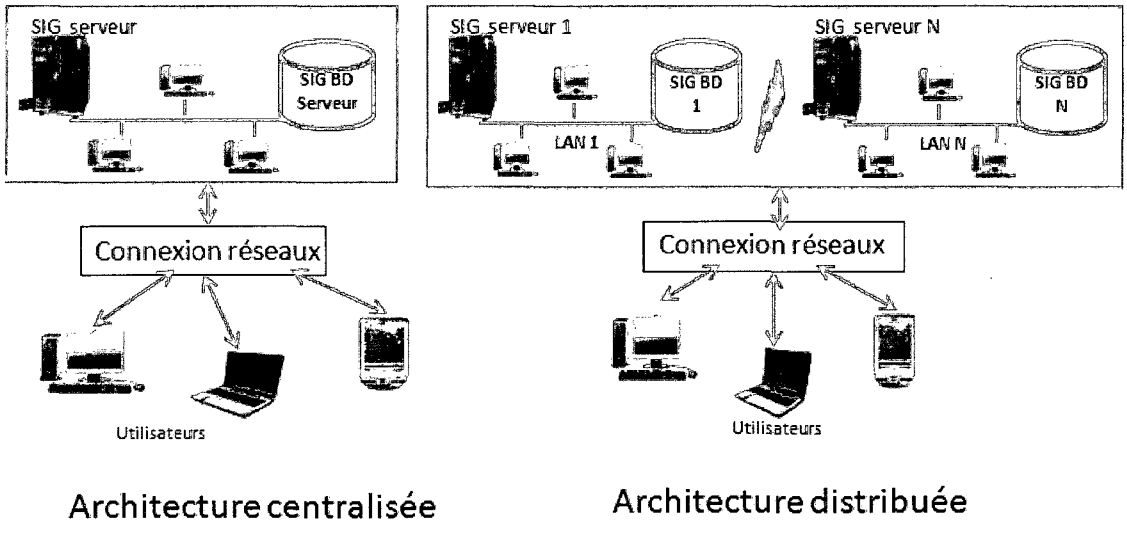


Figure 2.8 Les architectures de bases de données en SIG

2.2.3.3 L'analyse

L'analyse est aussi une fonctionnalité très importante dans les SIG. Cette fonction fournit des outils permettant aux utilisateurs d'analyser les données géographiques,

d'interroger les configurations spatiales observées ou de proposer des simulations d'organisation spatiale. Elle permet aussi de modifier les objets spatiaux et de mesurer leurs relations. Les différentes relations que l'on prend en compte en analyse spatiale sont par exemple la proximité (trouver les objets proches d'un autre), la topologie (objets jointifs, inclus, partiellement inclus, exclus, etc.) ou encore la forme (taille, type, etc.). Les données spatiales étant composées d'attributs géométriques et sémantiques (la description des objets géographiques), l'analyse spatiale doit permettre de combiner les propriétés géométriques et les propriétés sémantiques afin de réaliser une analyse complète. Parmi les autres fonctionnalités d'analyse spatiale et de modélisation d'un SIG on peut citer:

-**Le géocodage**: Pour certains projets de SIG, les données traitées ne sont pas directement fournies avec une localisation directe. Dans ce cas, les SIG doivent disposer des géocodeurs qui permettent de déterminer les coordonnées d'une adresse en utilisant des données géographiques de référence dont les coordonnées sont connues.

-**L'analyse de voisinage**: Ce sont une gamme de fonctions qui permettent d'analyser des relations entre un endroit précis et son voisinage.

-**Les requêtes spatiales**: Elles permettent de sélectionner, de localiser, ou de manipuler des objets répondant à un ou plusieurs critères.

-**La superposition des couches**: Les SIG utilisent des couches afin de superposer différents types d'informations. Chaque couche représente une catégorie d'informations telle que des routes ou des lignes de bus.

2.2.3.4 L'affichage

Les cartes géographiques sont les sorties finales d'un SIG. Un SIG affiche les données par superposition de couches thématiques pouvant être reliées les unes aux autres par la géographie. Les utilisateurs peuvent produire les cartes selon leurs objectifs à partir des différentes données géographiques disponibles. Pour ce faire, on retrouve

toujours et dans tous les SIG une palette de fonctions de gestion de l'affichage qui permettent d'effectuer au moins les fonctions suivantes:

- Fonction de lecture des données géographiques;
- Fonction d'impression des cartes;
- Fonction de requête;
- Fonction de localisation des objets géographiques (des points, des lignes, des polygones);
- Fonction visualisation à plusieurs échelle de grossissement;
- Fonction de déplacement latéral.

2.2.4 Les défis du SIG

"Les nouvelles technologies ont toujours été une force d'entraînement principale en sciences de l'information géographique, comme elles le sont dans la science et la société généralement" (Goodchild et Haining, 2003). Et c'est aussi le cas des SIG. Les nouveaux systèmes de télédétection permettent d'acquérir les données numériques de géo-référencé avec une haute résolution spectrale. Des organisations publiques et privées sont en train de collecter, de produire et de fournir des données numériques de la terre, des données socio-économiques d'un haut niveau de détail. Les dispositifs de localisation tels que les téléphones cellulaires et les GPS permettent de suivre les mouvements individuels dans l'espace. Cela fait augmenter énormément le volume et la complexité des données spatiales. Ils posent donc plusieurs défis à la gestion de base de données, à l'analyse et aussi à l'affichage en SIG.

2.2.4.1 Gestion de bases de données

Effectivement, les SGBDR ne répondent pas suffisamment aux besoins de la gestion des données spatiales actuelles. Parmi les SGDB récemment développés pour les données spatiales, le SGBDOO (Système de Gestion de Bases de Données Orienté-Objet) et le SGBDRO (Système de Gestion de Bases de Données Relationnel-Objet) sont les plus adaptés (Longley *et al.*, 2005). Le grand volume de données cause aussi des problèmes de performance pour les SGBD. En effet, le développement des SIG actuels doit mieux prendre en compte la capacité de distribution basée sur l'architecture distribuée. D'autres défis à relever pour la gestion de bases de données sont:

- L'intégration des données spatiales de sources et de formats multiples.
- La gestion des données temporelles.

2.2.4.2 L'analyse de données

L'augmentation du volume et de la complexité des données géographiques actuelles posent des difficultés aux méthodes traditionnelles de l'analyse spatiale. En effet, ces méthodes ne permettent pas de travailler avec les données ayant des relations spatiales. Les techniques du data mining peuvent répondre à ce défi. Pourtant, de nombreux travaux sont encore nécessaires pour modifier et améliorer les techniques existantes du data mining pour les adapter aux caractéristiques des données spatiales.

2.2.4.3 L'affichage de données

Les GIS passent de mode 2D(x,y) à 3D(x,y,z) puis 4D(x,y,z,t(time)). La recherche des techniques de visualisation pour les données 3D et 4D est donc une voie de recherche importante en géovisualistion. Le grand volume de données (spatiales et non spatiales) à visualiser cause également un défi pour l'affichage en SIG. De plus, l'intégration avec l'internet demande aux SIG de passer d'un affichage traditionnel à un affichage plus interactif.

CHAPITRE 3

REVUE DE LITTÉRATURE

Comme nous l'avons vu dans la partie précédente, le croisement des défis du data mining et des SIG offre un potentiel pour l'intégration de ces deux technologies. Cet état de l'art résume les deux approches de cette intégration. D'une part, les techniques du data mining peuvent être appliquées à la phase d'analyse d'un SIG; d'autre part, les SIG peuvent contribuer à de nombreuses phases du data mining (phase de prétraitement, phase d'évaluation et phase de déploiement).

3.1 Les techniques du data mining appliquées aux données spatiales

L'application des techniques du data mining aux données spatiales est connue sous le nom de "data mining spatial". Selon (Zeitouni, 1998), le data mining spatial est "l'extraction de connaissances implicites ou de relations spatiales ou d'autres propriétés non explicitement stockées dans les bases de données spatiales. Il permet de retrouver des régularités implicites et des relations entre données spatiales et/ou non spatiales, de constituer des bases de connaissances, d'optimiser des requêtes ou de réorganiser la base de données spatiale". Le data mining spatial est donc une extension du data mining traditionnel tenant compte des caractéristiques des données spatiales. Dans les sections suivantes, nous présenterons quelques caractéristiques des données spatiales. Nous focaliserons ensuite notre attention sur les techniques principales du "data mining spatial". Nous finirons par donner quelques exemples d'application de ces techniques.

3.1.1 Les caractéristiques des données spatiales

D'après (Buttenfield *et al.*, 2000) et (Bacao *et al.*, 2005), les données spatiales présentent trois caractéristiques principales qui causent des difficultés pour les méthodes d'analyse de données.

Les relations spatiales constituent la première caractéristique. Elles soulignent les relations et l'influence du voisinage entre les entités spatiales. La dépendance spatiale est connue comme la première loi en géographie "tout ce qui se passe à un endroit est lié à ce qui ce passe au voisinage et ce lien décroît avec l'éloignement" (Tobler, 1979). Cette loi signifie que les données spatiales ne sont pas indépendantes et que l'analyse des données spatiale doit prendre en compte des caractéristiques des objets du voisinage et des relations spatiales qui les relient. L'autocorrélation spatiale est une relation spatiale qui estime la corrélation d'une variable en référence à sa localisation dans l'espace ou dans le temps. Elle est directement liée avec la dépendance spatiale. Elle mesure le niveau de l'interdépendance entre les variables (les observations géographiques) et peut avoir une valeur positive ou négative. Une autocorrélation spatiale positive se traduit sur une carte par le regroupement géographique des valeurs similaires ou par le regroupement géographique de valeur dissimilaires dans le cas d'une autocorrélation négative. Une absence d'autocorrélation signifie que la répartition spatiale est aléatoire.

La deuxième caractéristique est l'hétérogénéité spatiale. Elle est liée à l'absence de stabilité sur le comportement des relations dans l'espace. L'analyse statistique classique d'une population est basée sur l'hypothèse que les éléments dans cette population ont des points communs, sur lesquels on peut établir des comparaisons et des régularités (Jayet, 2001). (Morency, 2006) a mentionné que les données spatiales présentent généralement une forte hétérogénéité qui implique que la valeur des observations varie dans l'espace.

La troisième caractéristique est la complexité des objets spatio-temporels. Ces derniers peuvent être définis comme un objet spatial dont la forme et/ou la position varient au cours du temps. La variabilité de ces objets dans le temps est complexe mais apporte des informations (voir (Hornsby et Egenhofer, 2002)). Cela pose des défis critiques pour l'analyse des données spatiales (Zeitouni, 2006).

3.1.2 Les techniques du data mining spatial

Les techniques du data mining traditionnel ne sont pas adaptées aux données spatiales car elles ne considèrent pas ce type de relation. Il est donc nécessaire de développer de nouvelles techniques ou d'améliorer les techniques existantes du data mining pour les adapter aux caractéristiques des données spatiales. (Ester *et al.*, 2001) ont proposé une plate-forme de base de données pour le data mining spatial. Leur plate-forme est essentiellement basée sur les relations de voisinage et le graphe de voisinage des objets.

A l'heure actuelle, plusieurs techniques du data mining spatial sont analogues à celles du data mining traditionnel. Nous présenterons brièvement quelques techniques importantes de data mining spatial dans les paragraphes suivantes.

3.1.2.1 Segmentation spatiale

La segmentation (clustering en anglais) est une méthode de classification automatique non supervisée bien connue en fouille de données et en statistique. Elle permet de regrouper des objets par classes homogènes de façon à ce que la similarité intra-classe soit maximale et la similarité inter-classe soit minimale. Néanmoins, (Zeitouni, 2006) a remarqué que dans le domaine spatial, la segmentation vise moins à classifier qu'à détecter les concentrations (par exemple, la détection des

zones accidentogènes en sécurité routière). Les principaux travaux actuels sur la segmentation spatiale portent sur l'optimisation des performances des algorithmes (Zeitouni, 2006). (Ester *et al.*, 1998) ont étendu les méthodes par partitionnement et hiérarchique pour les données spatiales en se basant sur l'extension de méthode DBSCAN et l'index spatial R*tree.

Dans plusieurs applications avec les données spatiales, les obstacles physiques peuvent influencer le résultat de segmentation. Considérons le cas de l'implantation d'un nouveau magasin, tel qu'illustré à la figure 3.1. Pour faire face à ces contraintes, (Tung *et al.*, 2001) ont introduit la méthode de segmentation COD(Clustering with Obstructed Distance) en prenant en compte des entités obstacles. Ils ont aussi développe méthode COE-CLARANS(Clustering with Obstacle Distance based on CLARANS) qui intègre les deux techniques COD et CLARANS(Clustering Large Applications based on RANdomized Search) pour améliorer l'efficacité de la segmentation.

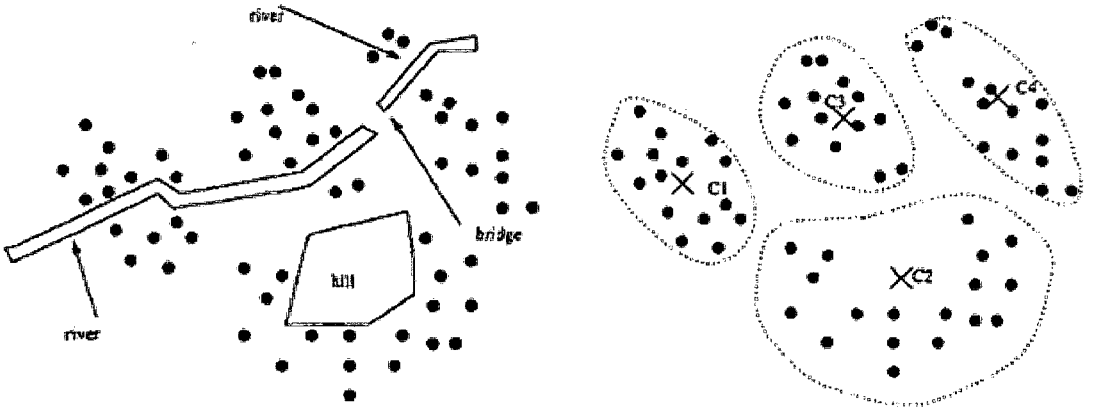


Figure 3.1 Segmentation avec et sans contraintes (Tung *et al.*, 2001)

Dans le cas des attributs non spatiaux ou des objets de forme linéaire ou surfacique,

(Ester *et al.*, 2001) ont proposé la méthode GDBSCAN (Generalized Density Based Spatial Clustering of Application with Noise) en généralisant la similarité et le prédicat de densité. Cette dernière se base sur la méthode DBSCAN. Elle généralise la notion de la densité des points à la notion de la densité des ensembles de points. Cette méthode permet donc de segmenter d'autres objets que les points (par exemple les polygones ou les lignes) dans les données spatiales . D'autres techniques de segmentation spatiale sont résumées dans (Han *et al.*, 2001).

La segmentation spatiale est souvent utilisée comme une étape de prétraitement pour d'autres tâches telles que la recherche d'associations entre segments ou la caractérisation au sein d'un segment.

3.1.2.2 Classification spatiale

L'objectif de la classification est d'attribuer un objet à une classe prédéfinie basée sur les valeurs des attributs de l'objet. Elle peut être utilisée pour prédire les classes de nouveaux objets ou pour expliquer les relations entre les propriétés de l'objet et sa classe. (Ester *et al.*, 2001) ont constaté que la classification spatiale doit prendre en compte non seulement des valeurs d'attributs des objets à analyser, mais également celles des objets voisins et des liens de voisinage. Parmi les techniques de classification, l'arbre de décision est la méthode la plus utilisée pour les données spatiales grâce à sa performance par rapport aux autres méthodes. (Fayyad *et al.*, 1996b) sont les premiers à utiliser les arbres de décision pour détecter les astres et les galaxies sur des images satellitaires. Cette méthode a montré son efficacité en traitant un grand volume de données.

Une amélioration de la méthode ID3 est proposée par (Ester *et al.*, 1997) en utilisant le concept de graphe de voisinage pour représenter les liens de voisinage. Cette

méthode considère les propriétés des objets ainsi que les attributs et les relations des objets voisins. Pourtant, (Zeitouni, 2006) a remarqué que cette méthode est limitée à une seule relation de voisinage et elle ne fait pas de distinction entre les thèmes.

Une autre méthode est présentée par (Koperski *et al.*, 1998) qui considère les attributs non spatiaux ainsi que les prédicats et les fonctions spatiales reliés par une relation spatiale à l'objet considéré. D'après (Zeitouni, 2006), cette méthode garantie une bonne classification mais présente une limitation au niveau du coût de prétraitement.

(Chelghoum *et al.*, 2002) ont proposé la méthode SCART (Spatial Classification and Regression Trees) qui se base sur la méthode CART. L'avantage de cette méthode est de prendre en compte l'organisation en couches thématiques et les relations spatiales, ce qui est essentiel dans les applications géographiques.

Un exemple de (Chelghoum et Zeitouni, 2004) démontre comment appliquer la méthode SCART à l'analyse de l'accidentologie en sécurité routière. Elle est utilisée pour classer les segments de route en deux classes: "segment non dangereux" (moins de 2 accidents) et "segment dangereux" (plus de 2 accidents) selon les attributs spatiaux (voir figure 3.2).
Dans l'arbre de la figure 3.2, la première condition de classification est " sens de circulation = double". A ce niveau, nous obtenons deux classes: la première classe correspond aux sections de route ayant un sens de circulation = double et la seconde son complément. La première classe est segmenté ensuite par la condition de distance avec les écoles. Dans cette classe, la condition de classification est une combinaison de la valeur "école", d'un attribut de la table voisine, de la relation spatiale "distance", de la comparateur "<=" et de la valeur de la relation spatiale "425 m". A partir de cet arbre, (Chelghoum et Zeitouni, 2004) ont obtenu 4 règles

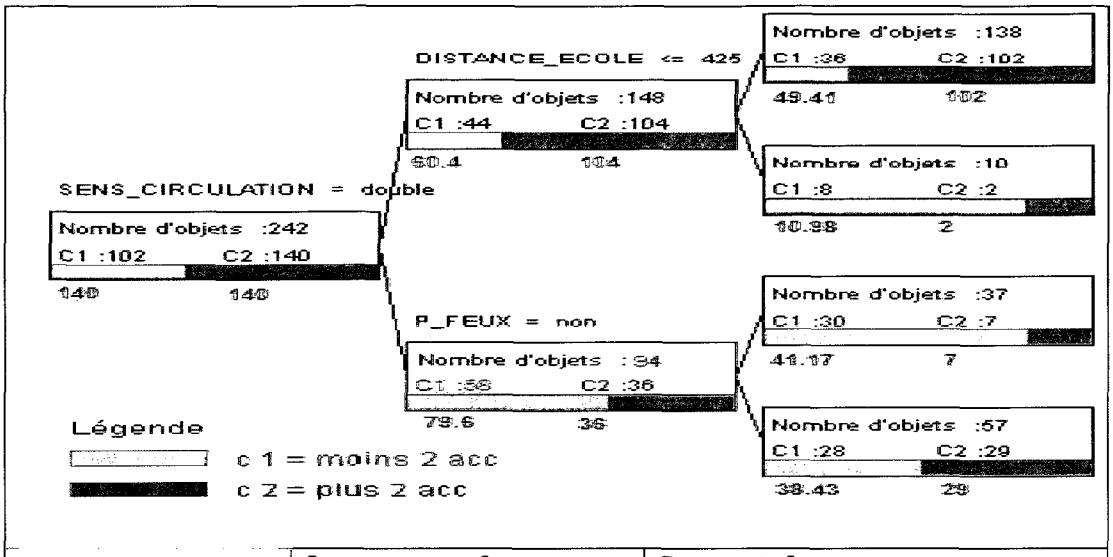
de décisions qui correspondent aux 4 feuilles finales dans l'arbre. De haut en bas, les règles peuvent être expliquées comme suit:

+ La première règle stipule qu'il y a plus de segments de route dangereux près des écoles (distance <= 425 m) et où le sens de circulation est double.

+ La deuxième règle dit que lorsque le sens de circulation est double et qu'on est loin des écoles (distance > 425 m) alors qu'il y a plus de segments de route non dangereux.

+ La troisième règle indique que si le sens de circulation est unique et qu'il n'y a pas de feu rouge alors on a une section de route non dangereuses.

+ La dernière règle montre qu'on a plus de sections de route non dangereuses lorsque le sens de circulation est unique et qu'il y a un feu rouge.



	Segment non dangereux		Segment dangereux	
	Nombre	%	Nombre	%
Segment non dangereux	66	**64.7%**	36	35.3%
Segment dangereux	38	27.15%	102	**72.8%**

Figure 3.2 Exemple d'arbre de décision spatial et matrice de confusion(Chelghoum et Zeitouni, 2004)

3.1.2.3 Règles d'association spatiales

D'après(Miller et Han, 2001), les règles d'association spatiales contiennent les prédicats spatiaux dans le précédent ou l'antécédent. Les règles peuvent être constituées par différents prédicats spatiaux. Quelques exemples de prédicats spatiaux incluent l'information spatiale (*close-to* et *far-away*), relation topologique (*overlap, disjoint*) et orientations spatiales (*right-of, west-of*) (Han et Kamber, 2006).

(Koperski et Han, 1995) ont proposé une méthode qui a été dérivée de la méthode des règles d'association. Elle permet de trouver des règles entre les objets en se

basant sur leurs relations de voisinage. Par exemple, la règle suivante est une règle d'association spatiale pour laquelle le premier prédicat est non spatial et le deuxième l'est.

$$is - a(X, cinéma) \Rightarrow close - to(X, Y) \wedge is - a(Y, restaurant)(70\%) \qquad (3.1)$$

Cette règle peut être expliquée comme suit : Dans l'endroit étudié, 70% des cinémas sont à proximité d'un restaurant. (Koperski et Han, 1995) ont aussi présenté la technique de recherche "top-down" en utilisant l'approximation spatiale pour découvrir des règles offrant un support et une confiance maximale. Cette approche a été mise en œuvre dans le système *GeoMiner*. Elle permet de générer des règles multi-niveau et d'optimiser la recherche des règles. Une limite de cette méthode est la production en grand nombre de règles bien connues, qui ne sont pas plus utiles qu'intéressantes.

Pour surmonter cette limite, une amélioration est proposé par (Bogorny *et al.*, 2008) pour éliminer les règles bien connues en géographie en utilisant une base de connaissance priori. Cette méthode a été mise en œuvre dans Weka-GDPM.

(Karasova, 2005) a appliqué les règles d'association pour découvrir l'influence éventuelle de certains objets géographiques sur l'occurence d'incidents. Exemple d'une règle obtenue est:

$$bars\ and\ restaurants => incidents\ (1.7\%;\ 40.0\%)$$

Cette règle signifie "Au cours de la période étudiée, un incident a eu lieu dans le voisinage de près de 40% des bars et des restaurants du centre-ville de Helsinki "

3.2 Le SIG est utilisé comme un outil de visualisation, d'analyse et d'interprétation des résultats du data mining

Le data mining a été utilisé avec succès dans de nombreux domaines tels que le marketing, la finance, l'environnement, la santé ou encore le transport. Du fait de l'importance croissante accordée au rôle de l'utilisateur pour l'extraction des connaissances, la visualisation est devenue un composant majeur dans le processus du data mining. Elle contribue à l'efficacité du processus de data mining en offrant aux utilisateurs des représentations intelligibles et en facilitant l'interaction (Han et Kamber, 2006). Le choix des techniques de visualisation dépend de la nature des données. (MacEachren et Kraak, 2001) ont estimé que 80% de l'ensemble des données numériques générées aujourd'hui comprennent des références géospatiales (par exemple, les adresses, les codes postaux, les coordonnées géographiques, etc.). Cependant, les techniques de visualisation traditionnelles du data mining n'ont pas été conçues pour tenir compte des caractéristiques des données géospatiales. Cela fait ressurgir le besoin d'intégrer les techniques de géovisualisation au processus d'extraction des connaissances. Dans cette approche, les SIG seront utilisés comme outil de géovisualisation. Les recherches sur cette approche sont encore rares. Nous discuterons ensuite de deux des recherches principales traitant de ce sujet: une de (MacEachren *et al.*, 1999) , l'autre de (Andrienko et Andrienko, 1999).

La géovisualisation n'est pas simplement un outil de visualisation. D'après (Wachowicz, 2001), elle traite de l'utilisation des visualisations géographiques pour explorer les données, puis générer des hypothèses et construire des connaissances. Selon (MacEachren *et al.*, 1999), les trois tâches majeures de la géovisualisation sont l'identification, la comparaison et l'interprétation. Nous présentons ici une approche d'intégration de la géovisualisation au processus du data mining proposé par (MacEachren *et al.*, 1999). Cette approche se décompose en trois niveaux:

conceptuel, opérationnel et mise en œuvre.

Niveau conceptuel: Il s'agit de déterminer les objectifs du processus de construction de la connaissance. Pour ce faire, il faut préciser:

- Quelle sont les types de données à exploiter (par exemple, données environnementales, données socio-démographiques, données chronologiques).
- Quelle sont les types de résultats requis (par exemple, groupe, règles d'association, génération d'hypothèses).
- Qui sont les utilisateurs qui s'intéressent au processus de construction de la connaissance (par exemple, l'expert du domaine).

A ce niveau, ni les techniques du data mining ni les techniques de visualisation ne sont décidées.

Niveau opérationnel: Ce niveau consiste à déterminer les méthodes appropriées du data mining et celles de la géovisualisation pour atteindre les objectifs du niveau conceptuel.

Niveau de la mise en œuvre: A ce niveau, le choix des outils du data mining et de la géovisualisation sont fait. Les méthodes du data mining sélectionnées dans le niveau conceptuel sont exécutes et les techniques de la géovisualisation sont ensuite appliquées aux résultats du data mining pour construire des représentations visuelles et interactives avec les utilisateurs. La mise en œuvre vise donc à intégrer différentes fonctionnalités dans un seul environnement informatique.

(Andrienko et Andrienko, 1999) ont intégré les techniques du data mining (classification et règles d'association) avec les techniques de visualisation cartographique. L'intégration est développée en se basant sur deux systèmes : Kepler pour exécuter les techniques du data mining et Descartes pour visualiser. Ils ont appliqué la visualisation cartographie à deux étapes du data mining : l'étape de prétraitement pour la sélection des attributs appropriés et l'étape de déploiement pour l'analyse et l'interprétation des résultats. Le système Kepler contient ses propres techniques

de présentation des résultats du data mining, il sera donc productif de faire un lien dynamique entre l'affichage de Kepler et celle de Descartes. (Andrienko et Andrienko, 1999) considèrent trois types de liens entre le data mining et la visualisation cartographique:

- De "géographie" à "mathématique": l'utilisation de cartes dynamiques, l'utilisateur arrive à certains résultats de géographie interprétables ou à des hypothèses et essaie alors de trouver une explication à ce résultat.
- De "mathématique" à "géographie": les résultats produits par les techniques du data mining sont ensuite analysés afin de les visualiser sur les cartes.
- Affichage liée: la visualisation non cartographique de résultats du data mining est affichée en même temps que les cartes. Les deux types d'affichage sont reliés visuellement et dynamiquement.

(Torun et Duzgun, 2006) présentent une application de l'intégration des techniques de géovisualisation aux résultats du data mining. Ils indiquent les endroits vulnérables dus au transport de pétrole et aux feux de pétrole/gaz à Istanbul Strait. Ils ont appliqué les méthodes de segmentation K-moyenne et ISODATA(Iterative Self-Organizing Data Analysis Techniques) pour respectivement segmenter les accidents de navires et les facteurs de vulnérabilité . Les facteurs de vulnérabilité sont divisées en 5 groupes. Les cartes produits par les techniques de superposition des couches et de "hot-spot" de SIG permettent d'interpréter les résultats. La figure 3.3 montre les chevauchements des zones vulnérables et des régions ayant une fréquence d'accidents de navire élevée. Trois accidents sont détectés sur cinq locations en utilisant l'analyse de données spatiales. En regardant la carte, (Torun et Duzgun, 2006) ont noté que deux régions au sud sont peuplés et ont eu des accidents graves.

Figure 3.3 Exemple de l'intégration de la géovisualisation aux data mining (Torun et Duzgun, 2006)

3.3 Les systèmes actuels de l'intégration entre le data mining et les SIG

Présentement, il y a deux groupes de recherche principaux qui contribuent au développement des systèmes d'intégration entre le data mining et les SIG. L'une du *Database Research Lab - Simon Fraser University* à Vancouver avec le système GeoMiner. L'autre groupe est l'equipe de l'extraction des connaissances du *German National Research Center for Information Technology* avec le système SPIN!. Le GeoMiner développé par (Han *et al.*, 1997) était le premier système qui intègre le data mining et les SIG. Il est une combinaison de DBMiner et de MapInfo. La figure 3.4 présente l'architecture générale du GeoMiner.

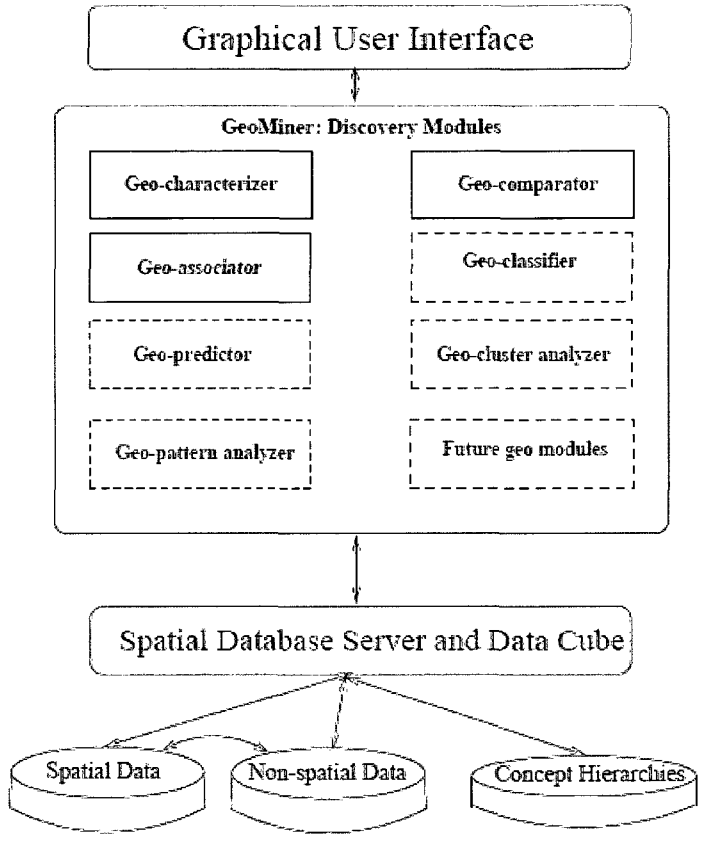


Figure 3.4 Architecture générale de GeoMiner (Han *et al.*, 1997)

Le système GeoMiner contient 3 modules principaux:

- le module de construction de cubes de données spatiales.

- le module OLAP(Online Analyticial Processing)pour analyser les données spatiales.

- le module de techniques du data mining telles que la segmentation, la classification et les règles d'association spatiale (voir figure 3.4).

Le SIG MapInfo est utilisé comme serveur de la base de données spatiale ainsi qu'en tant que interface graphique et de visualisation de cartes. Le système permet de visualiser les résultats sous forme d'une relation généralisée ou d'une carte. Cependant, la visualisation par carte présente dans l'etat actuel des fonctionnalités

basiques et ne tire pas le meilleur parti de MapInfo.

Le SPIN! (Spatial Mining for data of public interest) est un projet de recherche géré par Eurostat (May et Savinov, 2002). L'objectif du SPIN! est de développer un système spatial basé web qui intègre les techniques du data mining et d'un SIG dans une architecture ouverte et extensible. SPIN! est une combinaison des systèmes Kepler et Descartes. L'architecture de SPIN! est présentée dans la figure 3.5.

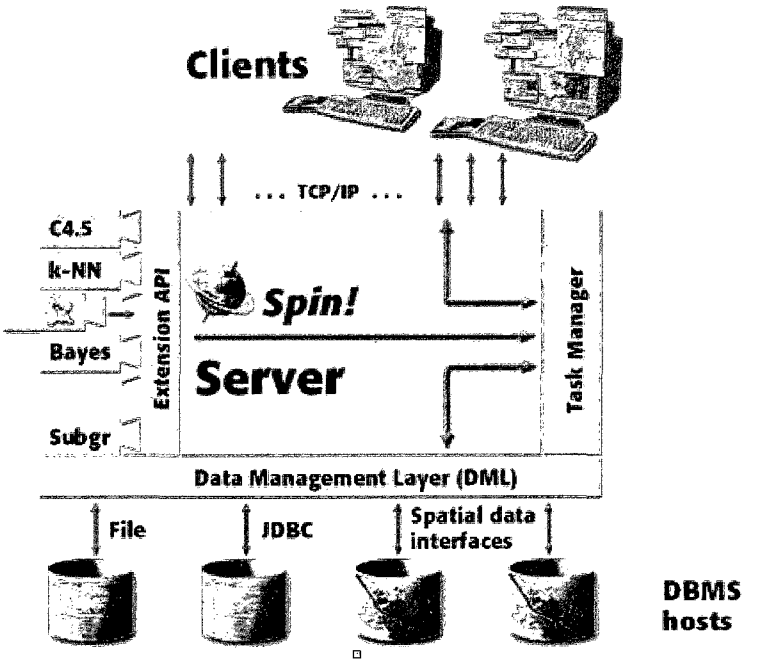


Figure 3.5 Architecture de SPIN! (May et Savinov, 2002)

Les fonctionnalités principales de SPIN! peuvent être regroupées en 5 niveaux:

- Niveau 1: Gestion des données: Il contient les fonctionnalités de l'accès aux données des sources homogènes et de prétraitement des données en se basant sur la plate-forme du data mining Kepler.
- Niveau 2: Affichage des cartes sur l'internet a l'aide de Descartes.
- Niveau 3: Cartographie thématique interactive pour visualiser les données statis-

tiques.

- Niveau 4: Détection des segments spatiaux.

- Niveau 5: Explication des segments et des phénomènes spatiaux.

3.4 Conclusion

Dans ce chapitre, nous avons présenté brièvement les travaux de recherche actuels portant sur l'intégration entre le data mining et les SIG. Les recherches de ce domaine sont encore récentes. Elles présentent donc un domaine de recherche potentiel. Avec l'augmentation continue du volume des données spatiales dans tous les domaines, cette intégration devient une recherche critique et a besoin des contributions de chercheurs provenant de différentes disciplines : géographie, statistiques, data mining, visualisation et bases de données.

Le développement des systèmes d'intégration entre le data mining et les SIG est encore limité. En effet, les deux systèmes principaux GeoMiner et SPIN! ne sont pour le moment utilisés que dans les laboratoires. En outre, les logiciels SIG tels que ArcGIS, MapInfo ont commencé à intégrer quelques méthodes du data mining dans leurs produits. Cependant, la plupart de ces méthodes font parties des techniques de segmentation. Les autres techniques telles que les règles d'association ou la classification ne sont pas encore intégrées dans ces logiciels. Cette limite explique le manque des études de cas empirique de cette intégration.

CHAPITRE 4

ÉTUDE DE CAS EN ANALYSE LE MARCHÉ DES MEUBLES AUX ÉTATS-UNIS

Notre cas d'étude est basé sur une collaboration avec Canadel, une entreprise de meuble canadienne. Notre objectif est de développer un système d'intégration des techniques de data mining avec les techniques de SIG pour analyser le marché des meubles aux États-Unis. En premier lieu, nous introduirons le contexte du marché des meubles américain et ainsi le contexte actuel de Canadel. Nous conduirons ensuite une revue de littérature concernant la segmentation du marché. Puis, nous proposerons notre méthodologie basée sur trois grandes étapes: l'étape de data mining, l'étape de SIG et l'étape de l'intégration. Les résultats obtenus et les conclusions seront discutés à la fin de ce chapitre. Ces résultats ont été publiés dans les conférences IESM'09 à Montréal, Canada (Le *et al.*, 2009b) et CIGI'09 à Tarbes, France en 2009 (Le *et al.*, 2009a).

4.1 Introduction

Dans cette section, nous présenterons d'abord les contraintes sur le marché des meubles américain qui justifient ce projet. Ensuite, nous analyserons le contexte de l'entreprise pour pouvoir proposer notre approche.

4.1.1 Les contraintes sur le marché américain

Dans le livre "Gestion du marketing", (Filion et Colbert, 1990) ont mentionné deux contraintes jouant un rôle majeur en déterminant les caractéristiques et les besoins sur tous les marchés: la concurrence et les variables du macro-environnement. Dans les parties suivantes, nous allons donc les analyser pour le marché des meubles américain.

4.1.1.1 La concurrence

Aux États-Unis, la concurrence est la première contrainte qui dirige le marché. Le marché américain des meubles est aujourd'hui très ouvert aux produits étrangers. Avec des importations provenant actuellement de plus de 120 pays, il est le marché le plus grand et le plus concurrentiel au monde. Le marché des meubles américain est le marché d'exportation le plus important pour l'industrie canadienne. Ceci est dû, entre autres, à une situation géographique particulière ainsi qu'aux conditions favorables créées par l'ALENA (Accord de Libre-Échange Nord-Américain) depuis 1994. En effet, plus de 95 % des meubles exportés du Canada le sont vers les États-Unis (Murillo, 2007).

Les exportations canadiennes de meubles vers les États-Unis ont connu une décennie de grand succès entre 1990 et 2000. Plus particulièrement, sur la période allant de 1990 à 1998, le Canada en était le premier fournisseur, suivi par la Chine et le Mexique. Depuis 2000, la part canadienne du marché des meubles américain a cependant baissé, tandis que la part chinoise (et plus généralement asiatique) ne cesse d'augmenter. En 2006, la Chine avait remplacé le Canada en tant que premier exportateur de meubles vers les États-Unis avec près de 50% de parts du marché. Le Canada se retrouve à la deuxième place avec une part de 16% (Murillo, 2007). La figure 4.1 illustre l'évolution des exportations sur le marché des meubles américain.

L'industrie canadienne du meuble fait face à la plus grande compétition qu'elle ait connue depuis 30 ans. Trouver les états potentiels et analyser la compétition sont des processus nécessaires pour toutes les entreprises canadiennes de meubles afin de défendre leur position sur ce marché.

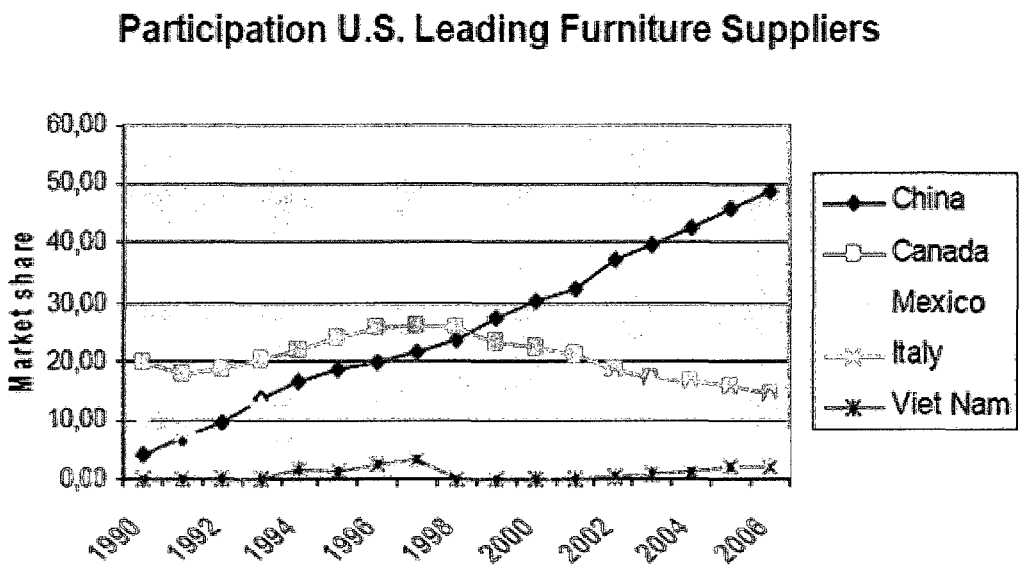


Figure 4.1 Participation des grands fournisseurs de fourniture aux États-Unis (Murillo, 2007)

4.1.1.2 Les variables du macro-environnement

Des variables incontrôlables affectent de façon continue la vie d'une entreprise. (Filion et Colbert, 1990) les distinguent en 5 groupes : l'environnement démographique, l'environnement culturel, l'environnement économique, l'environnement politico-légal et l'environnement technologique. Nous résumons ici quelques caractéristiques importantes du marché américain.

- ***L'environnement géographique et démographique***

  Les caractéristiques démographiques telles que l'âge, le sexe, la composition

ethnique des différentes communautés sont des dimensions qui influencent les comportements d'achat de la population. Les États-Unis sont composés de 50 états et se classent au quatrième rang des pays les plus vastes du monde avec une superficie de 9,4 millions de kilomètres carrés, derrière la Russie, le Canada et la Chine. En 2008, les États-Unis comptaient plus de 302 millions d'habitants et étaient le troisième pays le plus peuplé du monde. Ils sont aussi le pays ayant le nombre d'immigrants le plus élevé. Selon les statistiques de 2005, plus de 36 millions d'habitants sont nés à l'étranger et ont immigrés aux États-Unis. Cela a aussi amené une natalité plus forte que dans les autres pays développés (Bureau of Labor Statistics, 2009).

- ***L'environnement culturel***

  Les facteurs culturels, y compris les valeurs, les idées, les attitudes, les croyances et les activités de groupes de population influencent aussi le comportement d'achat des consommateurs. La culture anglo-saxonne est la base de la culture américaine. La langue officielle est l'anglais. Cependant, les différents immigrants apportent leur propre culture et font des États-Unis un creuset culturel (Ottaviano et Peri, 2006).

- ***L'environnement économique***

  Dans le monde des affaires, l'environnement économique prend une place très importante. Le niveau de revenu, le rythme de l'inflation et le taux de chômage influencent fortement le pouvoir d'achat des consommateurs. Étant la première puissance économique mondiale, les États-Unis se placent à la huitième place mondiale pour le produit national brut par habitant et à la quatrième place à parité de pouvoir d'achat. En 2007, le taux de chômage est relativement faible par rapport au taux moyen de chômage mondial, soit entre 3 et 5 % de la population active (U.S. Census Bureau, 2009). Les États-Unis sont les plus grands importateurs de produits manufacturés et les deuxièmes

exportateurs derrière l'Allemagne. Le Canada, la Chine, le Mexique, le Japon et l'Allemagne sont ses principaux partenaires commerciaux. Ces caractéristiques économiques rendent les États-Unis le marché d'exportation le plus important pour plusieurs pays.

4.1.2 Le contexte de l'entreprise

Fondée en 1982, Canadel est le plus important manufacturier d'ameublement de cuisine et de salle à manger au Canada. Après son entrée dans le marché américain en 1992, le marché américain devenait le marché principal de Canadel, qui représente plus de 85% des vente de l'entreprise. Depuis 2005, les ventes de Canadel ont commencé à baisser sur le marché américain même si la demande dans ce marché continuait d'augmenter.

Canadel, comme la plupart des entreprises des meubles, ne dispose pas de ses propres magasins et doit s'en remettre aux canaux de distribution existants. Les canaux de distribution traditionnels dans l'industrie du meuble passent par des représentants des fabricants, des grossistes ou des détaillants. Canadel ne fabrique que des meubles de luxe sur demande. La distribution est par conséquent exclusivement basée sur un réseau de représentants. Chaque représentant peut couvrir une ou plusieurs zones géographiques. Sur les zones pour lesquelles les gammes de meubles s'écoulent rapidement, il peut y avoir plusieurs représentants. Un représentant touche une commission sur le prix de vente par produit vendu. Leur rôle est de déterminer quels magasins de détail sont les plus susceptibles de réaliser des ventes optimales pour les produits de l'entreprise. Les ventes de l'entreprise dépendent totalement des magasins dans le réseau des représentants. Dans le contexte compétitif actuel du marché du meuble américain, la détermination des magasins cibles est importante pour les représentants comme pour l'entreprise.

4.1.3 Notre approche

Au niveau de la stratégie de l'analyse du marché du meuble américain, notre objectif vise à répondre aux demandes suivantes:

- La segmentation du marché selon certains critères.
- La détermination d'un positionnement compétitif et le positionnement du canal de distribution pour analyser la compétition actuelle sur le marché.

La segmentation consiste à décomposer un marché en groupes homogènes pour lesquels les consommateurs présentent des caractéristiques communes. Elle permet à l'entreprise de:

- trouver les facteurs de différenciation les plus importants pour le consommateur et d'identifier les groupes de clients potentiels;
- comprendre les caractéristiques des consommateurs afin de lancer un programme de marketing ou de mettre sur le marché un nouveau produit;
- d'évaluer la répartition des ventes des produits sur des différents segments;
- d'évaluer les performances des vendeurs pour chaque zone de vente.

La détermination d'un positionnement compétitif permet aux représentants:

- de localiser ses compétiteurs sur leurs territoires;
- de déterminer les magasins cibles;
- d'élargir leurs réseaux de magasins pour augmenter les ventes

Pour atteindre ces objectifs, nous proposons de développer un système basé sur le web qui intègre les techniques de segmentation du data mining au SIG pour lesquelles :

- le système web peut faciliter l'accès des utilisateurs;
- les techniques de segmentation du data mining permettent d'identifier les états potentiels pour l'entreprise;
- les résultats de la segmentation et du SIG sont intégrés afin de visualiser et d'analyser les différents facteurs sur ces segments;
- le SIG analyse la compétition du marché en permettant de visualiser la localisation des canaux de distribution de l'entreprise par rapport à ceux de ses compétiteurs.

4.1.4 Revue de littérature de la segmentation

Dans cette section, nous ferons l'état de l'art sur la segmentation du marché. La première partie traitera des avantages et des inconvénients de deux catégories de variables de la segmentation. Dans la deuxième partie, nous discuterons sur le choix entre les données primaires et secondaires dans un projet de segmentation d'un marché. La dernière partie portera sur deux approches de segmentation: l'approche a priori et l'approche a posteriori.

4.1.4.1 Les variables de la segmentation du marché

Pour les analystes, il ne suffit pas de décomposer le marché en segments mais de pouvoir les décrire. Le choix des variables de segmentation peut affecter l'efficacité

opérationnelles des différentes segmentations (Smadja, 1988). Selon (Wedel et Kamakura, 1999), les variables de segmentation du marché peuvent être catégorisées en deux groupes : les variables observables et les variables inobservables.

- ***Groupe des variables observables***

  Ce groupe comprend des variables géographiques, démographiques, ainsi que des variables socioéconomiques. La segmentation basée sur ces variables suppose que les personnes qui ont des caractéristiques démographiques et socioéconomiques similaires ont les mêmes habitudes d'achat. Les variables les plus souvent utilisées sont le revenu, l'âge, le sexe, le nombre de pièces dans la maison, la profession, la langue et la religion. Ces variables présentent de multiples avantages en segmentation de marché (Filion et Colbert, 1990). Premièrement, elles sont observables. Elles facilitent donc la description des consommateurs visés et ainsi la quantification du potentiel de vente. Leur deuxième avantage est la disponibilité des sources des données. En effet, ce type de données a souvent pour origine des sources gouvernementales ou des organismes produisant des statistiques. Ces données sont donc dignes de confiance, faciles d'accès et peu coûteuses. (Weinstein, 1994) a aussi noté que la segmentation sociodémographique fournit un bref aperçu du marché. De par ces avantages, ces variables sont largement utilisées en segmentation.

  Pourtant, ces variables présentent aussi des inconvénients, dont le principal est le manque de qualité. Les sources de données statistiques sont collectées à un intervalle fixe, souvent tous les cinq ans. Elles peuvent alors peut-être être dépassées au moment où va s'effectuer la segmentation. De plus, ces données sont aussi disponibles aux compétiteurs de l'entreprise. Une autre limite des variables socioéconomiques et démographiques est la difficulté de transposer le résultat de la segmentation aux actions marketing (Wedel et Kamakura, 1999).

- ***Groupe des variables inobservables***

  Les variables psychologiques, la personnalité, les styles de vie et les comportements d'achat sont de type inobservable. Ce type de segmentation suppose que les activités, les intérêts, les opinions et le style de vie sont des facteurs importants pour la consommation (Weinstein, 1994). Depuis leur proposition (à l'origine par (Mitchell et McGoldrick, 1963)), les études psychologiques en segmentation ont reçu une grande attention de la part des chercheurs en marketing. Ces études basées sur les travaux réalisés en psychologie ont montré les interrelations entre la personnalité de l'être humain et son comportement d'achat. Le plus grand avantage de ce type de segmentation est sa capacité à transposer le résultat dans la stratégie marketing(Weinstein, 1994). Néanmoins, les données telles que les styles de vie, la personnalité ou les intérêts sont souvent collectées par des sondages ou des enquêtes de terrain. Elles sont donc subjectives, très coûteuses et prennent beaucoup de temps à être collectées. (Majurin, 2001) a aussi souligné que ces variables sont difficiles à mesurer.

Conclusion

(Wedel et Kamakura, 1999) ont noté que le style de vie, le comportement d'achat ainsi que la personnalité peuvent fournir une base d'action particulièrement utile pour le développement de la publicité. Cependant, la plupart des recherches montrent que ces variables ne sont ni stables ni adaptées pour certains types de marchés. Le tableau 4.1 présente la comparaison de ces deux groupes de variables en se basant sur les critères d'évaluation de la segmentation. Dans la plupart des cas, plusieurs variables peuvent être utilisées ensemble, en utilisant les avantages de chacun, pour obtenir le résultat le plus efficace.

Tableau 4.1 Comparaison des variables observables et inobservables (Wedel et Kamakura, 1999)

Critères	Identifiabilité	Substantialité	Accessibilité	Stabilité	Capacité à être transformer en actions	Réactivité
Observable: Démographique, Géographique, Etc.	Très bien	Très bien	Très bien	Très bien	Faible	Faible
Inobservable: Psychologique, Comportement, Etc.	Modéré	Bien	Faible	Modéré	Bien	Faible

4.1.4.2 Les sources des données pour la segmentation

On ne peut pas réaliser la segmentation sans données. Dans cette partie, nous décrirons les avantages et les inconvénient des sources de données pour réaliser la segmentation. Les données utilisées pour la segmentation du marché peuvent être collectées soit à partir de sources primaires, soit à partir de sources secondaires.

- ***Les sources primaires***

  Les données primaires sont des données recueillies dans un but de recherche précis. Normalement, les données primaires ne sont pas disponibles telles quelles au moment du besoin et nécessitent un effort particulier de collecte par des sondages, des entrevues, des observations et des questionnaires, etc. Elles présentent les avantages suivants:

- Elles sont plus récentes et plus spécifiques que les données secondaires.
- Elles sont recueillies spécialement pour répondre à un besoin particulier, elles sont donc plus représentatives.

Pourtant, le temps et le coût pour effectuer les sondages et les requêtes sur un grand marché sont très élevés. De plus, les résultats obtenus par ces méthodes sont subjectifs et difficiles à évaluer. La collecte de données primaires est donc très coûteuse en temps et en ressources. Une fois que les données primaires sont utilisées, elles peuvent être stockées pour un usage futur. Elles deviennent alors des données secondaires pour un autre objectif.

- ***Les sources secondaires***

  En général, les données secondaires sont des données qui ont été collectées par d'autres. Elles sont donc déjà disponibles parce qu'elles ont été produites pour d'autres objectifs (Castleberry, 2001). Ces données peuvent être fournies par des moyens externes (sources gouvernementales) ou par des moyens internes comme les données historiques de l'entreprise (les données de facturation, les données de vente, etc). (Patzer, 1995) et (Aaker *et al.*, 2001) ont souligné l'importance des sources de données secondaires dans la recherche du marketing. Ils citent les avantages de l'utilisation des sources secondaires par rapport aux données primaires :

  - La collecte des données secondaires est moins coûteuse que celle des données primaires. Le temps nécessaire à la collecte des données secondaires est de loin inférieur à celui requis pour acquérir des données primaires.
  - Dans certains cas, les sources secondaires peuvent fournir des données plus précises que celles obtenues par la recherche primaire. Les données statistiques sur la population fournies par les gouvernements sont plus fiables que celles d'autres sources.

- Les données secondaires sont importantes pour définir le problème et établir des hypothèses avant de collecter des données primaires. Elles peuvent aussi aider à caractériser la population d'un marché.

Cependant, elles impliquent certains risques (Filion et Colbert, 1990):

- La qualité des données secondaires est parfois difficile à contrôler.
- Il faut comprendre ces données pour pouvoir sélectionner celles qui sont pertinentes et informatives pour le projet.
- Les données secondaires peuvent être périmées au moment auquel la segmentation est effectuée.

4.1.4.3 Les méthodes de la segmentation du marché

Les méthodes de segmentation peuvent être considérées dans deux grandes catégories : les méthodes dites de segmentation a priori et celles dites de segmentation a posteriori (Wedel et Kamakura, 1999) (voir aussi (Green, 1977) et (Filion et Colbert, 1990)) comme représenté dans la figure 4.2.

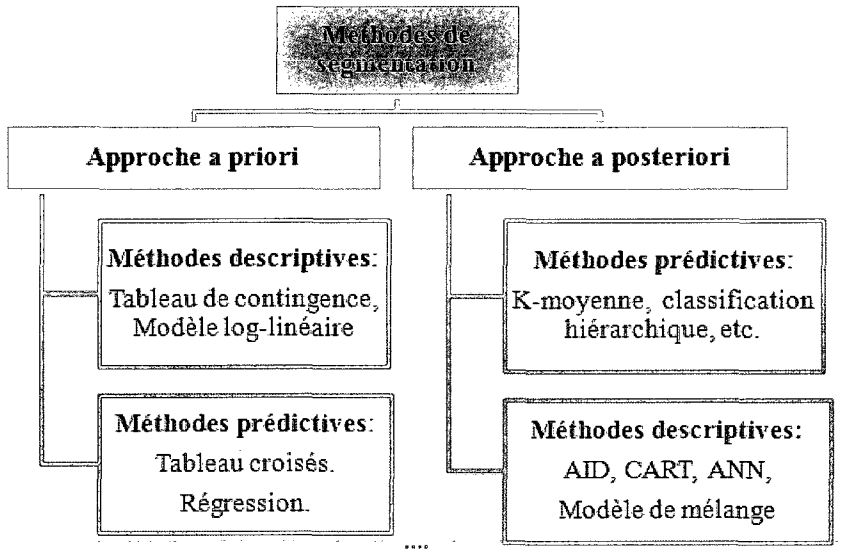


Figure 4.2 Catégoriser les méthodes de segmentation (Wedel et Kamakura, 1999)

Les sections suivantes vont détailler les éléments essentiels de cette classification.

- ***L'approche a priori***

  Dans l'approche a priori, le marché est segmenté selon des critères préexistants tels que l'âge, le sexe ou le statut économique et social. Les méthodes ont alors pour objectif de découvrir ou de décrire les caractéristiques des clients dans les segments déjà connus. Les méthodes de modèle log-linéaire, de tableau de contingence, des tableaux croisés et de la régression sont de type a priori. Cette approche est efficace dans le cas où les variables d'analyse sont bien définies. Cependant, il est parfois difficile d'identifier les variables à utiliser. Les méthodes a priori sont donc utilisées soit dans des cas particuliers, quand on dispose a priori d'une certaine connaissance du domaine, soit en combinaison avec des méthodes a posteriori.

- ***L'approche a posteriori***

  A l'inverse de l'approche a priori, les méthodes de cette approche ont pour

objectif de découvrir des segments du marché inconnus jusqu'a présent. Le nombre de segments et les caractéristiques de chaque segment sont déterminés par les données et les méthodes utilisées. En effet, les méthodes a posteriori sont plus utilisées que les méthodes a priori. Les méthodes les plus connues dans cette approche sont l'analyse discriminante, la segmentation par partitionnement, la segmentation hiérarchique, les réseaux de neurones, etc (Wedel et Kamakura, 1999). Dans leur revue de littérature des méthodes de segmentation, (Punj et Stewart, 1983) ont conclu que les méthodes de segmentation par partitionnement sont plus efficaces que les méthodes hiérarchiques en pratique car elles sont faciles à mettre en œvre et peuvent s'appliquer à divers types de données.

4.2 La méthodologie

Nous proposons un modèle comprenant trois étapes principales comme présenté à la figure 4.3. L'étape de data mining permet d'identifier les groupes d'états homogènes parmi les 50 états des États-Unis. L'étape de SIG consiste à utiliser les services de Google Map pour analyser les canaux de distributions sur une carte géographique. La dernière étape aboutit à l'intégration du data mining et du SIG pour produire les résultats finaux. Dans cette étape, la technique de géo-visualisation permet de visualiser et de supporter les résultats du data mining.

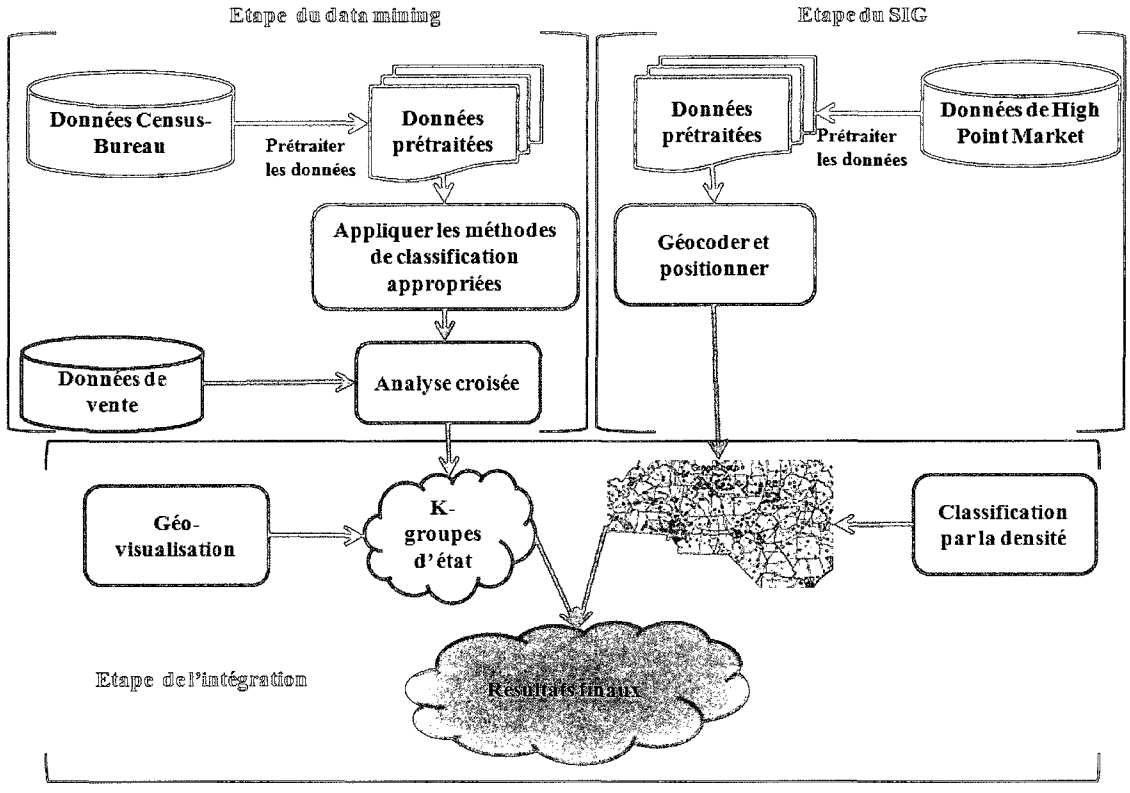


Figure 4.3 Représentation simplifiée de la méthodologie

4.2.1 L'étape de data mining

Cette étape est basée sur le processus du data mining, elle se compose donc en cinq tâches majeures ci-dessous:

- Identifier les variables et les sources de données
- Collecter les données.
- Pré-traiter les données.
- Appliquer les méthodes de segmentation du data mining.
- Analyse croisée avec les données de vente.

4.2.1.1 Identifier les variables et les sources de données

Cette étape vise à déterminer les variables et les sources de données pour la segmentation, ainsi que les sources de données pour l'analyse de compétition. Le choix des variables de segmentation et les sources de données sont basées sur les caractéristiques du marché, le niveau et l'objectif de la segmentation. Pour la segmentation du marché des meubles américain, les variables économiques et sociodémographiques sont les plus appropriée pour les raisons suivantes:

- D'après la revue de littérature, les caractéristiques économiques et sociodémographiques sont des contraintes importantes sur un marché (Filion et Colbert, 1990).
- La littérature confirme que le marché des meubles en général est fortement corrélé avec les critères de logement et de l'économie (Volpe et Peluso, 2007).
- Les experts de l'entreprise ont souligné aussi l'importance des critères économiques et sociodémographiques dans l'analyse du marché de haut niveau.
- En particulier sur le marché américain, les données statistiques démographiques et socioéconomiques sont librement accessibles sur le site web de l'U.S Census Bureau.

Nous utiliserons donc les données de l'U.S Census Bureau pour effectuer la segmentation. Le U.S Census Bureau est une source des données fiables et gratuites. Les entreprises peuvent non seulement bénéficier de données statistiques générales sur les États-Unis, mais aussi de données statistiques sur divers domaines (affaires, transports, commerce extérieur, etc.).

4.2.1.2 Collection des données

Les données du Census Bureau:

Le U.S Census Bureau est l'organisme de statistique le plus grand des États-

Unis. Il rassemble des statistiques sur le pays, les citoyens et l'économie tous les cinq ans en effectuant des recensements et des enquêtes. Au moment où nous réalisons ce projet, les données les plus récentes du Census Bureau datent de 2005. Pour chaque état, les données sont organisées en 4 groupes:

- Groupe démographique : chiffre la population par sexe, par classe d'âge, groupe ethnique et par situation familiale (type de famille, nombre d'enfants et leur âge).
- Groupe social : niveau d'étude, état matrimonial, lieu de naissance, langue parlée.
- Groupe économique : revenu personnel, revenu familial, profession, situation d'emploi, catégorie socioprofessionnelle.
- Groupe de logements : type de logement, nombre de pièces, nombre de chambres à coucher, nombre de véhicules, statut de propriétaire ou locataire, valeur du logement.

Figure 4.4 montre un exemple de données récupérées pour l'Alabama.

General Demographic Characteristics: 2005
Data Set: 2005 American Community Survey
Survey: 2005 American Community Survey
Geographic Area: Alabama

NOTE. Data are limited to the household population and exclude the population living in institutions, college dormitories, and other group quarters. For information on confidentiality protection, sampling error,

General Demographic Characteristics: 2005	Estimate	Margin of Error
Total population	4,442,558	*****
SEX AND AGE		
Male	2,142,901	+/-3,564
Female	2,299,657	+/-3,564
Under 5 years	292,375	+/-2,206
5 to 9 years	290,406	+/-8,159
10 to 14 years	318,043	+/-8,349
15 to 19 years	295,128	+/-3,500
20 to 24 years	301,972	+/-4,325
25 to 34 years	583,691	+/-5,884
35 to 44 years	640,564	+/-6,315
45 to 54 years	652,826	+/-3,926

Demographic | Social | Economic | Housing

Figure 4.4 Les données Census Bureau de l'Alabama

Les données de vente de l'entreprise:

Le système informatique actuel de l'entreprise ne permet pas d'intégrer d'autres sources de données. Nous devrons récupérer les données de vente en 2007 et 2008 sous le format Excel. Les variables dans l'ensemble de données originales incluaient les données facturées et le détail des magasins partenaires. Ces données ont été ramassées en vue d'opérer l'analyse croisée avec le résultat de la segmentation.

4.2.1.3 Pré-traiter les données

Les données collectées sont ensuite traitées pour pouvoir les exploiter par les algorithmes de data mining. Après une analyse initiale, nous procédons aux tâches ci-dessous:

Réduire l'espace des données:

Les données statistiques du Census Bureau contiennent plus de 150 variables.

Toutes ces variables ne sont pas utiles pour analyser le marché des meubles américain. En travaillant avec les experts du domaine, parmi 150 variables, nous en avons choisi 60 pouvant influencer le marché des meubles.

Transformer les données:

Les données du Census Bureau sont des données générales pour le marché américain, incluant le marché des meubles. La transformation a pour but de produire des données adaptées à l'objectif et aux algorithmes utilisés. En se basant sur les échelles définies par l'entreprise, nous réalisons des transformations telles que :

- Grouper la population selon des échelles d'âge, de salaire ou de nombre de pièce dans la maison.
- Créer des ratios de pièces dans la maison, de foyer, de salaire, etc.

La figure 4.5 illustre des données après la transformation.

Etat	Population	% de 35 a 60 ans	% famille en foyer sur nb foyer	....	% population dans la meme maison	Couple marié de 35 à 65 ans	% taux Employée
Alabama	4,442,558	35.29%	68.41%	....	82.70%	32.0%	7.83
Alaska	641,724	37.87%	67.39%	....	80.24%	34.7%	0.53
Arizona	5,829,839	32.02%	66.22%	....	76.76%	28.6%	5.75
Arkansas	2,701,431	34.26%	68.27%	....	80.31%	30.5%	4.94
....	....	....	....	....	....	....	....
West Virg	1,771,750	37.26%	67.52%	....	87.26%	33.9%	3.35
Wisconsin	5,375,751	36.87%	64.91%	....	84.12%	33.3%	6.92
Wyoming	495,226	37.41%	65.34%	....	80.68%	35.4%	0.66

Figure 4.5 Exemple de données du Census Bureau après le prétraitement

Intégrer les données:

Pour faciliter les prochaines étapes, il est nécessaire d'intégrer les données dans une seule base de données. Nous utilisons Microsoft Access pour stocker les données. Dans notre contexte, les données du Census Bureau ne changent

que tous les cinq ans, la méthode de l'intégration matérialisée est donc appropriée. Nous avons créé une base de données Access qui intègre les données traitées du Census Bureau, les données de vente ainsi que les données de High Point Market. Il est évident que les données venant des sources différentes n'ont pas une relation claire et définie. Nous créons de nouveaux tableaux pour relier les données de ces sources. Par exemple, le tableau État avec la clé ÉtatID pour connecter les données du Census Bureau, les données des magasins de l'entreprise et les données de High Point Market. La figure 4.6 illustre le schéma relationnel de notre base de données.

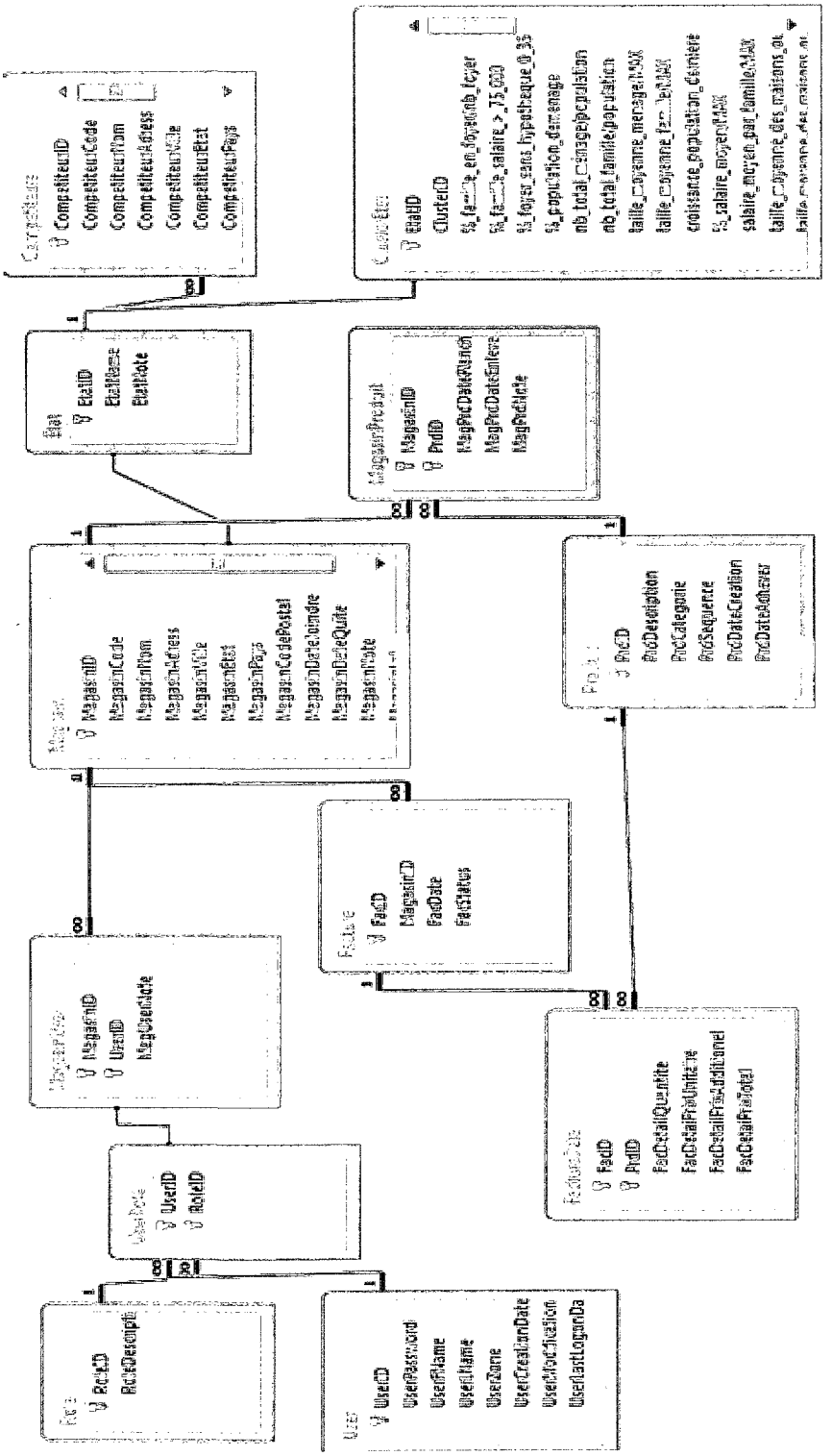


Figure 4.6 Schéma relationnel de base de données

4.2.1.4 Appliquer les techniques de segmentation du data mining

Dans notre étude de cas, l'objectif est d'identifier des groupes d'états homogènes d'un point de vue socioéconomique. Nous ne connaissons pas encore les critères, ni le nombre de segments. Les méthodes a posteriori sont donc les plus appropriées dans ce contexte. Le nombre de segments est un paramètre difficile à déterminer dans les méthodes par partitionnement. Pour trouver un nombre raisonnable de segment, plusieurs méthodes sont proposées telles que : les tableaux croisés, les tests de permutation et le ré-échantillonnage. Par contre, ces méthodes ne fonctionnent pas très bien en pratique avec des données volumineuses ou multidimensionnelles (Salvador et Chan, 2004). La combinaison de méthodes hiérarchiques et méthodes des k-moyennes est une approche efficace pour ce type de données (Singh, 1990).

Nous avons donc utilisé la méthode hiérarchique pour déterminer le nombre raisonnable de segment, puis la méthode des k- moyennes pour la détermination des segments. Pour exécuter ces méthodes, nous utilisons TANAGRA (TANAGRA, 2008), un outil open source de data mining. Les paragraphes suivants vont détailler ces méthodes.

Méthode de segmentation hiérarchique ascendante

La segmentation hiérarchique ascendante part d'une partition où chaque donnée représente un segment et à chaque itération les deux segments les plus proches sont fusionnés jusqu'à ce que tous les points se trouvent dans un seul grand segment. Cette méthode retourne le résultat sous la forme d'un arbre, souvent appelé dendrogramme (voir figure 4.7). Elle permet de choisir le nombre de segments de façon optimale par les indicateurs comme le ratio BSS (Between Sum of Square) et la valeur de GAP statistique. Le BSS pour la somme des carrés entre les moyennes ou aussi appelée la somme des carrés inter-groupes. Le ratio BSS mesure la dissimilarité entre les segments, tandis

que la valeur de GAP statistique mesure le compacité des segments (Tibshirani *et al.*, 2001). Le nombre de segment est approprié quand la valeur de ratio BSS et la valeur de GAP statistique sont maximales.

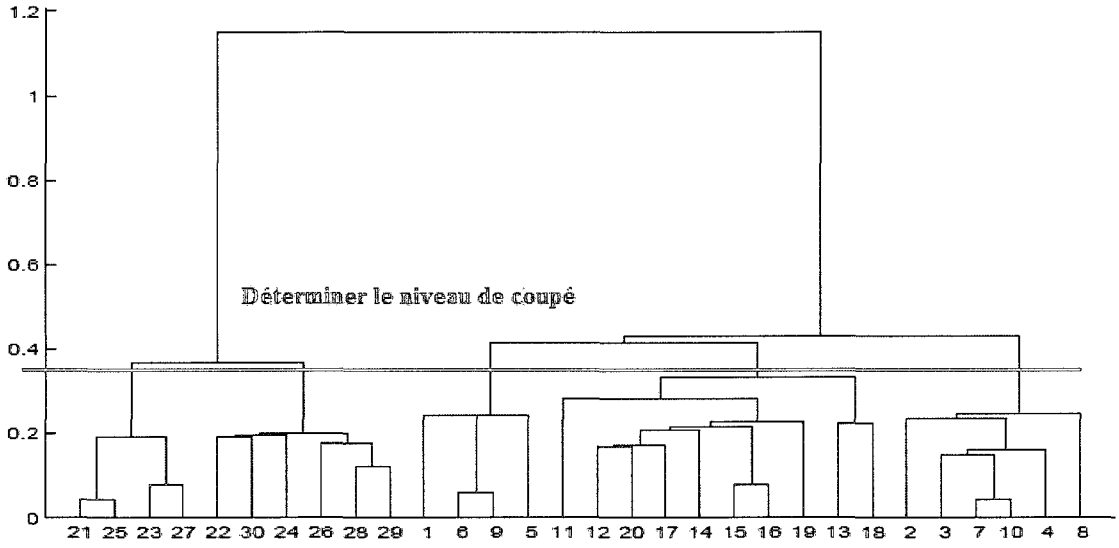


Figure 4.7 Dendrogramme d'une segmentation hiérarchique ascendante

Méthode des k-moyennes

Dans la méthode des k-moyennes de (MacQueen, 1966), le nombre de segments voulu (K) doit être connu, il est déterminé par la méthode de segmentation hiérarchique ascendante que nous avons expliquée avant. Cette méthode peut être décrite par les phases suivantes (K est donné):

- Phase d'initialisation: Partitionner l'ensemble de donnees en K sous-ensembles (non vide).
- Phase 1: Calculer des centres de K groupes de la partition courrants.
- Phase 2: Affectuer chaque individu au groupe dont le centre leur est le plus proche.

Réitérer les phases 1 et 2 jusqu'à ce que les segments soient stables, c'est-à-dire lorsqu'on ne constate plus de changement d'individus.

4.2.1.5 Analyse croisée avec les données de vente

[Berry et Linoff, 1997] ont remarqué que le résultat de la segmentation doit être intégré à d'autres méthodes pour en dégager la signification. La plupart des applications de la segmentation du marché s'arrêtent à la fin de la méthode de segmentation, c'est-à-dire à la constitution des groupes homogènes. Cependant, il est difficile de prendre une quelconque décision à partir de ce résultat partiel. Nous proposons ici un modèle intégrant le résultat de la segmentation du marché aux données historiques de ventes en vue d'aider une entreprise à prendre des décisions de développement à partir des groupes formés. En analysant les relations entre les données historiques de ventes et les groupes formés, le modèle aide à déterminer les endroits stratégiques. Il sert ensuite à établir le tableau de bord pour l'entreprise à partir des résultats du data mining.

4.2.2 L'étape de SIG

Parmi tous les SIG libres, nous avons choisi d'utiliser Google Map pour les raisons suivantes:

- Les API (Application programming interface ou en français interface de programmation applicative) de Google Map permettent de mettre en place divers services de géo-localisation.
- Les API de Google Map permettent d'ajouter des contrôles personnalisés pour modifier l'affichage de la carte.
- Les API de Google Map permettent d'extraire des données "géocodées " à partir d'une adresse.
- Google Map est facile à intégrer avec les applications web et compatible avec la plupart des navigateurs graphiques.

L'étape de SIG comprend trois tâches majeures qui seront abordées ci-dessous.

4.2.2.1 Collecter les données

En ce qui concerne les données pour l'étape de SIG, nous utiliserons les données fournies par High Point Market et les données de vente de l'entreprise. L'International Home Furnishings Market (salon des fournitures et accessoires d'ameublement de maison) de High Point est le plus grand salon de l'industrie du meuble du monde. Il attire tous les grands magasins des quatres coins du monde deux fois par année, en avril et en octobre. Les participants peuvent y obtenir la liste des magasins de l'industrie mondiale du meuble. Les données de vente sont utilisées pour extraire les adresses des magasins partenaires avec l'entreprise. Les données obtenues du High Point Market sont dans une liste, sous format Excel (voir figure 4.8). Cette liste contient le nom, l'adresse, le code postal, le pays, le type de magasins ainsi que les catégories de vente de 22076 magasins, dont 19190 magasins aux États-Unis en 2008.

stnCompI	stCompD	stCompE	stAddress	stAddress	stCity	stState	stNonUS	stZip	stCountry	stPhone	stFax	StStoreCatDesc	stStoreTy	stPricePo	stVolumeDesc
36569	1	1	1543 Queens Rd		Charlotte	NC		28207	United Sta	704.375.34	704.378.20	Furniture	Antiques	Upper End	Under .5 Million
50463	1 Contemp	gill@con	27 Mellichamp Dr Ste		Bluffton	SC		29910	United Sta	843.706.93	843.706.93	Furniture	Furniture S	Upper End	Under .5 Million
87769	1 Selling a	fleemarke	858 Lakewood Rd		Waterbury	CT		06704	United Sta	203.575.01	203.575.01	None Selected	Furniture S	None Sel	None Selected
72735	10 KA Inte	kayalar@	Dalboyu S	06180 Ste	Ankara				Turkey	+90.312.35	+90.312.35	Furniture	Retail Sto	None Sel	None Selected
42320	101 Home	Furnishing	1301 E Wade	Hampt	Green	SC		29651	United Sta	864.877.90	864.877.90	Furniture	Furniture S	Medium/H	.5 Million to 1 Million
31033	1212 Mart	ndaniels26	101 Shadow Valley P		High Point	NC	NC	27262	United Sta	336.889.12	336.889.65	Furniture	Antiques	Medium	Under .5 Million
52959	125 West	[illegible]@	1275 West	Suite 320	Annapolis	MD		21401	United Sta	410.295.39	410.295.39	Furniture	Catalog/M	Upper End	2 Million to 5 Million
80433	1502 Fabri	customers	2103 Dunmore Ct		High Point	NC		27263	United Sta	336.434.21	336.434.43	None Selected	Design Sh	Medium/H	5 Million to 10 Million
29150	1670 Hous	mschbeale	P O Box 1063		Southold	NY		11971	United Sta	631.765.20	631.765.22	Furniture	Furniture S	Upper End	.5 Million to 1 Million
81935	1743 Pine	nathy22@	1743 McDonald Ave		Brooklyn	NY	NY	11230-690	United Sta	718-382-11	718-645-44	None Selected	Departme	Medium/H	2 Million to 5 Million
59197	1840s Cor	AnnePorz	PO Box 119		Stevenson	MD		21153-011	United Sta	410.486-24	410.486.17	Design	Design Sh	Medium/H	5 Million to 10 Million
24693	1890 General	Store A	10894 NC Highway 1		Banner Elk	NC	NC	28604	United Sta	828.963.53	828.963.53	Accessory	Gift/Acc	Medium/H	Under .5 Million
74033	2 Burn Inc	[illegible]@	6817 W Lynx Ave		Milwaukee	WI	WI	53225	United Sta	414.727.12	414.727.12	None Selected	Catalog/M	Medium	1 Million to 2 Million
21672	2 Danes	2danes@	73 White Bridge Rd	S	Nashville	TN		37205	United Sta	615.352.60	615.352.60	Furniture	Furniture S	Upper End	2 Million to 5 Million
24633	2 Design	kmodes@	927 W Washington St		Eden	NC		27288	United Sta	336.635.21	336.635.21	Design	Design Sh	Medium	Under .5 Million
20536	2 Design	2design@	31 Bedrock Circle		Santa Cla	UT	UT	84765	United Sta	435.673.97	435.673.93	Furniture	Furniture S	Medium/H	.5 Million to 1 Million
83312	2 Modern	deanna@	21212 Binghampton	(	Huntingtor	CA		92646	United Sta	638.240.53	714.593.64	None Selected	Online Ret	None Sel	None Selected
85524	2 of a Kind	Design	660 Brown Trail		Asheboro	NC		27205	United Sta	336.301.0275		None Selected	Design Sh	Medium	Under .5 Million

Figure 4.8 Exemple de données de High Point Market

4.2.2.2 Pré-traiter les données

Nous effectuons les deux tâches suivantes:

Réduire l'espace des données:

Pour éliminer les données qui ne sont pas nécessaires dans les données de High Point Market, nous les filtrons et choisissons 6953 magasins de meubles aux États-Unis parmi 22076 magasins dans la liste. Nous avons donc diminué l'espace des données à traiter de 2/3.

Nettoyer les données:

Les données redondantes et aberrantes sont les erreurs les plus courantes dans les données de High Point. Ceci est généralement dû à des erreurs de saisie lors de l'entrée les données. Un cas commun est présenté à la figure 4.9: l'adresse est la même, mais avec une façon de l'écrire différente ("St" et "Street").Pour ce cas, nous utilisons des requêtes en Microsoft Access pour éliminer les magasins redondants selon leur nom et leur code postal.

nomclient	adresse2	Ville	Pays	Etat	ZIP
Bon-Ton	3585 S Church St	Whitehall, PA, USA	USA	PA	18052
Bon-Ton	3585 S Church St	Whitehall, PA, USA	USA	PA	18052
Bon-Ton	3585 S Church St	Whitehall, PA, USA	USA	PA	18052
Bon-Ton	3585 S Church St	Whitehall, PA, USA	USA	PA	18052
Bon-Ton	3585 S Church St	Whitehall, PA, USA	USA	PA	18052
Bon-Ton	3585 S Church St	Whitehall, PA, USA	USA	PA	18052
Bon-Ton	3585 S Church St	Whitehall, PA, USA	USA	PA	18052
Bon-Ton	3585 S Church St	Whitehall, PA, USA	USA	PA	18052
Bon-Ton	3585 S. Church Street	Whitehall, PA, USA	USA	PA	18052
Bon-Ton	3585 S. Church Street	Whitehall, PA, USA	USA	PA	18052
Bon-Ton	3585 S. Church Street	Whitehall, PA, USA	USA	PA	18052
Bon-Ton	3585 S. Church Street	Whitehall, PA, USA	USA	PA	18052
Bon-Ton	3585 S. Church Street	Whitehall, PA, USA	USA	PA	18052
Bon-Ton	3585 S. Church Street	Whitehall, PA, USA	USA	PA	18052

Est ce que c'est le même magasin

Figure 4.9 Exemple de données redondantes

4.2.2.3 Géocoder et localiser

Cette étape contient deux modules qui permettent de géocoder puis de localiser les magasins sur une carte géographique de Google Map(voir figure

4.10) .

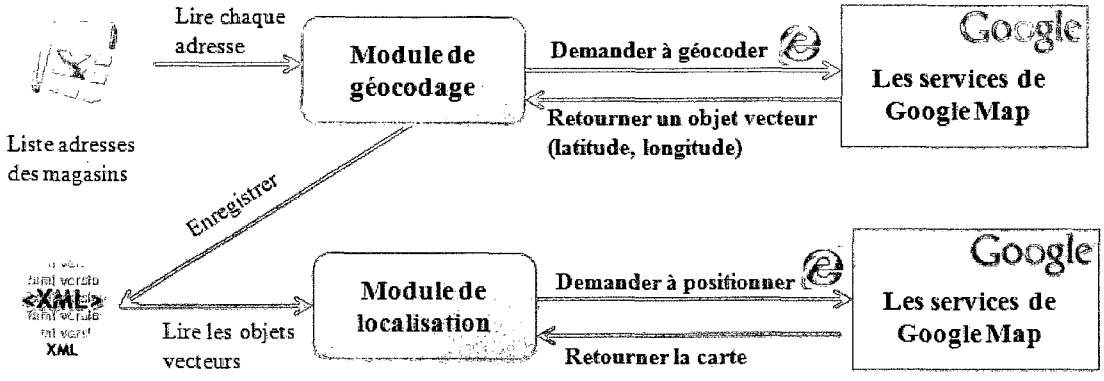
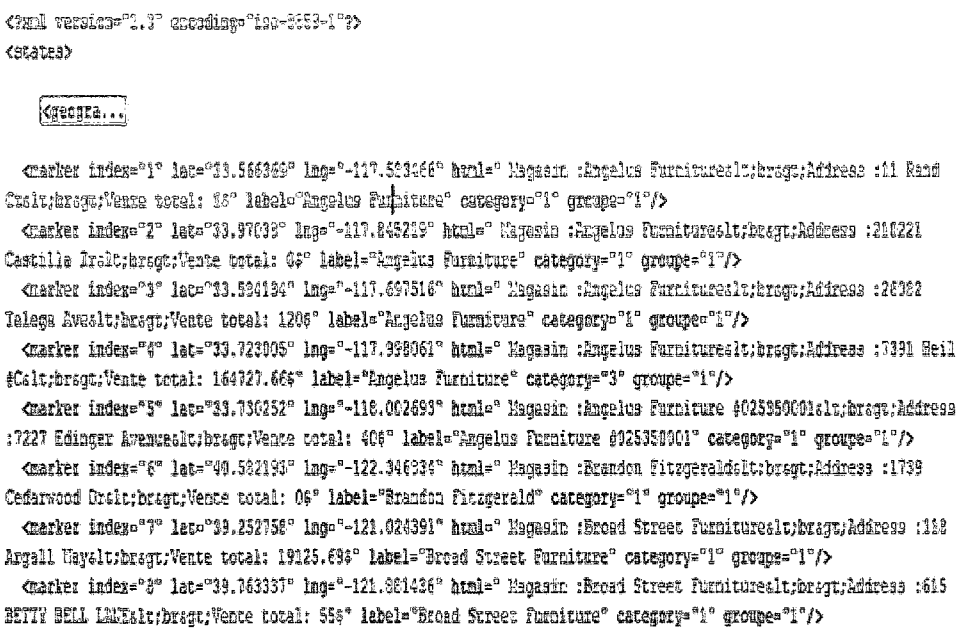


Figure 4.10 Modules de l'étape de SIG

Module de géocodage

Le service de géocodage de Google Map convertit une adresse en un objet vecteur des positions X,Y (latitude, longitude). En effet, la précision du positionnement est fortement liée à la manière dont est renseignée l'adresse géocodée. Le module de géocodage lit chaque adresse dans les la liste des magasins et l'envoie au service de Google Map. Les adresses géocodées sont ensuite enregistrées dans la base de données pour s'en resservir lors des prochaines utilisations. La figure 4.11 montre un exemple de données géocodées. Par exemple, le magasin Angelus Furniture à l'adresse "11 Rand Ct, Trabuco Canyon, CA 92679, USA" est associé au couple de coordonnées : latitude = 33.566369 et longitude = -117.583466



<?xml version="1.0" encoding="iso-8859-1"?>
<states>
<geogra...
<marker index="1" lat="33.566369" lng="-117.803466" html="Magasin :Angelus Furniture
Address :11 Rand C
Vente total: 18" label="Angelus Furniture" category="1" groupe="1"/>
<marker index="2" lat="33.97039" lng="-117.845219" html="Magasin :Angelus Furniture
Address :210221 Castilla Dr
Vente total: 0$" label="Angelus Furniture" category="1" groupe="1"/>
<marker index="3" lat="33.594194" lng="-117.697516" html="Magasin :Angelus Furniture
Address :26362 Telega Ave
Vente total: 1206" label="Angelus Furniture" category="1" groupe="1"/>
<marker index="4" lat="33.723005" lng="-117.998061" html="Magasin :Angelus Furniture
Address :7391 Heil #C
Vente total: 164727.66$" label="Angelus Furniture" category="3" groupe="1"/>
<marker index="5" lat="33.730252" lng="-118.002693" html="Magasin :Angelus Furniture 0025950901
Address :7227 Edinger Avenue
Vente total: 406" label="Angelus Furniture 0025950901" category="1" groupe="1"/>
<marker index="6" lat="49.582193" lng="-122.346934" html="Magasin :Brandon Fitzgerald
Address :1739 Cedarwood Dr
Vente total: 0$" label="Brandon Fitzgerald" category="1" groupe="1"/>
<marker index="7" lat="39.252758" lng="-121.024391" html="Magasin :Broad Street Furniture
Address :118 Argall Way
Vente total: 19125.69$" label="Broad Street Furniture" category="1" groupe="1"/>
<marker index="8" lat="39.763337" lng="-121.861436" html="Magasin :Broad Street Furniture
Address :615 BETTY BELL LANE
Vente total: 55$" label="Broad Street Furniture" category="1" groupe="1"/>

Figure 4.11 Exemple des données géocodées par le service de géocodage du Google Map

Module de localisation

Une fois toutes les adresses géocodées et enregistrées dans un fichier XML, le module de localisation permet de lire ce fichier et d'envoyer ces données vers le service de localisation de Google Map. Puis, il affiche la carte de localisation des magasins de l'entreprise et ceux de ses compétiteurs. Cette carte sert à visualiser le positionnement du canal de distribution de l'entreprise et à analyser la concurrence.

4.2.3 Étape de l'intégration du résultat de classification et du SIG

Les deux étapes précédentes peuvent être effectuées de manière indépendante. Cette étape est le pont entre les résultats du data mining et ceux du SIG. Il est évident que la visualisation joue un rôle très important pour la compréhension du résultat du data mining. Dans cette étape, nous utilisons la

fonctionnalité de superposition des couches du SIG pour visualiser le résultat de la segmentation puis analyser les attributs importants de chaque segment. Le résultat est ensuite croisé avec la carte de localisation des magasins de l'étape du SIG pour analyser la concurrence dans chaque segment.

Pour effectuer la superposition des couches, nous employons les données des couches d'information géographique des États-Unis (routes, frontières, point d'intérêt, etc.) provenant de TIGER/Line 2000 produites par le U.S. Census Bureau. Ces données permettent de déterminer les frontières des villes, des états, etc à divers niveaux de précision.

Cependant, une des limites actuelles de Google Map est qu'il ne peut pas encore travailler directement avec le format de données de TIGER/Line. Nous avons donc besoins de convertir les données de TIGER/Line en format XML. Le fichier XML présenté dans la figure 4.12 a été simplifié : il contient seulement les informations des frontières des états aux États-Unis.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<states>
  <state name="Alabama" colour="#000088" Population="4442558">...</state>
  <state name="Alaska" colour="#000088" Population="641724">...</state>
  <state name="Arizona" colour="#000088" Population="5829839">...</state>
  <state name="Arkansas" colour="#000088" Population="2701431">...</state>
  <state name="California" colour="#ff0000" Population="35278768">...</state>
  <state name ="Colorado" colour="#00ff88" Population="4562244">
    <point lat="37.0004" lng="-109.0448"/>
    <point lat="36.9949" lng="-102.0424"/>
    <point lat="41.0006" lng="-102.0534"/>
    <point lat="40.9996" lng="-109.0489"/>
    <point lat="37.0004" lng="-109.0448"/>
  </state>
  <state name ="Connecticut" colour="#00ff88" Population="3394751" >
    <point lat="42.0498" lng="-73.4875"/>
    <point lat="42.0511" lng="-73.4247"/>
    <point lat="42.0371" lng="-72.8146"/>
    <point lat="41.9983" lng="-72.8174"/>
    <point lat="42.0044" lng="-72.7638"/>
    <point lat="42.0360" lng="-72.7563"/>
```

Figure 4.12 Données simplifiées des frontières des états aux États-Unis

4.3 Résultats

Le système web permet aux utilisateurs de naviguer entre les résultats de chaque étape et d'interpréter ces résultats. Nous présenterons ensuite les résultats obtenus correspondant avec les trois étapes.

4.3.1 Résultats du data mining

4.3.1.1 Résultats de la segmentation

Le résultat de la méthode hiérarchique nous retourne un dendrogramme, tel qu'illustré à la figure 4.13. Ensuite, la valeur du ratio BSS et la valeur de GAP statistique sont calculées pour chaque niveau de coupe. Nous avons déterminé qu'un niveau de coupe formant 3 segments correspond à la valeur maximale de GAP, tandis qu'un niveau de coupe formant 4 segments correspond à la valeur maximale du ratio BSS. Par conséquent, le nombre de segments approprié peut soit être égal à 3 ou à 4.

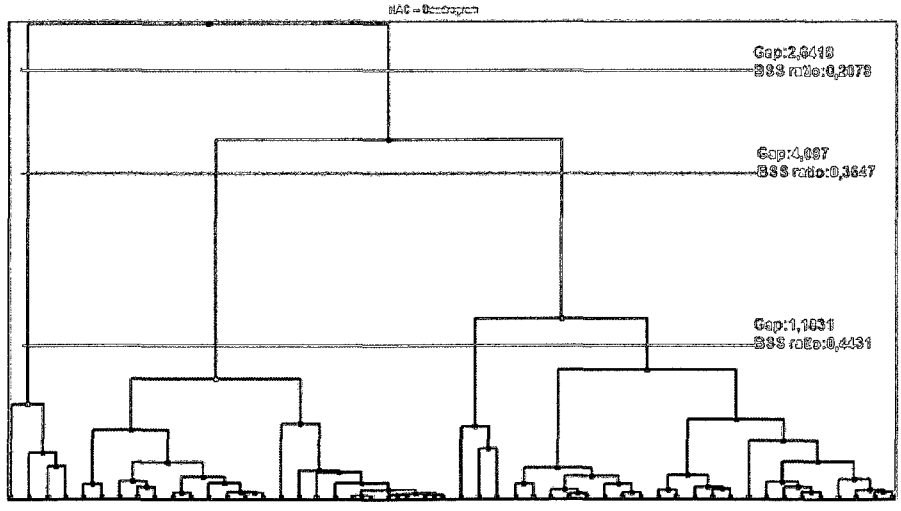


Figure 4.13 Résultat de la classification hiérarchique

Nous exécutons ensuite la méthode des k-moyennes avec un nombre de segment égal à 3 puis à 4. Après quelques itérations et confrontations avec les experts de l'entreprise, nous sommes arrivés à un partitionnement en trois groupes qui couvre l'ensemble du territoire des États-Unis (voir figure 4.14)

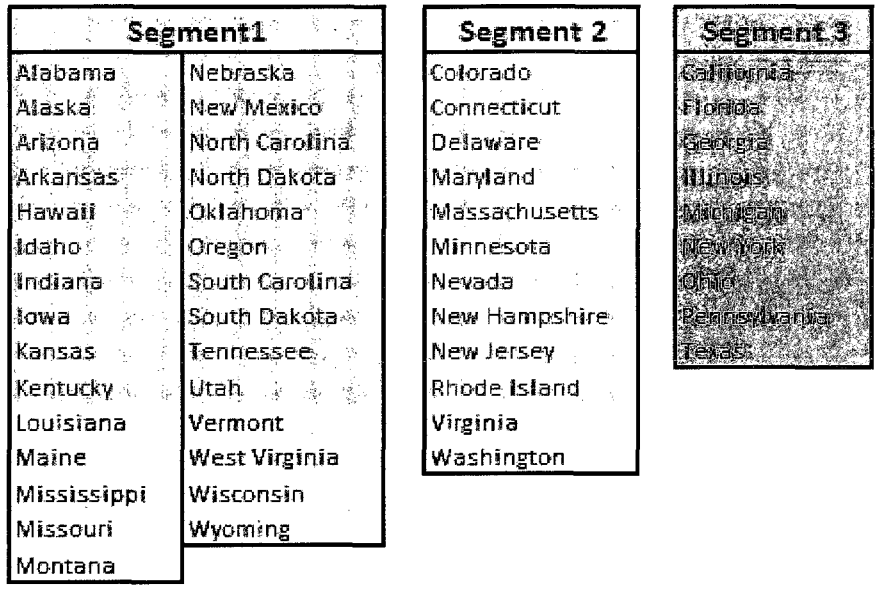


Figure 4.14 Résultat de la segmentation

À partir des données sociodémographiques du Census Bureau, nous arrivons à l'analyse suivante :

Groupe 1 : "Majorité de classe moyenne inférieure":

Le groupe 1 regroupe la majorité des états (29 états sur 50). Ces états couvrent un vaste territoire. Le salaire moyen par personne de la population de ces états est le plus faible des États-Unis, tout comme la part des familles aux salaires élevés. La majorité de la population y gagne peu d'argent : moins de 29% du salaire moyen max (sur l'ensemble des états) (voir figure 4.15). Ces caractéristiques nous font catégoriser le groupe 1 comme la "classe moyenne inférieure".

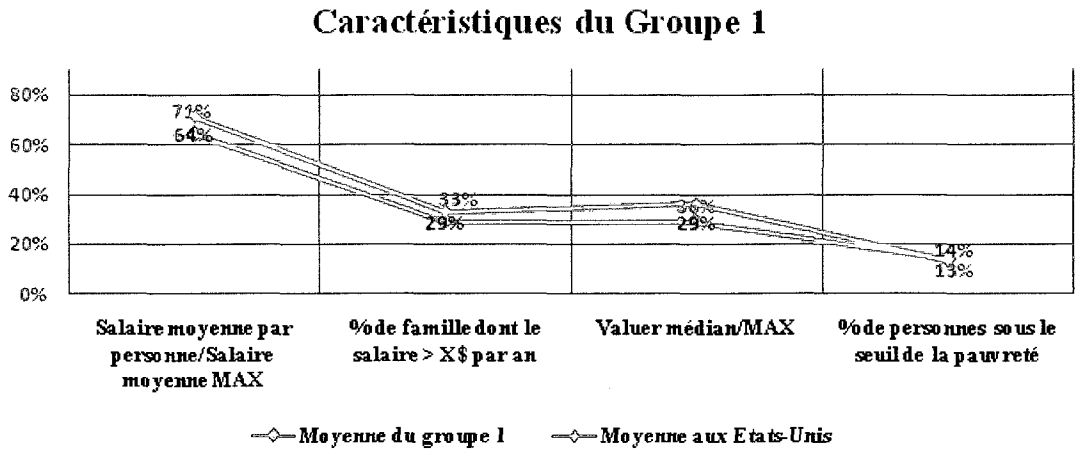


Figure 4.15 Caractéristiques du groupe 1

Groupe 2 : "Majorité de classe moyenne supérieure":

Le groupe 2 regroupe 12 états situés principalement à l'est. Le pourcentage de familles avec des salaires supérieurs à la moyenne y est plus élevé (43% au lieu de 33 %) que sur l'ensemble des États-Unis. La majorité des personnes dans ces états (>50% - médiane) gagnent plus que dans les autres états et le salaire moyen par personne y est plus élevé (voir figure 4.16). Ces caractéristiques nous font catégoriser ce groupe comme la "classe moyenne supérieure".

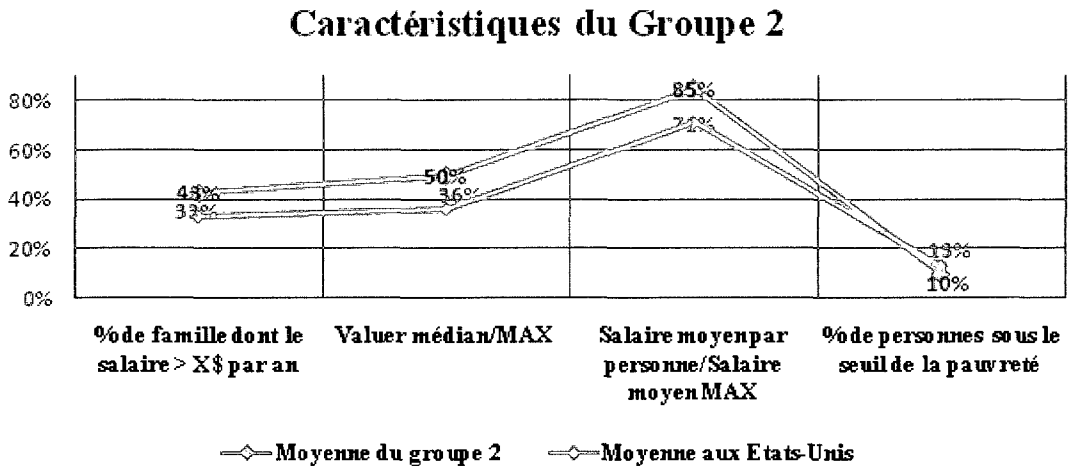


Figure 4.16 Caractéristiques du groupe 2

Groupe 3 : "Majorité de classe moyenne":

Le groupe 3 regroupe les états les plus peuplés. Avec 9 états principalement localisés autour des Grands lacs plus la Floride et la Californie, il est caractérisé par une majorité de foyers sans hypothèque et par de très grandes maisons. De plus, le nombre de foyers avec un bon revenu y est supérieur à la moyenne américaine (voir figure 4.17). Ces caractéristiques nous font catégoriser ce groupe comme la "classe moyenne".

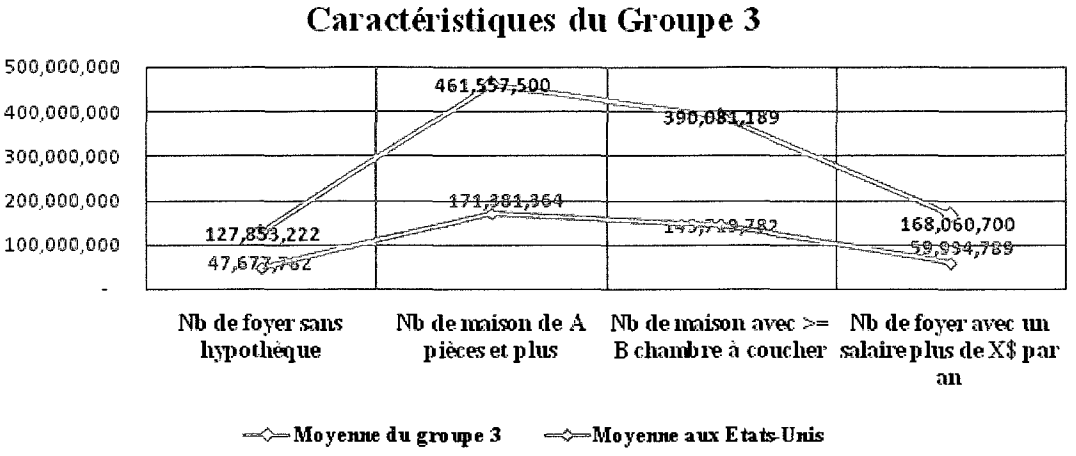


Figure 4.17 Caractéristiques du groupe 3

4.3.1.2 Résultat de l'analyse croisée avec les données de vente

Pour les dirigeants, les analyses des segments précédentes ne sont pas suffisantes pour déterminer le groupe potentiel. L'analyse croisée des trois segments avec les données historiques de vente de l'entreprise a pour but de soutenir cette détermination .

Nous constatons que les ventes du groupe 1 sont plus faibles même s'il regroupe 29 états. Par contre, le groupe 3 compte seulement 9 états mais les ventes y sont les plus élevées (près de 3 fois celles du groupe 1). Les 9 états du groupe 3 semblent les marchés potentiels (voir figure 4.18).

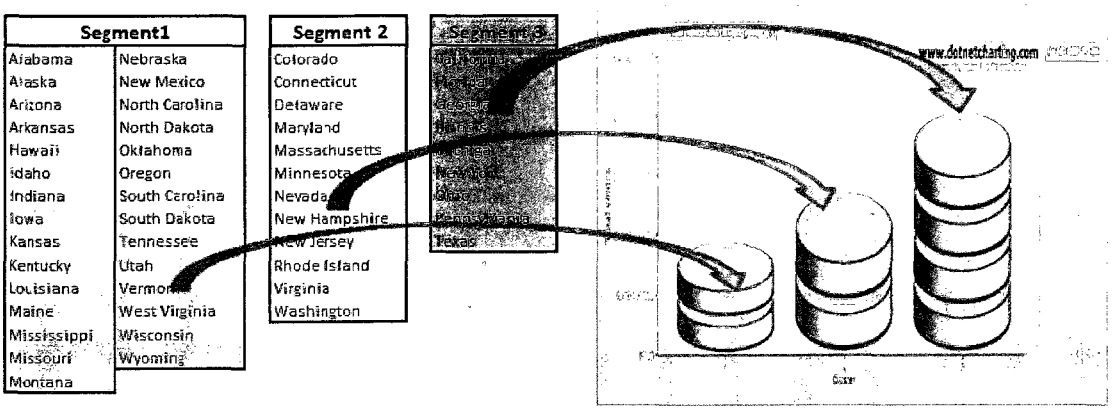


Figure 4.18 Résultat de l'analyse croisée avec les données de vente

En se basant sur les segments formés et les données de vente, nous construisons par la suite un tableau de bord qui mesure la performance de vente des groupes par des indicateurs de performance. L'indicateur ci-dessous est un exemple de ce tableau de bord.

- Quantités de chaque type de produit doivent être augmentées dans chaque catégorie

Cet indicateur est basé sur l'hypothèse que les états d'un même groupe peuvent atteindre le même volume de produits vendus. Pour chaque état, cet indicateur indique le volume d'augmentation que chaque catégories devrait atteindre pour de parvenir à égaler le meilleur état de son groupe (voir figure 4.19).

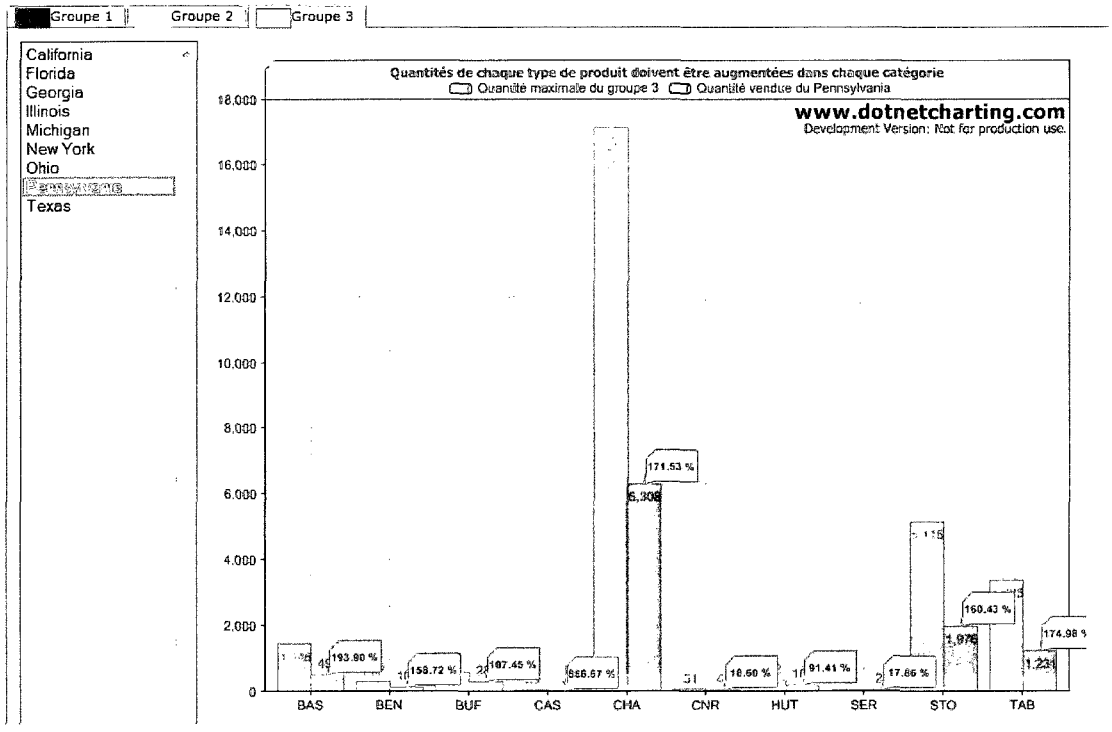


Figure 4.19 Augmentation désirée de quantité de chaque type de produit à Pennsylvania

4.3.2 Résultats du SIG

Les résultats finaux de l'étape du SIG sont les cartes. Dans notre cas, ce sont des cartes localisant les magasins de l'entreprise et de ses compétiteurs dans chaque état. Les cartes fournissent aussi des options pour filtrer les magasins selon certain critères. La figure 4.20 illustre la carte de localisation des magasins de la Californie.

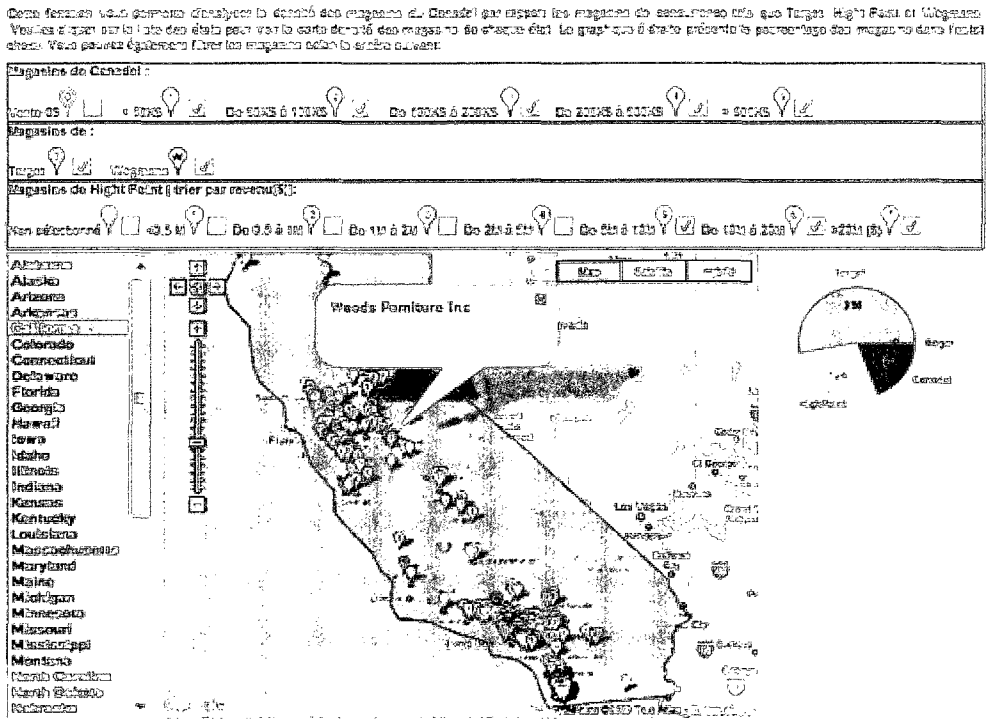


Figure 4.20 Carte de localisation des magasins de Californie

4.3.3 Résultats de l'intégration

En utilisant la superposition de couche pour visualiser les segments, nous obtenons la carte 4.21. Les 3 groupes trouvés sont représentés sur la carte en trois couleurs : le groupe 1 par le rouge, le groupe 2 par le bleu et le groupe 3 par le cyan. En regardant la carte des segments, nous constatons que les 9 états du groupe 3, sont principalement regroupés autour des Grands Lacs en incluant la Floride et la Californie. Le groupe 2 contient 12 états, surtout situés le long de la côte nord-est. Les 29 autres états du groupe 1 sont répartis sur le territoire.

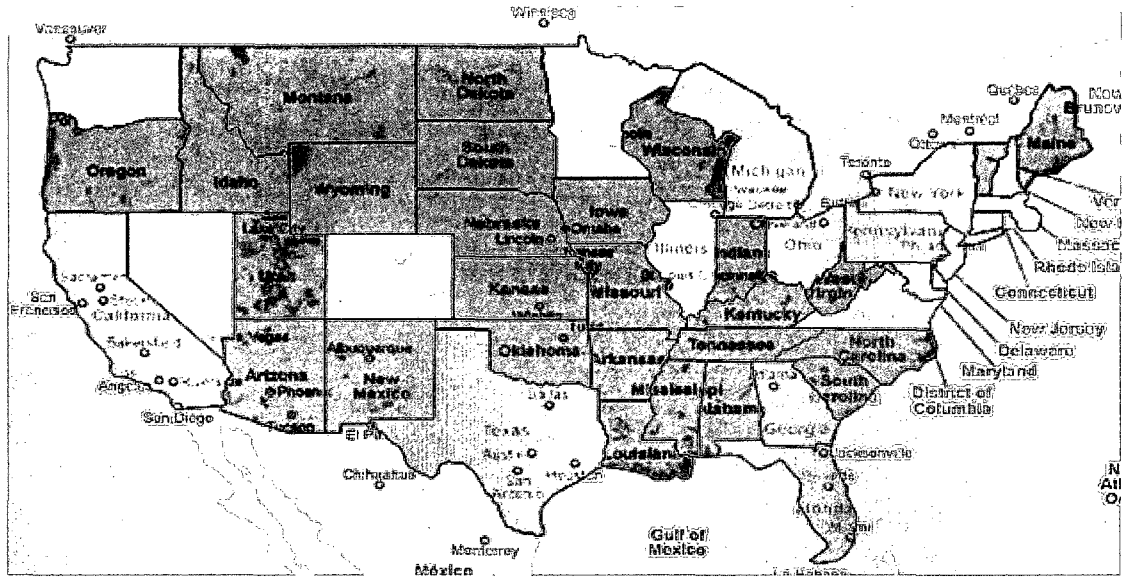


Figure 4.21 Carte des segments

La technique de superposition de couche du SIG sert aussi aux utilisateurs pour visualiser les différents facteurs des 3 segments par différentes couches. Par exemple, la figure 4.22 présente consécutivement les cartes des facteurs importants pour les 3 segments : population, éducation, revenus et logement. En analysant ces cartes, nous constatons que la plupart des états du groupe 3 sont densément peuplés, ce qui induit le fait que le nombre de maisons dans ces états est le plus élevé aux États-Unis. D'une autre part, le groupe 2 a le plus haut niveau d'éducation, ce qui implique que son revenu moyen soit supérieur par rapport aux autres groupes.

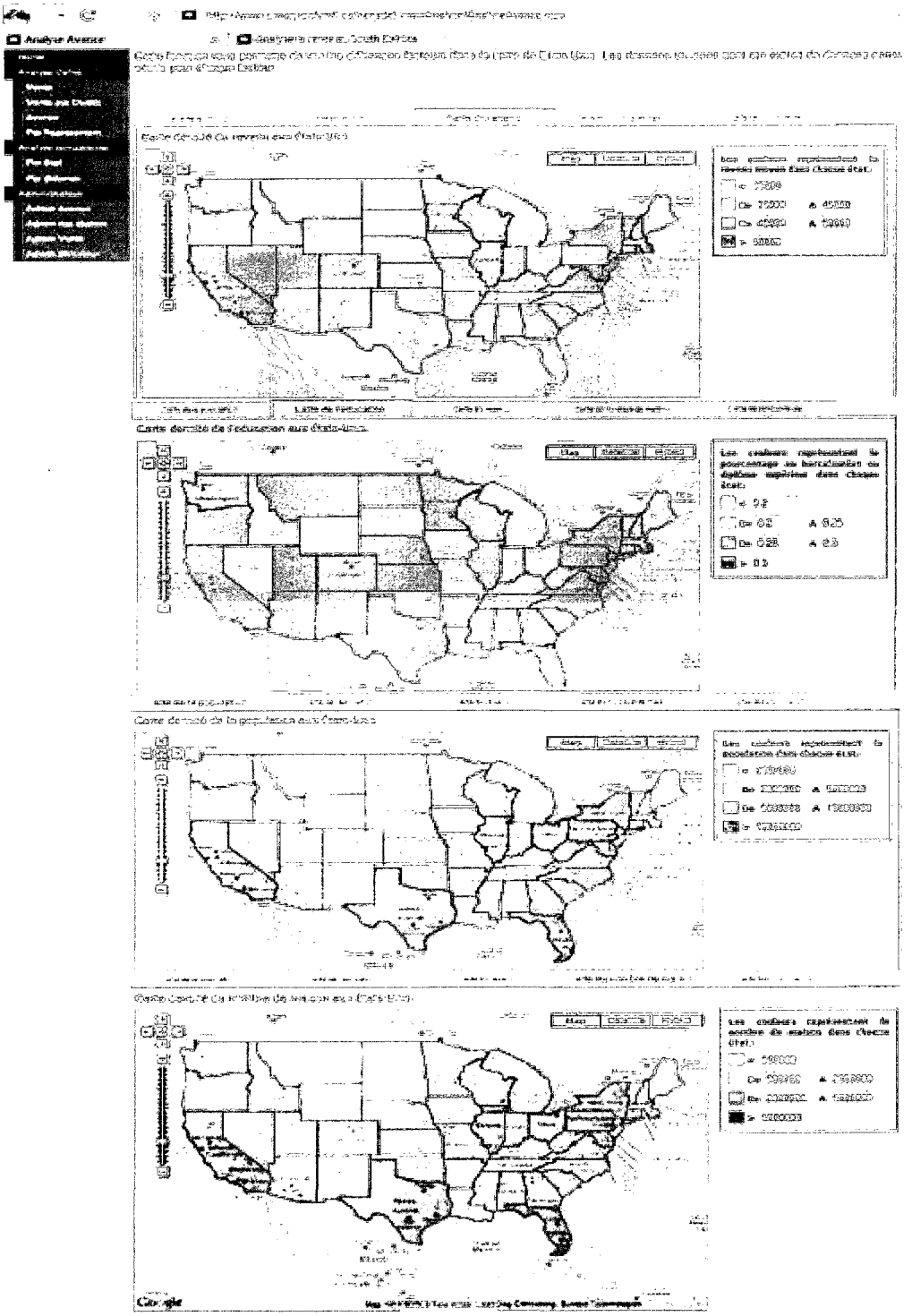


Figure 4.22 Les différentes couches de trois segments

En superposant les cartes de facteur et les cartes de localisation des magasins de l'étape du SIG, nous obtenons les cartes de localisation de chaque groupe. Ces cartes assistent de raffiner le résultat du data mining en déterminant les zones ayant la densité des magasins la plus élevé dans chaque groupe. Nous avons étendu cette fonctionnalité en permettant les utilisateurs de choisir le zone à analyser comme présenté dans la figure 4.23.

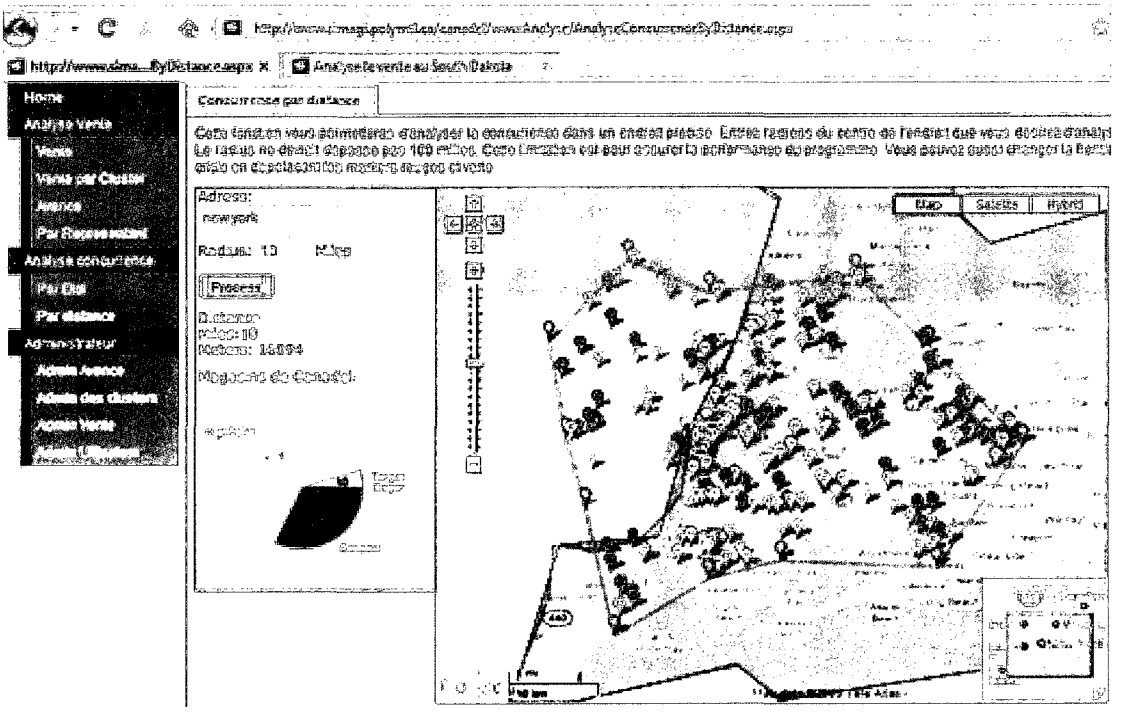


Figure 4.23 Carte de localisations des magasins par zone

4.4 Conclusion et perspective pour le cas d'étude

Les résultats obtenus nous confirment les avantages de l'intégration entre les techniques du data mining et les SIG. Cette intégration facilite aux utilisateurs la prise de décision à partir des résultats du data mining et du SIG. Cependant, dû à la contrainte temporelle, ce cas d'étude présente encore des

limitations:

- Nous n'utilisons que la géovisualisation dans cette intégration. Cependant, les méthodes de segmentation spatiale sont aussi intéressantes pour définir une localisation appropriée à l'implantation d'un nouveau magasin.
- Les étapes de prétraitement de données ont été effectuées manuellement et ont pris un temps considérable. Il faudrait dans le futur développer le module de prétraitement automatisé.
- Le niveau de précision de la segmentation est encore faible. En effet, la segmentation au niveau des états ne suffit pas afin d'établir des stratégies pertinentes pour l'entreprise
- La méthode des k-moyennes utilisée est la méthode classique des k-moyennes. Le résultat de cette méthode dépend beaucoup du choix initial des centres de segment. Il est préférable d'appliquer d'autres méthodes améliorées des k-moyennes.

Cette étude de cas envisage donc différentes voies futures de recherche:

- Une mise en direct avec les données de vente de l'entreprise pour suivre les changements du marché en temps réel.
- Une analyse de la corrélation entre le contenu d'un magasin et les ventes locales.
- La mesure de l'impact de l'introduction d'un nouveau produit sur les planchers de vente.
- D'appliquer la segmentation plutôt au niveau des villes qu'au niveau des états pour en raffiner le résultat.
- Intègre les techniques de segmentation spatiale pour la recherche de bonne localisation pour implanter un nouveau magasin.
- D'intégrer d'autres techniques du data mining telles que la prévision, les règles d'association aux SIG pour découvrir d'autres opportunités des marchés

et pour aider l'entreprise dans la prise de décision.

CONCLUSION

Les travaux présentés dans ce mémoire ont pour objectif de synthétiser les deux approches de recherche sur l'intégration entre le data mining et les SIG d'une part, et de démontrer les avantages de cette intégration par un cas d'étude d'autre part.

Dans la revue de littérature de la première approche, l'accent des recherches est mis sur l'amélioration des techniques existantes du data mining pour les adapter aux caractéristiques des données spatiales, plutôt que de proposer de nouvelles techniques. Les principaux travaux dans cette voie sont ceux de (Han *et al.*, 1997), de (Ester *et al.*, 2001) et de (Zeitouni, 2006).

Les recherches de la deuxième approche sont focalisées sur le développement de la méthodologie de l'intégration de la géovisualisation au data mining. (MacEachren *et al.*, 1999) ont présenté une approche d'intégration à trois niveau: conceptuel, opérationnel et mise en œuvre. (Andrienko et Andrienko, 1999) ont développé une plate-forme d'intégration de la géovisualisation au data mining en reliant le résultat du data mining avec la géovisualisation par trois façons.

A travers la revue des défis actuels du data mining et des SIG, nous constatons que l'intégration entre le data mining et les SIG révèle une voie de recherche potentielle pour les deux domaines. Pourtant, la recherche actuelle sur cette intégration que nous avons résumé dans le chapitre 3 est encore limité. Mis à part GeoMiner et SPIN!, nous ne dénombrons présentement pas d'autres prototypes opérationnels. Avec l'augmentation du volume et des nouveaux

types de données spatiales, la recherche sur cette intégration devient donc critique.

Dans le chapitre 4, nous avons mise en œuvre la deuxième approche par un cas d'étude de l'analyse du marché des meubles américain. Les résultats obtenus apportent une valeur ajoutée aux expérimentations d'application de la géovisualisation au data mining.

Par ce cas d'étude, nous remarquons que l'intégration entre le data mining et les SIG est un processus complexe. L'intégration dépend de l'objectif, des techniques de data mining utilisées et des capacités de SIG utilisées. Plus de recherches sur la méthodologie d'intégration sont donc nécessaires.

Notre recherche a confirmé le potentiel et les avantages de cette intégration. Elle pourrait continuer sur quelques perspectives suivantes:

- La plupart des systèmes actuels de l'intégration entre le data mining et les SIG sont basés sur les SIG commerciales. Ils sont donc difficiles à personnaliser. Pour répondre à ce défi, il demande plus de recherche sur l'architecture et la méthodologie de l'intégration entre les systèmes libres de data mining et ceux de SIG.
- Appliquer alternative les deux approches de l'intégration entre le data mining et les SIG dans d'autres projets en pratique pour évaluer et vérifier les avantages qu'elles peuvent apporter.

BIBLIOGRAPHIE

D.A. AAKER, V. KUMAR et G. DAY : Marketing research. John Wiley& Sons. *Inc. New York*, 2001.

G. ANDRIENKO et N. ANDRIENKO : Knowledge-based visualization to support spatial data mining. *Lecture notes in computer science*, pages 149–160, 1999.

F. BACAO, L. VICTOR et M. PAINHO : On the particular characteristics of spatial data and its similarities to secondary data used in data mining. *In GIS PLANET 2005, International Conference and Exhibition On Geographic Information*, 2005.

V. BOGORNY, B. KUIJPERS et LO ALVARES : Reducing uninteresting spatial association rules in geographic databases using background knowledge: a summary of results. *International Journal of Geographical Information Science*, 22(4):361–386, 2008.

US Government Bureau of LABOR STATISTICS : Economic statistics by geography, sector, and frequency. `http://www.census.gov/econ/www/`, Juin 2009.

B. BUTTENFIELD, M. GAHEGAN, H. MILLER et M. YUAN : Geospatial data mining and knowledge discovery. *UCGIS White Paper on Emergent Research Themes*, 2000.

S.B. CASTLEBERRY : Using secondary data in marketing research: A project that melds Web and Off-Web sources. *Journal of Marketing Education*, 23 (3):195, 2001.

N. CHELGHOUM et K. ZEITOUNI : Data mining spatial un probleme de data mining multi-tables. *RIST*, 14(2), 2004.

N. CHELGHOUM, K. ZEITOUNI, A. BOULMAKOUL, V. CEDEX et M. MOHAMMEDIA : A decision tree for multi-layered spatial data. *In Advances in Spatial Data Handling: 10th International Symposium on Spatial Data Handling*. Springer Verlag, 2002.

ESRI : Data models. `http://www.esri.com/`, November 2008.

M. ESTER, H.P. KRIEGEL et J. SANDER : Spatial data mining: A database approach. *Lecture Notes in Computer Science*, 1262:47–68, 1997.

M. ESTER, H.P. KRIEGEL et J. SANDER : Algorithms and applications for spatial data mining. *Geographic Data Mining and Knowledge Discovery*, 5 (6), 2001.

M. ESTER, H.P. KRIEGEL, J. SANDER et X. XU : Clustering for mining in large spatial databases. *KI*, 12(1):18–24, 1998.

U. FAYYAD, G. PIATETSKY, P. SMYTH *et al.* : From data mining to knowledge discovery in databases. *Communications of the ACM*, 39(11):24–26, 1996a.

U.M. FAYYAD, S.G. DJORGOVSKI et N. WEIR : Automating the analysis and cataloging of sky surveys. 1996b.

M. FILION et F. COLBERT : *Gestion du marketing*. G. Morin, 1990.

M.F. GOODCHILD et R.P. HAINING : GIS and spatial data analysis: Converging perspectives. *Papers in Regional Science*, 83(1):363–385, 2003.

P.E. GREEN : A new approach to market segmentation. *Business Horizons*, 20(1):61–73, 1977.

M.S. HACID et C. REYNAUD : L'intégration de sources de données. *Revue Information - Interaction - Intelligence (R I3)*, 2004.

J. HAN et M. KAMBER : *Data mining: concepts and techniques.* Morgan Kaufmann, 2006.

J. HAN, M. KAMBER et AK TUNG : Spatial Clustering Methods in Data Mining: A Survey. *Miller H, Han J. Geographic Data Mining and Knowledge Discovery. London: Taylor and Francis*, 2001.

J. HAN, K. KOPERSKI et N. STEFANOVIC : GeoMiner: a system prototype for spatial data mining. *In Proceedings of the 1997 ACM SIGMOD international conference on Management of data.* ACM New York, 1997.

K. HORNSBY et M.J. EGENHOFER : Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence*, 36(1):177–194, 2002.

H. JAYET : Économétrie et données spatiales: une introduction à la pratique. *Cahiers d'économie et sociologie rurales*, (58-59):105–129, 2001.

V. KARASOVA : *Spatial data mining as a tool for improving geographical models.* Thèse de doctorat, Helsinki University of Technology, 2005.

K. KOPERSKI, J. HAH et N. STEFANOVIC : An efficient two-step method for classification of spatial data. *In Proc. Symposium on Spatial Data Handling*, 1998.

K. KOPERSKI et J. HAN : Discovery of spatial association rules in geographic information databases. *Lecture Notes in Computer Science*, 951:47–66, 1995.

D.T. LAROSE : *Discovering knowledge in data: an introduction to data mining.* Wiley-Interscience, 2005.

T.T.H LE, B. AGARD et S. DEVEAULT : Application du data mining à la segmentation du marché des meubles aux États-Unis. *In 8ème Congrès International de Génie Industriel-CIGI'09*, 2009a.

T.T.H LE, B. AGARD et S. DEVEAULT : Decision support based on socio-demographic segmentation and distribution channel analysis in the US furniture market. *In International Conference on Industrial Engineering and Systems Management Ű IESMŠ 2009*, 2009b.

P. LONGLEY, M.F. GOODCHILD, D.J. MAGUIRE et D.W. RHIND : *Geographical information systems and science.* John Wiley & Sons Inc, 2005.

A.M. MACEACHREN et M.J. KRAAK : Research challenges in geovisualization. *Cartography and Geographic Information Science*, 28(1):3–12, 2001.

A.M. MACEACHREN, M. WACHOWICZ, R. EDSALL, D. HAUG et R. MASTERS : Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science*, 13 (4):311–334, 1999.

J.B. MACQUEEN : Some methods for classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* Western Management Science INST UNIV OF California Los Angeles, 1966.

AL MAJURIN : Industrial Segmentation: A Review. *Preliminary report for ABO Academy School of Business*, 203, 2001.

M. MAY et A. SAVINOV : An integrated platform for spatial data mining and interactive visual analysis. *In Third International Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields*, 2002.

H.J. MILLER et J. HAN : *Geographic data mining and knowledge discovery.* CRC Press, 2001.

VW MITCHELL et P.J. MCGOLDRICK : Life style concepts and marketing, toward scientific marketing. *Proceeding of the AMA Winter Conference*, pages 130–139, 1963.

C. MORENCY : Étude de méthodes dŠanalyse spatiale et illustration à lŠaide de microdonnées urbaines de la grande région de Montréal. *Les Cahiers scientifiques du transport*, (49):77–102, 2006.

L.M. MURILLO : Manufacturers-retailers: The new actor in the us furniture industry. characteristics and implications for the chinese furniture industry. *International Journal of Human and Social Sciences*, 1:3, 2007.

G.I.P. OTTAVIANO et G. PERI : The economic value of cultural diversity: evidence from US cities. *Journal of Economic Geography*, 6(1):9–44, 2006.

G.L. PATZER : *Using secondary data in marketing research: United States and worldwide.* Quorum Books, 1995.

G. PUNJ et D.W. STEWART : Cluster analysis in marketing research: review and suggestions for application. *Journal of Marketing Research*, pages 134–148, 1983.

D. PYLE : *Data preparation for data mining.* Morgan Kaufmann Pub, 1999.

D.B. RUBIN : Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

S. SALVADOR et P. CHAN : Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *In 16th IEEE International Conference on Tools with Artificial Intelligence*, 2004.

J. SINGH : A typology of consumer dissatisfaction response styles. *Journal of Retailing*, 66(1):57–99, 1990.

A. SMADJA : *Segmenter ses marchés: application pratique des techniques de segmentation dans le marketing.* Presses polytechniques romandes, 1988.

TANAGRA : free data mining software. `http://eric.univ-lyon2.fr/ ricco/tanagra/en/tanagra.html/`, November 2008.

R. TIBSHIRANI, G. WALTHER et T. HASTIE : Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 411–423, 2001.

W. TOBLER : Cellular geography. *Philosophy in geography*, pages 379–386, 1979.

A. TORUN et S. DUZGUN : Using Spatial Data Mining Techniques to Reveal Vulnerability of People and Places Due to Oil Transportation and Accidents: A Case Study Of Istanbul Strait. *In ISPRS Technical Commission II Symposium, Vienna*, pages 12–14, 2006.

S. TUFFÉRY : *Data mining et statistique décisionnelle: l'intelligence dans les bases de données.* Editions Technip, 2005.

AKH TUNG, J. HOU et J. HAN : Spatial clustering in the presence of obstacles. *In Data Engineering, 2001. Proceedings. 17th International Conference on*, pages 359–367, 2001.

USA U.S. CENSUS BUREAU : Economic statistics by geography, sector, and frequency. `http://www.census.gov/econ/www/`, Juin 2009.

Aurelio VOLPE et Mariano PELUSO : The kitchen furniture market in the united states. CSIL reports S34, Centre for Industrial Studies, Oct 2007. http://ideas.repec.org/p/mst/csilre/s34.html.

M. WACHOWICZ : GeoInsight: an approach for developing a knowledge construction process based on the integration of GVis and KDD method. *Geographic Data Mining and Knowledge Discovery*, pages 239–259, 2001.

M. WEDEL et W.A. KAMAKURA : *Market segmentation: conceptual and methodological foundations.* Springer, 1999.

A. WEINSTEIN : *Market segmentation: using demographics, psychographics and other niche marketing techniques to predict and model customer behavior.* Probus, 1994.

K. ZEITOUNI : Etat de lŠart sur lŠextension du data mining aux bases de données géographiques. *Rapport de Recherche, Laboratoire Prism, Université de Versailles*, 1998.

K. ZEITOUNI : *Analyse et extraction de connaissances des bases de données spatio-temporelles.* Thèse de doctorat, Université de Versailles-Saint Quentin en Yvelines, 2006.