POLYPUBLIE
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Titre: Title:	Conception et réalisation d'un outil de traitement et analyse des données spatiales pour l'aide à la décision : application au secteur de la distribution
Auteur: Author:	Gautier Daras
Date:	2017
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Daras, G. (2017). Conception et réalisation d'un outil de traitement et analyse des données spatiales pour l'aide à la décision : application au secteur de la distribution [Ph.D. thesis, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/2939/

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/2939/
Directeurs de recherche: Advisors:	Bruno Agard, & Bernard Penz
Programme: Program:	Doctorat en génie industriel

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal

https://publications.polymtl.ca

UNIVERSITÉ DE MONTRÉAL

CONCEPTION ET RÉALISATION D'UN OUTIL DE TRAITEMENT ET ANALYSE DES DONNÉES SPATIALES POUR L'AIDE À LA DÉCISION : APPLICATION AU SECTEUR DE LA DISTRIBUTION

GAUTIER DARAS
DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL
ET
UNIVERSITÉ GRENOBLE ALPES

THÈSE EN COTUTELLE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIÆ DOCTOR
(GÉNIE INDUSTRIEL)
DÉCEMBRE 2017

© Gautier Daras, 2017.

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée:

CONCEPTION ET RÉALISATION D'UN OUTIL DE TRAITEMENT ET ANALYSE DES DONNÉES SPATIALES POUR L'AIDE À LA DÉCISION : APPLICATION AU SECTEUR DE LA DISTRIBUTION

présentée par: DARAS Gautier
en vue de l'obtention du diplôme de: Philosophiæ Doctor
a été dûment acceptée par le jury d'examen constitué de:

M. TRÉPANIER Martin, Ph. D., président
M. AGARD Bruno, Doctorat, membre et directeur de recherche
M. PENZ Bernard, Doctorat, membre et codirecteur de recherche
M. LEBEL Luc, Ph. D., membre
M. WIDMER Marino, Doctorat ès Sc., membre externe

DÉDICACE

À ma femme, à mes filles, qui ont su pimenter ces travaux de thèse.

Je vous aime.

REMERCIEMENTS

Je tiens tout d'abord à remercier les membres du jury de cette thèse pour avoir accepté de participer à l'évaluation de mes travaux de recherche. Merci à Martin Trépanier d'avoir accepté de présider le Jury, à Marino Widmer et Luc Lebel d'en être membre, et à Catherine Morency d'être rapporteur du directeur des études supérieures.

Je veux aussi remercier tout particulièrement mes directeurs de recherche qui m'ont encadré avec bienveillance tout au long de mes travaux de thèse. Merci à tous les deux de m'avoir permis de mener mes travaux de recherches tout en me donnant les moyens de profiter de ma petite famille.

Merci Bernard de m'avoir donné envie de faire une thèse et de m'avoir proposé cette superbe opportunité de la réaliser en cotutelle. Merci pour ta disponibilité et ta réactivité tout au long de ces travaux de recherches, merci pour tes nombreux conseils tant au niveau de la recherche qu'au niveau des démarches.

Merci aussi à Bruno de m'avoir accueilli au Canada, de m'avoir guidé dans mes recherches tout en me laissant la liberté d'approfondir dans les domaines pour lesquels j'ai de l'intérêt. Merci aussi pour ton aide précieuse et de ton soutien moral lors des périodes de rédaction. Merci pour ta compréhension et tes conseils avisés concernant l'immigration en famille.

Je tiens aussi à remercier l'entreprise Maibec, le partenaire industriel de ces travaux de recherches. Tout particulièrement Dominic Boulanger et Florence Lepage, avec qui j'ai pu travailler en toute sérénité sur des problématique réelles. Merci pour vos conseils précieux et vos nombreux retours qui m'ont permis de mieux comprendre les besoins de votre entreprise.

Je souhaite par ailleurs remercier le FORAC pour la confiance qu'ils m'ont accordée pour réaliser ces travaux de recherches avec Maibec. Merci pour le support et pour les multiples conférences et autres moments d'échange qui m'ont permis d'apprendre et de progresser.

Merci aussi aux différentes structures d'encadrement, en particulier l'école Polytechnique et le laboratoire G-SCOP. La qualité des services administratifs, et de leur personnel, a toujours permis de réaliser efficacement les différentes démarches tout au long de la réalisation de cette thèse. Je tenais aussi à remercier les enseignants qui, au cours de mon cursus scolaire, ont su me donner confiance en moi, et m'ont guidé vers des sujets qui sont aujourd'hui au coeur de mes centre d'intérêts professionnels. Merci à Mme Pommier et Mme Bensalem pour m'avoir fait aimer les Mathématiques. Merci à Nadia Brauner de m'avoir fait découvrir la Recherche Opérationnelle et de m'avoir incité à me lancer dans cette voie.

Pour terminer je voulais remercier les membres de ma famille, mes parents et mes beaux parents pour leur soutien tout au long de cette thèse. Ma femme, Nathalie, et mes filles, Margaux, Anna, et Mathilde, sans qui cette thèse aurait surement duré un an de moins, mais grâce à qui ma vie prend tout son sens, je vous aime plus que tout.

RÉSUMÉ

Pour améliorer leurs performances, les entreprises du secteur de la distribution peuvent vouloir mettre en place des systèmes d'aide à la décision spatiale. La mise en place de ce type de systèmes reste aujourd'hui complexe. En effet, plusieurs verrous existent pour la conception, le développement et l'utilisation de ces outils, dont l'objectif est l'analyse des données spatiales, la création de connaissances, et leur utilisation dans un objectif d'optimisation.

Ainsi, l'objectif de cette thèse est de proposer des contributions méthodologiques et logicielles pour permettre la conception et le développement d'un outil de traitement et analyse de données spatiales pour l'aide à la décision. Les trois problématiques abordées sont relatives à la prise en compte des données spatiales à trois niveaux: (1) au niveau de la conception et du développement d'un système d'aide à la décision ; (2) au niveau de l'application d'outils d'analyse de données sur des données spatiales ; et (3) au niveau des approches d'optimisation relatives à la localisation.

La première contribution propose un cadre flexible pour le développement d'un système d'aide à la décision spatiale. Un ensemble d'étapes est proposé pour permettre de passer par les différentes phases requises pour la conception de ce type de système. La méthodologie permet la factorisation des différentes tâches dans le cas où plusieurs applications dédiées doivent être développées pour un même projet. De plus, une architecture propose des outils de manipulation des données spatiales, permettant de procéder à leur analyse et à la visualisation des résultats. Pour valider l'efficacité de notre cadre conceptuel, la conception d'un système spécifique, dédié à une application industrielle, est présentée. Le système développé est maintenant utilisé par une entreprise qui distribue des produits pour la construction immobilière à travers différents réseaux de distribution. Les possibilités offertes par la méthodologie sont illustrées grâce à une explication des différentes étapes de développement.

La deuxième contribution réside dans une approche de prétraitement des données spatiales pour permettre leur analyse. L'approche proposée permet d'assister et d'accélérer le prétraitement des données spatiales. Cette approche est composée de plusieurs étapes qui d'une part, automatisent en grande partie le processus de prétraitement des données, et d'autre part, permettent de réaliser ce prétraitement sans avoir besoin de connaissance en relation spatiale ou en Système d'Information Géographique. L'approche est validée par la création d'un outil dont l'utilisation apporte des améliorations significatives par rapport aux outils actuellement disponibles. Les prétraitements ont pu être réalisés plus rapidement, et ils ont permis d'améliorer sensiblement la qualité des analyses.

La troisième contribution réside dans l'utilisation conjointe du traitement des données spatiales et de la modélisation du problème à traiter pour permettre une approche d'optimisation plus réaliste que les modèles actuels de la littérature. La solution proposée permet la prise en compte des zones de chalandise et de la capture collaborative de la demande dans le traitement du problème de maximisation de la couverture. Pour permettre cela les capacités actuelles des Systèmes d'Information Géographique sont mises à profit pour intégrer les données spatiales relatives aux zones de chalandise. Les données traitées sont ensuites utilisées au travers d'une une modélisation adaptée à la problématique. La chaine de traitement est complètement validée, des données brutes jusqu'à la proposition des solutions de localisation.

Ces trois contributions ont été intégrées au sein d'un seul outil de traitement et analyse des données spatiales pour l'aide à la décision, le tout basé sur des outils logiciels libres. L'outil développé a été validé sur une application avec un partenaire industriel du secteur de la distribution.

ABSTRACT

Spatial decision support systems can assist companies in the distribution sector improve their performance through the analysis of spatial data, the creation of knowledge, and optimization purposes However, the implementation of this type of system remains complex today. Indeed, several barriers exist for the design, development and use of these tools.

The objective of this thesis is to propose methodological and software contributions to allow for the design and development of a spatial data processing and analysis tool for decision support. The issues addressed relate to the integration of spatial data characteristics at three levels: (1) in the design and development of a decision support system; (2) at the application of data analysis tools on spatial data; and (3) in localization optimization approaches.

The first contribution proposes a flexible framework for the development of a spatial decision support system. To validate the effectiveness of our methodology, the design of a specific system dedicated to an industrial application is presented. The developed system is now used by a company that distributes products for building construction through various distribution networks. The possibilities offered by the methodology are illustrated by a presentation of the different stages of development of the application

The second contribution is a pre-processing approach that allows for spatial data to be taken into account in data analysis. The proposed approach makes it possible to assist and accelerate the pre-processing of spatial data. The approach is then validated by the creation of a tool whose use brings significant improvements compared to the tools currently available. Pretreatments were processed faster and significantly improved the quality of the analysis.

The third contribution lies in the joint use of spatial data processing and modelling techniques in order to allow a more realistic optimization approach than the available models of the literature. The proposed solution allows the catchment areas and the collaborative capture of demand to be taken into account in dealing with the problem of maximizing coverage. The processing chain is completely validated, from the raw data to the proposal of the localization solutions.

These three contributions have been integrated into a single spatial data processing and analysis tool for decision support, all based on free software tools. The developed tool has been validated on an application with an industrial partner in the distribution sector.

TABLE DES MATIÈRES

DÉDICACE iii

REMERCIEMENTS iv

RÉSUMÉ vi

ABSTRACT viii

TABLE DES MATIÈRES ix

LISTE DES TABLEAUX xii

LISTE DES FIGURES xiii

LISTE DES SIGLES ET ABRÉVIATIONS xv

CHAPITRE 1 INTRODUCTION 1

CHAPITRE 2 REVUE DE LITTÉRATURE 4

2.1 Le besoin d'outil d'aide à la décision spatiale 4

2.1.1 Les contraintes liées à la conception de système d'aide à la décision spatiale 5

2.1.2 Synthèse 7

2.2 Les solutions actuelles pour développer des outils d'aide à la décision et pour prendre en compte des données spatiales. 7

2.2.1 Systèmes d'Information Géographique 8

2.2.2 Système d'aide à la décision et Extraction des Connaissances à partir des Données 10

2.2.3 Les défis liés à la conception de SDSS 11

2.2.4 Synthèse 12

2.3 Complexité de la prise en compte des données spatiales dans les outils d'analyse de données 13

2.3.1 La nécessité du prétraitement des données spatiales 13

2.3.2 Un prétraitement complexe et chronophage 15

2.3.3 Synthèse 17

2.4 Les outils d'optimisation de couverture 18

2.4.1 Approches d'optimisation de la couverture par modélisation 18
2.4.2 Prétraitement des données spatiales pour l'optimisation 20
2.4.3 Synthèse 20
2.5 Synthèse de la Revue de littérature 20

CHAPITRE 3 DÉMARCHE ET ORGANISATION 22
3.1 Les problématiques de recherches 22
3.2 Les approches de recherche 22
3.3 Intégration des problématiques scientifiques dans un contexte industriel . . . 24
3.4 Contributions 25

CHAPITRE 4 ARTICLE 1: CONCEPTUAL FRAMEWORK FOR SDSS DEVELOPMENT: APPLICATION IN THE RETAIL INDUSTRY 27
4.1 Introduction 27
4.2 Literature review 29
4.2.1 Design principles and requirements of development architecture . . . 29
4.2.2 Design and development methodologies 30
4.2.3 Synthesis 31
4.3 Framework presentation through a case study 33
4.3.1 The needs of a retail company 33
4.3.2 SDSS applications, development process, and tools required 34
4.3.3 SDSS development 39
4.3.4 Extension to other datasets 51
4.4 Conclusion 51

CHAPITRE 5 ARTICLE 2: A SPATIAL DATA PRE-PROCESSING TOOL TO IMPROVE ANALYSIS QUALITY AND TO REDUCE PREPARATION DURATION 53
5.1 Introduction 53
5.2 Literature Review 54
5.2.1 Spatial decision-making 54
5.2.2 Spatial Data mining: specificities and challenges 55
5.2.3 Spatial data pre-processing step 55
5.3 Necessity and complexity of pre-processing 59
5.3.1 Necessity of pre-processing through an example 60
5.3.2 The consequences of pre-processing choice on analysis 62
5.4 Specifications and technical aspects 64
5.4.1 Pre-processing steps 64

5.4.2 Implementation 67
5.5 Case study to evaluate improvements 74
5.5.1 Steps without the proposed approach 74
5.5.2 Steps with the implemented approach 76
5.5.3 Analysis quality improvement 78
5.6 Conclusions and perspectives 79

CHAPITRE 6 ARTICLE 3: REALISTIC MODEL FOR THE MAXIMUM COVERING LOCATION PROBLEM USING SPATIAL DATA PRE-PROCESSING 81
6.1 Introduction 81
6.2 Literature Review 83
6.3 Model 84
6.3.1 Initial dataset and realistic constraints 86
6.3.2 MCLP Model 87
6.4 Pre-processing data transformation 89
6.4.1 Simple transformations 89
6.4.2 Generalization and formalized constraints 95
6.4.3 Implemented Transformation 96
6.4.4 Proof of respect of the constraints 98
6.4.5 Spatial data pre-processing steps 100
6.5 MCLP resolution 103
6.5.1 Resolution algorithm and performance 104
6.5.2 Data structures and Evaluation function 106
6.6 Conclusion and perspectives 108

CHAPITRE 7 DISCUSSION GÉNÉRALE 110

CHAPITRE 8 CONCLUSION ET RECOMMENDATIONS 112
8.1 Les contributions et leurs limites 112
8.1.1 Traitement des trois problématiques: contributions et limites 112
8.1.2 Limite de l'approche globale sur un cas industriel 113
8.2 Perspectives 114
8.2.1 Perspectives de recherches relatives aux contributions 114
8.2.2 Perspective pour l'outil de traitement et analyse de données 115

LISTE DES RÉFÉRENCES 116

LISTE DES TABLEAUX

Table 4.1	Difficulties identified to transfer other research to various contexts .	32
Table 5.1	Spatial Request Processing Time	70
Table 5.2	Manipulation and Computing Time	78
Table 6.1	Pre-processing, computing and gain performances	105

LISTE DES FIGURES

Figure 2.1	Couches de données spatiales ([64])	8
Figure 2.2	Superposition des couches de données ([64])	9
Figure 2.3	Influence entre les parties prenantes au dévéloppement ([64])	11
Figure 2.4	Les relations entre éléments spatiaux ([18])	14
Figure 2.5	Durée des étapes du process d'ECD ([27])	16
Figure 2.6	Impact des données spatiales sur l'ECD ([100])	16
Figure 3.1	Interactions entre les parties prenantes	23
Figure 4.1	SDSS development workflow	34
Figure 4.2	Schematic architecture	36
Figure 4.3	SDSS architecture and development toolkit	37
Figure 4.4	Dedicated sales sum query	41
Figure 4.5	First visualization interface	42
Figure 4.6	Point in area cardinal query	43
Figure 4.7	Aggregated sales per area	43
Figure 4.8	Aggregated sales visualization interface	44
Figure 4.9	Filtering options	45
Figure 4.10	Polygon simplification	46
Figure 4.11	Table transformation to allow temporal queries	47
Figure 4.12	Configurable temporal query	47
Figure 4.13	User interface capabilities	49
Figure 4.14	Transfer to the Province of Ontario	51
Figure 5.1	Spatial Relations ([18])	56
Figure 5.2	Time spent in KDP phases ([27])	57
Figure 5.3	Impact of Spatial Data on the KDD Process ([100])	57
Figure 5.4	Gathered Datasets	61
Figure 5.5	Data Integration	62
Figure 5.6	Cardinality Computation	63
Figure 5.7	Alternative Spatial Relations	64
Figure 5.8	Pre-Processing Steps	66
Figure 5.9	Basic Database Structure	67
Figure 5.10	Possible Architecture for Implementation	68
Figure 5.11	Spatial Relation Computation Illustration	69
Figure 5.12	Distance to Nearest Query	70

Figure 5.13 Faster Distance to Nearest Query 70
Figure 5.14 Coverage computation requests 71
Figure 5.15 Database Modified Structure 72
Figure 5.16 Pre-Processing Intersection Areas 73
Figure 5.17 Catchment Area with QGIS 75
Figure 5.18 Target Layer Selection 76
Figure 5.19 Source Layer Selection 77
Figure 5.20 Random Forest and Treebag performances depending on input dataset 79
Figure 6.1 Approaches for MCLP resolution 82
Figure 6.2 Usual data transformation illustrated 85
Figure 6.3 Transformation process illustrated 85
Figure 6.4 Initial dataset illustrated 86
Figure 6.5 Trade areas separation 89
Figure 6.6 Transformation of intersections to demand points 90
Figure 6.7 Capture rate assignment 91
Figure 6.8 Two-trade area intersection transformation 1/2 91
Figure 6.9 Two-trade area intersection transformation 2/2 92
Figure 6.10 Converting intersections to demand points 93
Figure 6.11 Capture rates to assign for the intersections of three trade areas . . . 93
Figure 6.12 Sub-problem for the affect on capture rates 94
Figure 6.13 Generalized Transformation 95
Figure 6.14 Intersection points renumbering and capture rates assignment 97
Figure 6.15 Transformation steps for two-trade areas intersection 97
Figure 6.16 Transformation steps for three-trade areas intersection 98
Figure 6.17 Case study initial datasets 101
Figure 6.18 Illustration of the request for the duplication intersection 102
Figure 6.19 Request for intersection with division 103
Figure 6.20 Initial network and proposed change 106

LISTE DES SIGLES ET ABRÉVIATIONS

CD	Census Division
CHMC	Canada Mortgage and Housing Corporation
CSD	Census Subdivision
CSV	Comma-Separated Values
DSS	Decision Support System
ECD	Extraction de Connaissances à partir de Données
GIS	Geographic Information System
GPS	Global Positioning System
HTML	HyperText Markup Language
IDE	Integrated Development Environment
KDD	Knowledge Discovery from Database
KDP	Knowledge Discovery Process
RAM	Random Access Memory
MCLP	Maximum Covering Location Problem
NHS	National Household Survey
SDSS	Spatial Decision Support System
SIG	Système d'Information Géographique
URL	Uniform Resource Locator
XML	Extensible Markup Language

CHAPITRE 1 INTRODUCTION

La compétitivité grandissante dans le secteur de la distribution amène aujourd'hui les entreprises du secteur, dans un souci de pérennisation ou de croissance, à améliorer leur réseau de distribution pour être en meilleure adéquation avec la demande. Différentes stratégies sont envisagées par les décideurs pour cela. Parmi ces stratégies, l'analyse de la localisation est un des domaines de recherche importants. Comme Cliquet et al. [29] le décrivent, la connaissance de l'environnement spatial est un élément clé dans le secteur de la distribution. En effet, la localisation est intrinsèquement liée aux performances des activités des entreprises de ce secteur. Les décisions stratégiques comme les ouvertures de points de vente, les fermetures ou leur redimensionnement, doivent être prises en ayant des informations pertinentes sur le contexte local dans lequel elles vont s'appliquer.

Pour mieux comprendre l'influence de la localisation sur les performances, une approche possible est l'analyse des données historiques de ventes. Thompson et Walker [125] mentionnent l'intérêt croissant pour l'analyse et la compréhension des données spatiales dans le secteur de la distribution. Il y a quelques décennies, Herring et al. [63] faisaient face aux problèmes de collecte et de stockage des informations liées à l'emplacement. Aujourd'hui la problématique a changé et Bradlow et al. [22] avancent qu'il existe de nombreuses sources de données disponibles contenant des informations spatiales. Avec l'évolution des capacités de stockage et de traitement des données, et les coûts en baisse des technologies informatiques, de plus en plus d'organisations accumulent des données sur leurs activités (par exemple les chiffres de vente), qui incluent des caractéristiques spatiales (telles que des adresses ou des coordonnées géographiques). Parallèlement, il existe une quantité croissante de données disponibles sur des éléments susceptibles d'influencer le rendement des activités de ces organisations (les données sociodémographiques par exemples). Pour ces raisons, des recherches importantes, dans divers domaines, ont eu pour objectif d'extraire des informations pertinentes pour comprendre ce qui influence réellement une activité. Erskin et al. [42] et MacEachren et Kraak [80] mettent en avant la nécessité d'utiliser efficacement les données collectées pour prendre des décisions stratégiques et organisationnelles

Il existe déjà des systèmes d'aide à la décision, et des méthodologies et outils sont disponibles pour mettre ces systèmes en place. Cependant, la plupart des systèmes d'aide à la décision existants ne prennent pas en compte les spécificités des données spatiales [100]. Ainsi la mise en place d'un système d'aide à la décision spatiale reste un défi et il existe plusieurs barrières pour pouvoir exploiter le potentiel des données spatiales au travers d'un système d'aide à la

décision :

— Une première barrière identifiée réside dans la mise en place en elle-même d'un système d'aide à la décision spatiale. En effet, la mise en place de systèmes d'aide à la décision spatiale reste aujourd'hui une tâche à laquelle il n'y a pas encore de réponse méthodologique convenable ni d'outils informatiques adaptés [100].

— Une deuxième barrière réside dans la prise en compte des données spatiales. Pour Erskin et al. [42], la gestion des données spatiales et l'utilisation de celles-ci restent un défi. Pretorius et Matthee [100] insistent sur l'incapacité de nombreux outils à prendre en compte les données spatiales. Pour rendre les données spatiales compatibles avec des outils d'analyse, un prétraitement est aujourd'hui nécessaire dans la plupart des cas. Borgony et al. [18] mettent en avant la complexité et l'aspect chronophage du prétraitement des données spatiales.

— Une troisième barrière repose sur les modèles théoriques actuels d'aide à la décision. Dans le secteur de la distribution, l'optimisation de la couverture est, par exemple, un problème que beaucoup d'entreprises cherchent à résoudre. La plupart des modèles proposés ne sont cependant pas capables de prendre en compte les données spatiales. En conséquence, les données spatiales sont le plus souvent simplifiées en se reposant sur des hypothèses irréalistes. Les solutions qui en résultent sont donc questionnables [111].

Les travaux de recherche de cette thèse s'attaquent à ces trois barrières dans le but de répondre à une problématique générale: "Comment procéder à la conception et au développement d'un outil de traitement et d'analyse des données spatiales pour l'assistance à la prise de décision ?". Ainsi, trois approches sont proposées pour permettre l'exploitation du potentiel des données spatiales au travers d'un système d'aide à la décision spatiale (SDSS pour *Spatial Decision Support System*) :

— La première approche repose sur la proposition d'un cadre conceptuel qui permet la mise en place d'un système d'aide à la décision spatiale. Le cadre proposé permet une intégration de diverses méthodes et outils existants pour permettre la prise en compte des spécificités liées aux données spatiales. L'approche proposée sera validée par la conception et le développement d'un système d'aide à la décision spatiale pour un partenaire industriel. Nous étudierons ainsi le développement d'un système dédié à l'évaluation et à l'amélioration du réseau commercial d'une entreprise partenaire.

— La deuxième approche s'attarde sur les problématiques liées à la prise en compte des données spatiales dans les outils d'analyse de données classiques. Pour pallier ces pro-

blèmes, et rendre plus accessible l'analyse des données spatiales, nos travaux proposent une approche pour aider au prétraitement des données spatiales. L'approche a été mise en œuvre au travers du développement d'un outil correspondant. L'outil a pu être utilisé sur des données réelles. Il permet de gagner un temps important mais surtout de guider l'utilisateur pour réaliser le prétraitement qui rend les données spatiales compatibles avec les autres données.

— La troisième approche permet la prise en compte d'aspects réalistes dans le traitement d'un problème d'optimisation. Le problème considéré est celui de la maximisation de la couverture. Pour arriver à prendre en compte des aspects réalistes, des techniques de prétraitement des données et de modélisation sont combinées. Notre approche permet de prendre en compte les données relatives aux zones de couvertures, et autorise la couverture collaborative. Un algorithme de résolution est aussi proposé, et l'ensemble de l'approche a pu être mis en œuvre sur les données de notre partenaire industriel.

La combinaison des différentes approches et leur intégration ont permis de faire tomber les barrières au développement d'un système d'aide à la décision spatiale. Pour preuve, un système d'aide à la décision spatiale, dédié à un cas industriel, a pu être conceptualisé et développé au cours de cette thèse. Le développement global de ce système a permis la validation des différentes approches, ainsi que la réalisation du processus de conception et de développement au complet.

Le besoin de système d'aide à la décision, ainsi que les trois barrières à leur mise en place sont présentés dans l'état de l'art (2). La problématique, le contexte d'application et la démarche globale de nos travaux de recherche sont présentés dans le chapitre 3. Les réponses sur les barrières au développement des approches associées font ensuite l'objet du corps de cette thèse : le chapitre 4 met en avant les problématiques liées à la mise en place d'un système d'aide à la décision spatiale, et propose une approche illustrée et mise en application sur le cas industriel. Le chapitre 5 traite de la complexité de la prise en compte des données spatiales dans les outils d'aide à la décision. L'approche qui permet d'assister cette prise en compte des données spatiales est présentée et mise en application dans un outil associé. Le chapitre 6 se concentre sur la prise en compte d'aspects spatiaux réalistes dans un modèle d'optimisation pour un problème de maximisation de la couverture. La combinaison proposée de techniques de modélisation et de prétraitement est appliquée sur les données industrielles. Un algorithme de résolution est aussi proposé pour valider la chaine complète de traitement, des données initiales à la solution. Pour finir, une discussion générale sur les travaux de recherche de cette thèse est proposée (Chapitre 7), puis des conclusions et perspectives sont données (Chapitre 8).

CHAPITRE 2 REVUE DE LITTÉRATURE

Cette revue de littérature va dans un premier temps souligner le besoin grandissant d'outils d'aide à la décision spatiale dans le secteur de la distribution. Dans un deuxième temps, les outils actuels qui permettent la prise en compte des données spatiales, ainsi que les outils d'aide à la décision classiques seront présentés. Les limites des systèmes actuels, et les caractéristiques qu'ils doivent respecter seront aussi présentées. Dans une troisième partie, nous présenterons les difficultés rencontrées pour la prise en compte des données spatiales dans les outils d'analyse de données classiques. Enfin les modèles actuels d'optimisation de la couverture seront présentés et leurs limites seront mises en avant.

2.1 Le besoin d'outil d'aide à la décision spatiale

Dans le secteur de la distribution, les activités d'une entreprise sont particulièrement impactées par la localisation de ses distributeurs [29]. En effet, plusieurs chercheurs soulignent l'impact de la localisation dans différents domaines liés au secteur de la distribution. Aloitaibi [4] souligne le lien entre la localisation et le besoin des clients. Pour Fang et al. [45] la rentabilité dépend de l'emplacement, et pour Perieira et al. [94] la logistique est fortement influencée par la disposition géographique. Jain et al. [66] soulignent l'impact de la localisation sur la sélection des fournisseurs et pour Cliquet et al. [29], l'emplacement de la compétition, ainsi que la topographie de l'environnement (qui influe, entre autres, sur l'accessibilité) ont aussi une influence sur l'attractivité des points de vente.

Ainsi, de nombreuses problématiques liées à la localisation sont abordées dans les travaux de recherche. Arentze et al. [8] ont travaillé sur le développement et l'évolution des marchés. Des recherches récentes comme celle de Yingru et Lin [76] présentent un problème d'évaluation de l'impact d'un nouveau distributeur sur un territoire. Un autre aspect abordé par Nikolopoulos et al. [90] est la prévision de la demande tandis que DeBeule et al. [38] cherchent à comprendre l'interaction spatiale entre les magasins d'une même franchise. Hernandez et Bennison [62] présentent plusieurs problématiques de recherche liées à la prise de décisions de localisation dans le secteur de la distribution : ils soulignent l'impact de la subjectivité des décideurs d'une part et la simplicité des systèmes d'aide à la décision d'autre part (au sens où ce qu'ils prennent en compte est limité).

Pour traiter les problématiques liées à la localisation, une approche de plus en plus répandue repose sur l'analyse des données et en particulier l'analyse des données spatiales. Les

besoins d'utilisation de l'analyse de données s'est amplifiée avec l'évolution des technologies de l'information. En effet, les technologies de l'information étant présentes au cœur du fonctionnement des entreprises, elles permettent, entre autres, de stocker une grande quantité de données contenant de l'information potentiellement utile. La recherche de cette information a mené l' « Extraction des Connaissances à partir des Données » (ECD) au premier plan des recherches actuelles [60]. Pour l'analyse des données spatiales, alors qu'il y a quelques décennies, Herring et DeBinder [63] faisaient face aux problèmes de collecte et de stockage des informations relatives à la localisation, Bradlow et al. [22] expliquent qu'il existe aujourd'hui de plus en plus de sources de données qui contiennent des informations spatialisées. Pour Erskin et al. [42], à mesure que les organisations recueillent de grandes quantités de données géo-spatiales, il devient nécessaire d'utiliser efficacement les données collectées pour prendre des décisions stratégiques ou organisationnelles.

Ainsi de nombreux chercheurs, dans des secteurs variés comme l'environnement [114] ou la santé [44] développent des outils pour tirer profit des données spatiales. Dans le secteur de la distribution aussi, l'intérêt croissant pour les techniques d'analyse et de compréhension des données spatialisées suit la compétitivité croissante [125]. Très récemment, Bradlow et al. [22] insistent sur la croissance du rôle de l'analyse des données dans la distribution. Pour Reinartz et al. [104], les entreprises du secteur de la distribution doivent développer de nouvelles méthodes d'analyse de données spatiales pour pouvoir obtenir des informations utiles et des opportunités pour leur développement. Reinartz et al. [104] insistent aussi sur le fait que le défi ne réside pas dans l'accès aux données, mais dans la capacité de les transformer en informations pertinentes pour la prise de décision.

2.1.1 Les contraintes liées à la conception de système d'aide à la décision spatiale

Il y a dix ans, Keenan [70] avait identifié que les décideurs ont besoin d'une analyse de l'information spatiale. Plus récemment, Zhong et al. [135] confirment ce besoin en affirmant que les dirigeants aspirent à tirer parti des nouvelles approches d'analyse de données pour optimiser les opérations et élaborer des décisions stratégiques.

Cependant, la plupart du temps, les outils d'analyse de données spatiales issus de la recherche ne peuvent pas être utilisés directement par les décideurs dans les entreprises. En effet, l'analyse des données nécessite des prétraitements en amont, et des outils sont aussi nécessaires pour permettre une meilleure compréhension des analyses. Ainsi, pour pouvoir être utilisés, les modèles d'analyses de données spatiales doivent s'intégrer dans des systèmes d'aide à la décision spatiale (SDSS en anglais ; pour *Spatial Decision Support System*). Introduits par Armstrong et al. [9] et Densham [39] au début des années 1990, les SDSS doivent permettre la

réalisation de trois phases principales [72, 39, 71],: - Intégration et manipulation des données spatiales ; - Analyses des données spatiales ; - Visualisation des résultats d'analyse spatiale.

Par rapport à l'intégration et à la manipulation des données spatiales, Sugumaran [121] souligne la nécessité de créer un SDSS intelligent qui utilise les technologies disponibles et facilite l'interopérabilité des données et des systèmes spatiaux. Pour Otto [92], la qualité des données est essentielle pour les entreprises si elles veulent les utiliser pour leurs objectifs commerciaux. En ce qui concerne l'intégration des données spatiales, Densham [39] souligne la nécessité de fournir des mécanismes pour l'entrée et l'intégration de données spatiales de manière flexible. Cette intégration des données spatiales est une tâche complexe dont les particularités sont présentées plus en détails dans la section 2.3 de cet état de l'art.

Pour Ren et al. [105], avant l'apparition de l'analyse des données spatiales, les SDSS se basaient sur les connaissances des décideurs et sur des modèles mathématiques. Aujourd'hui les SDSS doivent pouvoir créer de l'information à partir des données qui leurs sont accessibles [105]. Pour Densham [39] les SDSS sont censés inclure des techniques analytiques propres à l'analyse spatiale et géographique et doivent pouvoir s'adapter aux besoins des utilisateurs à mesure qu'ils évoluent. Wanderer et Herle [130] avancent qu'il faut permettre à différents modèles d'analyse d'être intégrés dans le même système. Un cas particulier de modèle d'optimisation sur des données spatiales est présenté dans la section 2.4 de cette revue de littérature.

La visualisation des résultats d'analyse contribue à l'efficacité du processus d'analyse de données [60]. Le défi est de créer des fonctionnalités qui répondent aux besoins des utilisateurs et qui permettent aux analystes et aux preneurs de décisions de penser visuellement [61]. En effet, l'humain apprend plus facilement et efficacement si le support est visuel plutôt que textuel ou numérique [77]. La géo visualisation est un sous domaine de la visualisation qui aspire à tirer profit de l'aspect spatial des données. Pour MacEachren et Kraak [80], la géo visualisation intègre des approches de visualisation venues du calcul scientifique, de la cartographie, de l'analyse d'image, de l'exploration des données et des systèmes d'information géographique. MacEachren et Kraak [80] avancent que cet assemblage de techniques doit pouvoir produire de la théorie, des méthodes et des outils pour l'exploration visuelle, l'analyse, la synthèse, et la présentation de données géo spatialisées. Le but de la géo visualisation est de transformer des grandes quantités de données disparates et hétérogènes en de l'information, et par la suite, en du savoir (compréhension dérivée de l'information) [61]. Les techniques de géo visualisation aspirent à tirer parti des capacités cognitives visuelles telles que la reconnaissance de modèle, la classification, et l'interprétation des repères visuels [61]. Par rapport aux interfaces utilisateur, Khan et Khan [72] mettent en avant leur importance

pour explorer et visualiser les données dans des formats adaptés. Khan et Khan [72] déclarent que les utilisateurs s'intéressent à l'information simplifiée. Ils présentent diverses techniques de visualisation qui améliorent la compréhension des données. Evans et Sabel [44] présentent des techniques qui peuvent être incluses pour améliorer l'utilisabilité, telles que l'ajout ou la suppression d'informations sur une carte ou le filtrage des informations à afficher (par exemple uniquement les informations relatives à l'intérêt de l'utilisateur)..

Un SDSS doit ainsi permettre de prendre en compte les données spatiales, mais comme le mentionne plusieurs recherches, les trois phases : intégration, analyse et visualisation, ne sont pas cloisonnées et interagissent entre elles. Pour Chen et al. [25], c'est en couplant efficacement des représentations visuelles, une interaction flexible et des méthodes de calcul, qu'un outil offre plus de chance de trouver de l'information potentiellement pertinente. Pour Keenan [71], une synthèse appropriée des techniques de modélisation, des interfaces et des approches de base de données doit être développée.

2.1.2 Synthèse

Dans cette première partie de la revue de littérature, nous avons souligné qu'il existe un fort besoin d'outil d'aide à la décision spatiale, en particulier dans le domaine de la distribution. Ces outils doivent permettre la prise en compte des données spatiales à trois niveaux principaux : l'intégration des données, l'analyse et la visualisation, mais aussi que l'interaction entre ces niveaux soit possible. Aujourd'hui, selon la littérature, il n'existe pas d'approche générale qui permettent de passer par ces différents niveaux en tenant compte de l'aspect spatial, tant au niveau des cadres de développement théoriques qu'au niveau des solutions industrielles. Cependant, certains outils répondent à ces besoins de façons partielles : les outils d'aide à la décision non spatiale (DSS pour *Decision Support System*), et les systèmes d'information géographique (SIG). De plus, le développement d'outils d'aide à la décision non spatiale a été décrit dans de multiples recherches. La prochaine section se concentre sur les DSS et les SIG, leurs caractéristiques, et les particularités liées à leur mise en place et leur utilisation.

2.2 Les solutions actuelles pour développer des outils d'aide à la décision et pour prendre en compte des données spatiales.

Cette section se concentre sur les solutions actuelles d'aide à la décision non spatiale et sur les outils de prise en compte des données spatiales. Les limites des approches actuelles dans la mise en place de SDSS complets seront mises en avant. Pour commencer et permettre une

meilleure compréhension, les particularités des données spatiales sont d'abord introduites.

Les données spatiales peuvent être stockées au travers d'unités appelées couches, représentées sur la Figure 2.1 [64]. Une couche peut contenir des éléments similaires, par exemple les différentes routes d'une zone d'intérêt, ou bien la localisation de magasins. Toutes les informations disponibles sur les différents éléments d'une couche sont contenus dans une table de données associée à cette couche. Par exemple, sur la Figure 2.1, trois types de couches, avec sur la gauche, les tables contenant les informations, et sur la droite, les représentations graphiques de l'espace occupé par les éléments. Ces représentations peuvent prendre différentes formes : points, droites, polygone, etc.

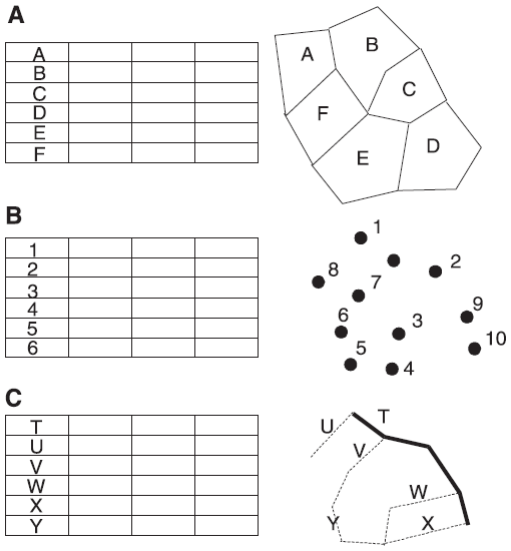


Figure 2.1 Couches de données spatiales ([64])

2.2.1 Systèmes d'Information Géographique

Les Systèmes d'Information Géographique, ou SIG, constituent une base pour la conception d'un SDSS. Les SIG peuvent faciliter les prises de décisions en permettant une intégration des données [64]. Les SIG peuvent contenir de nombreuses couches: routes, magasins, parcs, centres d'intérêt, etc. À partir de ces couches de données, les SIG peuvent créer des cartes suivant les besoins de l'utilisateur, par exemple en permettant de filtrer les informations auxquels l'utilisateur souhaite avoir accès. Contrairement aux cartes papier, les SIG sont capables de stocker, manipuler, et afficher une quantité d'information beaucoup plus grande et de manière dynamique. Les SIG sont aussi capables d'afficher plusieurs couches en même temps, et de placer précisément les éléments relativement par rapport aux autres. La Figure 2.2 illustre

cela en montrant la superposition de couches de la Figure 2.1. Ainsi, on voit instantanément à quelles zones de la Figure 2.1.A, les points de la Figure Figure 2.1.B appartiennent, même s'il n'y a aucun attribut commun dans les données.

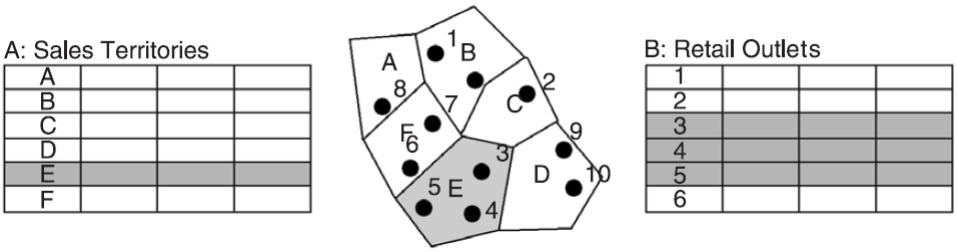


Figure 2.2 Superposition des couches de données ([64])

Un autre point fort des SIG et leur habilité à trouver des relations de localisation entre différents éléments (dans des couches différentes ou dans les mêmes): intersection, densité, distance, etc. Pour Smelcer et Carmel [116], les SIG permettent aux décideurs d'afficher géographiquement les données des tableaux, en tant que cartes, ce qui permet de faciliter la résolution des problèmes de décision spatiale que les décideurs rencontrent.

Les SIG ont souvent été utilisés comme des SDSS, mais de nombreuses limites ont été relevées. Pour Mendes et Themido [82] les SIG n'offrent pas la flexibilité d'autres outils d'analyse et ils ne permettent pas l'intégration d'outils d'analyse de données avancés. Pour West [132] la complexité des SIG rend leur utilisation problématique pour les utilisateurs finaux. Récemment, Vahedi et al. [127] ajoutent que les SIG ne sont pas conçus de manière à permettre aux utilisateurs de trouver et de comprendre des résultats d'analyse sans passer par des procédures souvent compliquées et difficiles à mettre en œuvre. Dans le même sens, pour Keenan [70], les SIG sont des éléments importants du SDSS, mais d'autres modèles et outils dédiés doivent être ajoutés pour créer un SDSS axé sur les besoins spécifiques d'un décideur.

Les SIG ne suffisent pas à la mise en place d'un SDSS, et peu de recherches se penchent sur les caractéristiques que doivent avoir les SDSS. Cependant, de nombreuses recherches existent sur les DSS non spatiaux et sur les processus d'ECD (KDP pour *Knowledge Discovery Process* en anglais). Les DSS non spatiaux permettent d'assister la prise de décisions mais ne permettent pas encore la prise en compte des données spatiales. Cependant, des cadres conceptuels existent pour le développement de DSS non spatiaux. Les processus d'ECD ont été le sujet de multiples recherches pour comprendre chacune des phases du traitement des données [27].

A partir de ces recherches, sur les DSS d'une part, et sur le processus d'analyse des données d'autre part, il est possible de mettre en avant les différentes caractéristiques requises à une

bonne conception et un bon développement.

2.2.2 Système d'aide à la décision et Extraction des Connaissances à partir des Données

De nombreux travaux ([117, 27, 47] par exemple) traitent du processus de l'ECD dans les systèmes d'aide à la décision. Ce processus de développement se compose généralement de six étapes principales. Étape 1 - Comprendre le domaine du problème : définir le problème, déterminer les objectifs du projet, identifier les personnes clés et la terminologie spécifique au domaine. Étape 2 - Collecte, sélection et nettoyage des données. Étape 3 - Préparation des données. Étape 4 - Sélection des méthodes d'analyses de données appropriées. Étape 5 - Évaluation des connaissances découvertes. Étape 6 - Utilisation des connaissances découvertes. Il est important de noter que ce processus n'est pas linéaire et que plusieurs retours en arrière sont généralement nécessaires.

Par rapport à l'étape 1, un point souvent soulevé dans la littérature est que la connaissance du domaine est essentielle au développement du DSS. Cette connaissance permet une meilleure compréhension du problème et augmente la qualité du système développé [1, 64]. Par rapport aux spécifications des DSS, Mendes et Themido [82] soulignent que les principaux responsables sont en mesure d'exprimer leurs besoins.

Au sujet des étapes 2 et 3, relatives à l'acquisition et préparation des données, Cios et al. [27] et Bogorny et al. [17] insistent sur le fait que la préparation des données prend beaucoup de temps (généralement plus de cinquante pour cent de la durée totale du projet). En outre, des problèmes apparaissent lors de la préparation des données spatiales [83, 49], telles que les incertitudes de mesure, l'unité des variables, etc.

Relativement à l'étape 4, sur la sélection des méthodes analyses, Sprague [117] déclare que pendant le processus de développement, les définitions et les objectifs du système changent fréquemment, et cela doit être pris en compte par l'utilisateur et le développeur. Le cadre de développement doit permettre des changements rapides et faciles. Loucks [79] a ajouté que pendant tout le processus de développement, les personnes impliquées en apprennent davantage sur ce qu'elles peuvent avoir, et leurs besoins peuvent évoluer au fur et à mesure qu'ils sont mieux informés.

Par rapport développement d'outil d'aide à la décision, Hess et al. [64] avancent que les différentes parties prenantes s'influencent entre elles, comme illustré dans la Figure 2.3. Ainsi la qualité des décisions proposées peut être améliorée grâce aux compétences des décideurs.

Il est important de noter que ces remarques ne tiennent pas compte de l'aspect spatial des

données. C'est généralement le cas dans les recherches sur les DSS ou l'ECD , qui n'intègrent pas encore les spécificités des données spatiales [100]. Dans ce sens, pour Ferretti et Montibeller [48], plusieurs défis persistent pour la conception de SDSS efficaces.

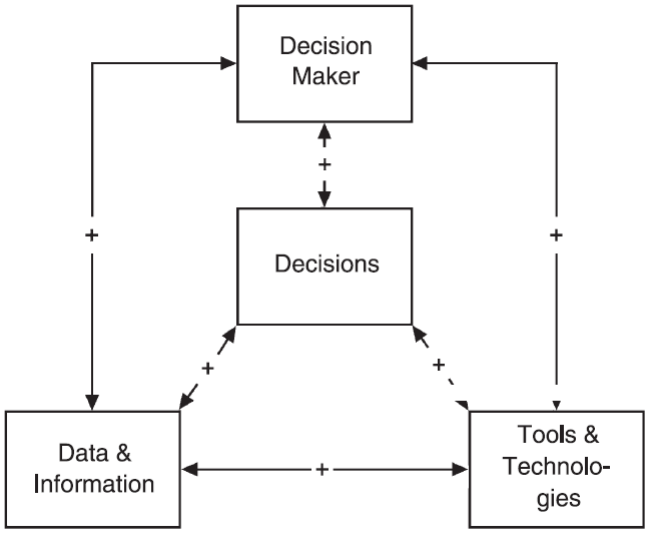


Figure 2.3 Influence entre les parties prenantes au dévéloppement ([64])

2.2.3 Les défis liés à la conception de SDSS

Par rapport au processus de développent des systèmes d'aide à la décision spatiale, Ren et al. [106] indiquent qu'un cadre est nécessaire pour gérer efficacement les données et la logique, et qu'il devrait contenir les différents niveaux d'abstraction de l'information. Par ailleurs, un SDSS peut fournir de l'aide sur plusieurs problématiques d'un décideur, auquel cas, des modèles dédiés à chacune des problématiques doivent pouvoir être intégrés au SDSS. Par rapport à l'ajout de modèle, Sprague [117] propose des fonctionnalités clés pour les systèmes d'aide à la décision : (1) la possibilité de créer de nouveaux modèles rapidement et facilement ; (2) la possibilité de cataloguer et de maintenir une large gamme de modèles ; et (3) la capacité d'interconnecter ces modèles avec des liens appropriés à travers une base de données.

Les approches actuelles ne permettent pas le développement d'outil d'aide à la décision spatiale et le manque de solutions commerciales pour la mise en place de SDSS conduit au développement de solutions SDSS dédiées à des cas d'application (par exemple, voir [130, 134]).

Pour Keenan [70], l'un de défis dans la conception d'un SDSS est de réaliser une synthèse

appropriée des techniques de modélisation et des approches d'interface et de base de données.

Sugurmaman [121] souligne aussi la nécessité de créer des SDSS capables d'utiliser la technologie pour faciliter l'interopérabilité avec les données spatiales. Pour Pretorius et Matthee [100] une méthodologie est essentielle pour que tout projet d'exploration de données réussisse. Pretorius et Matthee [100] ajoutent que : (1) il existe une différence entre les méthodologies utilisées pour l'exploration de données non spatiales et l'exploration de données spatiales ; et (2) aucune méthodologie uniforme et générique n'existe pour l'exploration de données spatiales.

2.2.4 Synthèse

Bien que le besoin en développement en SDSS soit avéré, il n'existe pas aujourd'hui de méthodologie ou de cadre pour procéder à leurs conception et développement. Cependant des recherches sur des systèmes similaires, les DSS et les SIG, permettent d'identifier les caractéristiques que doivent avoir les SDSS. De plus, certaines recherches spécifiques au SDSS ont relevé d'autres défis spécifiques aux données spatiales.

Les SIG ne sont pas adaptés aux décideurs de par leur complexité d'utilisation et l'impossibilité d'y intégrer des outils d'analyse de données spécifiques aux problématiques des décideurs.

Les méthodologies génériques qui portent sur le développement des DSS et les processus d'ECD, n'incluent pas la spécificité spatiale, telles que la préparation des données spatiales et les analyses présentées dans certaines recherches [17, 49] ainsi que la visualisation des résultats d'analyse spatiale. En conséquence, ces méthodologies de développement de DSS et le processus d'ECD sont difficilement applicables directement au développement d'un SDSS.

Il existe des cas d'application présentant des SDSS développés spécifiquement pour des problématiques particulières (comme [44, 114, 130] par exemple), mais ils ne sont pas adaptables à des problématiques autres. Les fonctionnalités abordées peuvent ne pas correspondre à d'autres besoins, où l'adaptation de ces SDSS peut être difficile ou impossible en raison des coûts technologiques ou des informations manquantes sur les technologies utilisées.

Les limites sur les approches pour la conception et le développement d'un SDSS ont été présentées dans cette section. Les caractéristiques à prendre en compte pour le développement de SDSS ont été extraites au travers de trois approches (1) en s'inspirant des domaines d'application voisins ; (2) en étudiant les cas d'application de développement de SDSS dédiés à des problématiques spécifiques ; et (3) en étudiant les recherches liées directement au développement de SDSS et à l'analyse de données spatiales.

2.3 Complexité de la prise en compte des données spatiales dans les outils d'analyse de données

L'exploration de données spatiales, telle que définie par Koperski et Han [75], consiste à extraire des connaissances implicites à partir de données spatiales. Jarupathirun et Zahedi [67] disent que l'exploration de données dans les SDSS permet de découvrir de nouvelles informations utilisables dans de nombreux domaines. La croissance importante du volume de données spatiales et l'utilisation généralisée des bases de données spatiales soulignent la nécessité de la découverte automatisée de la connaissance spatiale [113].

Il y a trente ans, Schmidt [112] avance que les décisions de localisation sont prises rapidement par des personnes sans expérience ni connaissance des problèmes. Les décisions sont prises subjectivement avec peu d'exigences et ne considèrent qu'une petite partie des options possibles. Dans le même temps, Herring et DeBinder [63] soutient que l'utilisation d'outils informatiques peut grandement améliorer le processus décisionnel de localisation.

Il y a quelques années, MacEachren et Kraak [80] notent que de nombreux problèmes dans les domaines scientifique et social ont un aspect spatial. Ils ajoutent qu'à cette époque la quantité de données disponibles contenant une composante spatiale augmente régulièrement. Cependant, en raison du manque de méthodologies appropriées pour les analyser, ces données sont rarement utilisées pour construire des connaissances utilisables dans la prise de décision.

Plus récemment, Mennis et Guo [83] déclarent que l'exploration de données spatiales est un domaine de recherche en vogue. Ce domaine de recherche est une extension de *Knowledge Discovery from Databases* (KDD) introduit par Fayyad et al. [47].

2.3.1 La nécessité du prétraitement des données spatiales

Bogorny et al. [18] avancent que bien que beaucoup de techniques d'exploration de données spatiales sont disponibles dans la littérature, la découverte de connaissances dans des bases de données spatiales réelles est une tâche ardue.

Pour Bogorny et al., il existe peu d'outils qui permettent directement la prise en compte des données spatiales. Pour Sikder [114] aussi, l'intégration de l'analyse des données dans les SDSS est une tâche complexe qui doit être effectuée avec précaution. En effet l'utilisation des données spatiales reste toujours un défi. Erskine et al. [42] et MacEachren et Kraak [80] soulignent la nécessité d'explorer ces sujets afin de simplifier la prise de décisions stratégiques ou organisationnelles. Keenan [70] affirme que l'utilisateur final d'un SDSS n'est pas directement concerné par les techniques spatiales : le système doit permettre à l'utilisateur de contrôler le processus de décision en interagissant avec les modèles ou la version personnali-

sée du système. Densham [39] a aussi souligné la nécessité de fournir des mécanismes flexibles pour l'entrée et l'intégration de données spatiale.

L'un des premiers aspects à prendre en compte lors de la préparation des données spatiales est que les différents ensembles de données doivent être compatibles les uns avec les autres. Cette compatibilité est nécessaire pour pouvoir visualiser les données ou pour étudier les relations spatiales entre les jeux de données [49]. Comme le déclare Mennis et Guo [83], les données spatiales proviennent souvent de différentes sources.

Un concept important à comprendre dans le prétraitement des données spatiales est celui des prédicats spatiaux. Bogorny et al.[18] définissent les prédicats spatiaux comme la matérialisation des relations spatiales, qui ne sont pas stockés explicitement dans les bases de données. Shekar et al. [113] indiquent que les données spatiales doivent être transformées en prédicats spatiaux afin d'être traitées par des outils classiques d'analyse de données.

Ces relations entre les éléments placés dans l'espace peuvent être calculées par des opérations spatiales. Ester et al. [43] et Bogorny et al. [18] désignent trois types de relations spatiales qui existent entre deux entités géospatiales : relations topologiques, relations à distance et relations direction / ordre (voir Figure 2.4, [18]).

Les relations topologiques caractérisent les types d'intersections qui existent entre les éléments [18] ; par exemple : « touche », "contient" ou "se chevauchent" (voir Figure 2.4.A).

Les relations à distance, représentées dans la Figure 2.4.B, peuvent être basées sur des métriques différentes, telles que la distance Euclidienne.

Les relations direction / ordre tiennent compte de la position des éléments les uns par rapport aux autres (par exemple, comme illustré dans la Figure 2.4.C).

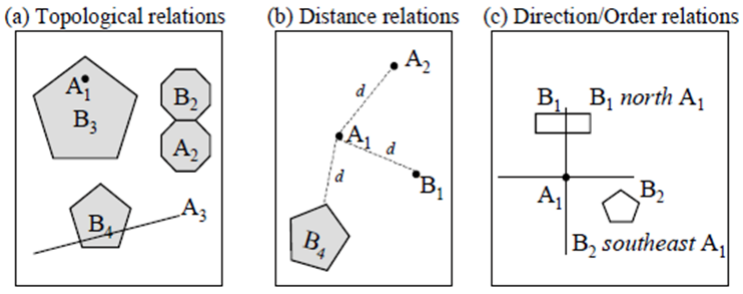


Figure 2.4 Les relations entre éléments spatiaux ([18])

Ainsi, les prédicats spatiaux peuvent avoir différents formats ; par exemple, ils peuvent être en valeurs booléennes pour les relations d'intersection (l'élément A coupe-t-il l'élément B ?) et ils peuvent être numériques dans le cas de la relation de distance.

La préparation des données spatiales doit être effectuée avec prudence, ce qui peut prendre beaucoup de temps, en particulier dans le calcul des relations spatiales. En effet, Flowerdrew [49] soutient que l'intégration des données spatiales n'est pas un processus clair et que, bien que les systèmes d'information géographique (SIG) semblent faciles à utiliser, cela reste un travail minutieux.

2.3.2 Un prétraitement complexe et chronophage

Comme il a été expliqué avant, pour pouvoir procéder à des analyses, une préparation des données spatiales est requise. Une caractéristique dont il faut avoir conscience est que cette préparation des données est une tâche fastidieuse qui requiert des compétences avancées. Pour Egenhofer et Herring [41], l'un des concepts fondamentaux nécessaire pour l'analyse des données spatiales est une compréhension formelle des relations géométriques entre objets spatiaux arbitraires. De plus, West [132] affirme que les SIG sont complexes à utiliser et que leur utilisation nécessite la maîtrise de notions cartographiques complexes qui semblent inaccessibles aux non initiés.

Vahedi et al. [127] ajoutent que les fonctions spatiales et leurs paramètres sont souvent difficiles à comprendre et à utiliser dans les SIG : les analystes ne sont souvent pas familiers avec les SIG et n'ont pas de formation spécifique dans ce domaine. Vahedi et al. ajoutent que l'absence d'une alternative rend obligatoire l'utilisation de SIG pour effectuer les calculs de relations spatiales. Bogorny et al. [18] mentionnent également que ceux qui effectuent l'analyse des données ne sont pas nécessairement des experts dans les bases de données spatiales. Dans le même sens, Appice et al. [7] notent que l'expertise requise pour prétraiter les données spatiales est souvent un obstacle à l'analyse des données spatiales.

Le prétraitement des données spatiales est pourtant nécessaire car la plupart des algorithmes d'exploration de données ne fournissent pas de fonctions de préparation ce type de données [3]. Selon Bogorny et al. [18], cette préparation est essentiellement une étape manuelle, souvent portée par un utilisateur qui n'a pas l'expertise requise. Pourtant la préparation de données est une étape importante et répétitive et la qualité des modèles découverts dépend de la qualité de l'ensemble des données d'entrée.

Dans les projets classiques d'ECD, la préparation des données est considérée comme l'étape la plus longue (prenant entre 45 et 60 pour-cents de la durée du projet) [27].

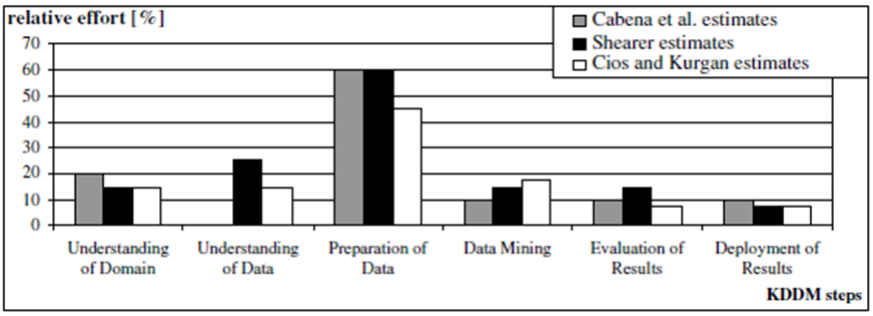


Figure 2.5 Durée des étapes du process d'ECD ([27])

En outre, Pretorius et Matthee [100] soutiennent que la phase la plus touchée par la composante spatiale dans le KDP est le processus de préparation de données (voir Figure 5.3 de [100]).

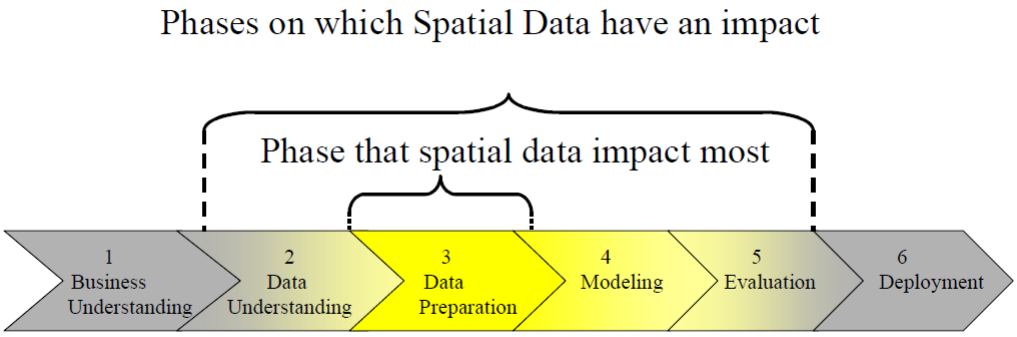


Figure 2.6 Impact des données spatiales sur l'ECD ([100])

Alors qu'il pourrait être envisagé de calculer "toutes" les relations spatiales, Clementini et al. [28] affirment que le stockage des résultats de toutes les relations spatiales est très coûteux en espace mémoire. Clementini et al. en déduisent qu'au lieu de stocker toutes les relations spatiales entre les éléments, il est plus efficace de les calculer lorsqu'elles sont nécessaires.

Un autre aspect à considérer, tel que mentionné par Mennis et Guo [83], est que plusieurs choix doivent être faits au cours de la préparation des données spatiales, par exemple, sur les paramètres choisis ou sur le type de relations à considérer. Alatrista et al. [3] soulignent les difficultés liées au prétraitement en cas de manque de connaissance dans le domaine d'étude.

Dans un grand nombre de recherches sur les analyses de données spatiales, les auteurs n'ont pas forcément conscience de la problématique du choix de la relation spatiale à prendre en

compte. Dans de nombreux cas d'application issus de la littérature, il y a très peu d'explications et de détails sur l'acquisition, le prétraitement et les types de relations spatiales utilisées [73], [101], [52], [129]. D'autres recherches donnent des détails, tels que celles de Evans et Sabel [44] ou Grabis et al. [58] qui tiennent compte des données spatiales, mais ne présentent pas explicitement ces ensembles de données et comment ils ont été rassemblés et transformés. Dans [130] et [5], certaines relations spatiales calculées sont mentionnées telles que la distance, la position, la densité ou la couverture, mais ils ne proposent pas d'explication sur leurs choix des relations spatiales sélectionnées. Toutefois, certains auteurs comme Roig-Tierno et al. [110] fournissent des explications sur les choix qu'ils ont fait.

Bogorny et al. [18] ont déjà identifié certains des problèmes induits par l'aspect spatial des données. Ils ont proposé un environnement avec des outils qui permettent la préparation de données spatiales. L'utilisation de leurs outils, cependant, nécessite une connaissance des SIG, et le calcul des relations spatiales n'est pas automatisé. En outre, leur approche n'offre pas d'assistance pour déterminer les relations spatiales à considérer. Dans une autre perspective, Anselin et al. [6] proposent un outil qui permet de s'extraire de la préparation des données, pour réaliser certaines analyses spatiales (par exemple de la corrélation spatiale), mais les analyses possibles sont limités, et l'outil n'est ni extensible ni personnalisable, ce qui contraint son utilisation.

2.3.3 Synthèse

Les données spatiales ont des particularités que les algorithmes d'analyse de données ne sont pas capables de prendre en compte actuellement. Pour permettre la prise en compte de ces données spatiales, un prétraitement complexe est nécessaire. Ce prétraitement requiert une maitrise des systèmes d'informations géographiques et une connaissance des relations spatiales. Dans le cas des données non spatiales, le prétraitement des données spatiales est déjà la partie la plus chronophage du processus d'extraction des connaissances à partir des données. Ce phénomène est aggravé par l'aspect spatial des données qui rend cette préparation des données encore plus couteuse en temps. Certains auteurs se sont penchés sur ces problématiques, mais les solutions proposées n'y répondent que partiellement : une connaissance des relations spatiales est souvent encore nécessaire, ou les solutions proposées limitent les analyses possibles.

2.4 Les outils d'optimisation de couverture

La croissance de la disponibilité des données spatiales soulève également des défis dans la prise de décision, en particulier dans l'utilisation, de manière appropriée, des techniques de modélisation qui tirent profit des capacités des Systèmes d'Information Géographique (SIG) [70, 71].

2.4.1 Approches d'optimisation de la couverture par modélisation

L'analyse de la localisation est un problème abordé depuis plus de trois décennies [112]. De nombreuses entreprises se basent sur l'expérience des décideurs et sur des critères subjectifs pour prendre leurs décisions de localisation [62]. D'autres approches plus scientifiques existent : de nombreux domaines de recherche dans le secteur de la distribution traitent des questions liées à l'emplacement. Parmi ces recherches, certaines aspirent à créer des approches qui trouvent par elles-mêmes des solutions intelligentes. Un des problèmes de localisation souvent abordé avec une approche de modélisation est celui du maximisation de la couverture (MCLP pour *Maximum Covering Location Problem*).

Le MCLP a été introduit pour la première fois par Church et Revelle [26]. Le MCLP consiste à positionner un nombre fini d'installations afin de maximiser la population couverte. Dans la première définition du MCLP, la population est représentée comme un ensemble de points de demande positionnés à des endroits précis, et l'ensemble des emplacements potentiels pour la mise en place des installations est discret (ensemble d'emplacements précis avec leurs coordonnées). Lorsqu'une installation est positionnée, les points de demande qui sont à une distance de cette installation inferieur à la distance de service sont considérés comme couvert. Dans la définition initiale, un point de demande ne peut être couvert que par une seule installation.

De nombreuse approches pour la résolution du MCLP ont été proposées quand les données spatiales n'étaient pas encore disponibles, ou que les outils pour les prendre en compte n'étaient pas encore opérationnels. Il en résulte que des hypothèses ont dû être faites à l'époque pour faire abstraction de certains aspects spatiaux, et rendre les données compatibles avec les techniques de modélisation disponibles.

Certaines recherches, telles que Berman et al. [15], soulignent que l'objectif de couverture MCLP considère des hypothèses irréalistes: c'est à dire que (1) les points de demande en dehors du rayon de couverture ne sont pas du tout couvert, et (2) La deuxième installation la plus proche d'un point de demande n'a aucun effet sur la couverture.

Par rapport à l'hypothèse classique du rayon de service, il faut savoir que l'estimation des

zones de services des installations est un sujet de recherche à part entière. Pour évaluer les zones de services, de nombreux aspects doivent être pris en compte et ce parce que de nombreux facteurs influent sur la performance ou la productivité d'un distributeur [21] [85], comme la concurrence [40] qui pourrait avoir une influence différente sur les zones de service en fonction des types de magasins [131] [38] [74]. L'aspect environnemental peut également être pris en compte [124] et les données démographiques sont également importantes [93]. Certaines recherches considèrent même qu'il existe différents niveaux de zones de services [11]. Ainsi ces différents éléments devraient être pris en compte pour le calcul de zones de services. Il existe aujourd'hui de nombreuses solutions pour calculer les aires de services, en utilisant les approches Voronoï comme dans Mendes et Themido [82], ou en utilisant le calcul du temps de déplacement dans un SIG comme dans Cui et al. [33]

De plus, beaucoup de recherches qui traitent le MCLP considèrent que les demandes sont situées à des points fixes [96, 107]. Pour Fotheringham et al. [111] la grande majorité des modèles d'allocation de localisation ont utilisé des données de demande agrégées. Ils affirment donc que leur fiabilité est discutable. De même, Current et Schilling [34] soulignent que l'agrégation des données clients est une perte d'informations et peut conduire à des solutions sous optimales. Matisziw et Murray [81] disent que la zone de service peut être interprétée de manière plus large qu'une distance de service.

Plusieurs travaux de recherche visent à intégrer des aspects réalistes dans le traitement du MCLP. Par exemple, Matisziw et Murray [81] autorisent les installations dans un espace continu et non plus discret. Berman et al. [13, 14] permettent une couverture partielle ou coopérative des demandes. Plastria et Vanhaverbeke [97] intègrent l'arrivée compétitive future. D'autres adaptations pour aborder le MCLP de manière plus réaliste, sont présentés dans Farahani et al. [46]. Pour la plupart de ces adaptations, l'approche réside dans l'ajout de contraintes mathématiques et dans la complexification du modèle. Pour les approches mathématiques sur l'analyse de localisation, Revelle et Eiselt [107] proposent une revue des modèles pour l'analyse de localisation.

Ces modèles aspirent à prendre en compte plus d'aspects réalistes, et sont de plus en plus complexes. Mendes et Themido [82] disent que la complexité croissante conduit à une meilleure précision du modèle, mais conduit à des exigences de robustesse et à un coût accru de mise en œuvre et de maintenance. Goodchild [55] avance également qu'il existe des arguments rationnels solides pour l'utilisation de modèle de localisation simple. Dans l'optique de conserver des modèles simples, tout en intégrant des aspects réalistes, il existe également d'autres approches [46]. Par exemple, pour Murray [88], les capacités croissantes des technologies permettent des meilleures solutions dans les problèmes de localisation, en particulier dans les

modèles d'optimisation.

2.4.2 Prétraitement des données spatiales pour l'optimisation

Par rapport aux autres approches qui visent à rendre les solutions plus réalistes, Suarez et al. [123] disent que les SIG ne doivent pas être considérés comme des outils de visualisation uniquement, et qu'ils peuvent être utilisés avec des modèles d'optimisation pour améliorer l'aide à la décision. Dans le même sens, Murray et al. [87] indiquent que les SIG ne doivent pas être utilisés pour fournir uniquement des données d'entrée de base, et que leurs capacités sont sous-estimées dans la conception des modèles. Dans plusieurs recherches [81, 87, 89] Murray et al. insistent sur le fait que les capacités des SIG dans les modèles de localisation offrent beaucoup de perspectives pour fournir des solutions plus réalistes et qu'elles doivent être mieux comprises et mises en place.

Dans ce contexte, Loranca et al. [78] utilisent le SIG prendre en compte les données relative à la demande, mais le processus semble difficile à reproduire ; par ailleurs, ils agrègent encore les données de la demande.

2.4.3 Synthèse

De nombreux modèles de résolution du MCLP, qui ont été proposés lorsque les SIG n'avaient pas encore leurs capacités actuelles, font des hypothèses simplificatrices fortes. Les approches qui aspirent à incorporer des aspects réalistes le font souvent au travers de la complexification du modèle d'optimisation. Des nouvelles approches sont envisagées pour permettre des solutions plus réalistes dans les problèmes d'optimisation, grâce à l'exploitation du potentiel des SIG. Actuellement, il n'y a pas encore d'approche théorique concrète pour aborder le MCLP par une approche basée sur le prétraitement des données.

2.5 Synthèse de la Revue de littérature

Les données spatiales sont de plus en plus présentes, les outils qui permettent de travailler avec sont matures, et les problématiques liées à la localisation sont nombreuses dans le domaine de la distribution. Pourtant le potentiel qui réside dans l'utilisation de ces données spatiales reste trop souvent inexploité pour de multiples raisons :

— **(1)** il est encore compliqué aujourd'hui de mettre en place des SDSS, leur conception et leur développement sont complexes et nécessitent des composants logiciels spécifiques ;

— **(2)** les données spatiales sont difficiles à prendre en compte dans la plupart des algorithmes d'analyse ;

— **(3)** les modèles d'optimisation collent rarement aux réalités des entreprises, il en résulte des solutions dont la qualité est questionnable.

Le constat de ce potentiel inexploité constitue le point de départ de nos travaux de recherche. Les travaux de recherches de cette thèse proposent des approches méthodologiques pour résoudre ces problématiques, tout en étant ancrés dans le réel de par leur développement et leur mise en application pour un partenaire industriel. Les problématiques, le cadre d'application, et la stratégie de recherches sont abordés dans le prochain chapitre.

CHAPITRE 3 DÉMARCHE ET ORGANISATION

3.1 Les problématiques de recherches

Les systèmes d'aide à la décision spatiale prennent de l'importance dans de nombreux secteurs, et tout particulièrement dans le secteur de la distribution. Cependant, le besoin de systèmes d'aide à la décision spatiale ne peut être comblé par les solutions méthodologiques existantes qui se heurtent aux problématiques relevées dans la revue de littérature. Pour permettre de combler ce besoin, il faut se pencher sur les différentes problématiques relatives à la conception des systèmes d'aide à la décision spatiale.

Des méthodologies existent pour concevoir des systèmes d'aide à la décision, mais elles ne tiennent pas encore suffisamment compte de l'aspect spatial de nombreuses données. La première problématique de recherche consiste donc en l'identification des éléments importants à prendre en compte pour permettre une bonne conception et un bon développement du système.

La plupart des outils d'extraction de connaissance à partir des données ne sont pas capables de tirer profit de l'aspect spatial des données sans un prétraitement complexe et chronophage pour les utilisateurs. La deuxième problématique de recherche se concentre donc à faciliter le traitement et l'exploitation des données spatiales.

De nombreux travaux aspirent à rendre plus réalistes les modèles d'optimisation relatifs à des problèmes de localisation, et les approches reposent souvent sur une complexification des modèles mathématiques. La troisième problématique de recherche porte sur l'intégration d'aspects réalistes par d'autres d'approches de modélisation.

Pour permettre une meilleure appréhension de ces trois problématiques, les approches de recherches ont été mise en place au sein d'une structure générale.

3.2 Les approches de recherche

Les trois problématiques énoncées sont souvent liées les unes aux autres dans le secteur industriel. Pour pouvoir exploiter des données spatiales, il faut passer par un processus complexe avant d'être en capacité de les utiliser pour une aide à la prise de décision. Chacune des étapes du processus apporte son lot de difficultés et peut avoir des répercussions sur la qualité finale de l'aide à la décision. Pour s'assurer que nos questions de recherche sont étudiées d'un point de vue global, nos travaux de recherche s'inscrivent dans la conception et le développement

complet d'un outil de traitement et d'analyse des données spatiales. Notre problématique générale de recherche peut donc se formuler ainsi :

"Comment procéder à la conception et au développement d'un outil de traitement et d'analyse des données spatiales pour l'assistance à la prise de décision ?"

Pour bien visualiser comment les trois problématiques de recherches s'inscrivent dans notre problématique globale, la figure 3.1 illustre les interactions entre les différentes parties prenantes.

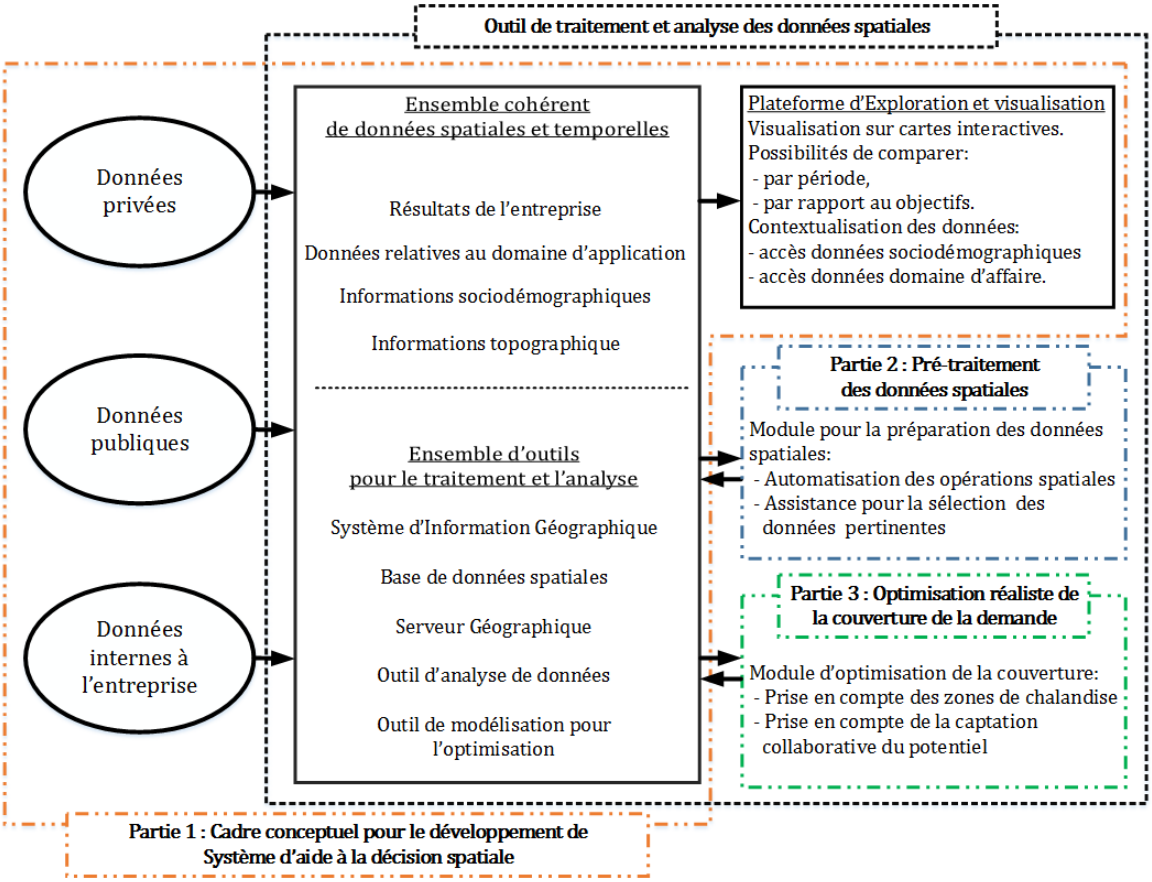


Figure 3.1 Interactions entre les parties prenantes

Relativement à la première problématique, comme il a été mentionné dans la revue de littérature, le développement d'un système d'aide à la décision spatiale requiert l'utilisation de plusieurs outils logiciels et de méthodologies au travers de multiples étapes. Il est apparu dans la revue de littérature que, bien qu'il existe de nombreuses approches pour le développement de système d'aide à la décision classique, l'aspect spatial de certaines données n'est pas idéalement pris en compte. Ainsi, avant même de pouvoir intégrer des modèles d'analyse de données ou d'optimisation, il est nécessaire de mettre en place un cadre pour rendre ces mo-

dèles utilisables. Notre première approche de recherche aspire à proposer un cadre conceptuel adaptable pour permettre de passer par les différentes étapes de mise en place d'un système d'aide à la décision spatiale. Cette approche s'inscrit dans la partie 1 de la figure 3.1

Par rapport à la deuxième problématique : l'intégration de modèles d'analyse de données spatiales au sein de la plateforme est une étape qui arrive naturellement lorsque les décideurs souhaitent obtenir de l'information pertinente à partir de leurs données spatiales. Cependant, bien que de nombreux modèles d'analyse de données soient disponibles, leur utilisation n'est pas initialement accessible avec des données spatiales. Pour pallier cette difficulté, une approche est proposée qui permet un prétraitement automatisé des données spatiales. L'approche proposée est mise en application au travers d'un outil d'assistance au prétraitement des données spatiales (voir Partie 2 de la Figure 3.1).

A propos de la troisième problématique, le problème d'optimisation qui nous intéresse est celui relatif à la maximisation de la couverture (MCLP). Le MCLP a été le sujet de nombreuses recherches, cependant il reste de nombreuses critiques quant à l'adéquation avec la réalité des solutions que les méthodes actuelles proposent. Dans le but de remédier aux critiques sur l'agrégation arbitraire des données relatives à la demande, une approche est proposée dans le chapitre 6. L'approche proposée est basée sur un mélange de techniques de prétraitement des données spatiales, avec des techniques de modélisation mathématique (voir Partie 3 de la Figure 3.1). Pour traiter pleinement la problématique une solution de résolution est proposée au travers d'un algorithme.

Ces travaux de recherche veulent apporter une forte contribution à la conception et au développement d'outils de traitement et d'analyse des données spatiales. Pour valider les contributions de recherches, et permettre une meilleure compréhension, les méthodologies proposées dans cette thèse ont été mises en application au sein d'un outil logiciel de traitement et analyse des données spatiales. Cet outil a été conçu et développé avec des briques logicielles libres et il a été mis à disposition de notre partenaire industriel. Par ailleurs, le traitement de cette problématique industrielle permet d'avoir accès à des données industrielles réelles et de développer un outil en collaboration avec des preneurs de décision du secteur de la distribution. Comme l'avance Hess et al. [64] la collaboration profite à chaque partie prenante du développement de l'outil d'aide à la décision.

3.3 Intégration des problématiques scientifiques dans un contexte industriel

L'entreprise partenaire est un leader canadien dans la production de lambris extérieurs de bois véritable. Il est aussi un important producteur de bardeaux de cèdre blanc en Amérique du

Nord. Notre partenaire distribue ses produits au Canada et aux États-Unis par l'intermédiaire de détaillants partenaires. Dans une optique d'amélioration de son fonctionnement, il envisage de modifier son réseau de distributeurs partenaires pour mieux couvrir la demande. Il dispose des données de ventes mensuelles et de l'adresse de chacun de ses distributeurs partenaires. Ces données brutes ne lui permettent pas en l'état de prendre des décisions éclairées relatives aux partenariats avec ses distributeurs.

Le premier besoin de notre partenaire était de pouvoir explorer ses données dans leur contexte spatial, c'est-à-dire au travers d'une carte et en ayant accès à des informations contextuelles (sociodémographique ou relatives à au domaine d'affaire). La recherche de solutions à ce besoin a permis d'apporter des éléments intéressants et complémentaires de la littérature sur la problématique du cadre conceptuel pour le développement d'outils d'aide à la décision.

Le deuxième besoin de l'industriel résidait dans la compréhension des facteurs spatiaux qui influencent les performances de son réseau de distribution. Des outils d'analyse de données pour extraire des facteurs de performances existent, mais ils ne sont pas capables de travailler directement sur des données spatiales [18]. La problématique du prétraitement des données spatiales a pu être abordée dans ce cadre. Par ailleurs, la recherche de réponses à ce besoin a permis d'identifier des éléments importants relatifs à l'intégration des modèles d'analyse dans un SDSS. Ces éléments ont permis d'illustrer concrètement des points relevés par la littérature.

Enfin, le troisième besoin de l'entreprise résidait sur la proposition de solutions d'amélioration des partenariats avec des distributeurs (ouverture ou fermeture de magasins pour maximiser de couverture). Les modèles d'optimisation de la couverture disponibles ne semblaient pas prendre en compte des composantes réalistes importantes, comme la compétition ou la notion de zones de chalandise. À partir de ce constat, et grâce à l'ensemble des données réuni et du système en place, il a été possible d'aborder la troisième problématique relative à une approche de modélisation associée au prétraitement des données apportant ainsi une approche novatrice par rapport à la littérature.

Le traitement des problématiques de recherches au travers du cas d'application industriel a permis de proposer en parallèle des contributions scientifiques et industrielles.

3.4 Contributions

La première contribution scientifique réside dans un cadre conceptuel pour le développement d'un SDSS. Cette contribution est validée au travers d'une contribution industrielle : la mise en place d'un d'un SDSS pour notre partenaire industriel. Le chapitre 4 présente le cadre

conceptuel et l'illustre au travers du cas d'application.

La deuxième contribution scientifique réside dans la proposition d'une approche pour automatiser le prétraitement des données spatiales. L'approche est validée au travers d'une seconde contribution industrielle, le développement d'un outil permettant de rendre possible l'analyse des données de notre partenaire. Le chapitre 5 présente cette contribution et illustre sa mise en en application.

La dernière contribution scientifique consiste en une proposition d'utilisation conjointe de modélisation mathématique et-de prépaiement des données spatiales pour permettre d'intégrer des aspects réalistes dans la résolution d'un problème d'optimisation. Cette contribution est aussi accompagnée d'une contribution industrielle ; un outil de résolution permet de proposer à de solutions d'amélioration de la couverture qui tiennent compte d'aspect réaliste. La contribution est présentée et illustrée sur notre cas d'application dans le chapitre 6.

Les trois contributions intégrées au sein d'une même structure ont permis de répondre à notre problématique de recherche générale. Des méthodologies sont proposées pour permettre la conception et le développement d'un outil de traitement et analyse des données spatiales pour l'aide à la décision.

CHAPITRE 4 ARTICLE 1: CONCEPTUAL FRAMEWORK FOR SDSS DEVELOPMENT: APPLICATION IN THE RETAIL INDUSTRY

Submitted to Business & Information System Engineering

Authors: G. DARAS, B. AGARD, B. PENZ.

Abstract. Spatial information is becoming crucial for strategic decision making, but accessing and understanding this information is not easy. Dedicated tools can support the decision process in many ways, such as visualization interfaces or data analyses. Numerous Decision Support System (DSS) development methodologies exist along with dedicated Spatial Decision Support System (SDSS). Unfortunately, for multiple reasons, these tools and methodologies are not easily adaptable for the development of other SDSS.

This paper proposes a framework for the development of a flexible SDSS that is built on open source software, allowing for low cost implementation. To support the efficiency of our approach, the design of a specific SDSS that is currently in use will be presented. This SDSS was developed for a company that distributes products through various retail networks. The multiple capabilities of the resulting SDSS will be revealed through an explanation of the different development steps. The complete framework is applied on a real data set that will be detailed in demonstration.

4.1 Introduction

For decades, public and private organizations from various domains have accumulated large amounts of information about their day to day activities. In many cases, collected data includes spatial references that may describe where an activity occurred. Nonetheless, the storage, management and use of spatialized data is still a challenge, and Erskine et al. [42] and MacEachren & Kraak [80] have highlighted the need to develop those topics in order to simplify strategic and organizational decision-making.

Researchers have focused on many of these topics, and as a result, more and more tools are available to valorize spatial information. Tools developed to achieve this goal are frequently called Spatial Decision Support System (SDSS) introduced by Armstrong et al. [9] and Densham [39] in the early 1990s. Then, Crossland et al. [32] state that SDSS allow for

improvement in Decision Maker performances. While many SDSS development projects were made, there are still challenges noted in the literature. Hernandez [61] claimed that a part of the challenge is to provide functionalities that meet the needs of decision makers and that facilitate adequate visualization of available data. Sugumaran [121] highlighted the need for building intelligent SDSS that uses available technologies and facilitates the interoperability of spatial data and systems. For Keenan [71], an appropriate synthesis of modeling techniques, interfaces and database approaches needs to be developed.

Knowledge about spatial environments may be a key element in improving understanding and performance in many fields, in particular for the retail sector [29, 40] that is facing many challenges [104]. The growing interest in techniques for the analysis and understanding of spatialized data follows growing competitiveness [125]. More recently, Bradlow et al. [22] insist on the growth of the role of data analysis in retailing. Keenan [70] identified that top managers need spatial information analysis and Mendes & Themido [82] outlined that top managers are able to express their needs. Unfortunately, most managers do not have the time and interest to learn how to use complex spatial data analysis systems or to develop their own mathematical models. In many situations, a lack of commercial solutions leads to the development of dedicated SDSS solutions (as in [130, 134] for example). Unfortunately, the developments recommended in that research are difficult or impossible to adapt to different case studies. Table 4.1 in section 4.2.3 exposes the main reasons for inadaptability that are encountered.

The contribution of this research is not in the development of a new theory, but in the proposition of a new conceptual framework that is easily adaptable for many SDSS development. The proposed approach allows for an integration of various existing methods and tools. For better clarity, demonstrations will be based on a real case study. While SDSS might focus on many research areas related to retailing, we will consider the development of a SDSS dedicated to the evaluation of the commercial network of our partner company.

The next section (4.2) presents relevant research in the domain. It focuses on the design principles and development architecture requirements as well as methods. Section 3 exposes the contribution of the paper. In section 4.3.1, the context of the case study is introduced. Section 4.3.2 explains the approach proposed in this paper, the architecture and selected tools will be highlighted. Then, in section 4.3.3, a step by step explanation, based on the case study, enables the proposed approach to be clearly illustrated. In this section, the needs of the case study are formalized. This shows how to meet those needs in the proposed framework along with the result of this implementation. Section (4.3.4) aims to demonstrate the SDSS adaptability by transferring previously developed applications to another geographic area.

The final section (4.4) concludes the paper and proposes future research directions.

All steps of the contribution have been validated in the case study with real data from the researchers' partner along with various public and private data. For reasons of confidentiality, all financial parameters included here have been transformed or masked.

4.2 Literature review

Technologies linked to SDSS development evolve rapidly and it is possible to identify several papers focusing on design principles and development architecture requirements (section 4.2.1). Some design and development methodologies from related fields: Decision Support System (DSS) and Knowledge Discovery from Databases (KDD) are also presented (section 4.2.2).

4.2.1 Design principles and requirements of development architecture

According to many researchers [72, 39, 71], SDSS consist in three main fields: **(1)** spatial and non-spatial data integration and manipulation, **(2)** spatial and non-spatial data analysis, and **(3)** spatial and non-spatial data and results visualization.

(1) As mentioned by Otto [92], data quality is critical for enterprises if they want to meet their business requirements. Regarding spatial data integration, Densham [39] highlighted the need to provide mechanisms for the input and integration of spatial data in a flexible manner. With regards to the spatial aspect of data treatment, Evans & Sabel [44] pointed out the efficiency of the PostGreSQL-PostGIS solution that allows spatial queries on the data to be made.

(2) On spatial data analysis, SDSS are supposed to include analytical techniques that are unique to both spatial and geographical analysis (including statistics) [39] and they must be able to adapt to the needs of the users as they evolve. Wanderer & Herle [130] advanced that it must be profitable to allow different models of analysis to be plugged into the same framework. Zhang et al.[134] insisted on examining the performance aspect if models are applied on the user's computer. Vatsavai et al. [128] pointed out that, in a classic fat server-light clients approach, the server could be overburdened by data access and spatial analysis while in the rich client-light server approach, the same limitations could occur on the client side.

(3) Regarding the visualization part, Lloyd [77] advanced that SDSS often use Geo-visualization techniques because humans learn more easily and efficiently when the support is visual rather

than when it is textual or numerical. Khan & Khan [72] said that the users are interested in simplified information and presented various visualization techniques that improve the understanding of the data. Evans & Sabel [44] presented techniques that could be included to improve usability, such as the addition or removal of information on a map, or a focus on specific aspects. Knezic & Mladineo [73] presented a multi-level approach that allows for a change in spatial granularity. Since those particular interactions are not easy to develop, the architecture must include libraries that allow for those interactions such as Openlayers, which is presented as an efficient solution in [2].

Two other points that are not specific to SDSS are accessibility and cost. On accessibility, Granell et al. [59] argued that web-based tools, including web-based SDSS, do not need any software to be installed, or mastered, so that they can be used by anybody, anywhere. Rinner et al.[109] added that there is a growing availability of free web services, which do not need knowledge of digital maps in order to be used (such as Google Maps [57]).

Concerning costs, many researchers state that it is one of the advantages of open source technologies [108, 86, 44]. Moreover, large open source communities are devoted to helping and sharing ideas [44].

There is little research linked to SDSS development process and methodologies. However, research has focused on similar domains such as DSS (Decision Support System) or KDD (Knowledge Discovery in Databases) applications. In the next section, the most accepted methodologies will be presented. Identified weakness when dealing with custom SDSS development will be indicated.

4.2.2 Design and development methodologies

Many papers ([117, 27, 47] for example) have identified the process of extracting information from data in decision systems. This development process is composed of six main steps. Step 1 - Understanding the problem domain: defining the problem, determining the project goals, identifying key people and domain-specific terminology. Step 2 - Data collection, selection and cleaning. Step 3 - Data preparation. Step 4 - Selection of appropriate Data mining methods. Step 5 - Evaluation of the discovered knowledge. Step 6 - Use of the discovered knowledge. This process is not linear and multiple steps back are usually necessary.

A common point found in the literature is that domain knowledge is essential to DSS development. It allows for better comprehension of the problem and increases the quality of the developed system [1, 64].

Cios et al. [27] and Borgony et al. [17] insisted that the data preparation is very time con-

suming (generally more than fifty percent of the total project duration). Moreover, other problems will appear when dealing with spatial data preparation [83, 49], such as measurement uncertainty, biased sampling, varying area unit, etc.

On the DSS development, Sprague [117] said that during the development process, the system definitions and objectives change frequently, and this has to be taken into account by the user and the developer. The development framework must permit quick and easy changes. Loucks [79] added that during the whole development process, people involved learn more about what they can have, and their needs may evolve as they become better informed.

4.2.3 Synthesis

According to what was found from the state of the art, many methodologies in close fields have been proposed and various SDSS have been developed. Those contributions may be separated into two categories: on one side, generic methodologies, and on the other side, case-base specific applications.

Generic methodologies mainly focus on DSS development and KDD Process but do not include spatial specificity, such as spatial data preparation and analyses that have been presented in some research [17, 49]. Then, it is not possible, or it is difficult, to apply them to the development of a SDSS.

Case study papers present SDSS developed specifically for a situation, but proposed functionalities may not correspond to someone else needs. Even if the need is the same, adapting those SDSS may be difficult or impossible because of the technology costs or missing information about involved technologies. Most of the time, authors explain the functionalities of the tools developed but not how they were developed, what kind of databases, what kind of objects, or which software were used (cf Table 4.1). In other cases, those elements are clearly explained but the functionalities are too specific and cannot be adapted easily in another situation (cf Table 4.1).

Benoit & Clarke [12] said that simple models are unlikely to succeed in many applied market situations and that customized solutions will lead to a better capture of complex situations. While many methodologies look like the generic one presented by Cios et al. [27], they often aim at one model project. Yet, many companies might need several decision tools. The data preparation phase is tedious and if not made in an adaptable way, it might have to be re-done for every situation. Furthermore, this adaptability is even more critical considering the evolving nature of custom SDSS as decision makers clarify their needs.

SDSS tools and methodologies are currently developed in many application domains such as

Table 4.1 Difficulties identified to transfer other research to various contexts

Reference	Metho-dology	Frame-work	Case Study	Main Focus	Difficulties identified to transfer to various contexts
[44]			x	Open source SDSS for health and environment	There is no possibility to implement advanced model analysis. There is no tool proposed for development. There is no spatial data treatment tool for advanced spatial data manipulation.
[114]			x	SDSS in environmental assessment	The proposed architecture is global. There is no open source tool proposed.
[110]	x		x	Retail site location decision process	The proposed process is dedicated to making a decision about retail location. There is no tool proposed for development.
[17]		x		Data preparation	The proposed architecture is global. There is no open source tool proposed. The user interface is not taken into account.
[52]		x	x	Design and implementation of web-based platform	There is no open source tool proposed.
[86]		x	x	Open source Spatial Information System	There is no spatial data treatment tool for advanced spatial data manipulation. There is no possibility to implement advanced model analysis.
[59]		x	x	Reusable geospatial services	Not on dedicated SDSS development.
[109]		x	x	Web framework for deliberation in spatial decision making	There is no spatial data treatment tool for advanced spatial data manipulation. There is no possibility to implement advanced model analysis.
[73]	x	x	x	DSS based on GIS	There is no open source tool proposed.
[68]		x		Web DSS for e-retail industry	The proposed architecture is global. There is no open source tool proposed.
[25]			x	Interfaces interaction visualization techniques	There is no global architecture proposed. There is no open source tool proposed.
[2]		x		Comparison of open source web-based GIS	There is no possibility to implement advanced model analysis.
[130]			x	Multi criteria SDSS	There is no open source tool proposed.
[134]		x		Web-based SDSS framework	There is no spatial data treatment tool for advanced spatial data manipulation.
[136]	x	x		Knowledge-based approach for SDSS development	There is no open source tool proposed.
[117]	x			DSS development framework	There is no open source tool proposed. There is no spatial data treatment tool for advanced spatial data manipulation.
[79]	x			DSS development & implementation	There is no open source tool proposed.
[47]	x			KDD process	There is only methodology. There is no tool proposed. There is no user interface.
[27]	x			KDD process	There is only methodology. There is no tool proposed. There is no user interface.

environment [52, 59, 130], healthcare [86], agriculture [114] and others [110, 68, 109] and may be useful in many other research and professional domains.

Table 4.1 makes a synthesis of the difficulties encountered when trying to adapt other SDSS to our case study. The difficulties identified in those papers relatively to the flexibility are identified. The focus is not to be exhaustive, but rather to highlight the need for a flexible tool that offers easy adaptability to various contexts.

Considering these elements, the focus of this paper is to propose a conceptual framework, based on a SDSS architecture and a development approach that could be adapted to various distinct applications. Those may serve as a foundation for the development of any SDSS application. The proposed architecture is based only on open source software to allow for a low cost implementation.

The SDSS was effectively developed, and is currently in use, by the partner company.

4.3 Framework presentation through a case study

The dedicated SDSS and how it was implemented is presented in this section. For better clarity, the following demonstrations will be based on a real case study. This section details the case study, the architecture of the SDSS and its main functionalities, and shows its adaptability. Each time it is necessary, we explain why the proposed solution is flexible and can be used in various contexts.

4.3.1 The needs of a retail company

Our industrial partner is a Canadian leader in lodge building material, based in Quebec (Canada), with a strong presence in North America, mostly on the East Coast. The partner manufactures different products and distributes them through about 700 dealers that are present in about 50 commercial networks. Considering growing competition, a consolidation and a better understanding of the market and commercial performances is needed.

The first need is therefore to assess the performance of the current commercial network, and then to evaluate the impacts of any changes in this network. Many questions have arisen immediately that concern the location of the dealers on the territory, as well as individual and global performances.

Furthermore, individual and global performance depends on the context. It is easier to run a restaurant downtown in a large, touristic city than in the suburbs of a small industrial one. The relevant local environmental context of each retailer needs to be extracted, in parallel

to the retailer's historic sales, before any evaluation is done. The SDSS presented here will allow for all of that information to be extracted locally for each dealer and globally for the network.

4.3.2 SDSS applications, development process, and tools required

Our proposed approach of SDSS development process (composed of six main steps: A to F) is represented in Figure 4.1. This figure presents a simplified view of the workflow, that do not aspire to be a theoretical methodology for SDSS development. This workflow will on one hand, structure the selection of the necessary tools to develop the SDSS. On the second hand, it support the development timeline of our case study in section 4.3.3.

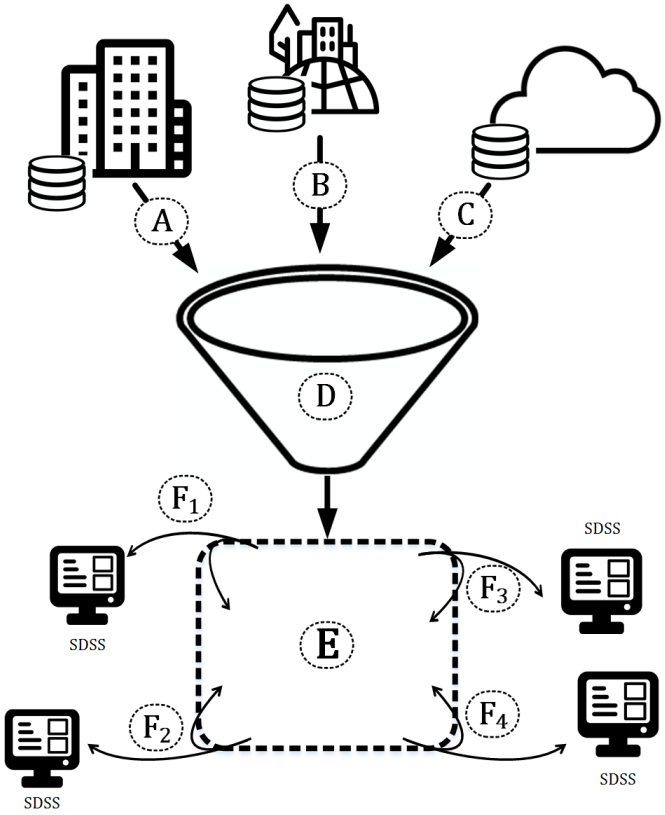


Figure 4.1 SDSS development workflow

Steps A, B and C consist of data collection. Many sources may be concerned as soon as those sources provide useful information for the case study. In each case study, the collected data vary and this step needs to be carefully observed. It is a critical aspect and bad data collection will simply lead to no answer, or worse, to wrong answers in the final steps.

Currently, many datasets about socio-demographic data, statistics, roads, circulation, and

so on exist and can be downloaded from different places. It is important to carefully look around at what are the possible sets of data and to be sure that the data is clean, structured and understood. Synthetic data sets are usually poor and may not be treated. Disaggregated data sets are much richer but also more difficult to treat.

For the case study presented here, among the different datasets used, the following are the most important:

A - The company provided raw data on a retailer's sales per month. The initial dataset concerned the province of Quebec (Canada). For each retailer, the following data was available: name, address, partner representative, retail network, and the sales of each month.

B - As the sales are influenced by socio-demographic characteristics, specific points of interest and construction activities, the following data were included:

— socio-demographic data from the national census bureau [118],

— localizations of golf courts [54], ski resorts [115] and large lakes [118],

— statistics on construction: building permits [118] and housing starts [31].

Some of these datasets had to be purchased from Canadian Mortgage and Household Corporation (CMHC) and Statistics Canada.

C - Geographical territory description.

The province of Quebec is subdivided into areas of various sizes. Different levels of aggregation are available; here the "Census Subdivision" (CSD) was selected. It is a level that satisfies both the level and accuracy that is expected for the case study and conforms to the datasets available in A and B.

Depending on the project's needs, other area sizes might be chosen, such as Province or Census Division (CD). As the future steps propose to develop a generic data treatment and tools, the adaptation to other areas will be easy. All of these datasets cannot be imported as they are in the SDSS database. Data treatments and integrations presented in part D focus on this.

D - Data treatment and integration.

As mentioned in the literature review, non-spatial and spatial data treatment and integration are tedious and complicated tasks. Also, with the goal of developing flexible SDSS, those tasks may be repeated several times and must be imported properly.

Moreover, the different datasets in A, B and C may be in different formats and may contain various problems. For each dataset, a specific engine may be required; those engines have to identify errors and adapt the data format to the SDSS database in E. Among the engines used for our case study, the following are crucial:

— address Geocoding (convert address to geographical coordinates),

— duplicate removal,

— area simplification (to improve interface responsiveness),

— XML, HTML and CSV data parsing,

— inconsistent data homogenization (punctuation, accents, etc).

This step will allow for clean datasets in the input for the development of the SDSS (step F).

E - A global architecture based on flexible tools.

This section describes the architecture and tools recommended to develop a flexible SDSS. In order to avoid server or client overload, two distinct categories of users were considered, with different needs and permission to access the server (see Figure 4.2).

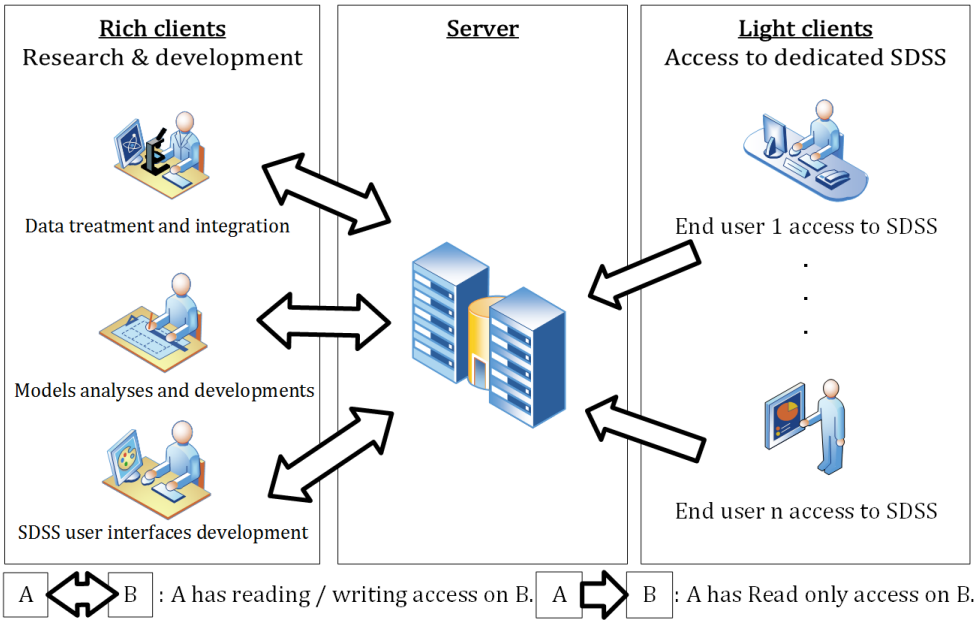


Figure 4.2 Schematic architecture

The first category of users is called Rich clients. Here, we may find users that will proceed to the analysis, modeling and development stages. Those users can read as well as write on the databases located on the server; besides, most analyses and computation run on their own

computers. The second category is called Light clients, they have read only access to the databases. They use the tools and interfaces developed by the Rich clients. For them, the server supports the computing; it enables access to the visualization interfaces through web browsers. It allows flexible access with any light device, such as small computers, tablets, and smartphones.

The global architecture, presented in Figure 4.3, shows a set of efficient tools selected for the data integration and for the development process.

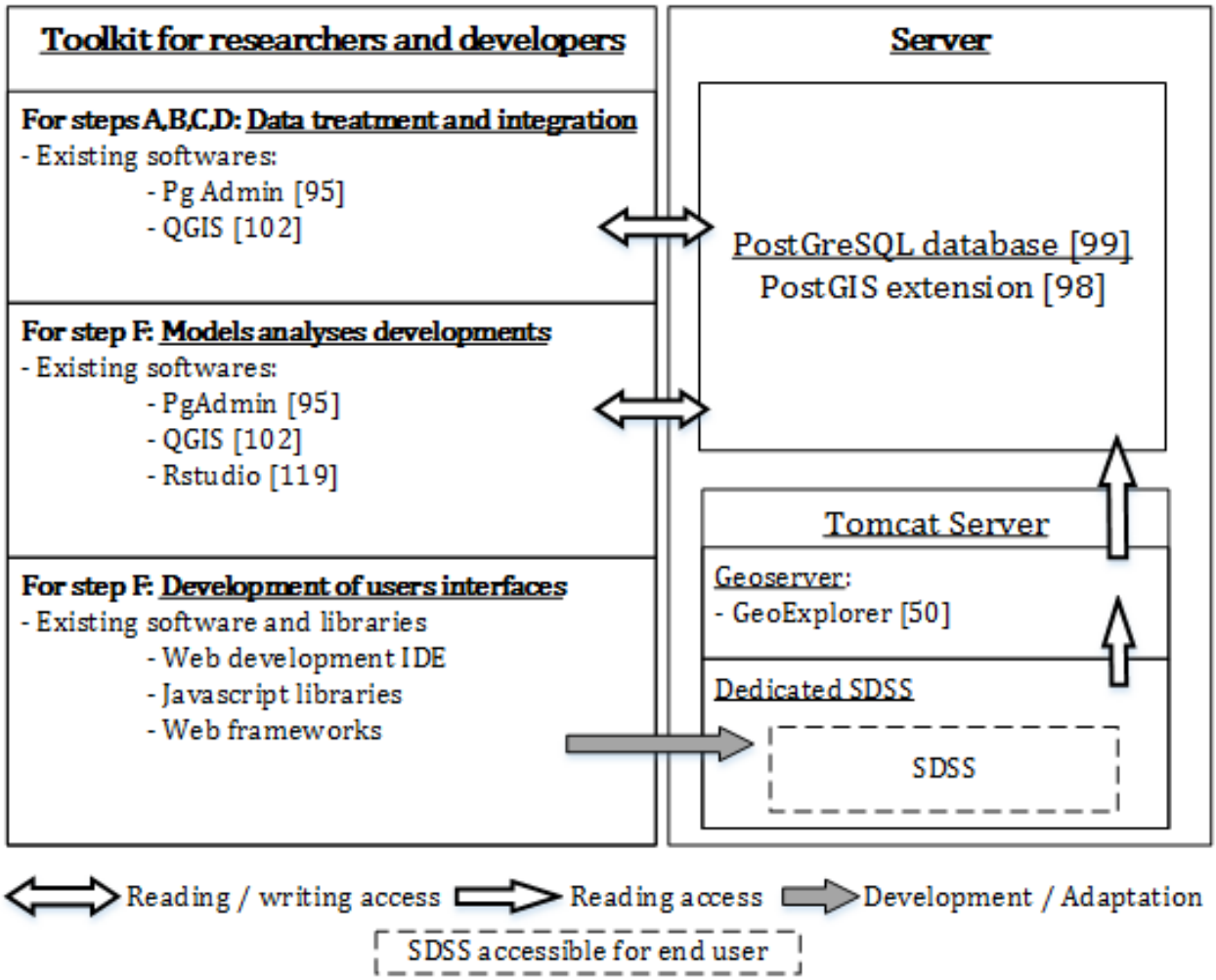


Figure 4.3 SDSS architecture and development toolkit

The tools proposed to achieve SDSS development have been selected according to the following elements:

— They were presented as efficient in the recent literature [134, 44, 2].

— They have native interfaces that facilitate interaction and integration.

Many of those tools can be found from the boundless geo suite [20].

The architecture is composed of two main parts, **(1)** the toolkit for the researchers and developers and **(2)** the server.

(1) The toolkit will allow three main tasks to be performed, data treatment and integration, model analysis development, and user interface development:

— For data treatment and integration, PgAdmin [95] allows for the database to be administrated and for manipulations on stored data to be performed. A GIS software is included for the manipulation and the treatment of the spatial data. For these tasks, QGIS software [102] is proposed. In addition, different generic data treatment engines are developed during the SDSS development process.

— For the model analyses and development, as previously mentioned and pointed out in the literature [130], the data might be used in multiple models. PgAdmin and QGIS might be used to perform basic analysis, but are not sufficient to process advanced modeling tasks. To overcome this, the R programing language [103], easily usable through the RStudio IDE [119], is proposed. As in other steps, models developed here are generic and reusable for multiple SDSS applications

— For the user interface development, many open source tools and libraries exist. As it is neither our focus nor our area of expertise, no specific software is proposed in our approach. However, for information, the used tools for each web interfaces development are cited in next section (4.3.3). Once again, flexibility in interface development must be thoughtful to allow codes reusability and adaptability for future development (as in section 4.3.4).

(2) The server contains the database and a web server. For the data storage, a lot of databases exist on the market. Few of them allow spatial data management. PostGreSQL [99] has been selected due to the PostGIS [98] extension that is available and allows to work easily with spatial data. A Tomcat web server [126] stores a GeoServer [51] and the SDSS Websites. GeoServer will allow easy access to spatial content through web services and/or through GeoExplorer [50].

F - Development of specific functional tools.

In A, B and C, we select relevant data for the case study. D proposes engines to prepare the data. E presents the a global architecture and existing software that supports the SDSS.

F focuses on the development of specific tools adapted for the case study. Rich clients (researchers and developers) have to design the tools and interfaces for the Light clients (end users).

The framework proposed in this paper enables extreme flexibility for adapting the tools to any case study. The following section details the development of various functional applications, based on the case study presented in section 4.3.1. For example: retailers' visualization, network identification, sales, socio-demographic parameters around each retailer, localization of the relevant points of interests, cover optimization and area segmentation are some of the proposed functionalities.

The case study validates the proposition and shows results based on real data. Once again, the framework is very flexible and the user simply needs to change the input data A, B, C and develop any functional tool to adapt to its situation. The development of functional tools is detailed in the following section (4.3.3), and the adaptability of the whole system is demonstrated in section 4.3.4.

4.3.3 SDSS development

The proposed approach was tested for the development of a real application. The developed dedicated SDSS were developed with a step-by-step validation with our industrial partner. The SDSS were considered approved when our industrial partner validated their efficiency and their usefulness for the business. Several SDSS are currently used by our partner.

When toolkit and servers (Figures 4.1 and 4.3) are set up, it is possible to plug in various dedicated applications. Several applications developed for the case study are presented next. For each application *Fi*, the need, the methodology, and the resulting tools are exposed.

F1 - Geographic assessment of dealers' sales

Need

As for many other companies, an initial assessment of the dealers' sales was provided by tables and charts. One aspect that is not easily understandable through that format is the geographical performance. Geographical assessment is a valuable insight to better understand the strengths and the weaknesses of a distribution network.

A first application was developed that aimed to present retailers' yearly performances. To allow for a quick analysis, the sales performances are presented through a colorimetric scale. If needed, detailed sales are available through their original format by clicking on each desired

retailer.

Methodology

First, data acquisition (step A) and integration (step D) were processed.

Step A: for each retailer, the input data is available in a tabular format with name, address (number and street), city, postal code, company representative, retail network and the sales of each month.

Step D consists in address geo-coding. This means converting the addresses to corresponding geographical coordinates. To get geographical coordinates from a proper address, a Google geocoding tool is available [56]. Google service parameters have to respect some rules, such as the address element (number, street, postal code, city) specifically formatted. Then a request is sent and the answer is formatted according to our system: geographical coordinates are extracted, and translated in the right projection in the database.

The model developed for this first tool consists in basic computing realized through a simple SQL request. An example is presented here that corresponds to the computation of yearly sales.

Monthly sales are saved in dedicated columns (see Table A in Figure 4.4). A yearly sales column is created and computed (see Figure 4.4.B) resulting in a modified dataset (see Figure 4.4.C). The resulting dataset is then easily used at any moment. For example, it is possible to offer a visual representation of the performance of each retailer. Here, a colorimetric scale was chosen and it is easy to set specific bounds to identify a different level (strong, neutral, weak) as represented in Figure 4.5.A.

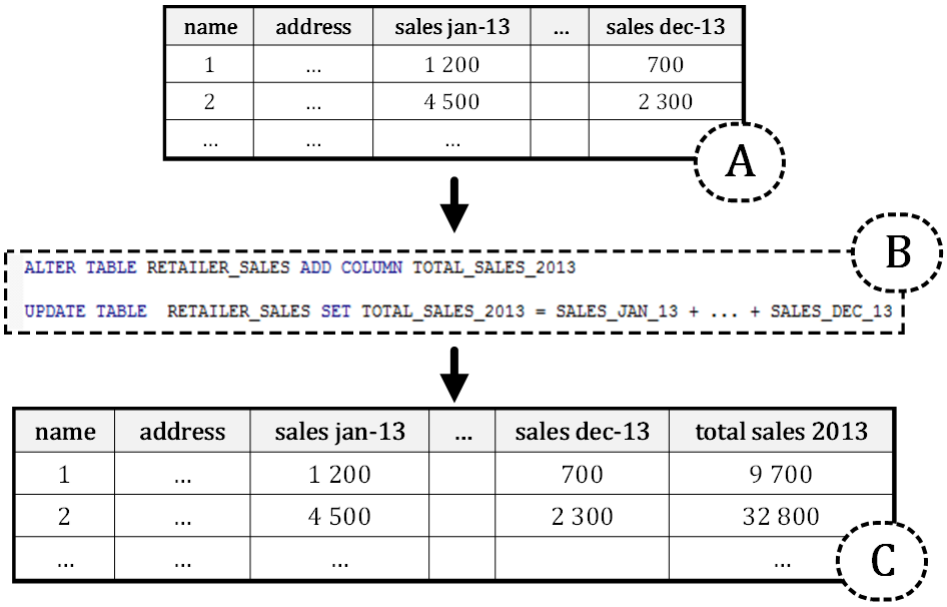


Figure 4.4 Dedicated sales sum query

Results

For this first application, geoexplorer [50] offers tools that allow easy layer management and styling (Figure 4.5.A). The main interface is presented in Figure 4.5.B . Figure 4.5.C, shows that the detailed information may be available after user interaction (a left click on the desired retailers).

This interface permits the easy visual representation of a retail network as well as a visual representation (green, white, red) of actual dealers' performances over any selected geographic area.

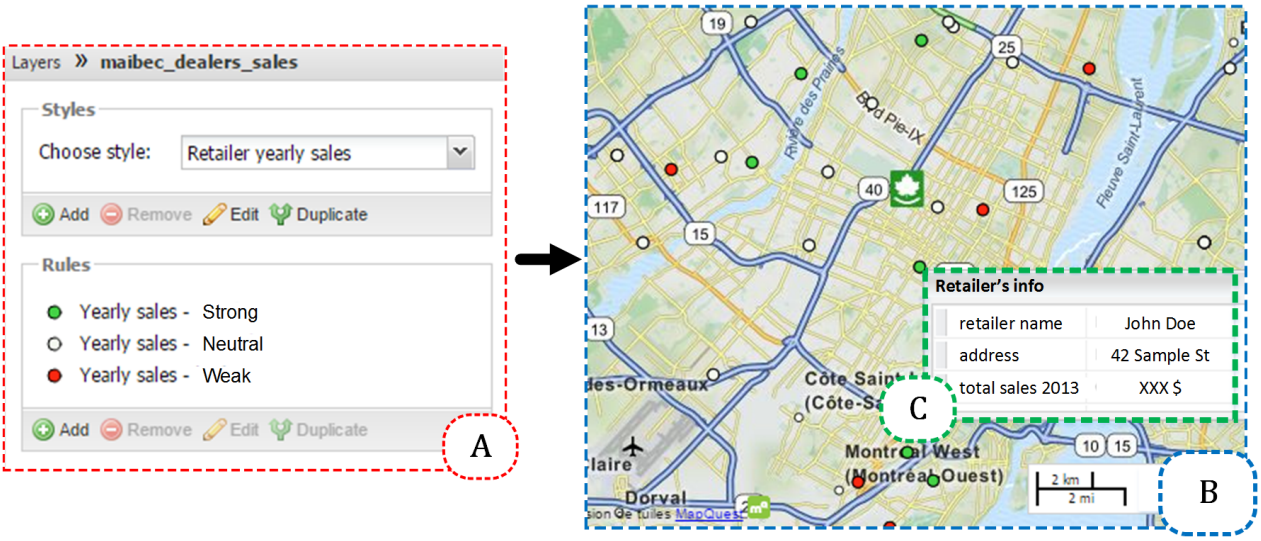


Figure 4.5 First visualization interface

F2 - Sales aggregation per area

Need

While assessing each retailer's performance, decision makers may also need aggregated information for specific areas.

Methodology

Step C consists in area file acquisition. First, the aggregation level has to be chosen. For this case study, the Census Subdivisions (CSD) were selected for multiple reasons: CSD areas cover the province of Quebec. Their size is adapted to obtain insightful data, and this fits with most other census datasets.

The CSD file limit consists of sets of polygons (defined by a list of geographical coordinates). For the Province of Quebec, the dataset is composed of 1285 census subdivisions, and for each one, the name, an identifier, and the associated polygon.

In order to perform some spatial data aggregation, spatial data queries must be processed.

```
1  SELECT Area.id, count(PointOfInterest) AS Cardinal
2      FROM Area LEFT JOIN PointOfInterest
3      ON st_contains(Area.geometry,PointOfInterest.geometry)
4      GROUP BY Area.id
```

Figure 4.6 Point in area cardinal query

Spatial queries (allowed by spatial database) permit, for example, to extract the number of specific elements in a designated area (for example, see query in Figure 4.6). It also permit to get the sum of the sales for all the retailers in the area (illustration in Figure 4.7).

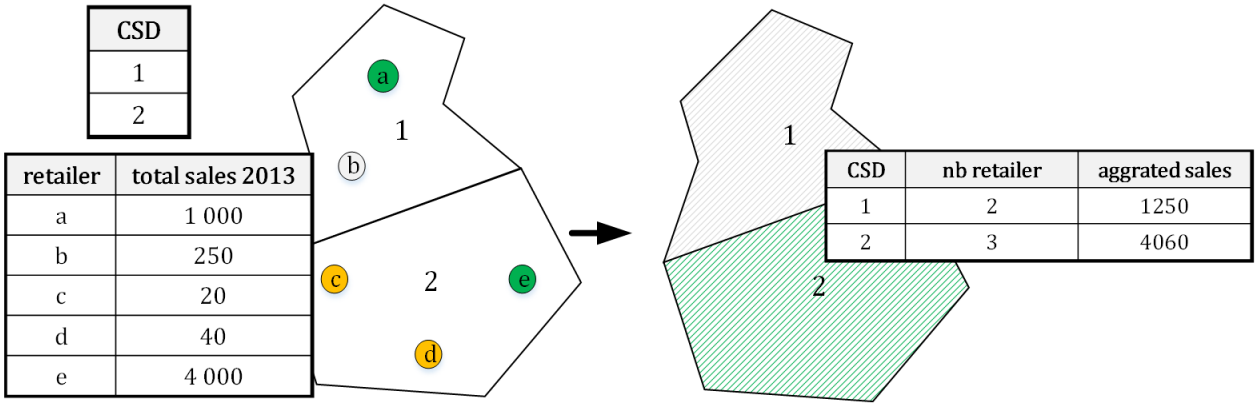


Figure 4.7 Aggregated sales per area

Results

A second interface was developed using the geoexplorer tool. The resulting interface is presented in Figure 4.8. Areas could be colored with a colorimetric scale (the same way as presented in *F1*) for any valuable information available from the retailers' dataset (monthly sales, yearly sales, etc.). The detailed information is accessible through basic interaction (Figure 4.8.B) for each area.

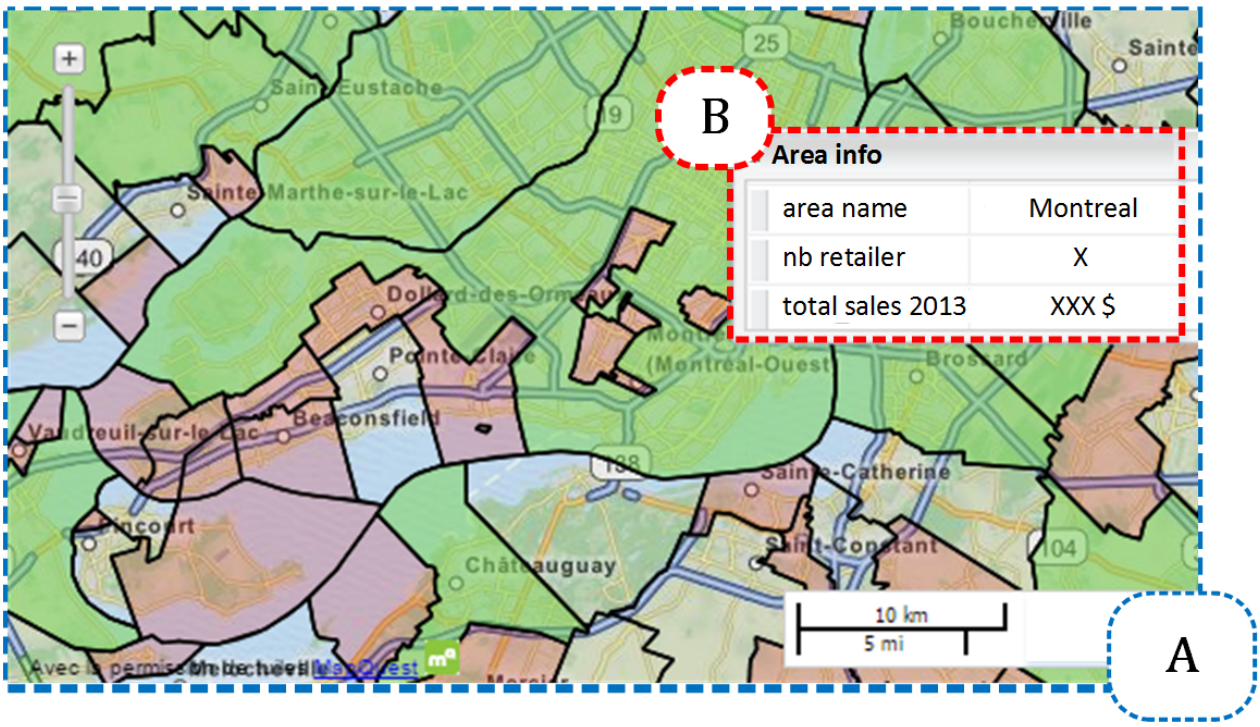


Figure 4.8 Aggregated sales visualization interface

This new interface allows for an aggregated evaluation to be obtained for any parameter from the retailers for each area. Then the colorimetric representation provides an easy visual representation of the results. Many area indicators could be represented (sum of the sales, average sales, etc.) that would allow different areas to be compared.

F3 - Filtering capabilities

Need

As our partner does not distribute its products itself, it makes partnerships with different retail companies. Thus, it may be pertinent to assess the retail network depending on each specified retail company. Furthermore, representatives are in charge of different retailers, and therefore it might also be of interest to get access to all retailers from a selected representative.

To meet these needs, a filtering tool was developed. It allows decision makers to focus on certain parts of the distribution network.

Methodology

Data related to retailers was provided by the partner company for each retailer, the representative in charge, and the retail network company. Step A consists of formatting and

importing that data.

In step F, a dedicated interface was developed. To facilitate web development, the Yeoman development stack is used [133]. Yeoman provides a local development environment that speeds up web interface development. Yeoman also allows for dependencies to be managed in the javascript libraries used in the project: Bootstrap [19] for web application design, Openlayers [91] for map visualization and interaction.

Partner networks and company representatives might change with time, thus the dedicated filters were developed generically to adapt to the retailer's current data. When the SDSS retrieves the retailer's data, it automatically extracts the existing representatives and retail networks to propose the different option in dedicated filters. This generic filtering is adaptable to any categorizing data (status, representative, purchasing group, etc). For numerical data, an algorithm that extracts bounding values has been developed, allowing, for example, for retailers to be filtered by turnover.

The filters on the interfaces are generated dynamically after the extraction of corresponding information.

Results

The resulting interface is shown in Figure 4.9. While initially the filters are not set up (Figure 4.9.A), all the retailers appear on the map (Figure 4.9.B). When some filters are selected (Figure 4.9.C), only the corresponding retailers appear (Figure 4.9.D).

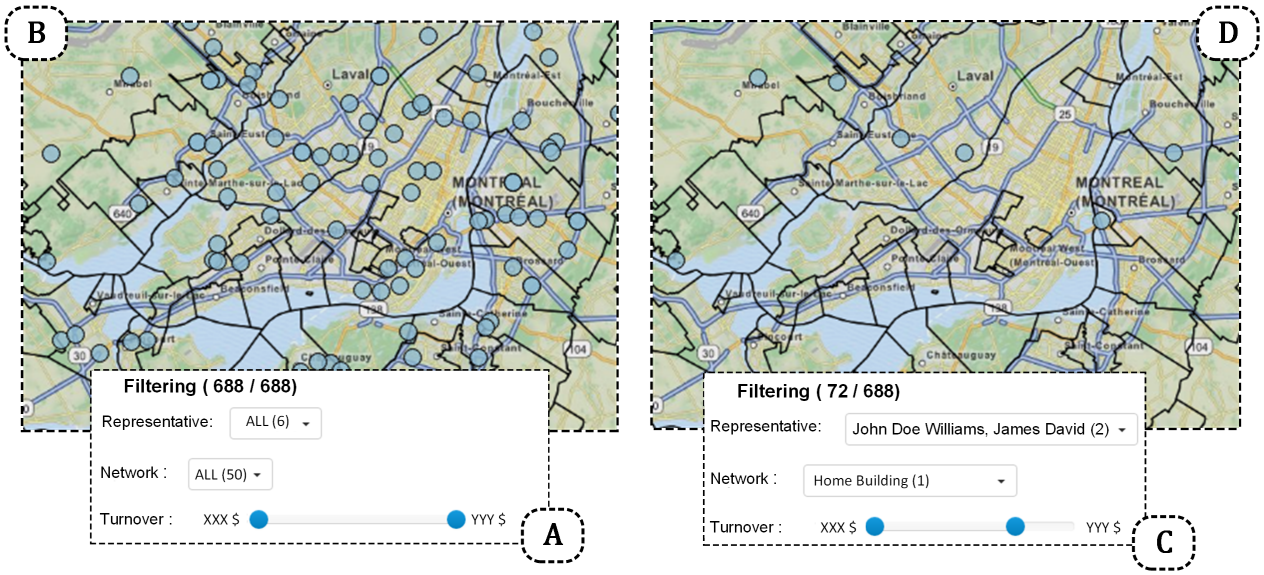


Figure 4.9 Filtering options

Filtering options allow a more precise analysis of the network. At any moment, it is possible to focus on a specific subpart of the network.

F4 - Dynamical area polygon styling

Need

Area limit files take a lot of memory and computing power for rendering on a web browser. In order to apply various styles dynamically and to get improved user experience, CSD limits file were simplified for the visualization interfaces while original files were kept in the database to perform analysis.

Methodology

The polygon simplification process is allowed by the Geographic Information System (GIS) QGIS. Figure 4.10, illustrate the simplification process of Montreal and Laval limits polygon, allowing for the corresponding file to go from 168kbytes to 6kbytes. Resulting interfaces will conserve the same information, but allow for much faster rendering and interactions. This simplification process was applied on the whole province CSD limits file.

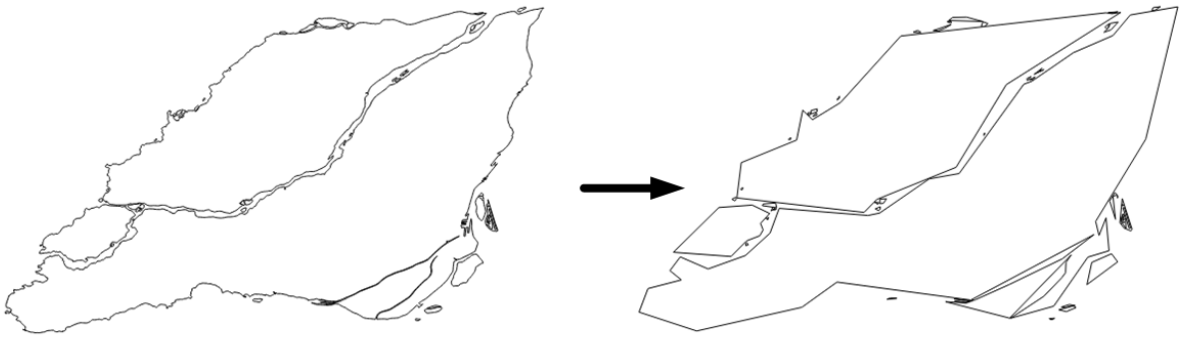


Figure 4.10 Polygon simplification

Results

Limiting files using less memory and computing power after simplifications, it is then possible to fluidly visualize the desired metrics through dynamic coloration of the areas.

F5 - Visualization of different sales performance metrics

Need

Sales assessments are not based on the only metric included at this time (yearly sales); other

data is required. Pairwise period sales comparisons or period sales to objective comparisons are metrics that are helpful to correctly assess sales performances.

The need to visualize those metrics through detailed or aggregate format (respectively retailers and areas) is held and the corresponding capabilities were developed.

Methodology

Sales metrics are based on many different periods and computing periodical sales requires a dedicated query. The data has been transformed to a format allowing for temporal queries. The sales data was transformed to have one record for each month per retailer. The data transformation is presented in Figure 4.11, allowing a configurable temporal query to be processed, such as in Figure 4.12.

name	address	sales jan-13	sales feb - 13	...	sales dec - 14
J.D Materials	...	150 $	220 $	...	330 $

name	address	sales month	sales amount
J.D Materials	...	janv-13	150 $
J.D Materials	...	feb-13	220 $
...	...	...	...
J.D Materials	...	dec-14	330 $

Figure 4.11 Table transformation to allow temporal queries

```
1  SELECT  RetailerID, SUM(
2              when Sales_date >= Period_Start and Sales_date < Period_End
3                  then Sales_Amount
4                  else 0 end)
5  FROM ReatailerSales
6  GROUP BY RetailerID
```

Figure 4.12 Configurable temporal query

For this, the development phase consisted essentially of developing the configuration interface and the association of each metric with a colorimetric scale. Indeed, as it is not practical to define a colorimetric scale for each metric, a dedicated tool that automatically computes color gradients from the data was developed and integrated in our web interfaces toolbox.

Results

Available sales metrics for areas and retailers are now as follows:

— Period sales: monthly, quarterly, yearly, or at year to date.

— Period sales comparison: monthly, quarterly, yearly, or at year to date.

— To objective comparison: monthly, quarterly, yearly, or at year to date.

The displayed metrics are directly configurable from the user interface.

F6 - Improved user experience by providing access to contextual information

Need

Still with the aim of improving spatial understanding of sales performances, the next application focuses on interaction and visualization capabilities.

The colorimetric scales provide a first assessment of the studied metrics. The focus now is to have quick access to the evaluated metrics. Moreover, charts have been integrated to provide enhanced visual evaluation.

Methodology

In addition to the OpenLayers [91] and Bootstrap [19] libraries, ChartsJS [24] is added for chart management. Interfaces were developed to conform to the input requirements of each library.

To allow for future adaptations, all new interfaces were developed taking into consideration flexibility. For example, retailers' styling is decomposed into multiple steps: retrieve coloring variable, get associated data, define corresponding color scale, and render the retailer dedicated point. If a change happens in the selected colored variable, the information is automatically propagated to the next steps.

Results

Selected metrics are now directly accessible when hovering over the desired retailer (Figure 4.13.D). A clicking interaction now allows for information that is relevant to the partner's decision makers.

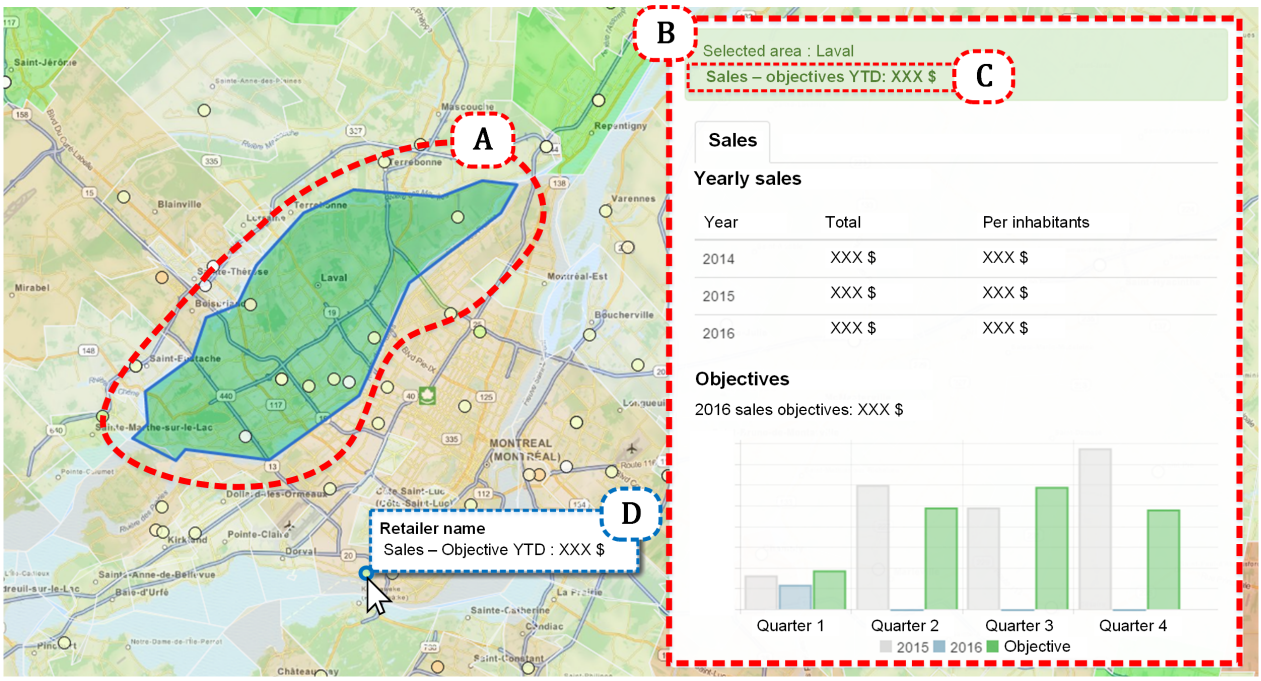


Figure 4.13 User interface capabilities

In Figure 4.13, selected areas are now highlighted Figure (4.13.A), and the current observed metric of the selected area is also presented at the top (Figure 4.13.C) of the information panel (Figure 4.13.B). As one important assessment metric is a comparison to the objectives, a bar chart shows sales of the past and current years in parallel with fixed objectives.

F7 - Investigation of environmental data

Need

Local environment is an important factor in the performance of the retail industry. Some areas are more favorable than others, depending on local socio-demographic factors or other parameters.

According to our partner, the three main fields where we must focus our attention were construction data, socio demographic data, and some points of interest, such as competitors, ski resorts and golf clubs.

While a future objective is to process KDD methods on those datasets, the partner was actually interested in having the abilities to access to that information through a dedicated application.

Furthermore, those datasets are huge and it is pertinent to determine for each case study

what information is relevant. Here, median income (Census Data) and construction building dates (National Household Survey (NHS) data) were some of the selected data collected in the application.

Methodology

The data acquisition on socio-demographic information was done from two main sources, the Census Data [118] and the NHS [118]. Those two datasets can be found in the adapted aggregation level (CSD for this application), so only a basic data treatment had to be processed before integration.

Construction data was acquired from the Census Database and the Canadian Mortgage and Household Corporation (CMHC)[31]. Those datasets had to be cleaned and transformed before being integrated in the database.

For the different points of interest, such as competitors, ski resorts and golf clubs, no available dataset was found. Nevertheless, that information is available through the web: competitors' websites often offer store location webpages, and ski resorts and golf clubs are listed along with their address on dedicated websites.

As the need to acquire data from websites appeared, tools to process data acquisition form websites were gathered and integrated in our toolkit. For example, http queries library (Java.net.URL) to automatize webpage access and source downloads (in html or other format). HTML parsing packages were also used (Jsoup [69]). Finally, when needed, addresses were converted to geographical coordinates using the tools developed previously (in *F1*). Thus, it was possible to gather the needed data and to group it into dedicated dataset layers. Here, the main development consists of the integration of the access to the data through adapted interfaces.

Results

The current interface allows for many sources of data to be investigated:

— number and amount of building permits (monthly per CSD),

— housing starts number by type (quarterly per CSD),

— Census Data (detailed per CSD),

— National Household Survey Data (detailed per CSD),

— ski resorts, golf club and competitor locations.

The research can be done easily by accessing the detailed information in each area (details are accessible by clicking on the desired area). It is also possible to add elements on the map by selecting the dedicated layer: competitors, ski resorts and golf clubs.

4.3.4 Extension to other datasets

This section shows how the developed SDSS is easily adaptable to other data sets. Here, a complete transfer to the Province of Ontario is exemplified. For this transfer, the partner provided sale and retailer data for the Province of Ontario. According to the proposed method (step A) the input data has been transformed, cleaned and processed with the tools already developed (address geo-coding (in *F1*), table transformations (in *F5*), etc.). Then, that data was integrated into the database. Step C: Ontario CSD limits file were downloaded, and a simplified version was computed with QGIS and integrated into the database (as in *F4*). Data queries for sales aggregation (in *F2*) or periodical queries (in *F2* and *F5*) were developed generically. Then, everything that was needed to provide data to the user interface was already available. The interface between the new data and the flexible interface from *F6* was configured.

All together, those steps took a day and a half for the transfer, allowing for our partner to visualize and investigate sales in Ontario through the same type of interfaces (Figure 4.14)

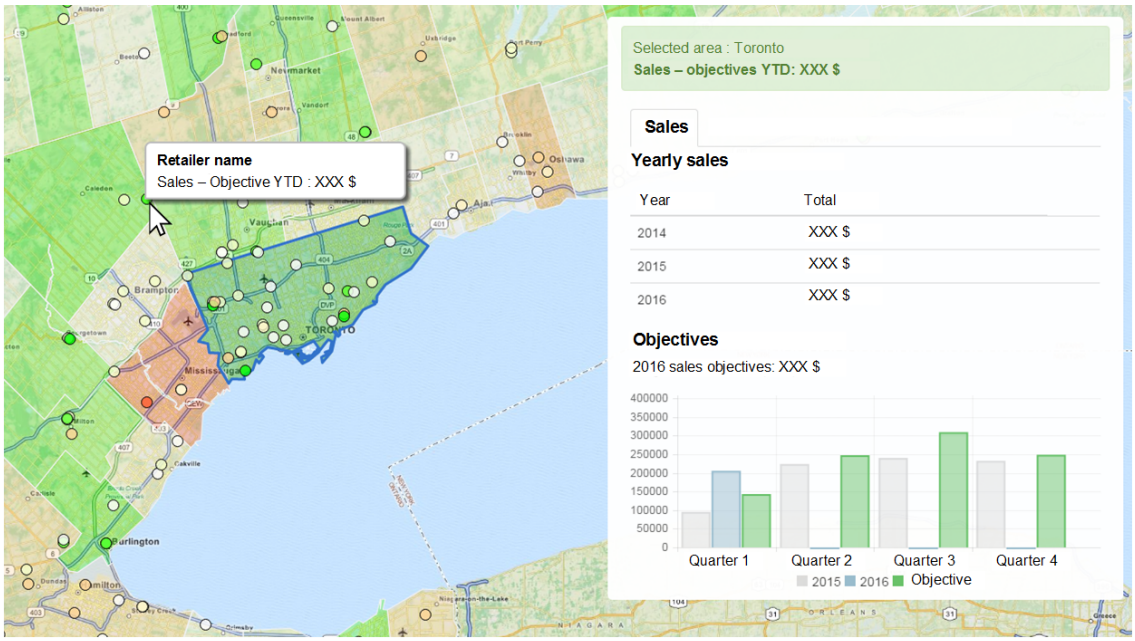


Figure 4.14 Transfer to the Province of Ontario

4.4 Conclusion

As it can be seen in the literature review, and particularly in Table 4.1, SDSS may be useful for many applications. Moreover, there is continuous growing volume of geospatial data

stored in various public or private domains that makes possible new areas of developments.

Besides, actual state of the art offers only partial solutions to these needs: on one side, generic methodologies are available and on the other side, case-base specific applications, but no reproducible solution.

The proposed approach allows the development of many SDSS in an integrated framework. This proposal relies on a conceptual framework composed of a SDSS architecture and a development approach. The architecture relies on two categories of clients (Rich clients and Light clients). Tools are proposed for the Rich clients so they may easily manipulate the data in the SDSS for any analysis and/or develop applications for the Light clients. Moreover, the proposed development approach makes it possible to factorize different tasks of the development in the case where several dedicated applications have to be developed for a same business

To summarize, the proposed framework offers:

— An integrated solution for the manipulation of geospatial data,

— A centralized system to integrate various sources of data,

— An open solution, independent of any proprietary system,

— A flexible solution that adapts to any case study, and

— A powerful solution for the case study included.

The framework has been validated on a real case study, and we presented a detailed explanation to show the development of a specific functional SDSS.

Future developments will focus mainly on the development of various tools that allow for the analysis of data in various case studies. Data mining and Operations Research techniques will be adapted and integrated within the SDSS. Evaluation methods to assess local and global performances adapted to the case study will be developed; for example, metrics to evaluate the sales potential among the territory. This will lead to maximizing demand coverage under constraints that are specific to this application.

The developed SDSS presented in the case study is currently in use by our industrial partner, and the results from our future analysis will be plugged into the current platform.

CHAPITRE 5 ARTICLE 2: A SPATIAL DATA PRE-PROCESSING TOOL TO IMPROVE ANALYSIS QUALITY AND TO REDUCE PREPARATION DURATION

Submitted to Computers & Industrial Engineering

Authors: G. DARAS, B. AGARD, B. PENZ.

Abstract. Spatial data analysis allows for a better understanding of environmental effects on the performance of an organization's activities. One of the first steps required to process such an analysis is to gather all of the spatialized data corresponding to the elements that might influence the activities. Then, a series of treatments must be processed on those datasets to make them ready to be used in classical data mining tools.

Those pre-processing steps are complex and time consuming tasks that may require advanced Geographic Information System (GIS) skills. Moreover, the choices involved in this process influence the quality of analysis results.

With the aim of addressing those issues, we developed a tool that automatizes several steps of spatial data pre-processing tasks. To allow for reproducibility, the specifications of our approach, tools, architectures and techniques required are presented in detail

To support the effectiveness of our approach, a case study is presented that focuses on an evaluation of the processing time that is saved and the improvement of the quality of analysis.

5.1 Introduction

Understanding the effects of the spatial environment on the performance of an activity is a real advantage for many organizations in the public and private sectors. With the changing capabilities and costs of technologies, more and more organizations are accumulating data on their activities (such as sales metrics), which include spatial characteristics such as addresses or geographical coordinates. At the same time, there is an increasing amount of data made available on elements that potentially influence the performance of these organizations' activities. For these reasons, significant research in various fields has aspired to extract relevant information to understand what actually influences an activity. Recently, Mennis and Guo [83] said that spatial data mining is a trending area. Spatial data mining, as defined by

Koperski and Han [75], consists of extracting implicit knowledge from spatial data. This research field is an extension of the Knowledge Discovery from Databases (KDD) introduced by Fayyad et al. [47]. However, many existing data mining algoritms are not able to take advantage of the spatial aspect of the data. Thus, spatial components have to be prepared to be taken into consideration, but this preparation of spatial data is a complex and tedious task.

The aim of this research is to present a tool that automates the pre-processing of the spatial data, removing the GIS skills requirement and allowing for improvement in the analysis quality and savings in processing time.

The next section (5.2) presents the elements of the literature related to spatial data analysis and pre-processing to allow a better understanding of the problems that arise from the consideration of spatial data. Section 5.3 first presents the specificities related to the preparation of spatial data, then it focuses on how the choices made in this pre-processing may influence additional analysis quality. Section 5.4 presents the specifications of our approach. Technical aspects related to the implementation of our solution are also presented. Section 5.5 permits an evaluation of the improvements provided by our tool. For this, a case study with real data shows the pre-processing tasks with and without our tool and how it performs. Finally, the limitations and perspectives of our research are discussed.

5.2 Literature Review

5.2.1 Spatial decision-making

Thirty years ago, Schmidt [112] revealed that localization decisions were made quickly by people without experience or knowledge of the issues involved. Decisions were made subjectively with few requirements and considering only a small portion of existing options. At the same time, Herring and DeBinder [63] argued that the use of computer tools could greatly improve the localization decision-making process. A few years ago MacEachren and Kraak [80] noted that many problems in the scientific and social fields had a spatial aspect. They add that at that time, the amount of data available with a spatial component was steadily increasing. However, due to the lack of appropriate methodologies to analyze them, these data were rarely used to construct valuable knowledge. More recently, according to Thompson and Walker [125], there has been a growing interest in spatial data analysis that is associated with the need for competitiveness in enterprises. In the same direction, Keenan [70] identifies a need for spatial information among high-level managers. As mentioned by Fang et al. [45] a customer's location matters for profitability, which leads to extensive research on

neighboring topics such as covering problems, which sometimes take spatial data into account ([46]). Many researchers are focused on developing tools to take advantage of spatial data. The tools developed for this purpose are called Spatial Decision Support Systems (SDSS), introduced by Armstrong et al. [9] and Densham [39] in the early nineties.

5.2.2 Spatial Data mining: specificities and challenges

Many SDSSs are based on spatial data mining which, according to Miller [84], makes it possible to get interesting models on elements or events that are distributed spatially. These models can be based on the spatial properties of these elements, or on the spatial relations that exist between the elements, in addition to the non-spatial attributes in traditional data mining. While many SDSS projects have been carried out, there are still several challenges mentioned in the literature. Erskine et al. [42] and MacEachren and Kraak [80] indicate that the storage, manipulation and use of spatial data needs to be studied to simplify strategic or organizational decision-making. Similarly, Sugumaran [121] points out the need to build SDSS capable of using technology to facilitate interoperability between spatial data and SDSS. According to Keenan [71], an approach that brings together interfaces, modelling techniques and appropriate databases needs to be developed. Pretorius and Matthee [100] indicate that there was no uniform and generic methodology for spatial data mining. Alatrista Salas et al. [3] state that classical data mining algorithms take in input data stored in tables, and that they do not directly take into account spatial information stored in other formats, such as geometries. For Bogorny et al. [18], there are few tools that directly manage spatial data. In most cases, as mentioned by Pretorius and Matthee [100], spatial data must be prepared by hand before being imported in order to take advantage of it.

5.2.3 Spatial data pre-processing step

Spatial data has several particularities (Section 5.2.3) that induce complexity in pre-processing (Section 5.2.3). Few works of research mention their method of pre-processing spatial data (Section 5.2.3) and there is currently no solution to automate this pre-processing (Section 5.2.3).

Spatial data particularities

One of the first aspects to be taken into account when preparing spatial data is that the different datasets must be compatible with each other. This compatibility is necessary in order to be able to visualize the data or to study the spatial relations between the datasets

[49]. As Mennis and Guo [83] state, spatial data often comes from different sources. Bogorny et al. [18] define spatial predicates as the materialization of spatial relations, which are not explicitly stored in databases. Shekhar et al. [113] say that spatial data must be transformed into spatial predicates in order to be processed by conventional data analysis tools. Spatial predicates can be in several formats; for example, they can be in boolean values for intersection relations (Does element A intersect element B?) and they can be numerical in the case of Euclidean distance relation. These relations between elements positioned in space can be computed by spatial operations. Ester et al. [43] and Bogorny et al. [18] denote three types of spatial relations that exist between two geospatial entities: topological relations, distance relations, and direction / order relations (see Figure 5.1, [18]). In (Bogorny et al., 2005) [18], topological relations characterize the types of intersections that exist between the elements; for example, touches, contains or overlaps (See Figure 5.1.A). Distance relations, shown in Figure 5.1.B, can be based on different metrics, such as Euclidean distance. The direction / order relations take into account the position of the elements in relation to others (for example, as illustrated in Figure 5.1.C).

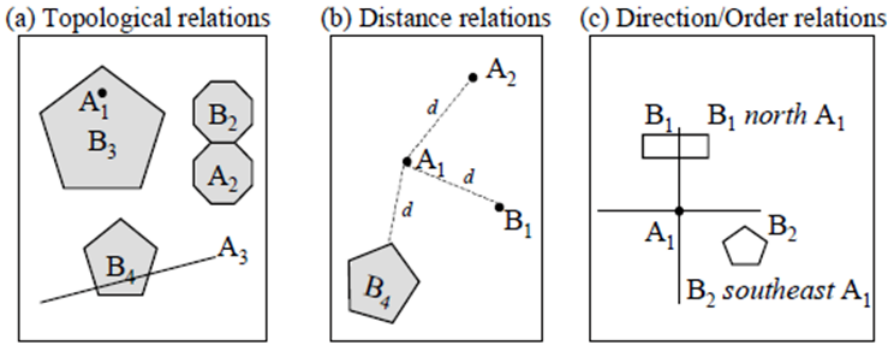


Figure 5.1 Spatial Relations ([18])

Complexity and time cost of spatial data pre-processing

Preparation of spatial data must be carried out with caution, and this can be time-consuming, particularly in the calculation of spatial relations. Indeed, Flowerdrew [49] argues that the integration of spatial data is not a clear process and that, although Geographic Information Systems (GIS) may seem easy to use, it remains careful work. In many studies on KDD Process (KDP) or on Knowledge Discovery and Data Mining (KDDM) , the data preparation phase is considered one of the most time consuming (see Figure 5.2 from [27]).

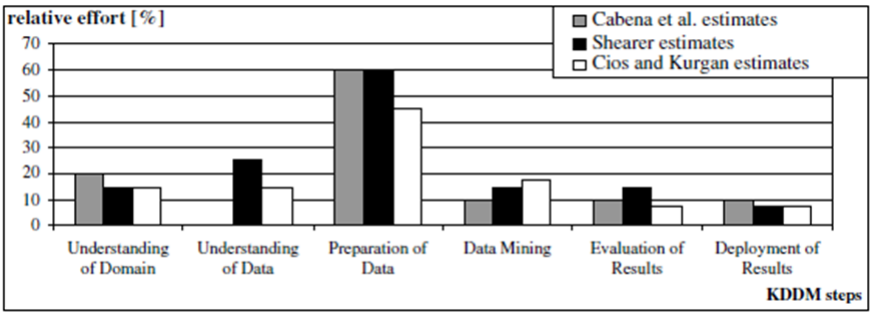


Figure 5.2 Time spent in KDP phases ([27])

In classical KDD projects, the preparation of the data is considered the most time consuming step (taking between 45 to 60 percent of the project duration). In addition, Pretorius and Matthee [100] argue that the phase most impacted by the spatial component in KDP is the data preparation process (see Figure 5.3 from [100]).

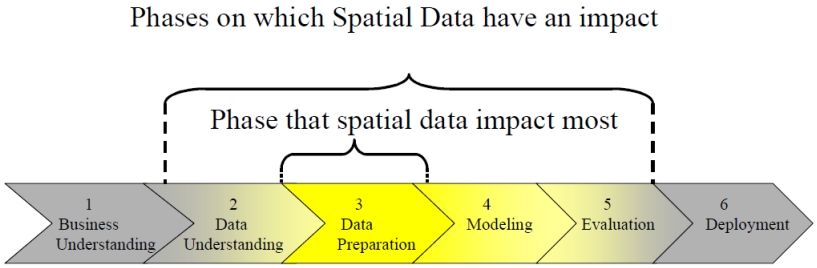


Figure 5.3 Impact of Spatial Data on the KDD Process ([100])

Clementini et al. [28] claim that storing the results of all spatial relations is very expensive in memory space. They deduce that instead of storing all the spatial relations between elements, it is more practical to compute them when they are needed. However, it must be known that to process the spatial relation computations, one must have a complete understanding of the techniques involved. Indeed, as Egenhofer and Herring [41] mention, fundamental concepts for the realization of spatial analysis have to be mastered, such as those relating to the geometrical relations between spatial elements. In the same direction, West [132] says that GIS are complex to use, and that their use requires the mastery of complex cartographic notions that may seem inaccessible to the uninitiated. Vahedi et al. [127] add that spatial functions and their parameters are often difficult to understand and use in GIS. Among other things, analysts are often unfamiliar with GIS and have no specific training in this

area. They add that the lack of an alternative makes the use of GIS to carry out spatial relations computations mandatory. Bogorny et al. [18] also mention that frequently, those who perform the data analysis are not necessarily experts in spatial databases. Appice et al. [7] note that the expertise required to pre-process spatial data is often an obstacle to perform spatial data analysis. Bogorny et al. [18] add that the preparation to make the spatial data ready for data mining algorithms is long and must be done by hand. For Gibert et al. [53], this preparation of spatial data, in addition to being time-consuming, must be repeated for each application. Another aspect to consider, as mentioned by Mennis and Guo [83], is that several choices must be made during this preparation; for example, on the chosen metrics or on the type of relations to be considered. Alatrista Salas et al. [3] underline the difficulties involved in pre-processing when there is a lack of knowledge in the area of study.

Pre-processing in spatial data analysis case studies

Many research present case studies containing spatial data analysis. In many of these, there are no explanations or details about the acquisition, the pre-processing and the types of spatial relations that are used (Knezic and Mladineo (2006) [73], Previl et al. (2003) [101], Ghaemi et al. (2009) [52], Vlachopoulou et al. (2001) [129]). Other research gives some detail, such as Evans and Sabel (2012) [44] or Grabis et al. (2012) [58] who take spatial data into account, but does not explicitly present those datasets and how they were gathered and transformed. In Wanderer and Herle (2015) [130] and Andrienko et al. (2001) [5] some computed spatial relations are mentioned such as distance, position, density, or coverage, but there is no explanation about why those relations where chosen. Less frequently, there is also research like Roig-Tierno et al. (2013) [110] which explains which relation they account for and why it was chosen.

Solutions to improve spatial data pre-processing

Bogorny et al. [18] already identified some of the problems induced by the spatial specificity of the data. They proposed an environment with tools that enable the preparation of spatial data. The use of their tools, however, requires knowledge of GIS, and the computation of spatial relations is not automated. Moreover, their approach does not offer assistance to determine the spatial relation to consider.

From another perspective, Anselin et al. [6] propose a tool that enables one to extract from data preparation, but this is only an introduction to the analysis of spatial data and it is neither extensible nor customizable.

To our knowledge, there is no solution that allows spatial data pre-processing without specific knowledge of GIS. Moreover, there is also no solution that will guide the choice of the spatial relation that has to be accounted for. To overcome this lack of an efficient solution, our research focuses on the automatization of the spatial data pre-processing. Since spatial data preparation relies on complicated concepts, the next section focuses on the explanation of those. The effects that the choices involved in this preparation can have on the results of the analysis are also presented.

5.3 Necessity and complexity of pre-processing

To illustrate the problems associated with spatial data pre-processing, the following section focuses on the real case of a partner company for which we develop a SDSS for a retail perspective. The company works in construction materials and distributes its products through third-party retailers. As mentioned by Cliquet et al. [29] and Dubelaar et al. [40] in the particular case of the retail sector, knowledge of the environment can be a major competitive advantage for improving performance. In this case, this company wants to better understand the effects of the environment on its sales performance in each area where it has partnerships with retailers.

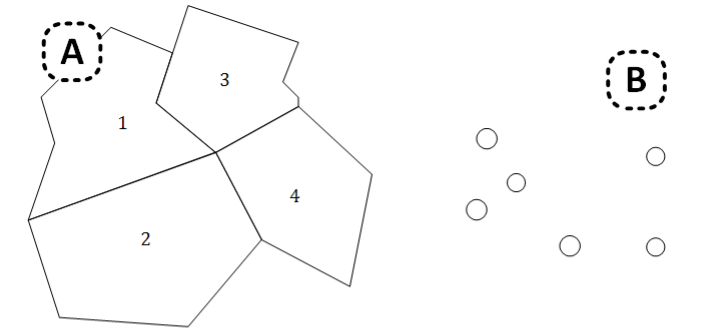
During the "Understanding of the Domain" step (presented in many KDP research, such as in Cios et al. (2007) [27]), managers of the company indicated what kind of information they want to access, and what elements might influence their sales performance. Following that information, we were able to gather several datasets. Managers wanted to be able to have different geographic abstraction levels to analyse their sales in a map representation. Thus, the boundary files of two division granularity levels of their sales territory were collected: Census Division (CD) and Census Subdivision (CSD). Socio-demographic datasets from census data associated with those areas were also gathered (From Census of Population and National Household Survey). Then, the available datasets on potentially influencing elements were acquired: golf clubs; ski resorts; architects (company's partners); lakes; competitors (the location of seven companies' retail stores); building permits; housing starts.

As mentioned before, in most cases, spatial data is inadequate for the application of conventional data mining algorithms. Since this notion is sometimes difficult to understand, the next section attempts to explain it on a dataset inspired by our case study. In Subsection 5.3.1, the need for pre-processing spatial data is justified. The second subsection (5.3.2) shows the complexity of the choices that are involved during this pre-processing and why these choices have consequences on further analysis.

5.3.1 Necessity of pre-processing through an example

In GIS, a set of data related to a type of element is called a layer. A layer is composed of geometrically similar elements: a set of points (for example, to define precise locations), a set of polygons (for example, to define areas), or a set of lines (for example, roads). Furthermore, a data table is associated with each layer, which contains information associated with each element. It has to be noted that depending on their characteristics, different elements of the same layer may not have the same influence. For example, in the retail industry, as mentioned by Ismail El-Adly [65], Teller and Reutterer [124] and Clulow and Reimers [30], the attractiveness and convenience of a retail store depends on its characteristics. Figure 5.4 provides a representation of several spatial data layers. The geographical areas (shown in Figure 5.4.A), as well as the lakes (Figure 5.4.E) are stored in the data as polygons. Other data can be associated with these polygons, such as the number of inhabitants per area, for example. The data associated with locations, such as retailers (Figure 5.4.B) or ski resorts (Figure 5.4.D) are stored as points, which can be associated with other data in the corresponding tables (such as turnover for each retailer).

The first step consists of collecting the available data. An area dataset corresponds to the geographical boundaries associated with socio-demographic data (number of inhabitants per area, for example) as shown in Figure 5.4.A. The company has sales data (turnover among others) of each of its partner's stores, as well as their locations (Figure 5.4.B). Competitor locations must be taken into account and as their influence depends on characteristics such as their company, the corresponding data is stored here (Figure 5.4.C). The salesmen of the company suspect that sales are influenced by the store's proximity to ski resorts, as well as to lakes. Data corresponding to these elements is then collected and stored in the corresponding layers (Figure 5.4.D and 5.4.E).



Area	Surface	Population
1	45	1 200
2	18	4 500
...	...	...

Retailer	Turnover
1	3 000
2	2 400
...	

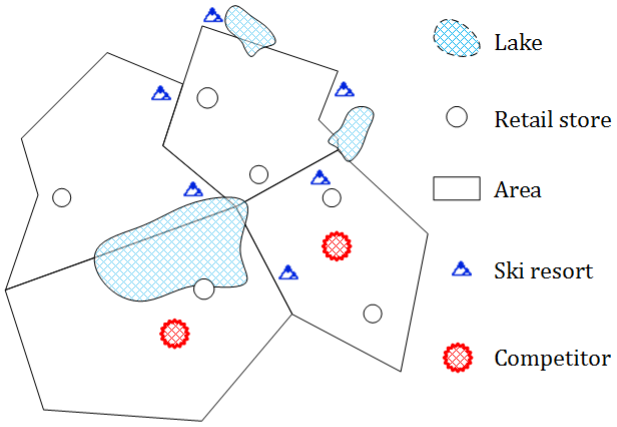
Competitor	Company
1	A
2	B
...	

Ski resort
1
2
...

Lake
1
2
...

Figure 5.4 Gathered Datasets

These different datasets, which are still dissociated, do not allow the application of most conventional data mining algorithms. It is necessary to "merge" them within a table in which the spatial relations between the different datasets will be computed, as in Figure 5.5.



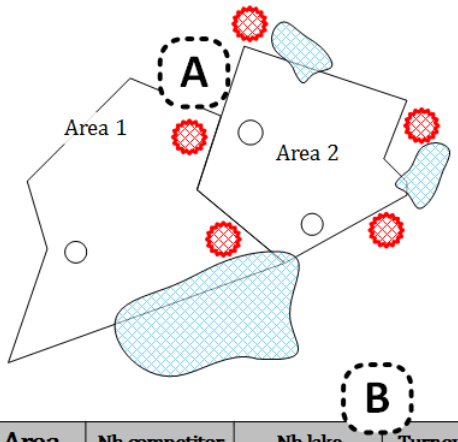
Area	Surface	Population	Spatial predicates					Total turnover
1	45	1 200	...	...	...	...	...	12 000
2	18	4 500	...	...	...	...	...	8 000
...	...	...	...	...	...	...	...	...

Figure 5.5 Data Integration

With the combinatory between the number of spatial relations and the amount of data, it is not possible to compute all existing spatial relations. It is therefore necessary to select the relevant spatial relations before importing them into a table on which data mining algorithms can be applied. The next section explains why the choices involved in pre-processing have consequences on further analysis results.

5.3.2 The consequences of pre-processing choice on analysis

Depending on the spatial relations selected for the environmental analysis, the results can vary. To illustrate the influence of the choice of the relation on the result of the analysis, we concentrate only on two areas, with the situation represented in Figure 5.6.A. If the study focuses on the effect of a competitor and lakes on sales, and the chosen relation is the number of elements per area, the resulting dataset will correspond to the Table in Figure 5.6.B.



Area	Nb competitor	Nb lake	Turnover
1	2	1	12 000
2	0	3	8 000

Figure 5.6 Cardinality Computation

However, it is also possible to choose other spatial relations such as the coverage of areas by a competitor's catchment areas (spatial relation illustrated in Figure 5.7.A). For lakes, one can imagine taking into account the amount of coastal kilometers in each area (spatial relation illustrated in Figure 5.7.B). In a case where those latter relations are selected, the resulting data will look like the Table in Figure 5.7.C. The differences between the Tables in Figure 5.6.B and 5.7.C suggest that further analysis carried out on them will not provide the same interpretations of the effects of lakes and ski resorts on performance. This simple example shows that during spatial data preparation:

— Choices among available data relations have to be made

— Those choices have consequences on further analysis results

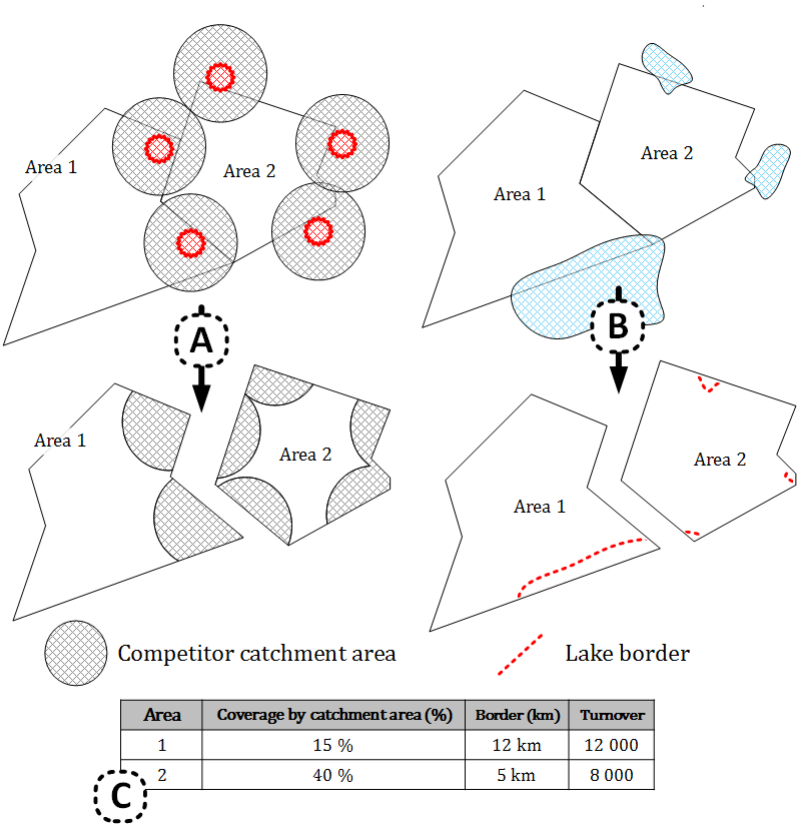
Moreover, as mentioned in the literature review, the computation of these relations is a time-consuming task, requiring advanced GIS skills.

Our tool, presented in the next section, aims to solve the problems linked to the spatial data preparation:

— Complexity of choice

— Time consuming characteristics

— GIS skills requirements



Area	Coverage by catchment area (%)	Border (km)	Turnover
1	15 %	12 km	12 000
2	40 %	5 km	8 000

Figure 5.7 Alternative Spatial Relations

5.4 Specifications and technical aspects

As the tool developed guides users through several steps, a first scheme that shows that sequence of steps is proposed and each step is then detailed. Next, the prerequisites necessary for the implementation of this solution are mentioned, a possible architecture is proposed and optimizations to improve usability are presented.

5.4.1 Pre-processing steps

Global scheme

Our approach, which aspires to automate the pre-processing of spatial data as much as possible, assumes that the data has been retrieved and stored in a database. As a result of the process presented in Figure 5.8, a dataset will be formatted and made available for the application of data mining algorithms. In Figure 5.8, the tasks in full lines are those that are fully automated; those in dotted lines have to be done by the user.

Descriptions of the steps presented in Figure 5.8

Step 1 corresponds to the selection of a Target Layer. The Target Layer is the one that contains a variable that the user wants to study (for example, the areas).

Step 2 is for selecting the Target Variable of the Target Layer; the choice of the Target Variable depends on the available data and the type of information sought. In our case study, the Target Variable could be related to sales, for example: the sum of a retailer's turnover per area.

Step 3 consists of selecting a Source Layer to study its influence on the Target Variable. A Source Layer is a layer of data that is suspected to have an influence on the Target Variable. In our case study, there are several potential Source Layers to consider, such as the competitors' location or the ski resorts.

Step 4 makes it possible, if necessary, to select a characteristic of the Source Layer by which it has to be categorized; this characteristic is called the Categorization Variable. For example, the user could categorize the competitor layer by the company variable (data presented in Figure 5.4.C).

Step 5 consists of a computation of spatial relations. Several relations are computed: for example, the competitor cardinality by area or the distance to the nearest competitor. Depending on whether or not the user has decided to select a Categorization Variable in step 4, two situations may arise:

(1) If the user did not wish to categorize the Source Layer, each spatial relation is computed with all elements of the Source Layer taken into account.

(2) In the case where the user chose a Categorization Variable for the Source Layer, each spatial relation is calculated with each category of the Source Layer. Concretely, in our example in Figure 5.4, the area is the Target Layer, the competitor is the Source Layer and the company is the Categorization Variable. Since the company has two possible values (A or B), the spatial relations will be calculated twice. They will be calculated once with only the competitors from company A, and once with only the competitors from B.

Step 6 consists of evaluating the correlation between the computed relations and the Target Variables. For each of the spatial relations computed in the tool, a correlation score with the Target Variable is computed and indicated to the user. This correlation score allows the

user to have indications on which spatial relations potentially influence the Target Variable the most.

Step 7 allows the user to select the spatial relations he wants to save. For example, if the user chooses to keep the number of competitors per area, the associated dataset will be stored in memory. These steps for computing relations and evaluating the correlations can be repeated for each Source Layer that the user wishes to study.

Step 8 allows for the dataset export, corresponding to the spatial relations of interest when the user has gone through the Source Layers he selected.

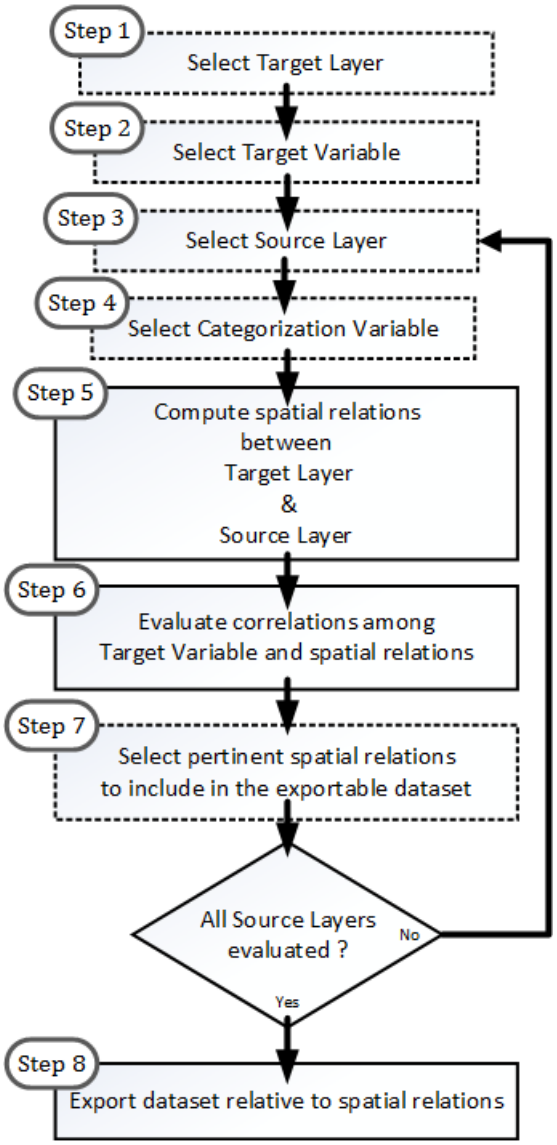


Figure 5.8 Pre-Processing Steps

An implementation of a tool that allows this process to be carried out is presented in the next section.

5.4.2 Implementation

The implementation of this approach is only feasible under certain conditions. There are prerequisites in the database that are presented in Section 5.4.2. In addition, to facilitate reproduction, the tools that we used and the architecture of our approach are mentioned in Section 5.4.2. Our implementation process is then presented in Section 5.4.2 and techniques to improve computation times are explained in Section 5.4.2. It must be noted that to improve clarity, queries on the data presented in this section have been simplified to make them easily understandable.

Database prerequisites

To allow the tool to retrieve the Target and Source potential Layers automatically, they must be stored in designated places, as in Figure 5.9, which presents a basic structure of the database. In Figure 5.9 the datasets are separated into two categories that are stored in two Schemas (which are PostgreSQL partition of the database). The Target Layers Schema contains potential Target Layers, such as different granularities of a territory division (like Census Division and Subdivision for example). The Source Layers Schema contains the potential Source Layers, such as Lakes, Retailers, or all other gathered datasets of interest. Furthermore, it is necessary to have a database that offers the possibility to make spatial relation computations and also offers interfacing possibilities with the other tools that are used in the implementation.

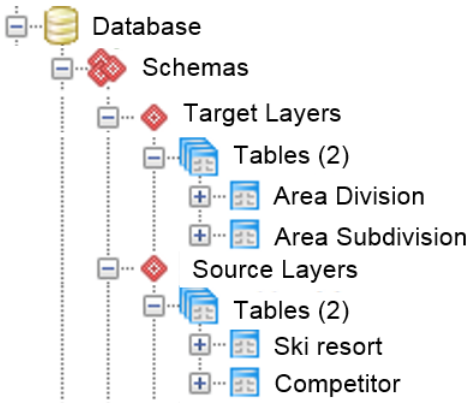


Figure 5.9 Basic Database Structure

Tools used in a working implementation

In a previous study, Daras et al. [35] proposed an architecture composed of open source tools to conduct spatial data analysis and visualization. A representation of the part of the architecture relating to the preparation of the data is presented in Figure 5.10.

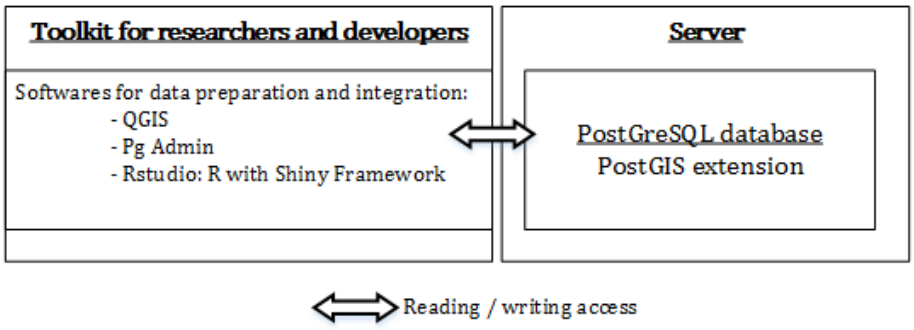


Figure 5.10 Possible Architecture for Implementation

The database used is PostgreSQL with the PostGIS extension allowing spatial data to be processed. The analysis language used in our approach is R, through the RStudio development environment. The spatial data pre-processing application was developed with the Shiny Framework. The application connects to the database and retrieves the list of potential Target and Source layers. In our implementation, the interfacing tools between the R language and the PostgreSQL databases allow this to be done easily.

Working implementation process

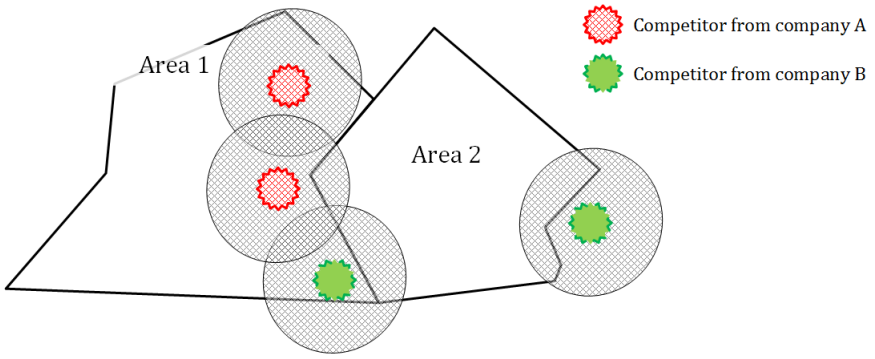
The solution automatically extracts available Target and Source Layers, and the associated Target Variables and Categorization Variables. Then, the user has to select the Target Layer, the Target Variable, the Source Layer (and maybe a Categorization Variable) from a dedicated interface. Next, the spatial relations computations are performed, Figure 5.11 illustrated the three computed spatial relations with the Areas set as Target Layer, Competitors set as Source Layer, and the Category set as Categorization Variable:.

— Number of elements of the Source Layer inside (relation called `Card_A` and `Card_B` in Figure 5.11, for cardinality of competitors of Category A and B)

— Distance to the closest element of the Source Layer (relation called `Dist_A` and `Dist_B` in Figure 5.11, for distance to the closest competitors of Category A and B)

— Surface covered by the catchment areas of the Source Layer elements (relation called `Cover_A` and `Cover_B` in Figure 5.11, for covering relation (illustrated in Figure 5.7.A)

with competitors of Category A and B)



Area	Card_A	Card_B	Dist_A	Dist_B	Cover_A	Cover_B
1	2	1	0	0	30%	10%
2	0	0	3 km	2 km	3%	15%

Figure 5.11 Spatial Relation Computation Illustration

These computations are carried out by means of requests transmitted from our tool to the database. Then, when the results are received from the database, correlation scores between the Target Variable and the relations are computed with the cor function natively present in the R-language. The choice of relations to export is then left to the user.

To improve the user experience when using the tool, it is important to maintain reasonable computation times. Choices made to keep those processing times acceptable are presented in the next section.

Optimization of spatial request and data structure

To maintain acceptable processing times, arrangements in the data structures and also optimization of the spatial relations requests have been selected. While the computation of the cardinality is not time-expensive, the distance to the nearest relation (request in Figure 5.12) and the coverage relation can take a lot of time to be computed (see Table 5.1 which presents computation time to process the requests on different datasets from our case study). For the distance to the closest, the classical query is shown in Figure 5.12. In this request, for each element of the Target Layer, the distance to all of the elements of the Source Layer is computed and then the element that has the shortest distance is conserved.

```
SELECT DISTINCT ON (TARGET_AREA.ID)
                TARGET_AREA.ID
                ST_DISTANCE(
                        ST_CENTROID(ST_ENVELOPE(TARGET_AREA.GEOM)),
                        SOURCE_POINT.GEOM)
FROM TARGET_AREA, SOURCE_POINT
ORDER BY TARGET_AREA.ID,ST_DISTANCE(
                                ST_CENTROID(ST_ENVELOPE(TARGET_AREA.GEOM)),
                                SOURCE_POINT.GEOM);
```

Figure 5.12 Distance to Nearest Query

Table 5.1 Spatial Request Processing Time

Computation time in seconds	98 areas 225 points	98 areas 726 points	1285 areas 225 points	1285 areas 726 points
Cardinality	0.6	1.5	2.2	4.5
Distance to first: basic	4	12	14	46
Distance to first: optimized	0.3	1	1.5	4
Area coverage: basic	31	433	494	193
Area coverage: optimized	0.2	21	4	21

It must be noted that PostGIS automatically stores spatial relations that allow faster access to some interesting data. Taking advantage of that, compared to the query in Figure 5.12, the query in Figure 5.13 divides the computation time by almost ten on the different datasets of our case study, as is presented in Table 5.1.

```
SELECT TARGET_AREA.ID,
       (SELECT TARGET_AREA.GEOM <-> SOURCE_POINT.GEOM
        FROM SOURCE_POINT
        ORDER BY TARGET_AREA.GEOM <-> SOURCE_POINT.GEOM LIMIT 1)
FROM TARGET_AREA ORDER BY TARGET_AREA.ID;
```

Figure 5.13 Faster Distance to Nearest Query

The coverage area computation is an even more time-consuming computation. In fact, this computation is divided into several time-consuming parts (illustrated in Figure 5.14).

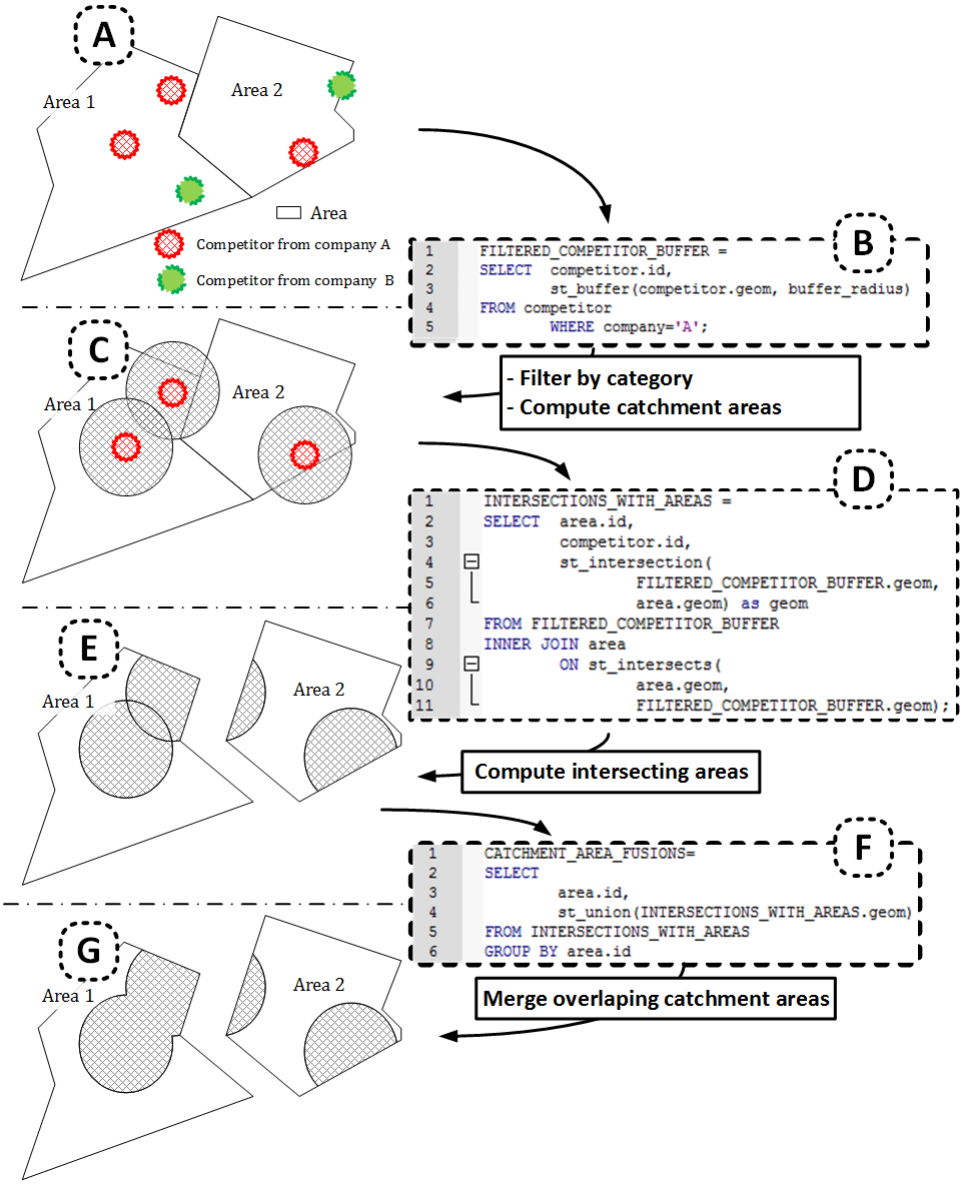


Figure 5.14 Coverage computation requests

First, if necessary, filter the elements of the Source Layer for each category. Then for each resulting layer the associated catchment areas have to be computed (request in Figure 5.14.B, resulting in data like Figure 5.14.C). Next, the intersection between the catchment areas and the Target Layer has to be computed (request in Figure 5.14.D resulting in data like in Figure 5.14.E). Finally, merge intersecting areas to avoid counting those areas twice (request in Figure 5.14.F, resulting in data like in Figure 5.14.G).

It should be noted that in the case of the selection of a Categorization Variable, these steps must be executed as many times as there are categories. Processing those different requests

might take up to more than 400 seconds on our datasets (494 seconds with 1285 areas and 225 points (Table 5.1)). One way to improve this computation time is to pre-process some computations when the data is integrated in the database. Those new datasets are integrated into the database, which must be now be structured as in Figure 5.15.

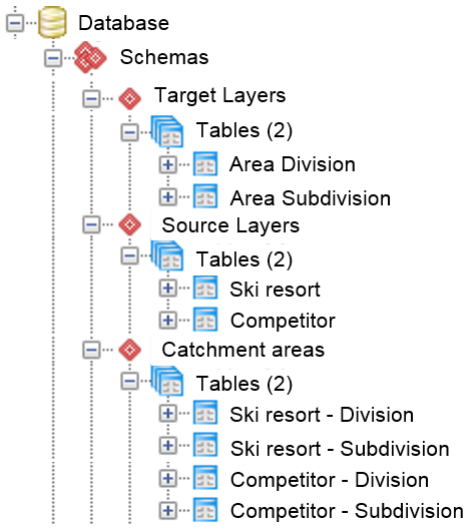


Figure 5.15 Database Modified Structure

For each possible pair (Target Layer, Source Layer), a layer of intersections of the Target Layer with the catchment areas of the Source Layers are computed and stored (transformation represented in Figure 5.16.A). These new layers consist of all intersection areas associated with the identifier of the corresponding Target and Source (resulting in Table INTERSECTIONS in Figure 5.16).

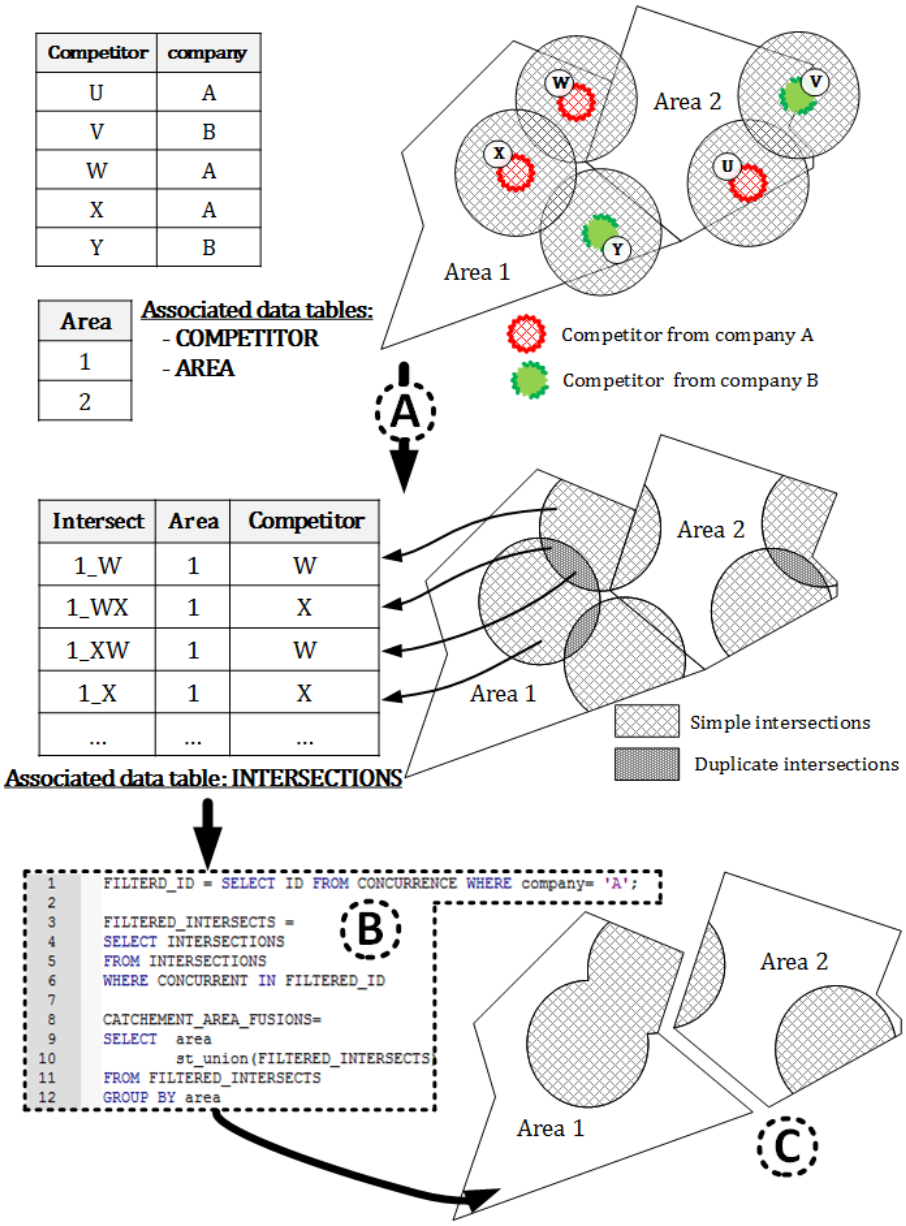


Figure 5.16 Pre-Processing Intersection Areas

Computations of intersections in those layers are only processed when a new element is added to a a Source or a Target Layer or when a new Source Layer or a new Target Layer is imported into the database.

Thereafter, when the coverage for a given category is desired, it is sufficient to make the request in Figure 5.16.B, producing the results (in Figure 5.16.C) identical to that in Figure 5.14.G, but with a shorter computation time (detailed processing times are in Table 5.1, row Area coverage: Basic for the request in Figure 5.14, and row Area coverage: Optimized for

the request in Figure 5.16.B).

The processing times of the different requests, processed on different layers of the case study dataset, are mentioned in Table 5.1. It is noteworthy that the optimization of the queries makes it possible to divide the processing time on our different datasets by nine or more. For reference, computations were performed on a Windows server equipped with an Intel (R) Xeon (TM) processor 3.73GHz (2 Processors) with 8.00GB RAM.

In the next section, a case study focuses on the improvements made by the proposed tool.

5.5 Case study to evaluate improvements

First, to support the efficiency of the proposed pre-processing tool, the next subsections describe the steps to be taken with (5.5.1) and without it (5.5.2). Second, the quality improvement in further analysis is presented (Section 5.5.3).

5.5.1 Steps without the proposed approach

With or without our tool, the first step remains the same and consists of data gathering and the homogenization of the spatial projections (to avoid errors in spatial relation computations).

In order to avoid repetition, only the steps that are required to study a single spatial relation for a single Source Layer are shown below (coverage of the catchment areas of a competitor). Also, to limit the redundancy of our remarks, a case without categorization of the Source Layer is presented. In the case with categorization, it would have been necessary to repeat the different steps as long as there are categories.

First, the different layers have to be loaded into a GIS. Then, various successive operations have to be performed (represented in Figure 5.17, composed of screenshots realized in QGIS, an Open Source GIS). After importing the data, a new layer composed of the buffers representing the catchment areas of the competitor is required (see Figure 5.17.B). This layer must then be "cut" according to the boundaries of the geographical areas, resulting in another layer (Figure 5.17.C). Then, to avoid double counting intersections between catchment areas, these must be merged into a new layer (see Figure 5.17.D).

It must be noted that each sub-step of Figure 5.17 has to be initiated through human interactions and requires each intermediate layers to be saved on the computers (resulting in additional time consuming manipulations that are not presented here, to conserve readability).

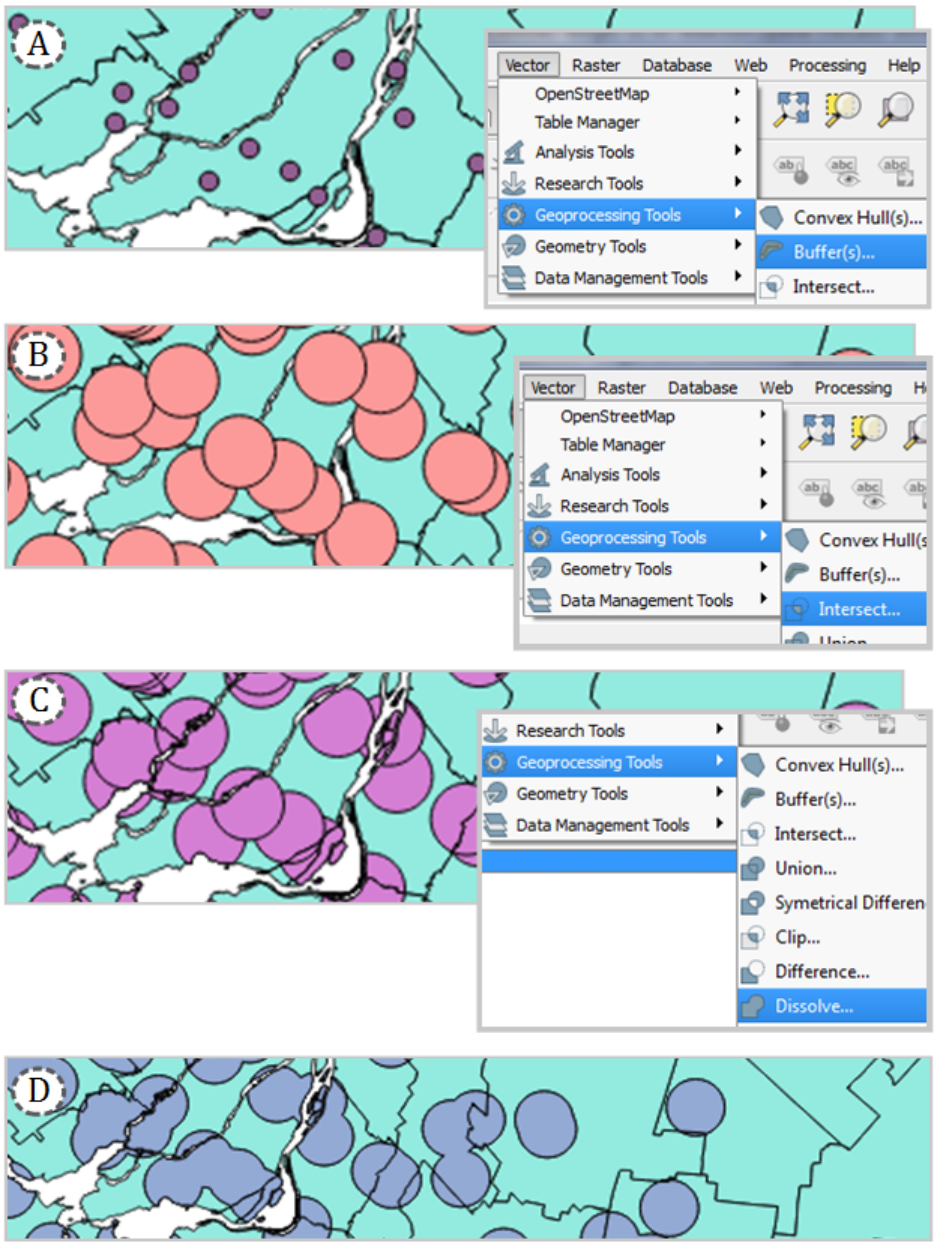


Figure 5.17 Catchment Area with QGIS

After obtaining the layer in 5.17.D , it is necessary to compute the superficies of the catchment areas that cover each of the Target areas. Finally, it is necessary to export the resulting dataset in an appropriate format that allows further analysis.

5.5.2 Steps with the implemented approach

The dataset must be stored in the corresponding schemas in the database. The user launches the application, selects the Target Layer (in Figure 5.18.A the user selects "DIVISION" which is a Census geographic unit), then the various possible Target Variables are proposed (in Figure 5.18.B the user select the sales per division for the year 2015). The user has access to different statistical information about the variable selected. The user can validate his choice by clicking on the corresponding button (Figure 5.18.C). The interface then switches to the selection of the Source Layer (Figure 5.19). After selecting one of the layers, the user can decide to set a Categorization Variable and launch the spatial relations computation (in Figure 5.19.A the user seek to evaluate Architects influence on sales). The different spatial relations are computed (for all categories if a Categorization Variable has been defined, which is not the case in Figure5.19). Then, the correlation scores between the spatial relations and the Target Variable are presented in a table (Figure 5.19.B). The interface in Figure 5.19.B, allows for the user to sort or filter spatial relations, and to add the ones he wants to a dataset of interest. When the user has evaluated the Source Layers selected and the desired spatial relations have been selected, the user can export the corresponding dataset by clicking on the dedicated button (Figure 5.19.C).

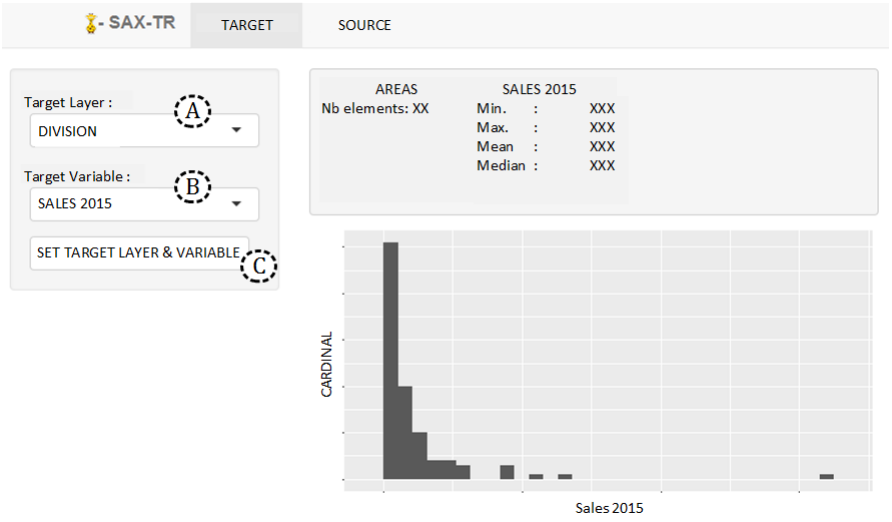


Figure 5.18 Target Layer Selection

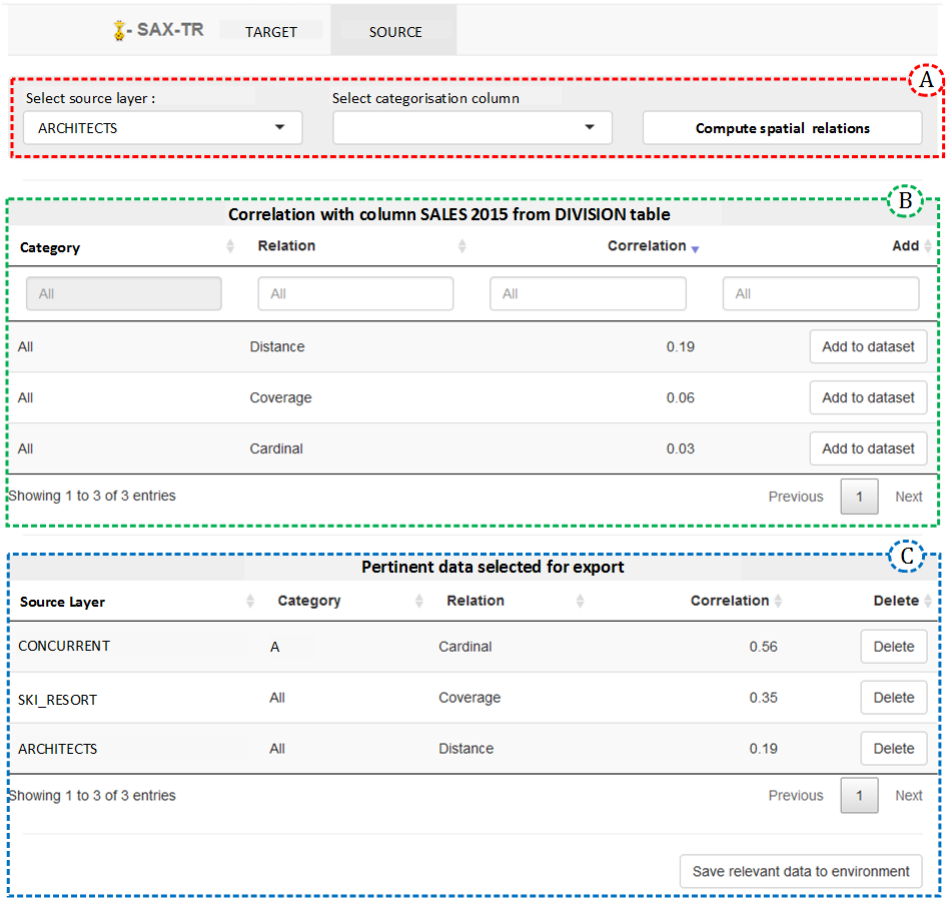


Figure 5.19 Source Layer Selection

The data is then available to the user in the RStudio environment. This example shows the reduction in the number of manipulations that the user must master and process to carry out the spatial data pre-processing.

To provide an idea of each sub-step duration, the manipulation to compute each included spatial relation was performed by a QGIS expert user who was familiar with the four different datasets. Resulting manipulation and computational times are presented in Table 5.2 (as QGIS does not offers any tools to monitor computation time, those were computed manually).

Table 5.2 Manipulation and Computing Time

Computation time in seconds	98 areas 225 points	98 areas 726 points	1285 areas 225 points	1285 areas 726 points
Cardinality				
Manipulation	≈ 20	≈ 20	≈ 20	≈ 20
Computing	≈ 1	≈ 1	≈ 2	≈ 2
Without tool total	≈ 21	≈ 21	≈ 22	≈ 22
With tool	0.6	1.5	2.2	4.5
Distance to first				
Manipulation	≈ 30	≈ 30	≈ 30	≈ 30
Computing	≈ 1	≈ 1	≈ 2	≈ 2
Without tool total	≈ 31	≈ 31	≈ 32	≈ 32
With tool	0.3	1	1.5	4
Area coverage				
Manipulation	≈ 90	≈ 90	≈ 90	≈ 90
Computing	≈ 4	≈ 20	≈ 8	≈ 15
Without tool total	≈ 94	≈ 110	≈ 98	≈ 105
With tool	0.2	21	4	21
All three relations				
Without tool	≈ 146	≈ 162	≈ 152	≈ 159
With tool	1.1	23.5	7.7	29.5

While those durations do not include the time required to gather and transfer resulting datasets into analysis software, it can already be observed that it is divided by at least five (on the task of computing three relations with four datasets of our case study).

However, the time saved is not the only improvement brought about by our tool, as illustrated in the next subsection.

5.5.3 Analysis quality improvement

The aim of this section is to support the idea that analysis quality differs depending on the data input: datasets with and without the data extracted with our tool.

In this example, the analysis goal is to predict, for each Census Subdivision, a potential score. Two forecasting models, Random Forest and Treebag, are used to compare predicted performances depending on the input datasets. Three input datasets were built:

— The first dataset is composed of all non spatial data (such as socio demographic data or construction data)

— The second dataset is also composed of all non spatial data and for each Source Layer,

the cardinal is added (number of Competitors, number of golf club, etc.)

— The third dataset also contains all non spatial data but for each Source Layer, it was the most pertinent spatial relation that was added to the dataset. As it can be observed in Figure 5.19 the different spatial relations have different correlation scores with the Target Variable. Here, for each Source Layer, the most correlated were added to the dataset.

The two models were then trained and evaluated with cross validation techniques, using the Caret package in R . This training and evaluation was performed three times for each model (one time for each dataset),resulting in the performances shown in Figure 5.20. As can be seen, for those two prediction models, compared to the first dataset (Env Data = None in Fig 5.20) the accuracy increases when cardinal data is included (Env Data = Cardinal in Fig 5.20), and increases more when it is the most correlated spatial relation that is added in the input dataset (Env Data = Pertinent in Fig 5.20)

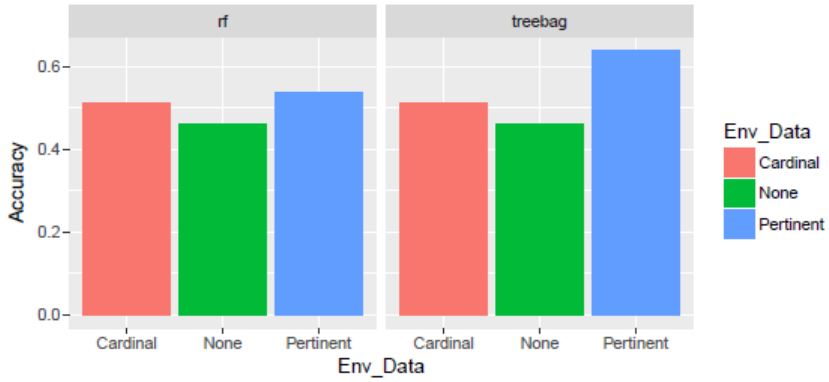


Figure 5.20 Random Forest and Treebag performances depending on input dataset

5.6 Conclusions and perspectives

Numerous studies denounce the complexity and time-cost of spatial data pre-processing. Few studies have tried to address these issues and have proposed methodological approaches or frameworks to facilitate pre-processing. Although these studies aim to simplify spatial data pre-processing, they do not provide solutions to the need for knowledge of GIS, or to the difficulty of choosing the spatial relations to be taken into account.

From this observation, our research proposes a tool that automates most of this process. The tool developed makes spatial data analysis available to users who do not have GIS knowledge. In addition to this, the realization of pre-processing can be carried out faster and with guided

choices for the selection of spatial relations.

To allow for reproducibility, this research presents the specifications for the tool : a process to follow, decomposed into several steps, which has been described in detail. Open-source tools are used in order to enable a cost-effective implementation of our approach.

For now, our implementation only works with a polygons layer as the Target and with a points layer as the Source. It might be useful to allow other types of data for the Target Layer (such as points or lines) and for the Source Layer (such as polygons or lines). Moreover, other spatial relations could be implemented and it could be of interest to allow for the spatial relations to be weighted by the numerical characteristics of the elements of the Source Layer.

CHAPITRE 6 ARTICLE 3: REALISTIC MODEL FOR THE MAXIMUM COVERING LOCATION PROBLEM USING SPATIAL DATA PRE-PROCESSING

Submitted to International Journal of Production Economics

Authors: G. DARAS, B. AGARD, B. PENZ.

Abstract. With the aim of proposing a realistic resolution approach to the Maximum Covering Location Problem (MCLP), in this paper we suggest a combination of data pre-processing and modelling techniques. The realistic constraints that are taken into account are presented, and the models and data pre-processing are explained through several examples.

In order to complete the entire treatment chain, from the initial data to the solutions, we also propose a relatively fast resolution algorithm. The proposed algorithm is based on dedicated data structures and evaluation functions. The resulting approach provides improvement for the case study instances in short processing times.

6.1 Introduction

Many problems facing science and society have a geospatial aspect [80]. While a few decades ago Herring et al. [63] were facing the problems of collecting and storing location related information, today Bradlow et al. [22] have advanced that there are more data sources available that contain spatial information. As organizations collect large amounts of geospatial data, there is actually a need to effectively use the collected data to make strategic and organizational decisions [42].

While the retail industry is facing many challenges [104], location analysis is one of its most important research areas. As Cliquet et al. [29] describe, knowledge about the spatial environment is a key element in the retail sector. Similarly, Thompson et al. [125] mention the growing interest for the analysis and understanding of spatialized data in the retail sector.

Location analysis is a broadly studied research field in geospatial research, and one problem inherently linked to location analysis and the retail industry is the Maximum Covering Location Problem (MCLP). Introduced by Church et al. [26], the MCLP consists of locating a fixed number of facilities in order to maximize the population covered within their ser-

vice distance. Population, represented as the demand point, is considered covered when it is within a service distance from a facility. As presented in Figure 6.1, there are two main approaches for solving the MCLP.

In a real application, initial datasets often correspond to those presented in Figure 6.1.A: the demand data and the facility location. Another dataset that might be available or computable is the trade area of facilities, corresponding to the area where customers might come from.

In the first approach, presented in Figure 6.1.B, initial demand data are transformed through a basic process into demand points; this transformation results in information loss [34], and then the models might have additional constraints to allow for a more realistic solution. Mendes et al [82] state that increasing complexity leads to robustness requirements and an increased cost of implementation and maintenance.

In the second approach used, for example, by Murray et al. [87] and presented in Figure 6.1.C, the initial dataset is preprocessed to take into account realistic aspects in the MCLP resolutions; thus it allows simple models to be kept for a resolution.

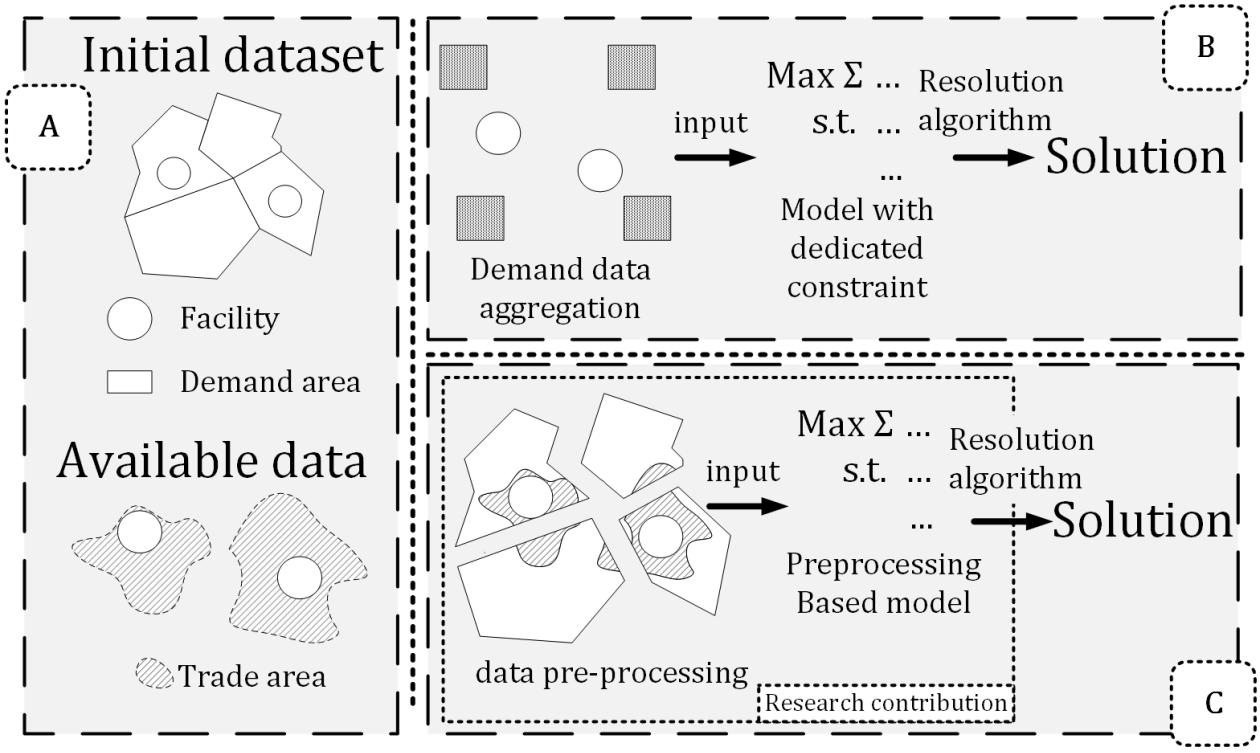


Figure 6.1 Approaches for MCLP resolution

This research aims to propose a data pre-processing transformation and an associated simple

model for the MCLP that take into account trade area data and allows for collaborative capture of the demand. The originality of the approach consists of the cross-use of advanced spatial database capabilities and adapted modelling techniques.

The paper is organized as follows: first, in the literature review, we present the MCLP, the associated problems and the research openings. Our modelling approach and the pre-processing of the data are presented in detail in Section 6.3 and 6.4. To complete the treatment chain, in Section 6.5 we present an algorithm that has been implemented for the proposed model that relies on data structures that allow for a rapid evaluation function. The whole treatment chain has been applied to the real dataset from our industrial partner with the aim of improving their demand coverage. Our industrial partner is a Canadian leader in building material, based in Quebec (Canada), with a strong presence in North America, mostly on the East Coast.

6.2 Literature Review

First introduced by Church et al. [26] a few decades ago the MCLP was inspired by the Location Set Covering Problem. Today, MCLP models are used in many research problems such as healthcare service [120] or logistics optimization in agriculture [10]. The MCLP basically consists of locating a finite number of facilities in order to maximize the covered population. A population is considered covered when it is within a service distance from a facility. In the first definition of the MCLP, population is represented as demand points at a specified location, and the potential location for setting facilities are discrete. In the initial definition, a demand point can be only covered by one facility. Some research such as Berman et al. [15] point out that the MCLP coverage objective makes unrealistic assumptions: (1) demand points outside the coverage radius are not covered at all. (2) The second-closest facility to a demand point has no bearing on coverage. Many research that treats the MCLP consider that demands are located at fixed points [96, 107]. As for Fotheringham et al. [111] the vast majority of location allocation models used aggregated demand data; they therefore state that their reliability is questionable. Similarly, Current et al. [34] point out that customer data aggregation is a loss of information and can lead to sub-optimal solutions. Matisziw et al. [81] say that service area can be interpreted in a broader way, and that the uniform demand assumption must be relaxed. Several work of research aim to integrate a realistic aspect into the initial MCLP. For example, Matisziw et al. [81] set their facilities wherever they wanted to within a continuous space. Berman et al. [13, 14] allow partial or cooperative coverage of demands. Plastria et al. [97] integrate future competitive arrival. For other MCLP adaptations that aim to be more realistic, see [46]. Mendes et al. [82] say that

increasing complexity leads to better model accuracy, but leads to robustness requirements and an increased cost of implementation and maintenance. Goodchild [55] also advances that there are strong rational arguments for the use of a simple location model. There are also other approaches that aim at making solutions more realistic. Blanquero et al. [16] studied a covering location problem that assumes the demand is distributed along networks.

While not specific to MCLP, Suarez et al. [122] use GIS to better understand their location models results. Murray et al. [87] mention that GIS should be used to interpret demand in a different format than points. As Suarez et al. [123] advance, GIS should not be only considered a visualization tool, it could also be used together with optimization models to improve decision-making aids. Similarly, Murray et al. [87] state that GIS should not be used to only provide basic input data, and that its capabilities are underestimated in model buildings. In several works of research [81, 87, 88, 89] Murray et al. insist that GIS, when used in the building of location models, offers a lot of perspective in providing more realistic solutions and that it must be better understood and solved. In this context, Loranca et al. [78] use GIS to take demand data into account in the input data. Unfortunately, their process seems hard to reproduce; furthermore, they still aggregate demand data.

From this literature review, the following assumptions remain:

— Location models should not be over-sophisticated [55, 82],

— Aggregate demand data as a point is inaccurate [34, 111],

— Coverage assumptions of classical models are unrealistic [15, 81].

Based on these assumptions our research aims at proposing a more realistic, yet simple, model that is allowed by a data pre-processing approach. In addition to taking into account trade area data, our proposed approach allows collaborative coverage to be taken into account, depending on those trade areas. For that, we take advantage of recent spatial database and GIS capabilities. Starting from data on potential facility trade areas and regional demands, we propose an improved model (Section 6.3) which is made possible via pre-processing the initial dataset (Section 6.4).

6.3 Model

In the usual approach, as has already been presented (Figure 6.1.B), the initial data transformation is often a basic transformation, as presented in Figure 6.2, consisting of transforming demand area into demand points at the position of their centroid, and to set links among each (facility, demand point) couple that are close to each other (i.e. within a service distance). This transformation, not taking into account trade areas, induces information loss,

particularly on the distance between "real" customers and a facility location. Furthermore, on the collaborative capture aspect, many models do not allow for it to be taken into account. Otherwise, the models that allow this aspect to be integrated generally contain complex constraints.

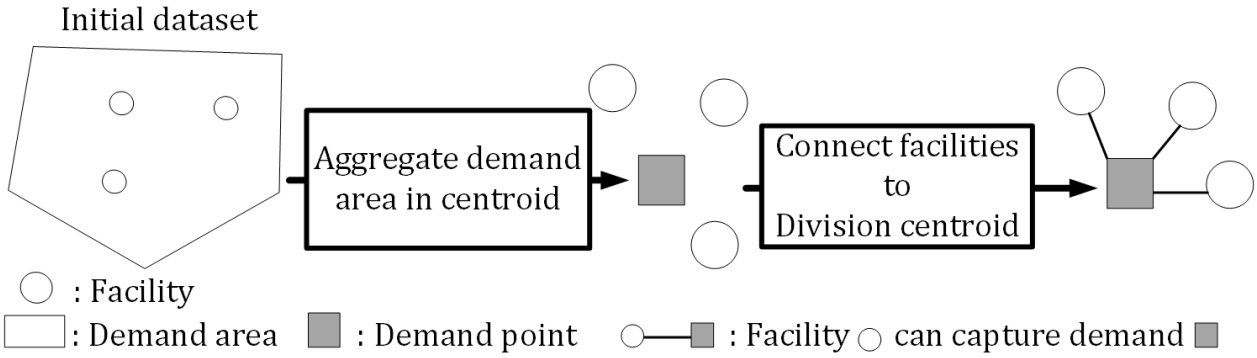


Figure 6.2 Usual data transformation illustrated

In our approach, the combination of modelling techniques and data pre-processing allows the trade area to be taken into account and the collaborative capture of the demand. The proposed pre-processing transformation of the initial dataset to the input data for our model is illustrated in Figure 6.3. It basically consists of (1) transforming demand area into demand points; (2) affecting the capture rate of each couple (facility, demand point). It must be noted that, contrary to the usual approach, there is no location associated with demand points. The proximity aspects are included in the fact that they will be linked to a facility only if they come from its trade area.

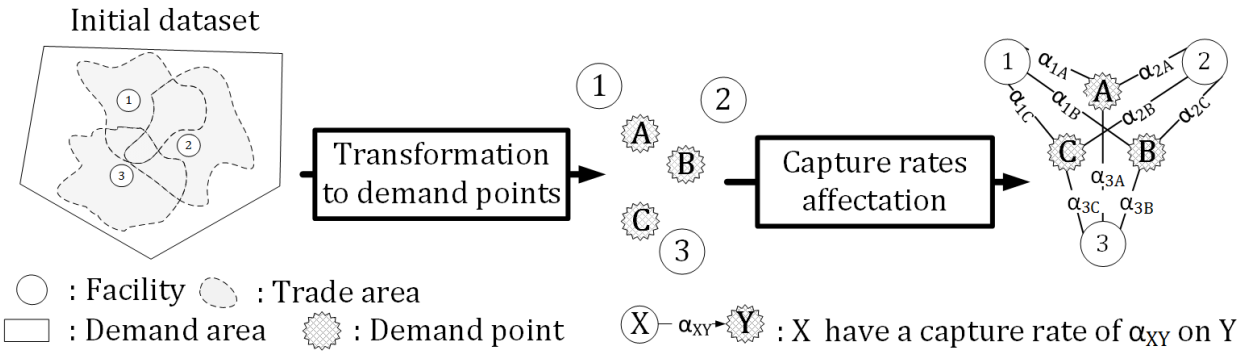


Figure 6.3 Transformation process illustrated

To begin with, the initial dataset and the constraints we impose on our transformation are presented in Section 6.3.1.

6.3.1 Initial dataset and realistic constraints

In most cases, demand data are spatial data that can be represented as in Figure 6.4. The demand data corresponds to a geographical area that is partitioned into several division areas. Each division (1, 2, 3 and 4 in Fig. 6.4) has associated data corresponding to its demand, such as the expected turnover, for example.

Each facility (a, b and c in Fig 6.4) is positioned at a given location and has an initial state: open or close. Additional data corresponding to the trade areas of the facility might be available as represented by the hatched areas in Fig. 6.4. As can be observed on Figure 6.4, trade areas might only partially cover demand area; furthermore, several trade areas might overlap. Our approach allows for those characteristics to be considered.

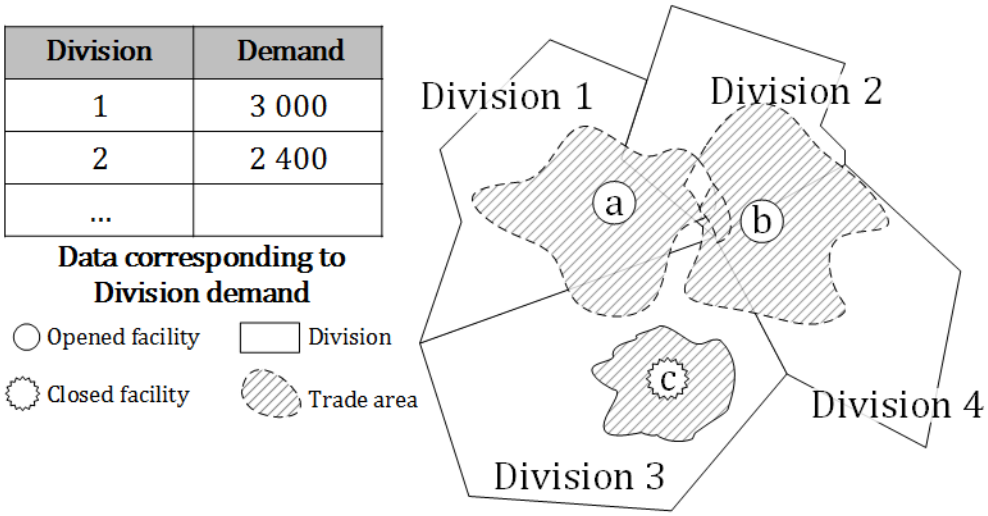


Figure 6.4 Initial dataset illustrated

Trade area computation is a research area itself (such as in [33]) that relies on the multiple factors that influence attractiveness [93] and on the interactivity among nearby facilities [131]. As the focus of this paper is not on trade area computation, and as simple methods exist (such as distance buffer) to compute approximate trade areas, we do not present further information on that subject. Nevertheless we do consider trade area data available for the application of our approach.

From that initial dataset, the demand potentially captured by a facility on a given division depends on the facility trade area, and on the presence of other facilities in the neighbourhood.

As each facility has an initial state - open or closed - they have associated opening and closing costs.

Based on those dataset, the constraints that we impose on our transformation are the following:

— constraint (A): a facility can only capture the demand that is within its trade area. In other words, if a trade area covers only ten percent of a division, it will capture at most ten percent of the demand of that division.

— constraint (B): a demand area that is covered by one or more trade areas (that overlap) cannot be captured to more than one hundred percent. That is, even if a division is covered by many trade areas, the maximum captured demand is equal to the division's demand.

— constraint (C): a facility opening can only increase the overall capture within its trade area. This means that, regardless of the open facilities, the opening of a new facility in a division can only increase total capture on this division.

To allow for the respect of those constraints, we propose transforming overlapping areas into demand points and to affect capture rates among those demand points and facilities. The model that takes this input is presented next.

6.3.2 MCLP Model

As mentioned before, the initial dataset is transformed to feed the model: demand areas are converted into demand points and capture rates are affected to each couple (facility, demand point). Our proposed model is the Graph model G.

Let $G = (V, E)$ with the vertices set $V = F \cup P$ be composed of :

— Facilities $i \in F$

— Demand points $j \in P$

Edges set E is composed of e_{ij}, $\forall i \in F, j \in P$.

Initial data

— $demand(j)$: demand of demand point $j \in P$.

— o_i: opening cost of facility $i \in F$.

— c_i: closing cost of facility $i \in F$.

— α_{ij}: capture rate of facility $i \in F$ over demand point $j \in P$.

— δ_i: equals 1 if facility $i \in F$ is initially opened, 0 otherwise.

Parameters

— $NbOpening$: maximum number of opening,

— $NbClosing$: maximum number of closing,

— $NbFinalOpen$: final number of open facilities.

Variables

— $y_i = 1$ if facility $i \in F$ is open, 0 otherwise.

— $x_{ij} = 1$ if demand point $j \in P$ is captured by facility $i \in F$, 0 otherwise.

The resulting model is composed of the following objective and constraints:

$$\max \sum_{\substack{i \in F \\ j \in P}} x_{ij} \times \alpha_{ij} \times demand(j)$$

$$\text{s.t.} \quad x_{ij} \leq y_i \qquad \forall i \in F\ ,\ \forall j \in P \tag{6.1}$$

$$\sum_{i \in F} x_{ij} \leq 1 \qquad \forall j \in P \tag{6.2}$$

$$\sum_{\substack{i \in F \\ \delta_i = 0}} y_i \leq NbOpening \tag{6.3}$$

$$\sum_{\substack{i \in F \\ \delta_i = 1}} 1 - y_i \leq NbClosing \tag{6.4}$$

$$\sum_{i \in F} y_i = NbFinalOpen \tag{6.5}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i \in F\ ,\ \forall j \in P \tag{6.6}$$

$$y_i \in \{0, 1\} \qquad \forall i \in F \tag{6.7}$$

The objective function aims at maximizing overall capture. Constraint 6.1 allows for the demand point to be covered only by the open facility. Constraint 6.2 limits the number of facilities that can cover a demand point to one. This constraint, associated with our transformation that will be presented in the next section, allows for the constraints to be respected ((B) and (C)). Constraints 6.3, 6.4 and 6.5 ensure that the limits on the opening and closing of facilities are respected. Those limits on facility openings and closings depend on the application case. In our case study, managers have an upper limit on facility openings and they need to conserve the same number of open facilities in the end.

To be able to start from the initial dataset previously presented (Section 6.3.1), and then to

get the model input data (facilities and capture rates) that respect the imposed constraints, a specific data transformation is requested. How the imposed constraints impact the transformation to obtain the input data, and what we propose as an implementable transformation, are presented in next section.

6.4 Pre-processing data transformation

To get from the initial dataset to data that can be provided to the model, transformations have to be made, and those transformations must respect the imposed constraints (Constraints A, B and C). To allow for a better understanding of the transformations and how the constraints impact them, they are first introduced in several small examples in Section 6.4.1.

To begin our explanation, we define an intersection I as an overlap of one or more trade areas with a demand area. The set of facilities that are part of the intersection I is named F_I. The first illustration is on the case where there is no overlap among trade areas, but several demand divisions, as is presented in Fig 6.5.

6.4.1 Simple transformations

No overlap among trade areas The first operation consists of separating trade areas into intersections depending on the division in which they are (see Fig 6.5).

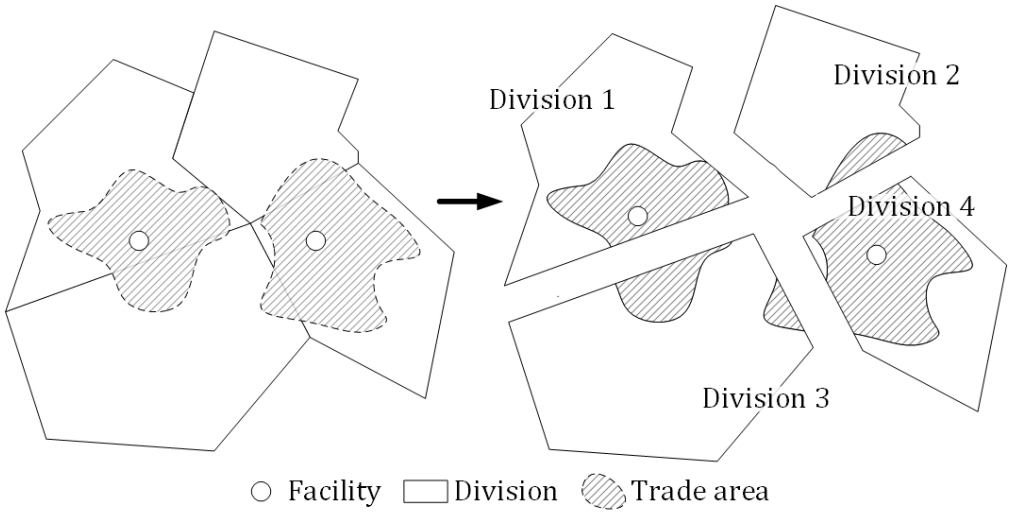


Figure 6.5 Trade areas separation

Then, every intersection I between a division and a trade area is transformed into $|F_I|$

demand point(s) (in this case: one demand point). Figure 6.6 presents the corresponding transformation for Division 1 and 3 of the previous example.

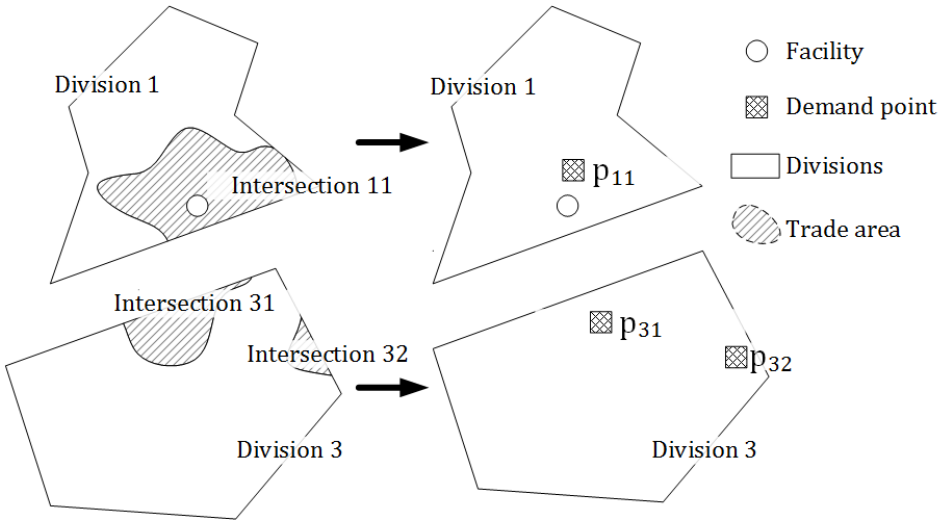


Figure 6.6 Transformation of intersections to demand points

To respect Constraint (A), the demand is defined as the ratio between the area of the trade area (for example, in squared meters) within the division and the area of the division (in the same unit). In our example, the demand points have an associated demand such as:

$$demand(p_{11}) = \frac{area\,(Intersection\,11)}{area\,(Division\,1)} \times demand(Division\,1)$$

$$demand(p_{31}) = \frac{area\,(Intersection\,31)}{area\,(Division\,3)} \times demand(Division\,3)$$

$$demand(p_{32}) = \frac{area\,(Intersection\,32)}{area\,(Division\,3)} \times demand(Division\,3)$$

As mentioned before, no location is associated with demand points. The proximity aspect is taken into account by the fact that demand points are only linked to the facility related to the associated trade areas. Each facility $i \in F$ will have, only for each demand point $j \in P$ that comes from its trade area, an associated capture rate α_{ij}. To respect constraint (A) there is no capture rate between a facility and a demand point that is not part of its trade area. Then, to respect Constraint (B), for each demand point that is part of a facility trade area, the capture rate has to be inferior or equal to one (as shown in Fig 6.7). The assignment of the capture rates is illustrated further.

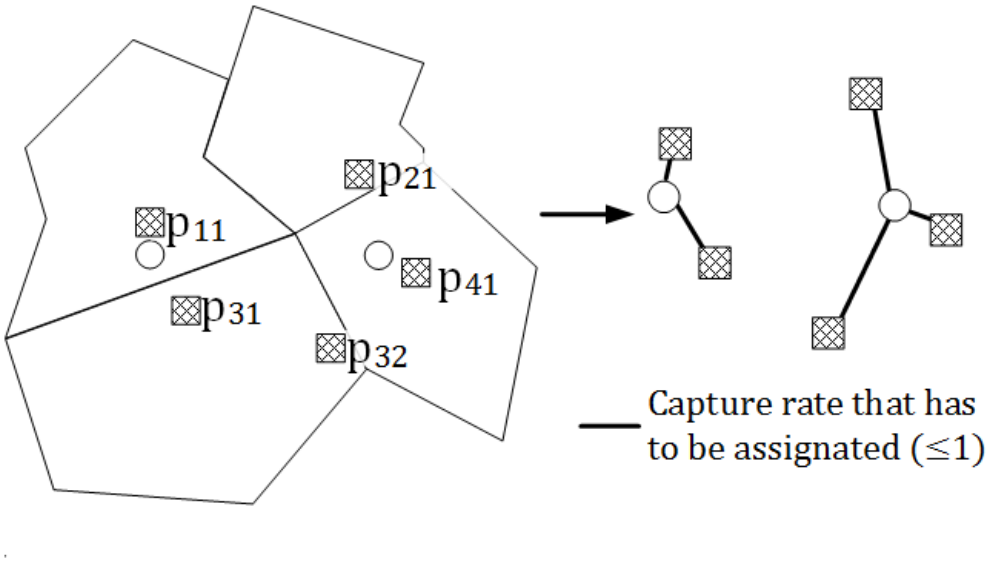


Figure 6.7 Capture rate assignment

For this basic case, this simple transformation respects the constraints. For more complex cases, the imposed constraints make the transformation more complex.

Two-trade areas intersections As in the previous example, non overlapping trade areas are transformed into a single demand point. In a case where two trade areas overlap, the resulting area is converted into two demand points as illustrated in Fig 6.8. This duplication of demand points allows for assigning capture rates that allow the collaborative capture of overlapping trade areas to be modelled.

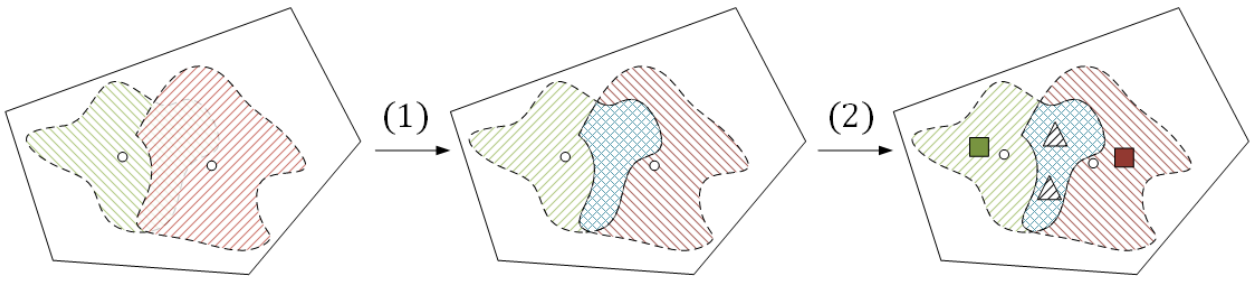


Figure 6.8 Two-trade area intersection transformation 1/2

Then, the capture rates between demand points and facilities are assigned as shown in Figure 6.9.

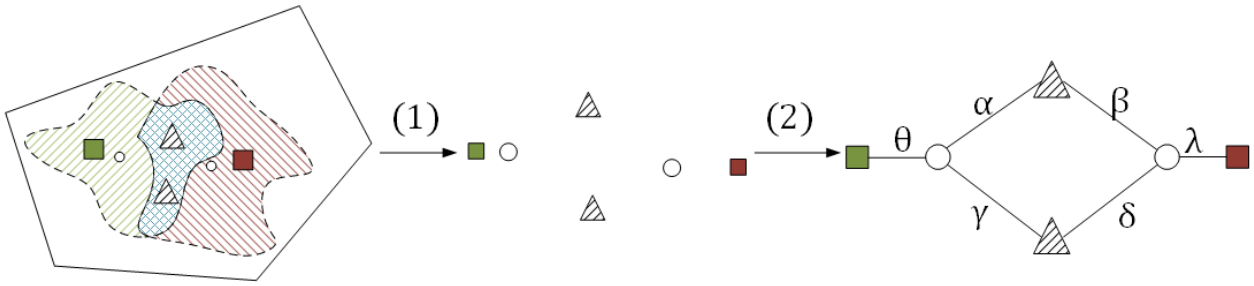


Figure 6.9 Two-trade area intersection transformation 2/2

For this particular case, relying on the second model constraint, which implies that a demand point can only be covered by one facility, the constraints correspond to:

— Constraint (B), overall capture rate inferior or equal to 1:

$$\theta \leq 1,\ \lambda \leq 1\ ,\ \alpha + \delta \leq 1 \text{ and } \gamma + \beta \leq 1$$

— Constraint (C), increasing overall capture rate when adding a facility:

$$\alpha + \gamma < \alpha + \delta \text{ and } \beta + \delta < \beta + \gamma$$

Values that respect Constraints (B) and (C) could be: $\theta = \lambda = 1,\ \alpha = \delta = 0.5,\ \beta = \gamma = 0.1$. If those are the affected values, then the overall capture of the two trade area intersection is either $0.5 + 0.1 = 0.6$ (if only one facility is open) or $0.5 + 0.5 = 1$ (if the two facilities are open). This second example introduces how the imposed constraints impact the assignment of capture rates. Before the generalization and the formalization of these constraints (Section 6.4.2), one final, more complex cas,e with the intersections of three trade areas is presented.

Three trade areas intersections As can be observed in Figure 6.10, each intersection is converted into as many demand points as it has trade areas. Each demand point is given a demand of

$$demand(demand\ point) = \frac{area\,(intersection)}{area\,(Division)} \times demand(Division)$$

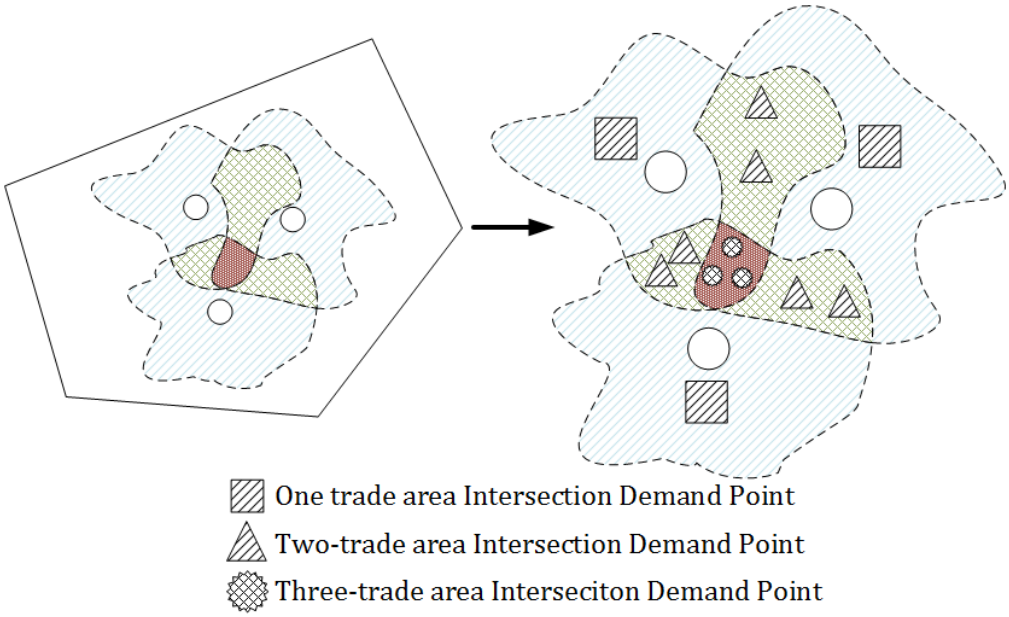


Figure 6.10 Converting intersections to demand points

This results in the graph presented in Fig. 6.11, where capture rates need to be assigned.

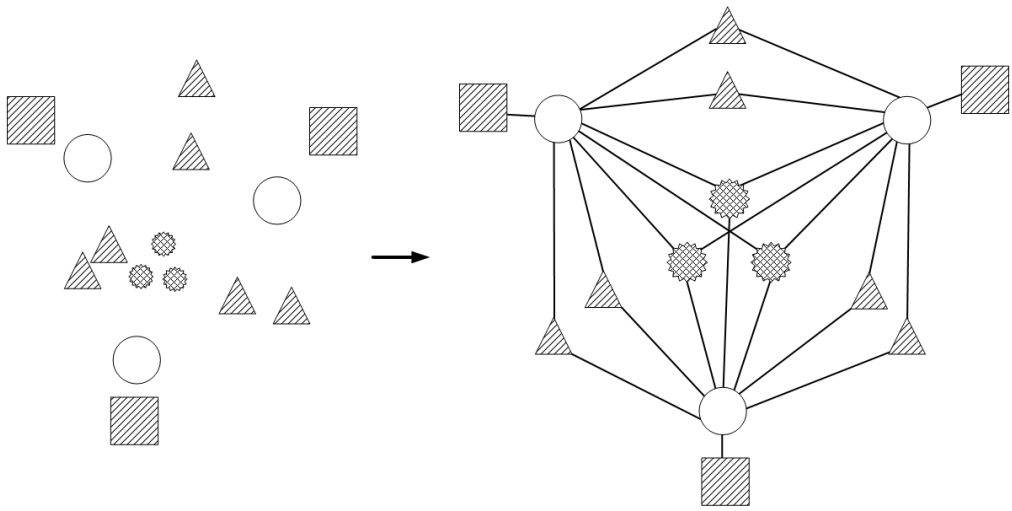


Figure 6.11 Capture rates to assign for the intersections of three trade areas

As the resulting graph is symmetrical, respecting the constraints can allow for the reduction to two particular cases in Figure 6.12.

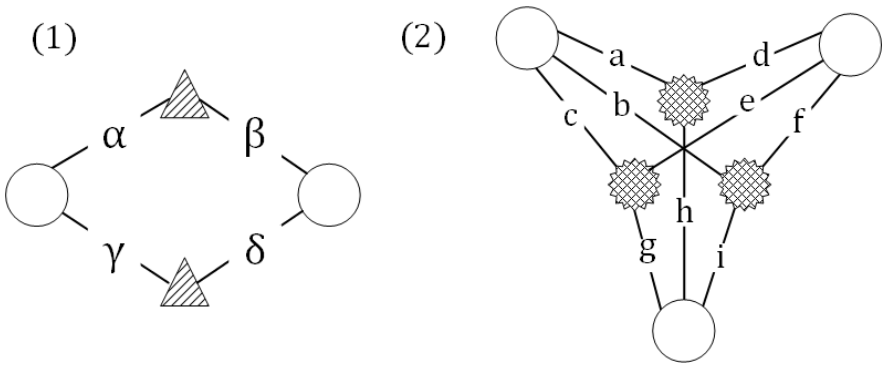


Figure 6.12 Sub-problem for the affect on capture rates

Then, still relying on the model constraint 6.2, which implies that a demand point can only be covered by one facility, the resulting sub-problems to respect the constraints (B) and (C) can be formalized. For Figure 6.12.(1), it is the same as in previous example with the intersection of the two trade areas (Figure 6.9. For Figure 6.12.(2)(for readability reasons, symmetrical constraints are not presented here) we have :

— Constraint (B) will be in the form of:

$$a + i + e \leq 1$$

Under the assumption that a demand point is at most captured by one facility, this constraint, and the symmetric ones, ensure that the maximal overall capture rates cannot exceed 1.

— Constraint (C) will be in the form of:

$$a + b + c < a + b + g$$

This constraint and the symmetric one ensure that two facilities have an overall capture rate that is superior to any one facility in the intersection.

$$a + b + g < a + f + g$$

This constraint ensures that any pair of facilities in the intersection have a lower overall capture rates than the three facilities have together.

In this example, values that respect the Constraints (A) and (B) can be: $a = e = i = 0.3$, $b = c = d = f = g = h = 0.2$. With those values, the maximal capture rates, if three facilities are open, is 0.9 (0.3+0.3+.03), for two open it is 0.7 (0.3+0.3+0.1), and for only one it is 0.5 (0.3+0.1+0.1). Our proposed transformation and how the constraints impact it are now presented. The next section generalizes the transformation and the impact of the constraints on it.

6.4.2 Generalization and formalized constraints

The constraints have to be respected on each intersection I. For each intersection I identified, the transformation to the demand points have to be made, and the capture rates have to be assigned. Figure 6.13 illustrates this process for an intersection I in which n facilities' trade areas overlaps; the resulting constraints are explained

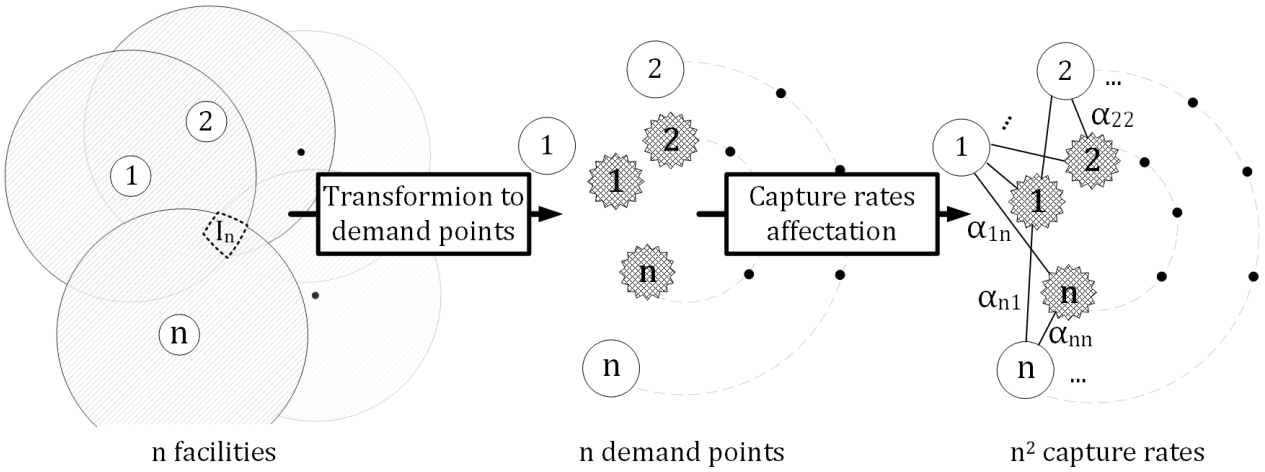


Figure 6.13 Generalized Transformation

In this general case, the resulting constraints can be formalized as follows. For each intersection I composed of one or more trade area with a demand area, we define:

— F_I: set of k facilities whose trade areas take part in intersection I.

— P_I: set of k demand points from intersection I.

— α_{ij}: capture rate of $i \in F$ on $j \in P$, is defined when i and j are related.

As a reminder, $x_{ij} = 1$ if demand point $j \in P$ is captured by $i \in F$, 0 otherwise a demand point can only be captured by one facility:

$$\sum_{i \in F_I} x_{ij} \leq 1 \ , \ \forall j \in P_I \tag{6.8}$$

For all intersection I of one or more trade areas with a division, and for all S, subset of F_I, we define $Capt(S)$ as the maximal demand captured by S on I under constraint (6.8).

$$Capt(S) = \sum_{j \in P_I} \max_{i \in S} \alpha_{ij} \tag{6.9}$$

Equation (6.9) can be explained as follows: as each demand point is captured by at most one facility, if each demand point is captured by the facility of S that have the best capture rate on it, then the overall capture is maximal. With this given definition, it is now possible to formalize constraint (B) and (C).

Constraint (B) formalized: overall capture rate on an intersection ≤ 1.

$$\forall I,\, \forall S \subset F_I,\, Capt(S) \leq 1 \tag{B}$$

Constraint (C) formalized: an opening increases intersection overall capture:

$$\forall I,\, \forall S \subset F_I, \forall i \in F_I \backslash S, Capt(S) < Capt(S \cup \{i\}) \tag{C}$$

Constraints (B) and (C) are now formalized, a transformation that respects those constraints is presented in next section.

6.4.3 Implemented Transformation

Demand points and capture rates To respect Constraint (A) : $\forall I$, demand values of demand point are:

$$\forall j \in P_I,\; demand(j) = \frac{area\,(I)}{area\,(division)} \times demand(division)$$

Then to allow for a rapid transformation for each Intersection I, demand points p_I and facilities F_I are renumbered to have respectively $P'_I = \{1..|P_I|\}$ and $F'_I = \{1..|F_I|\}$ as illustrated in Figure 6.14. Then, for each couple ($i \in F'_I$,$j \in F'_I$) of a same intersection I,the capture rate α_{ij} is set :

$$\forall i \in F'_I,\, \forall j \in P'_I,$$
$$\text{if } i + j \equiv 0 \mod |F_I| + 1 \text{ then } \alpha_{ij} = \frac{1}{|F_I|} \text{ else } \alpha_{ij} = \frac{1}{|F_I| + \varepsilon_{ij}},\, \varepsilon_{ij} > 0$$

The corresponding assignment of capture rates for a given intersection I is illustrated in Figure 6.14

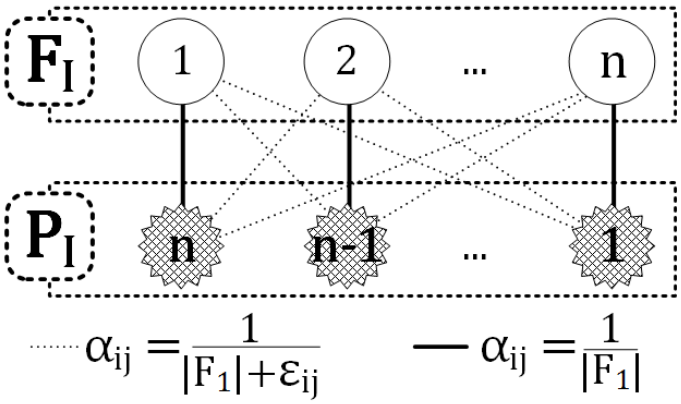


Figure 6.14 Intersection points renumbering and capture rates assignment

The ε_{ij} value is defined depending on the application domain characteristics and on facility attractiveness. For example, setting high ε values result in weak demand capture on intersections where only a few facilities are open. Setting little ε results in high demand capture on intersections even if a facility is alone to cover it.

The implemented transformation illustrated on two trade area intersections with ε =1

Figure 6.15 shows the implemented transformation with a two-trade area intersection and ε =1. The capture rate is either $\frac{1}{|F_I|}$ or $\frac{1}{|F_I|+\varepsilon}$ which in this case, with two facilities part of the intersection and ε =1, respectively correspond to $\frac{1}{2}$ and $\frac{1}{3}$.

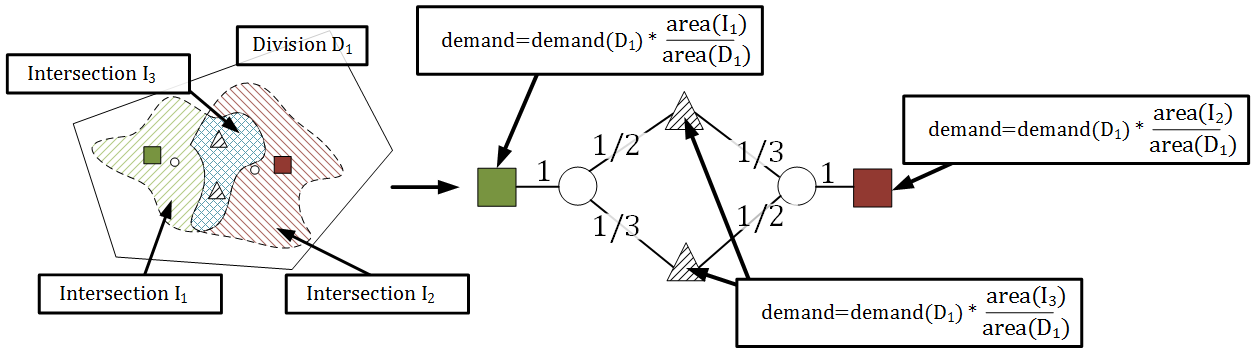


Figure 6.15 Transformation steps for two-trade areas intersection

The implemented transformation illustrated on three trade area intersections with ε =1

Figure 6.16 shows the implemented transformation with a three-trade area intersection (Fig. 6.16.(1)) and the capture rates assignment with ε =1 (Fig. 6.16.(2)). As before, the capture rate is either $\frac{1}{|F_I|}$ or $\frac{1}{|F_I|+\varepsilon}$. In this case, with three facilities part of the intersection and ε =1, the corresponding capture rates are respectively equal to $\frac{1}{3}$ and $\frac{1}{4}$.

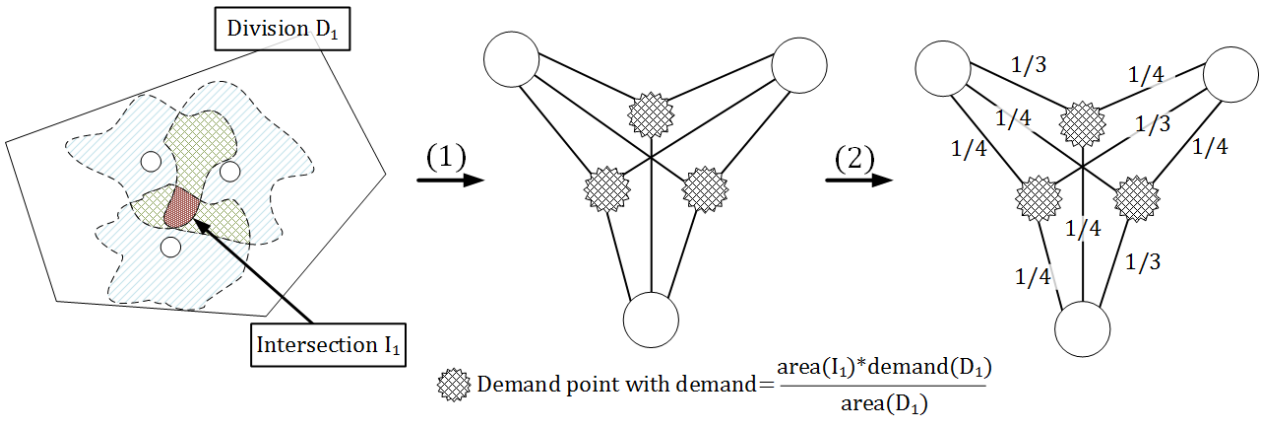


Figure 6.16 Transformation steps for three-trade areas intersection

The proof that the proposed implementation of the transformation respects the imposed constraints is given in the next section.

6.4.4 Proof of respect of the constraints

This section, while not needed for understanding nor for reproducibility, presents proof that the proposed implementation of the transformation respects the imposed constraints.

Proposition 1. ***Constraint (B)****, i.e.* $\forall I, \forall S \subseteq F_I, Capt(S) \leq 1$*, is respected.*

Proof. By definition, for each intersection I, facilities that belong to F_I cannot have a capture rate on demand points of P_I superior to $\frac{1}{|F_I|}$.

$$\forall I, \forall S \subseteq F_I, \forall i \in S, \forall j \in P_I, \alpha_{ij} \leq \frac{1}{|F_I|} \tag{6.10}$$

As a demand point is captured by at most one facility, there is at most $|P_I| = |F_I|$ edges x_{ij}

set to 1.

$$\sum_{\substack{i \in S \\ j \in P_I}} x_{ij} \leq |F_I| \tag{6.11}$$

So, from (6.10) and (6.11):

$$Capt(S) = \sum_{\substack{i \in S \\ j \in P_I}} x_{ij} \times \alpha_{ij} \leq |F_I| + \frac{1}{|F_I|} = 1$$

□

Proposition 2. ***Constraint (C)*** *is respected, i.e.*

$$\forall I, \forall S \subset F_I, \forall i \in F_I \backslash S, Capt(S) < Capt(S \cup \{i\})$$

Proof. for each facility i of an Intersection I, there is one and only one demand point j^* for which the capture rate is maximal (and equal to $\frac{1}{|F_I|}$).

Existence: $\alpha_{ij^*} = \frac{1}{F_I}$ for $j^* = |F_I| - i + 1$. Then j^* exists.

Unicity: by definition:

$$\forall j' \in [\![1, .., |P_I|]\!] \backslash \{|F_I| - i + 1\}, \alpha_{ij'} = \frac{1}{|F_I| + \varepsilon_{ij'}}, \varepsilon_{ij'} > 0 \tag{6.12}$$

Then, for each facility i of an intersection I, there is a demand point j^* for which this facility is the only one to have the maximal capture rate.

$$\forall I, \forall i \in F_I, \exists j^* \in P_I \mid \alpha_{ij^*} > \alpha_{i'j^*}, \forall i' \in F_I \backslash \{i\} \tag{6.13}$$

So, from (6.8) and (6.13), as the demand point is only captured by one facility:

$\forall I, \forall S \subset F_I, \forall i \in F_I \backslash S, \exists j^* \mid$

$$\alpha_{ij^*} > \sum_{i' \in S} x_{i'j^*} \times \alpha_{i'j^*}$$

So for any facility subset S of any intersection I, if a demand point i of $I \backslash S$ is added to S, the overall capture rate increases.

$\forall I, \forall S \subset F_I, \forall i \in F_I \backslash S, \exists j^* \mid$

$$Capt(S \cup \{i\}) \geq Capt(S) + \alpha_{ij^*} - \sum_{i' \in S} x_{i'j^*} \times \alpha_{i'j^*} > Capt(S)$$

□

From that section, it is proven that the proposed data transformations allow for the model input data to respect realistic constraints. The next section is focus on the data pre-processing that corresponds to those transformations to allow for reproducibility.

6.4.5 Spatial data pre-processing steps

In this section, we present the initial dataset and the different operations to perform to get the dataset adapted to the resolution algorithm presented in Section 6.5. In previous work for this case study, we had to set-up a Spatial Decision Support System framework that allows for multiple models to be integrated. This research, and another focusing on spatial data pre-processing automation [37], results in tools that use a previously set-up framework [36], and are integrated in the case study dedicated SDSS. However, the entire framework is not required for this pre-processing. This approach is made with PostgreSQL for database and the PostGIS extension to allow for spatial treatment. It must be noted that pre-processing is also possible with other Geographic Information System such as QGIS.

Starting dataset for pre-processing

Figure 6.17 presents the starting dataset of our case study. It is composed, on the one hand, of the trade areas dataset that consist of trade area shapes and IDs equal to the corresponding facility IDs (see Figure 6.17.A). On the other hand, there is the demand dataset that is composed of the demand area shapes and demands (illustrated in Fig 6.17.B).

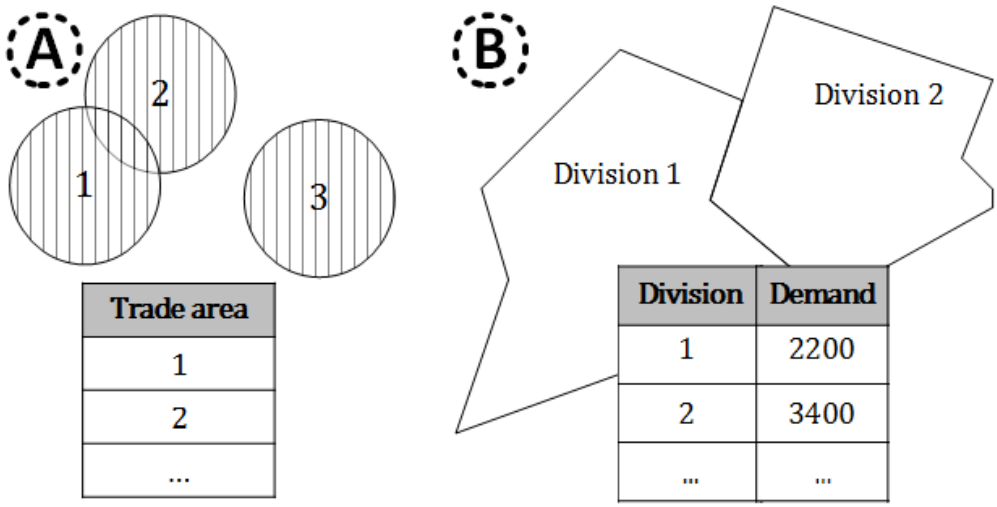


Figure 6.17 Case study initial datasets

Duplicate overlapping trade areas

This step consists first of finding and identifying the areas that overlap (Figure 6.18.U). Second, it consists of the duplication of each previously obtained overlap in as many records as there are trade areas that are part of the overlap (Figure 6.18.V). The resulting dataset, as illustrated in the given example, contains for each record the area shape, the original trade area ID, and the overlap ID.

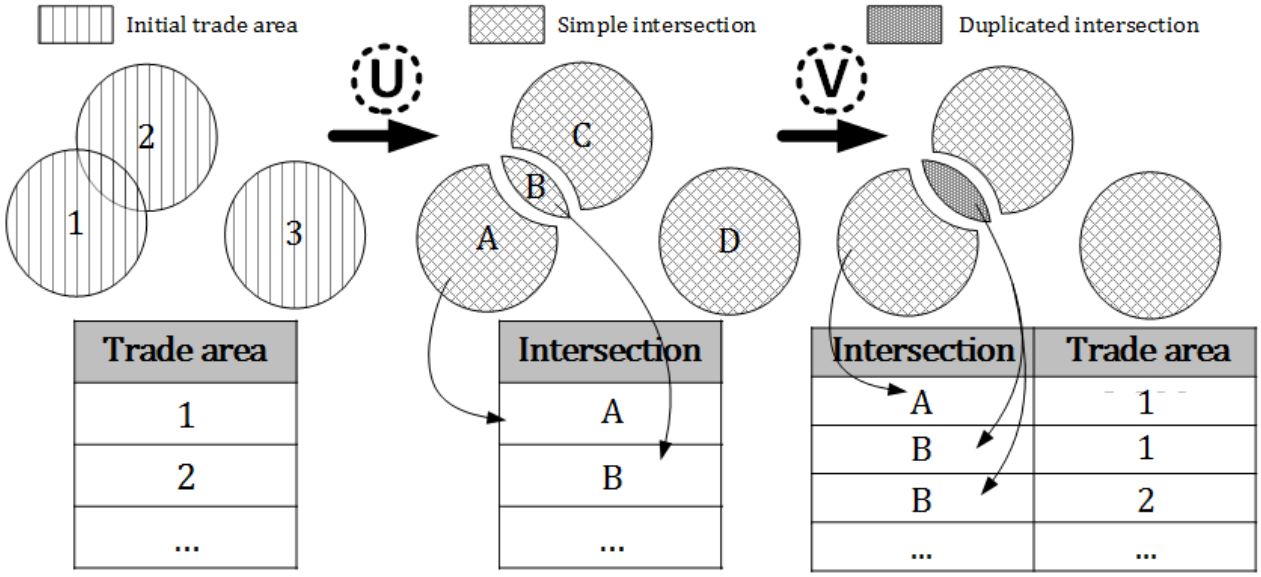
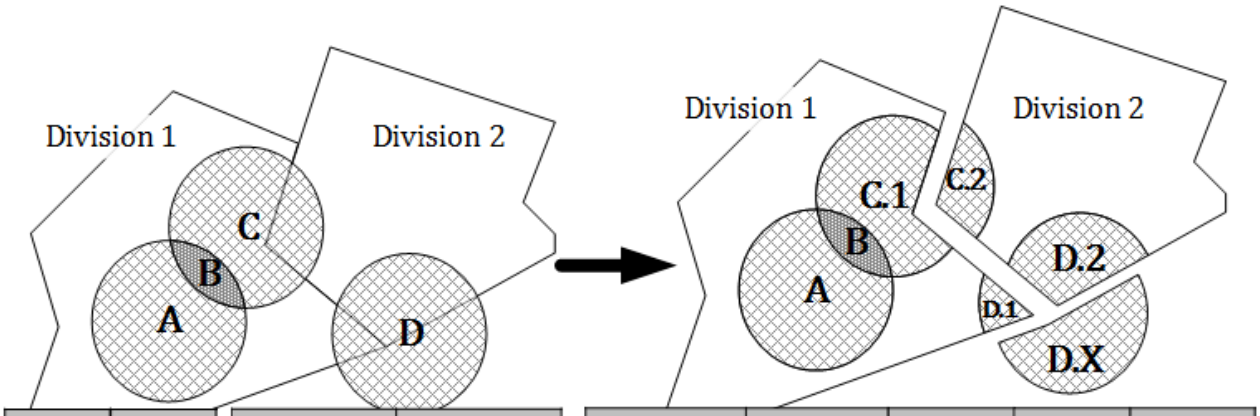


Figure 6.18 Illustration of the request for the duplication intersection

Intersection with demand areas

In the second step, we compute intersections among the demand divisions and the dataset resulting from the previous step, as illustrated in Fig. 6.19. For each record, a unique ID is given, and shapes and other information are saved (as illustrated in Figure 6.19).



Division	Demand
1	2200
2	3400
...	...

Overlap	Trade area
A	1
B	1
B	2
...	...

Intersection	Trade area	Division	Demand	Point ID
A	1	1	2200	1
B	1	1	2200	2
B	2	1	2200	3
C.1	2	1	2200	4
C.2	2	2	3400	5
...	...	...	...	...

Figure 6.19 Request for intersection with division

The resulting dataset contains every piece of information that is required to prepare the input data for the optimization model. Indeed, in addition to the saved information, the shape allows for the area ratio to be computed among divisions and intersections that will give the demand point value. To complete the treatment chain, the next section (6.5) presents the algorithm that proposes solutions starting from the dataset obtained.

6.5 MCLP resolution

In this section, we present our algorithm that aims at improving demand coverage. In our case study, our partner has an initial retail network that is already in place. Managers need to close and open a given number of facilities. We initially developed two algorithms, one with a naïve approach, and one with a Variable Neighbourhood Search (VNS) approach. As the naïve approach performs better on any of our case study instances, here we only present the naïve approach.

The algorithm starts from existing solution (with initially opened and closed facilities) and intends to improve the demand capture by proposing facility closings and openings. Section 6.5.1 focus on the main algorithm and presents the results obtained on our datasets. Then

the hypothesis on which our resolution approach relies and how it allows for a fast evaluation function are presented in Section 6.5.2.

6.5.1 Resolution algorithm and performance

Starting from an existing solution, and with *NbOpening* and *NbClosing* parameters, the main algorithm (Algorithm 1) opens and closes the corresponding number of facilities. The approach consists of: first finding the *NbClosing* worst facilities and closing them: the algorithm evaluates the impact of closing for each opened facility, picks the worst one, and closes it (with the *Close* function that also updates the affected variables), and reproduces those steps *NbClosing* times.

Second, finding the *NbOpening* best facilities and opening them: in the same way, the algorithm evaluates for each currently closed facility the impacts of opening, picks the best and opens it (with the *Open* function that also updates the affected variables), and reproduces this procedure *NbOpening* times.

Algorithm 1 Main Algorithm

$NbClosed = 0$; $NnOpened = 0$;
while $NbClosed < NbClosing$ **do**
  $IdWorstFacility = null$; $WorstCapture = \infty$
  for $i \in \{ F \mid \delta_i = 1 \; AND \; y_i = 1 \}$ **do**
    $Gap = evaluateGap\{i, -1\}$
    if $Gap < WorstCapture$ **then**
      $IdWorstFacility = i$; $WorstCapture = Gap$
    end if
  end for
  $Close\{IdWorstFacility\}$; $NbClosed+ = 1$
end while
while $NbOpened < NbOpening$ **do**
  $IdBestFacility = null$; $BestCapture = 0$
  for $i \in \{ F \mid y_i = 0 \}$ **do**
    $Gap = evaluateGap\{i, 1\}$
    if $Gap > BestCapture$ **then**
      $IdBestFacility = i$; $BestCapture = Gap$
    end if
  end for
  $Open\{IdBestFacility\}$; $NbOpened+ = 1$
end while

Table 6.1 presents the performances of our approach on the full case study dataset, and

on subsets with low and high facility density. The low density dataset (illustrated further in Figure 6.20) is composed of the 13 Census Subdivisions (CSD) and 112 facilities, the transformation results in 1002 demand points created. The high density dataset corresponds to the two CSD of Laval and Montreal, with the 42 facilities that are on it. The pre-processing transformation on the high density dataset results in 3998 demand points. Finally, the full dataset is composed of the 1285 CSD of the province of Quebec, and the 1068 facilities that are in Quebec. With this dataset, the transformation results in 30346 demand points.

In addition to the capture gain, Table 6.1 presents the processing time for the algorithm, and the data pre-processing time. It must be noted that the data pre-processing has to be done only one time, then the resolution algorithm can be processed as many times as needed.

Table 6.1 Pre-processing, computing and gain performances

NbOpening	Pre-processing time (in seconds)	Resolution time (in milliseconds)	Initial capture	Final capture	Gain (%)
Low density dataset: 112 facilities (68 initially opened), 13 divisions, 1002 demand points					
5	12	8	101488	111047	9.41
10		9		111481	9.84
20		14		111729	10.09
60		27		111864	10.22
High density dataset: 42 facilities (25 initially opened), 2 divisions, 3998 demand points					
5	61	8	2557276	2696519	5.44
10		8		2704182	5.74
15		13		2707438	5.87
20		14		2707438	5.87
Whole dataset: 1068 facilities (692 initially opened), 1285 divisions, 30346 demand points					
10	765	70	7553634	7757788	2.70
50		183		7948531	5.22
200		473		8090671	7.10
600		1200		8093069	7.14

Figure 6.20 illustrates the concreteness of our approach, as it shows the low density subset and the proposed solution through a visualisation interface.

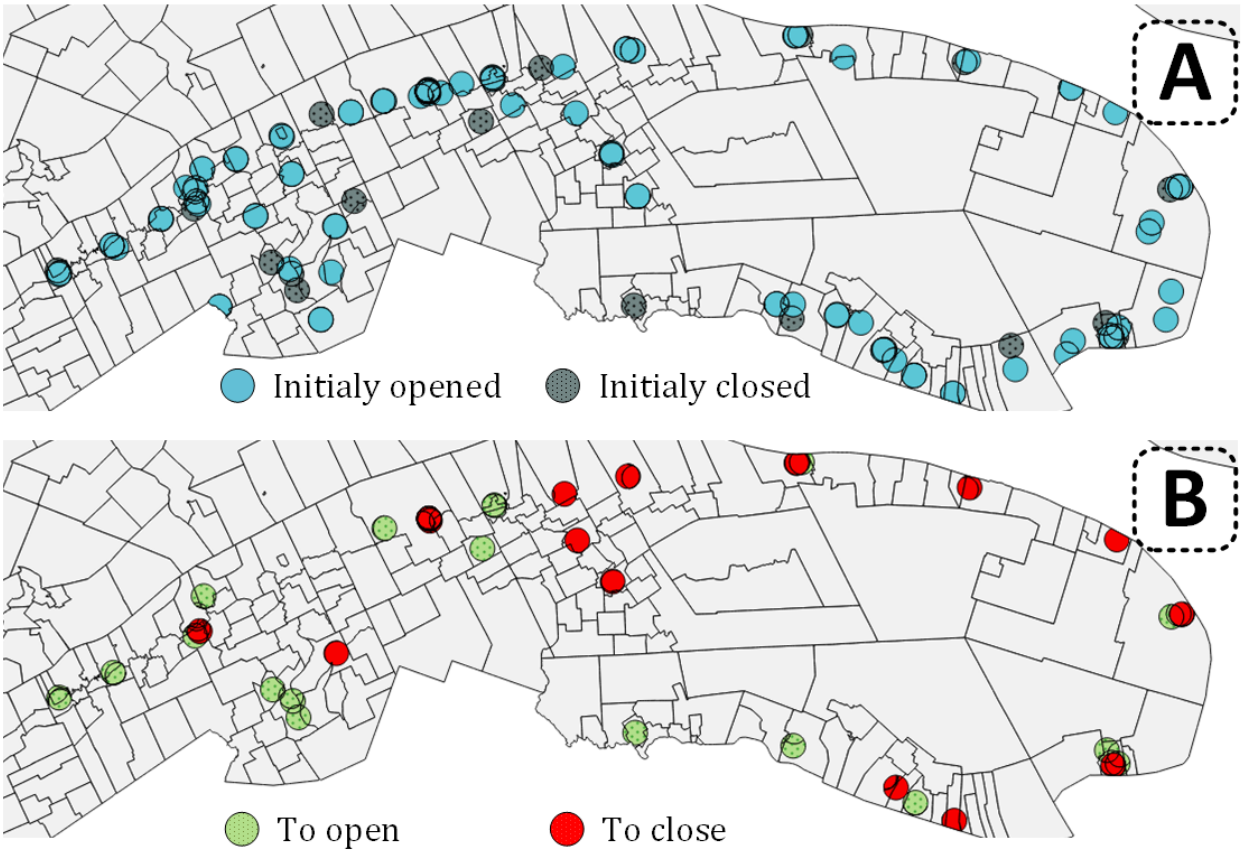


Figure 6.20 Initial network and proposed change

As can be seen in Table 6.1, the algorithm is processed in relatively short time. This is due to the developed data structure and the associated capture evaluation function. Those aspects are developed in the next section.

6.5.2 Data structures and Evaluation function

With the transformation proposed with $\varepsilon = 1$, the resulting capture rates are then, for all intersections I, for all facilities $i \in F_I$, and for all demand points $j \in P_I$,

$$\text{if } i + j \equiv 0 \mod |F_I| + 1 \text{ then } \alpha_{ij} = \frac{1}{|F_I|} \text{ else } \alpha_{ij} = \frac{1}{|F_I| + 1}$$

In this transformation, ε is defined in a way where, for each demand point, one and only one facility has a "best" capture rate, and the others all have the same "low" capture rates. Our approach takes advantage of that particularity. When a transformation respects it, the

overall capture of an intersection can be computed by only knowing the number of open and close related facilities (it is not necessary to know which ones are open and which ones are closed for a given intersection).

Indeed, as it is a maximization objective, in an intersection where there are k facilities opened for n demand points, there are k demand points that have their "best matching" facility opened. The $n-k$ remaining demand points will then be captured at their "low" capture rates. If $k=0$, the overall capture is null, otherwise, for an intersection with k facility opened on n, the resulting overall capture equals:

$$Capt(I) = \left(\frac{k}{n} + \frac{(n-k)}{n+\varepsilon} \right) \times demand(I)$$

To take advantage of this aspect, the following data structures are set up:
- (1) The intersection object that has the corresponding intersection ID I_{id}, its high and low capture rates hcr and lcr, the demand $Demand$, the total number of related facilities n, the number of related open facilities in the current state $nbOpen$, and its current overall capture $Capture$.
- (2) The $Intersections$ hash table that has intersection ID as keys, and the corresponding intersection object as value.
- (3) The $Facilities$ hash table which has facility IDs has keys, and values are list of intersections IDs related to the facility.

From those data structures, it is possible to have functions that allow for evaluating the consequences of a facility state change in the overall capture. A pseudo-code corresponding to the gap evaluation function for a given intersection is proposed in Algorithm 2. it takes in parameters the intersection id and, $Change$ that equals 1 for opening, and -1 for closing.

Algorithm 2 $IntersectionGapEvaluation(id, Change)$

$I = Intersection.\{id\}$
if $I.nbOpen + Change = 0$ **then**
  return $-I.Capture$
else
  $nbHigh = I.nbOpen + Change$
  $nbLow = I.n - (nbOpen + Change)$
  $CaptureRate = nbHigh \times I.hcr + nbLow \times I.lcr$
  $Gap = CaptureRate \times I.Demand - I.Capture$
  return Gap
end if

This function is then used in the algorithm 3 that allows for evaluating the overall capture gap for a facility opening or closing.

Algorithm 3 $evaluateGap\{F_{id}, OpenOrClose\}$

$Gap := 0$
for all Intersection ID $I_{id} \in Facilities(F_{id}).value$ **do**
  $Gap+ = IntersectionGapEvaluation\{I_{id}, OpenOrClose\}$
end for

6.6 Conclusion and perspectives

Realistic aspects such as trade areas and collaborative capture of demand are often not taken into account in MCLP Models. Our approach proposes a solution by combining modelling techniques, and spatial data pre-processing. Moreover, taking advantage of our proposed data transformation, a relatively fast resolution model is proposed to complete the whole process, from initial data to the real proposition for the improvement of capture.

In this paper, the model and the resolution algorithm are specific to the case study needs (i.e. respect a maximum number of openings), in further research, the model and the data structure could be extended to other problematic areas such as cost constraint.

As mentioned before, a VNS research has been developed but does not provide a good performance compared to the naïve approach on the case study instances. This does not mean that it would be the same on other datasets; thus, using VNS on other datasets to assess it for its efficiency could be of interest. Moreover, other approaches could be tested on the existing data structures.

Furthermore, the proposed data transformation with a generic ε value for all intersections might be over simplistic; this should be the interest of further research to allow for a more realistic solution.

Otherwise, the combination of data-preprocessing and modelling techniques could be applied to many problems that are location related. As the spatial component is more and more integrated in the data, and as GIS capabilities are constantly evolving, researchers in this area will have several opportunities to seize. Other covering problems, such as the Location Set Covering Problem could be addressed through the same approach.

As many problems were defined at a time when GIS capabilities were not available many realistic aspects have been introduced through modelling techniques. An interesting work

might be to review several location related problems and to identify which constraints could be replaced by a data-preprocessing approach.

CHAPITRE 7 DISCUSSION GÉNÉRALE

Les travaux de recherches contenus dans cette thèse se sont concentrés sur trois problématiques relatives à la conception et au développement d' un outil d'aide à la décision spatiale. Les différentes problématiques ont été traitées au sein d'une approche globale aspirant in fine à répondre au besoin d'un partenaire industriel. Ce partenariat a permis de travailler sur des données et des problématiques industrielles réelles. Pour chacune des problématiques scientifiques abordées, une contribution a pu être proposée, et intégrée au sein de l'approche globale pour permettre la conception et le développement d'un outil de traitement et d'analyse des données spatiales pour l'aide à la décision.

Comme on a pu le voir dans la revue de la littérature, les SDSS peuvent être utile pour de nombreuses applications, mais leur conception reste complexe. Pourtant, les recherches actuelles ne proposent que des réponses partielles au besoin d'un : d'une part des méthodologies génériques et d'autre part des applications spécifiques difficilement reproductibles. Pour répondre à cette problématique, et faciliter la conception de SDSS, la première contribution repose sur un cadre conceptuel composé d'une architecture SDSS et d'une méthodologie de développement. L'architecture repose sur deux catégories de clients (clients riches et clients légers). Des outils sont proposés pour les clients riches afin qu'ils puissent facilement manipuler les données spatiales et les analyser et/ou développer des applications pour les clients légers. La méthodologie de développement proposée permet de factoriser les différentes tâches du développement. Une étude de cas réelle est présentée avec des explications détaillées pour illustrer l'application du cadre conceptuel pour le développement d'un SDSS fonctionnel spécifique.

Par ailleurs, de nombreuses études dénoncent la complexité et l'aspect chronophage du prétraitement des données spatiales. Cependant, peu de recherche se penche sur la résolution de ces problématiques, et à notre connaissance, aucune recherche n'adresse la problématique de l'exigence de connaissance en Système d'Information Géographique. Pour répondre à ces problématiques, la deuxième contribution repose sur une approche générique pour automatiser en grande partie le processus de prétraitement des données spatiales. L'approche a pu être mise en place au travers d'un outil qui rend le prétraitement des données spatiales rapide et accessible à des utilisateurs n'ayant pas de compétences particulières en Systèmes d'Information Géographique (SIG). En conséquence, l'application d'algorithme d'analyse de données devient plus accessible avec des données spatiales. L'approche a pu être exploitée pour fournir des données prétraitées permettant de réaliser des analyses de données pour un partenaire

industriel. L'approche proposée a aussi été intégrée au sein d'un outil global est utilisé sur les données d'un partenaire industriel. Pour permettre la reproductibilité, les différentes étapes du processus à suivre sont décrites en détail. De plus, des outils open-source sont utilisés et mentionnés afin de permettre une reproduction de notre approche à coût réduit.

Enfin, relativement à l'application d'outils d'optimisations, peu de recherches récentes tirent profit des capacités des Système d'Information Géographique. Pour illustration, dans le cas du problème de maximisation de la couverture, des aspects tels que les zones de service et la captation collaborative de la demande sont rarement pris en compte. Pour palier cela, la troisième contribution de cette thèse propose une utilisation combinée de techniques de traitement des données et de modélisation pour ainsi permettre d'intégrer la captation collaborative et les zones de service dans le traitement du problème de maximisation de la couverture. De plus, en profitant de la transformation de données proposée, un modèle de résolution rapide a été conçu pour compléter l'ensemble du processus, depuis les données initiales jusqu'à la proposition réelle de solution pour l'amélioration de la couverture.

Les trois contributions sont issues d'une démarche cohérente. Cette démarche a permis de faire tomber les barrières à la mise en place d'un système d'aide à la décision spatiale qui avait été identifiées dans la littérature.

L'approche globale a permis la mise en commun des trois contributions au sein d'un outil de traitement et analyse des données spatiales, répondant ainsi à notre problématique générale de recherche.

Ainsi, pour le cas d'application de notre partenaire industriel, la mise en oeuvre des contributions a permis de développer un système d'aide à la décision spatiale. L'utilisation et l'adaptation des contributions proposées dans cette thèse pourront permettre une meilleure exploitation des données spatiales pour assister les processus décisionnels.

CHAPITRE 8 CONCLUSION ET RECOMMENDATIONS

Un besoin de solution d'assistance à la prise de décision spatiale a été identifié dans la revue de littérature. L'absence d'approche scientifique pour répondre à ce besoin est le point de départ des travaux de cette thèse qui se concentre sur la problématique générale de la conception et du développement d'un outil de traitement et analyse des données spatiales pour l'aide à la décision. Pour répondre à cette problématique, les travaux de recherche de cette thèse se sont penchés sur trois sous problématiques relatives à la mise en place d'un système d'aide à la décision spatial : (1) la proposition d'un cadre conceptuel de conception et développement, (2) le prétraitement des données spatiales, et (3) la proposition d'une approche de modélisation pour le problème de maximisation de la couverture qui intègre des aspects spatiaux réalistes. Des solutions ont été proposées au travers des trois contributions présentées dans cette thèse. Les contributions, leurs limites, ainsi que les perspectives de recherches associées sont présentées dans cette section.

8.1 Les contributions et leurs limites

8.1.1 Traitement des trois problématiques: contributions et limites

La première contribution repose sur la proposition d'un cadre conceptuel composé d'une architecture et d'une méthodologie de développement. L'architecture propose des outils pour manipuler les données spatiales, pour pouvoir procéder à leur analyse et à la visualisation des données et des résultats d'analyse. La méthodologie de développement proposée permet de factoriser différentes tâches dans le cas où plusieurs applications dédiées doivent être développées pour un même projet. Le cadre conceptuel a été validé sur une étude de cas réelle, et une explication détaillée a été présenté pour illustrer le développement d'un SDSS fonctionnel.

L'approche proposée à des limites, l'architecture proposée ne tient pas compte des contraintes qui peuvent apparaitre dans des cas de développement académique ou industriel : - Passage à l'échelle, l'approche architecturale proposée de tient pas compte de contraintes de ressources en cas d'utilisation à grand échelle (nombre d'utilisateurs, quantité de données, etc.) - Les aspects liés à la sécurisation des données ne sont pas abordés, certains secteurs scientifiques ou industriels pourraient avoir à assurer la confidentialité et l'intégrité des données. - Il peut exister des contraintes d'utilisation d'un logiciel propriétaire pour la réalisation des analyses, les interfaces avec le système proposé seraient à mettre en place.

La deuxième contribution propose une approche qui automatise la majeure partie du processus de prétraitement des données spatiales. L'outil développé met l'analyse des données spatiales à la disposition des utilisateurs qui n'ont pas de connaissances approfondies en SIG. En plus de cela, la réalisation des prétraitement peuvent être effectuées plus rapidement et avec des choix guidés pour la sélection des relations spatiales d'intérêt.

Cette approche a pour l'instant encore des limites qui peuvent être explorées: - L'implémentation ne fonctionne qu'avec une couche de polygones comme cible et avec une couche de points comme source. Il pourrait être utile d'autoriser d'autres types de données pour la couche cible (comme les points ou les lignes) et pour la couche source (par exemple, des polygones ou des lignes). - D'autres relations spatiales pourraient être mises en œuvre et il pourrait être intéressant de permettre que les relations spatiales soient pondérées par les caractéristiques numériques des éléments de la couche source. - La méthode pour permettre de guider le choix de la sélection pourrait être l'objet de recherche dédiée de manière à améliorer sa pertinence.

La troisième contribution propose une approche pour intégrer des aspects réalistes dans la modélisation du problème de maximisation de la couverture. La captation collaborative de la demande, et la prise en compte des zones de services sont rendus possibles par la combinaison de prétraitements des données et de techniques de modélisation.

Les limites de notre approche résident en partie dans sa spécificité, le modèle et l'algorithme de résolution sont spécifiques aux besoins de l'étude de cas (c'est-à-dire respecter un nombre maximal d'ouvertures). Dans le cadre de recherches supplémentaires, le modèle et la structure de données pourraient être étendus à d'autres domaines problématiques tels que la contrainte de coûts. Par ailleurs, un algorithme de recherche par voisinage a été développé mais ne fournit pas une bonne performance par rapport à l'approche naïve sur d'étude de cas. Cette approche par voisinage pourrait être évaluée sur d'autres ensembles de données

8.1.2 Limite de l'approche globale sur un cas industriel

L'outil a été validé sur un cas d'application industriel, cette approche permet des avantages comme l'accès à des données réelles et la confrontation à des problématiques concrètes. Cependant cette approche a aussi des limites, en effet en se restreignant à un seul cas, il est possible que certaines problématiques n'aient pas été rencontrées. Par exemple, relativement à la mise en commun des différents outils entre eux, notre approche de prend pas en compte des limitations industrielles qui pourraient être rencontrées comme des restrictions architecturales ou logicielles. Dans le cas où il serait envisagé que l'outil soit intégré dans un système d'information existant, la problématique de la compatibilité de nos solutions pourrait se

présenter.

Relativement au type d'aide à la décision proposée, l'approche globale se restreint aussi au cas particulier du partenaire industriel, lié à l'amélioration de la couverture de la demande. De nombreuses autres problématiques industrielles peuvent être approchées par l'optimisation, et les modèles associés pourraient impliquer d'autres contraintes relativement au type de données à récupérer, et à prendre en compte.

Le cas d'application industriel ne nécessitant pas la prise en compte d'une quantité de données importante (i.e. qui nécessiterait un stockage distribué) ; les problématiques liées à la gestion de données distribuées ne sont pas abordées. Relativement à la complexité des modèles d'analyse et d'optimisation, notre approche ne prend pas en compte des capacités de calcul distribuées qui pourraient permettre d'améliorer les performances.

8.2 Perspectives

8.2.1 Perspectives de recherches relatives aux contributions

Par rapport au cadre conceptuel, plusieurs perspectives de recherche sont envisageables relativement aux limites de la solution proposée. Ainsi des recherches pour intégrer des aspects comme le passage à l'échelle, la sécurité et l'intégrité des données, où des contraintes logicielles pourraient être réalisées. Par ailleurs, il pourrait être envisagé de comparer l'efficacité de l'approche proposée dans d'autres cas d'application. Enfin, permettre la prise en compte de flux de données en temps réels, et de leur analyse pour la prise de décision automatique, soulèverait de nombreuses problématiques tant architecturales que méthodologiques qui seront à traiter dans le futur des SDSS.

Relativement à l'approche pour le prétraitement des données spatiales, les recherches peuvent aussi se porter sur les limites de notre solution. Une étude des relations spatiales à intégrer serait pertinente, et la prise en compte de différent type de données pourrait être intéressante. L'étude de la pertinence des différentes fonctions de corrélation serait aussi intéressante. Un autre aspect qui pourrait être abordé dans l'approche est relatif à la prise en compte des couples de couches de données spatiales (par exemple l'étude de l'influence des éléments de la couche A qui respecte une condition par rapport aux éléments de la couche B).

Pour la troisième contribution, les paramètres du modèle devraient être l'objet de recherches supplémentaires pour permettre des solutions qui tiennent mieux compte des réalités du terrain, par exemple en tenant compte de l'attractivité des distributeurs. De plus la combinaison des techniques de prétraitement et de modélisation des données pourrait être appliquée à de nombreux problèmes liés à l'emplacement. Comme beaucoup de problèmes ont été définis

à un moment où les capacités SIG n'étaient pas disponibles, de nombreux aspects ont été introduits grâce à des techniques de modélisation. Un travail intéressant pourrait être d'examiner plusieurs problèmes liés à l'emplacement et d'identifier quelles contraintes pourraient être remplacées par une approche de prétraitement des données.

8.2.2 Perspective pour l'outil de traitement et analyse de données

Relativement aux limites présentées, des pistes de recherches sont envisageables dans un premier temps, comme l'application à d'autres secteurs industriels, ou à d'autres problèmes d'analyse de données ou d'optimisation.

Les différentes contributions prises ensemble permettent de traiter un problème dans son intégralité, mais des manipulations des données sont encore nécessaires entre les différentes étapes. L'automatisation de la chaine complète de traitement est un sujet qui mérite d'être étudié, de manière à simplifier le processus pour qu'il nécessite moins de compétence technique et puisse être complété plus rapidement.

Enfin à plus long terme, les décisions prises réellement par les décideurs devraient pouvoir être tracées et évaluées, de manière à autoriser une auto-évaluation de l'outil: les décisions proposées ont-elles apporté les résultats attendus ? Il peut être envisagé que les outils se calibrent en fonction de ces résultats de manière à améliorer leur efficacité dans l'assistance à la prise de décision.

LISTE DES RÉFÉRENCES

[1] A. Adejuwon et A. Mosavi, “Domain driven data mining – application to business”, International journal of computer science, vol. 7, no. 4, pp. 41–44, 2010.

[2] S. Agrawal et R. D. Gupta, “Development and comparison of open source based web gis frameworks on wamp and apache tomcat web servers”, ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XL-4, pp. 1–5, 2014.

[3] H. Alatrista Salas, S. Bringay, F. Flouvat, N. Selmaoui-Folcher, et M. Teisseire, “A spatial-based kdd process to better understand the spatiotemporal phenomena.” 2013.

[4] Y. Alotaibi, “Business process modelling challenges and solutions: a literature review”, Journal of Intelligent Manufacturing, vol. 27, no. 4, pp. 701–723, 2014.

[5] N. Andrienko, G. Andrienko, A. Savinov, H. Voss, et D. Wettschereck, “Exploratory analysis of spatial data using interactive maps and data mining”, Cartography and Geographic Information Science, vol. 28, no. 3, pp. 151–166, 2001.

[6] L. Anselin, I. Syabri, et Y. Kho, “Geoda: An introduction to spatial data analysis”, Geographical Analysis, vol. 38, no. 1, pp. 5–22, 2006.

[7] A. Appice, M. Ceci, A. Lanza, et F. A. Lisi, “Discovery of spatial association rules in geo-referenced census data: A relational mining approach”, Intelligent Data Analysis, vol. 7, no. 6, pp. 541–566, 2003.

[8] T. Arentze, A. Borgers, et H. Timmermans, “A knowledge-based system for developing retail location strategies”, Computers, Environment and Urban Systems, vol. 24, no. 6, pp. 489–508, 2000.

[9] M. Armstrong, S. De, P. J. Densham, P. Lolonis, G. Rushton, et V. Tewari, “A knowledge-based approach for supporting locational decisionmaking”, Environment and Planning B: Planning and Design, vol. 17, no. 3, pp. 341–364, 1990.

[10] Z. Babić et T. Perić, “Optimization of livestock feed blend by use of goal programming”, International Journal of Production Economics, vol. 130, no. 2, pp. 218–223, 2011.

[11] X. Bai, W. Shang, W. Yin, et J. Dong, “A service-oriented solution for retail store network planning”, pp. 482–489, 2007.

[12] D. Benoit et G. P. Clarke, “Assessing gis for retail location planning”, Journal of Retailing and Consumer Services, vol. 4, no. 4, pp. 239–258, 1997.

[13] O. Berman, Z. Drezner, et D. Krass, “Continuous covering and cooperative covering problems with a general decay function on networks”, Journal of the Operational Research Society, vol. 64, no. 11, pp. 1644–1653, 2012.

[14] O. Berman et D. Krass, “The generalized maximal covering location problem”, Computers & Operations Research, vol. 29, no. 6, pp. 563–581, 2002.

[15] O. Berman, Z. Drezner, et D. Krass, “Generalized coverage: New developments in covering location models”, Computers & Operations Research, vol. 37, no. 10, pp. 1675–1687, 2010.

[16] R. Blanquero, E. Carrizosa, et B. G.-Tóth, “Maximal covering location problems on networks with regional demand”, Omega, vol. 64, pp. 77–85, 2016.

[17] V. Bogorny, P. Martins Engel, et L. O. Alvares, “A reuse-based spatial data preparation framework for data mining.” dans Proceedings of the 17th International Conference on Software Engineering and Knowledge Engineering (SEKE’2005), 2005, Electronic Article, pp. 649–652.

[18] ——, “Spatial data preparation for knowledge discovery. in: I ifip academy on the state pf software theory and practice - phd colloquim. porto alegre, brazil (2005)”, 2005, Conference Proceedings.

[19] Bootstrap, (accessed 13 July 2016). En ligne: `http://getbootstrap.com/`

[20] BoundlessGeo, (accessed 13 July 2016). En ligne: `http://boundlessgeo.com/`

[21] O. Bouzaâbia et S. Boumaiza, “Le rôle de la performance logistique dans la satisfaction des consommateurs: Investigation dans la grande distribution”, La Revue Gestion et Organisation, vol. 5, no. 2, pp. 121–129, 2013.

[22] E. T. Bradlow, M. Gangwar, P. Kopalle, et S. Voleti, “The role of big data and predictive analytics in retailing”, Journal of Retailing, vol. 93, no. 1, pp. 79–95, 2017.

[23] Caret, (accessed 19 July 2017). En ligne: https://cran.r-project.org/web/packages/caret/index.html

[24] Chart.js, (accessed 13 July 2016). En ligne: http://www.chartjs.org/

[25] J. Chen, A. M. Maceachren, et D. Guo, "Supporting the process of exploring and interpreting space-time multivariate patterns: The visual inquiry toolkit", Cartography and Geographic Information Science, vol. 35, no. 1, pp. 33–50, 2008.

[26] R. Church et C. ReVelle, "The maximal covering location problem", Papers in Regional Science, vol. 32, no. 1, pp. 101–118, 1974.

[27] K. Cios, W. Pedrycz, R. Swiniarski, et L. Kurgan, The knowledge discovery process, in: Data Mining A knowledge discovery approach. Springer, 2007, pp. 9–24.

[28] E. Clementini, P. Felice, et P. Oosterom, "A small set of formal topological relationships suitable for end-user interaction." dans Third International Symposium on Advances in Spatial Databases - SSD'93, vol. 692, 1993, Conference Paper, pp. 277–295.

[29] G. Cliquet, A. Fady, et G. Basset, Management de la distribution, 2e éd. Dunod, 2006.

[30] V. Clulow et V. Reimers, "How do consumers define retail centre convenience ?" Australasian Marketing Journal (AMJ), vol. 17, no. 3, pp. 125–132, 2009.

[31] CMHC, (accessed 13 July 2016). En ligne: https://www.cmhc-schl.gc.ca/

[32] M. Crossland, B. Wynne, et W. Perkins, "Spatial decision support systems: An overview of technology and a test of efficacy", Decision Support Systems, vol. 14, no. 3, pp. 219–235, 1995.

[33] C. Cui, J. Wang, Y. Pu, J. Ma, et G. Chen, "Gis-based method of delimitating trade area for retail chains", International Journal of Geographical Information Science, vol. 26, no. 10, pp. 1863–1879, 2012.

[34] J. R. Current et D. A. Schilling, "Analysis of errors due to demand data aggregation in the set covering and maximal covering location problems", Geographical Analysis, vol. 22, no. 2, pp. 116–126, 1990.

[35] G. Daras, B. Agard, H. Cambazard, et B. Penz, “Development of business spatial analysis tools: methodology and framework”, dans 2015 IFAC Symposium on Information Control in Manufacturing – INCOM 2015. Ottawa (Ontario), Canada., May 11-13, 2015 2015, Conference Paper.

[36] G. Daras, B. Agard, et B. Penz, “Conceptual framework for sdss development: Application in the retail industry”, Submitted., 2016.

[37] ——, “A spatial data pre-processing tool to improve the quality of the analysis and to reduce preparation duration”, Submitted., 2017.

[38] M. De Beule, D. Van den Poel, et N. Van de Weghe, “Assessing the principles of spatial competition between stores within a retail network”, Applied Geography, vol. 62, pp. 125–135, 2015.

[39] P. Densham, “Spatial decisions support systems.” dans Geographical Information Systems: Principles and Applications, Wiley, éd. Wiley, 1991, Book Section, pp. 403–412.

[40] C. Dubelaar, M. Bhargava, et D. Ferrarin, “Measuring retail productivity: what really matters ?” Journal of Business Research, vol. 55, no. 5, pp. 417–426, 2002.

[41] M. J. Egenhofer et J. Herring, “Categorizing binary topological relations between regions, lines, and points in geographic databases.” dans Third International Conference on Principles of Knowledge, Representation and Reasoning - KR 1992, 1992, Conference Proceedings.

[42] M. Erskine, D. Gregg, J. Karimi, et J. Scott, “Business decision-making using geospatial data: A research framework and literature review”, Axioms, vol. 3, no. 1, pp. 10–30, 2013.

[43] M. Ester, H.-P. Kriegel, et J. Sander, Knowledge Discovery in Spatial Databases. Springer Berlin Heidelberg, 1999, vol. 1701, pp. 61–74.

[44] B. Evans et C. E. Sabel, “Open-source web-based geographical information system for health exposure assessment”, International Journal of Health Geographics, vol. 11, pp. 1–11, 2012.

[45] K. Fang, Y. Jiang, et M. Song, “Customer profitability forecasting using big data analytics: A case study of the insurance industry”, Computers & Industrial Engineering, vol. 101, pp. 554–564, 2016.

[46] R. Z. Farahani, N. Asgari, N. Heidari, M. Hosseininia, et M. Goh, “Covering problems in facility location: A review”, Computers & Industrial Engineering, vol. 62, no. 1, pp. 368–407, 2012.

[47] U. Fayyad, G. Piatetsky-Shapiro, et P. Smyth, “From data mining to knowledge discovery in databases”, AI Magazine, vol. 17, no. 3, pp. 37–54, 1996.

[48] V. Ferretti et G. Montibeller, “Key challenges and meta-choices in designing and applying multi-criteria spatial decision support systems”, Decision Support Systems, vol. 84, pp. 41–52, 2016.

[49] R. Flowerdrew, “Spatial data integration”, Geographical information systems, vol. 1, pp. 375–387, 1991.

[50] GeoExplorer, (accessed 13 July 2016). En ligne: `http://suite.opengeo.org/opengeo-docs/geoexplorer/`

[51] GeoServer, (accessed 13 July 2016). En ligne: `http://geoserver.org/`

[52] P. Ghaemi, J. Swift, C. Sister, J. P. Wilson, et J. Wolch, “Design and implementation of a web-based platform to support interactive environmental planning”, Computers, Environment and Urban Systems, vol. 33, no. 6, pp. 482–491, 2009.

[53] K. Gibert, J. Izquierdo, G. Holmes, I. Athanasiadis, J. Comas, et Sanchez-Marre, “On the role of pre and post-processing in environmental data mining.” dans The iEMSs: International Congress on Environmental Modeling and Software Integrating Sciences and Information Technology for Environmental Assessment and Decision Making. Vol. 3, 2008, pp. pp. 1937–1958.

[54] Golf-clubs, (accessed 13 July 2016). En ligne: `http://www.legolfquebecois.com/`

[55] M. F. Goodchild, “Ilacs: A location-allocation model for retail site selection”, Journal of retailing, vol. 60, no. 1, pp. 84–100, 1984.

[56] Google, “Google maps geocoding api”, (accessed 13 July 2016). En ligne: `https://developers.google.com/maps/documentation/geocoding/start`

[57] ——, “Google maps api”, (accessed 13 July 2016). En ligne: `https://developers.google.com/maps/`

[58] J. Grabis, C. Chandra, et J. Kampars, “Use of distributed data sources in facility location”, Computers & Industrial Engineering, vol. 63, no. 4, pp. 855–863, 2012.

[59] C. Granell, L. Díaz, et M. Gould, “Service-oriented applications for environmental models: Reusable geospatial services”, Environmental Modelling & Software, vol. 25, no. 2, pp. 182–198, 2010.

[60] J. Han, M. Kamber, et A. K. H. Tung, “Spatial clustering methods in data mining: A survey”, 2001. En ligne: `http://www-faculty.cs.uiuc.edu/~hanj/pdf/gkdbk01.pdf`

[61] T. Hernandez, “Enhancing retail location decision support: The development and application of geovisualization”, Journal of Retailing and Consumer Services, vol. 14, no. 4, pp. 249–258, 2007.

[62] T. Hernández et D. Bennison, “The art and science of retail location decisions”, International Journal of Retail & Distribution Management, vol. 28, no. 8, pp. 357–367, 2000.

[63] B. E. Herring et J. G. DeBinder, “A computer-aided modeling system for geobased data”, Computers & Industrial Engineering, vol. 5, no. 3, pp. 141–152, 1981.

[64] R. Hess, R. Rubin, et L. West, “Geographic information systems as a marketing information system technology”, Decision Support Systems, vol. 38, no. 2, pp. 197–212, 2004.

[65] M. Ismail El-Adly, “Shopping malls attractiveness: a segmentation approach”, International Journal of Retail & Distribution Management, vol. 35, no. 11, pp. 936–950, 2007.

[66] R. Jain, A. R. Singh, H. C. Yadav, et P. K. Mishra, “Using data mining synergies for evaluating criteria at pre-qualification stage of supplier selection”, Journal of Intelligent Manufacturing, vol. 25, no. 1, pp. 165–175, 2012.

[67] S. Jarupathirun et F. M. Zahedi, “Exploring the influence of perceptual factors in the success of web-based spatial dss”, Decision Support Systems, vol. 43, no. 3, pp. 933–951, 2007.

[68] X. Jin, “A supply chain optimization dss web-services-based for e-retail industry.” dans Power Engineering and Automation Conference (PEAM), 2011 IEEE. Wuhan: IEEE, 2011, Book Section, pp. 229–232.

[69] Jsoup, (accessed 13 July 2016). En ligne: `https://jsoup.org/`

[70] P. Keenan, "Spatial decision support systems: A coming of age", Control and Cybernetics, vol. 35, pp. 9–27, 2006.

[71] ——, Using a GIS as a DSS generator. Hyderabad: ICFAI University Press, 2004, pp. 97–113.

[72] M. Khan et S. S. Khan, "Data and information visualization methods, and interactive mechanisms: a survey", International Journal of Computer Applications, vol. 34, pp. 1–14, 2011.

[73] S. Knezic et N. Mladineo, "Gis-based dss for priority setting in humanitarian mine-action", International Journal of Geographical Information Science, vol. 20, no. 5, pp. 565–588, 2006.

[74] H. Konishi, "Concentration of competing retail stores", Journal of Urban Economics, vol. 58, no. 3, pp. 488–512, 2005.

[75] K. Koperski et J. Han, "Discovery of spatial association rules in geographic information databases", Advances in Spatial Databases, vol. 951, pp. 47–66, 1995.

[76] Y. Li et L. Liu, "Assessing the impact of retail location on store performance: A comparison of wal-mart and kmart stores in cincinnati", Applied Geography, vol. 32, no. 2, pp. 591–600, 2012.

[77] R. Lloyd, Spatial Cognition: Geographic Environments, série The GeoJournal Library. Springer Netherlands, 1997, vol. 39.

[78] M. B. B. Loranca, R. G. Velázquez, M. E. Analco, M. B. Díaz, G. M. Guzman, et A. S. López, "Experiment design for the location-allocation problem", Applied Mathematics, vol. 05, no. 14, pp. 2168–2183, 2014.

[79] D. P. Loucks, "Developing and implementing decision support systems: A critique and a challenge", Journal of the American Water Resources Association, vol. 31, no. 4, pp. 571–582, 1995.

[80] A. M. MacEachren et M.-J. Kraak, "Research challenges in geovisualization", Cartography and Geographic Information Science, vol. 28, no. 1, pp. 3–12, 2001.

[81] T. C. Matisziw et A. T. Murray, “Siting a facility in continuous space to maximize coverage of a region”, Socio-Economic Planning Sciences, vol. 43, no. 2, pp. 131–139, 2009.

[82] A. Mendes et I. Themido, “Multi-outlet retail site location assessment”, International Transactions in Operational Research, vol. 11, no. 1, pp. 1–18, 2004.

[83] J. Mennis et D. Guo, “Spatial data mining and geographic knowledge discovery—an introduction”, Computers, Environment and Urban Systems, vol. 33, no. 6, pp. 403–408, 2009.

[84] H. J. Miller, Geographic Data Mining and Knowledge Discovery. Blackwell, 2007, pp. 352–366.

[85] M. Y. Mohamad, “A gis application for location selection and customers’ preferences for shopping malls in al ain city ; uae”, American Journal of Geographic Information System, 2015.

[86] R. Moreno-Sanchez, G. Anderson, J. Cruz, et M. Hayden, “The potential for the use of open source software and open specifications in creating web-based cross-border health spatial information systems”, International Journal of Geographical Information Science, vol. 21, no. 10, pp. 1135–1163, 2007.

[87] A. T. Murray, D. Tong, et K. Kim, “Enhancing classic coverage location models”, International Regional Science Review, vol. 33, no. 2, pp. 115–133, 2009.

[88] A. T. Murray, “Advances in location modeling: GIS linkages and contributions”, Journal of Geographical Systems, vol. 12, no. 3, pp. 335–354, 2010.

[89] A. T. Murray, M. E. O’Kelly, et R. L. Church, “Regional service coverage modeling”, Computers & Operations Research, vol. 35, no. 2, pp. 339–355, 2008.

[90] K. I. Nikolopoulos, M. Z. Babai, et K. Bozos, “Forecasting supply chain sporadic demand with nearest neighbor approaches”, International Journal of Production Economics, vol. 177, pp. 139–148, 2016.

[91] OpenLayers, (accessed 13 July 2016). En ligne: `http://openlayers.org/`

[92] B. Otto, “Data governance”, Business & Information Systems Engineering, vol. 3, no. 4, pp. 241–244, 2011.

[93] Y. Pan et G. M. Zinkhan, “Determinants of retail patronage: A meta-analytical perspective”, Journal of Retailing, vol. 82, no. 3, pp. 229–243, 2006.

[94] A. A. Pereira Klen, R. J. Rabelo, A. C. Ferreira, et L. M. Spinosa, “Managing distributed business processes in the virtual enterprise”, Journal of Intelligent Manufacturing, vol. 12, no. 2, pp. 185–197, 2001.

[95] PgAdmin, (accessed 13 July 2016). En ligne: `http://www.pgadmin.org/`

[96] F. Plastria et E. Carrizosa, “Optimal location and design of a competitive facility”, Mathematical Programming, vol. 100, no. 2, pp. 247–265, 2004.

[97] F. Plastria et L. Vanhaverbeke, “Discrete models for competitive location with foresight”, Computers & Operations Research, vol. 35, no. 3, pp. 683–700, 2008.

[98] PostGIS, (accessed 13 July 2016). En ligne: `http://postgis.net/`

[99] PostgreSQL, (accessed 13 July 2016). En ligne: `https://www.postgresql.org/`

[100] J. Pretorius et M. Matthee, “The impact of spatial data on the knowledge discovery process.” dans Proceedings of the Conference on Information Technology in Tertiary Education, Pretoria, South Africa, 18-20 September 2006, Conference Paper.

[101] C. Previl, M. Thériault, et J. Rouffignat, “Combining multicriteria analysis and gis to help decision making processes in portneuf county (quebec, canada).” dans Proceedings of 2nd Annual URISA Public Participation GIS Conference. URISA Summer Conference, Portland Oregon, July 20-22 2003, Conference Paper, pp. 529–554.

[102] QGIS, (accessed 13 July 2016). En ligne: `http://www.qgis.org/en/site/`

[103] R, (accessed 13 July 2016). En ligne: `http://www.r-project.org/`

[104] W. Reinartz, B. Dellaert, M. Krafft, V. Kumar, et R. Varadarajan, “Retailing innovations in a globalizing retail market environment”, Journal of Retailing, vol. 87, pp. S53–S66, 2011.

[105] C. Ren, W. Wang, et H. Luo, “Research on the gis-based business decision system”, dans 2010 18th International Conference on Geoinformatics, 2010, Journal Article, pp. 1–6.

[106] S. Ren, X. D. Zhang, et X. P. Zhang, “A new generation of decision support systems for advanced manufacturing enterprises”, Journal of Intelligent Manufacturing, vol. 8, no. 5, pp. 335–343, 1997.

[107] C. S. ReVelle et H. A. Eiselt, “Location analysis: A synthesis and survey”, European Journal of Operational Research, vol. 165, no. 1, pp. 1–19, 2005.

[108] S. Rey, “Show me the code - spatial analysis and open source”, Journal of Geographical Systems, vol. 11, no. 2, pp. 191–207, 2009.

[109] C. Rinner, C. Keßler, et S. Andrulis, “The use of web 2.0 concepts to support deliberation in spatial decision-making”, Computers, Environment and Urban Systems, vol. 32, no. 5, pp. 386–395, 2008.

[110] N. Roig-Tierno, A. Baviera-Puig, J. Buitrago-Vera, et F. Mas-Verdu, “The retail site location decision process using gis and the analytical hierarchy process”, Applied Geography, vol. 40, pp. 191–198, 2013.

[111] F. A. S., P. J. Densham, et A. Curtis, “The zone definition problem in location allocation modeling”, Geographical analysis, vol. 27, no. 1, 1995.

[112] C. G. Schmidt, “Location decision-making within a retail corporation”, Regional Science perspectives, vol. 13, pp. 60–71, 1983.

[113] S. Shekhar, P. Zhang, Y. Huang, et R. R. Vatsavai, “Trends in spatial data mining”, dans Data Mining: Next Generation Challenges and Future Directions. AAAI/MIT Press, 2003.

[114] I. U. Sikder, “Knowledge-based spatial decision support systems: An assessment of environmental adaptability of crops”, Expert Systems with Applications, vol. 36, no. 3, pp. 5341–5347, 2009.

[115] Ski-resort, (accessed 13 July 2016). En ligne: `http://www.maneige.com/`

[116] J. Smelcer et E. Carmel, “The effectiveness of different representations for managerial problem solving: Comparing tables and maps”, Decision Sciences, vol. 28, no. 2, pp. 391–420, 1997.

[117] R. H. Sprague, “A framework for the development of decision support systems”, MIS Quarterly, vol. 4, no. 4, pp. 1–26, 1980.

[118] Statistics-canada, (accessed 13 July 2016). En ligne: https://www12.statcan.gc.ca/

[119] R. Studio, "R studio", (accessed 24 October 2016). En ligne: https://www.rstudio.com/

[120] Q. Su, Q. Luo, et S. H. Huang, "Cost-effective analyses for emergency medical services deployment: A case study in shanghai", International Journal of Production Economics, vol. 163, pp. 112–123, 2015.

[121] V. Sugumaran, "Web-based spatial decision support systems (websdss): Evolution, architecture, examples and challenges", Communications of the Association for Information Systems, vol. 19, pp. 844–875, 2007.

[122] R. Suárez-Vega et D. R. Santos-Peñate, "The use of gis tools to support decision-making in the expansion of chain stores", International Journal of Geographical Information Science, vol. 28, no. 3, pp. 553–569, 2013.

[123] R. Suárez-Vega, D. R. Santos-Peñate, et P. Dorta-González, "Location models and gis tools for retail site location", Applied Geography, vol. 35, no. 1-2, pp. 12–22, 2012.

[124] C. Teller et T. Reutterer, "The evolving concept of retail attractiveness: What makes retail agglomerations attractive when customers shop at them ?" Journal of Retailing and Consumer Services, vol. 15, no. 3, pp. 127–143, 2008.

[125] A. Thompson et J. Walker, "Retail network planning — achieving competitive advantage through geographical analysis", Journal of Targeting, Measurement and Analysis for Marketing, vol. 13, no. 3, pp. 250–257, 2005.

[126] Tomcat, (accessed 13 July 2016). En ligne: http://tomcat.apache.org/

[127] B. Vahedi, W. Kuhn, et A. Ballatore, "Question-based spatial computing—a case study." dans Geospatial Data in a Changing World. Lecture Notes in Geoinformation and Cartography 1., B. Springer, éd., 2016, Book Section, pp. 37–50.

[128] R. R. Vatsavai, S. Shekhar, T. E. Burk, et S. Lime, "Umn-mapserver: A high-performance, interoperable, and open source web mapping and geo-spatial analysis system." dans Geographic Information Science: 4th International Conference, GIScience 2006, Munster, Germany, September 20-23, M. Raubal, H. J. Miller, A. U. Frank, et M. F. Goodchild, éds., vol. 4197. Springer Berlin Heidelberg, 2006, Book Section, pp. 400–417.

[129] M. Vlachopoulou, G. Silleos, et V. Manthou, “Geographic information systems in warehouse site selection decisions”, International Journal of Production Economics, vol. 71, no. 1-3, pp. 205–212, 2001.

[130] T. Wanderer et S. Herle, “Creating a spatial multi-criteria decision support system for energy related integrated environmental impact assessment”, Environmental Impact Assessment Review, vol. 52, pp. 2–8, 2015.

[131] F. Wang, C. Chen, C. Xiu, et P. Zhang, “Location analysis of retail stores in changchun, china: A street centrality perspective”, Cities, vol. 41, pp. 54–63, 2014.

[132] L. A. West, “Designing end-user geographic information systems”, Journal of Organizational and End User Computing, vol. 12, no. 3, pp. 14–22, 2000.

[133] Yeoman, (accessed 13 July 2016). En ligne: `http://yeoman.io/`

[134] C. Zhang, T. Zhao, et W. Li, “The framework of a geospatial semantic web-based spatial decision support system for digital earth”, International Journal of Digital Earth, vol. 3, no. 2, pp. 111–134, 2010.

[135] R. Y. Zhong, S. T. Newman, G. Q. Huang, et S. Lan, “Big data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives”, Computers & Industrial Engineering, vol. 101, pp. 572–591, 2016.

[136] X. Zhu, R. Healey, et R. Aspinall, “A knowledge-based systems approach to design of spatial decision support systems for environmental management”, Environmental Management, vol. 22, no. 1, pp. 35–48, 1998.