Automatisation du processus de prétraitement des données spatiales

GAUTIER DARAS[1, 2, 4, 5], BRUNO AGARD[1, 3], BERNARD PENZ[4, 6]

[1] FORAC, CIRRELT, Département de Mathématiques et Génie Industriel, École Polytechnique de Montréal, Montréal (QC), H3T 1J4, Canada

[2] gautier.daras@polymtl.ca
[3] bruno.agard@polymtl.ca

[4] Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP, F-38000 Grenoble, France

[5] gautier.daras@grenoble-inp.fr
[6] bernard.penz@grenoble-inp.fr

***Résumé* – L'analyse des données géospatiales peut permettre de mieux comprendre les effets de l'environnement sur les performances d'une activité. Avant de pouvoir procéder à une telle analyse, les données relatives aux éléments géospatialisés susceptibles d'influencer l'activité doivent être rassemblées. Il faut ensuite effectuer différents prétraitements afin de pouvoir utiliser ces données au travers des outils classiques d'analyses de données. Ces prétraitements sont aujourd'hui des tâches qui en plus de demander des compétences techniques complexes, sont très gourmandes en temps et peuvent introduire des aléas dans les analyses qui suivront. Dans le but de remédier à ces problèmes, nous proposons une approche qui permet d'automatiser en grande partie le processus de prétraitement des données spatiales. En plus de présenter les spécifications de notre approche, les outils open source permettant son implémentation sont mentionnés. Enfin, l'utilisation de l'outil est présentée au travers d'un bref cas d'étude qui montrera son efficacité en mettant en parallèle les étapes nécessaires au prétraitement des données sans l'utilisation de notre outil.**

***Abstract* – Spatial data analysis might allow a better understanding of environmental effect on the performances of an organization's activities. One of the first steps required to process such analysis is to gather all the spatial data corresponding to the elements that might influence the activities. Then a series of treatment must be processed on those datasets to make them ready for the use of classical data mining tools. Those pre-processing steps are complex tasks that in addition to require advanced GIS skills are time consuming. Moreover, the choices involved in this process might influence further analysis results. With the aim of addressing those issues, our approach automatizes several steps of the spatial data pre-processing. In addition to our approach specifications, tools to implement it will be mentioned. To support the effectiveness of our approach, a brief case study will focus on the steps required with and without our solution.**

***Mots clés* - ECD, data mining, intégration des données, données spatiales, SIG.**

***Keywords* - KDD, Datamining, Data integration, spatial data, GIS**

1 INTRODUCTION

La compréhension des effets de l'environnement sur les performances d'une activité est un réel avantage pour de nombreuses organisations dans les domaines public et privé.

Avec l'évolution des capacités et du coût des technologies, de plus en plus d'organisations accumulent des données relatives à leurs activités, incluant des caractéristiques spatiales, par exemple des adresses ou des coordonnées GPS.

En parallèle, il y a de plus en plus de données rendues accessibles à propos d'éléments qui, potentiellement, influencent les performances des activités de ces organisations. Pour ces raisons, de nombreuses recherches dans des domaines variés aspirent à extraire de l'information pertinente pour comprendre ce qui influence concrètement les résultats d'activité (par exemple dans [Roig-Tierno et al., 2013]).

[Mennis and Guo, 2009] avancent qu'au sein de leurs recherches, le datamining spatial est un domaine qui prend de l'ampleur. Le datamining spatial, comme le définissent [Koperski and Han, 1995], consiste en l'extraction de connaissances implicites à partir de données géospatiales. Ce domaine de recherche est une extension du Knowledge Discovery from Databases (KDD) introduit par [Fayyad et al., 1996] dans les années quatre-vingt-dix.

Cependant, l'ajout de la composante spatiale aux données n'est pas sans conséquence sur le processus d'extraction de connaissance à partir des données.

En effet, une des étapes qui compose ce procédé, la préparation des données, est très affectée par cette composante spatiale [Bogorny et al., 2005]. Pour cause, une grande partie des algorithmes d'analyse de données ne sont pas capables d'interpréter la signification de coordonnées GPS ou d'autres données spatialisées. De nombreuses recherches s'accordent à dire que la préparation des données spatiales est une tâche complexe et laborieuse.

L'objectif de cette recherche est de proposer une approche qui automatise en grande partie le processus de prétraitement des données spatiales. La section suivante (2) présente les éléments de la littérature sur ce sujet. Pour permettre une meilleure compréhension de la problématique, la section 3 présentera dans un premier temps les spécificités liées à la préparation des données spatiales, et dans un second temps en quoi ce prétraitement peut induire des aléas dans la suite de l'analyse des données.

La section 4 présentera notre approche dans sa globalité et chaque étape de manière plus détaillée. Les aspects techniques et logiciels liés à la mise en place de notre solution seront aussi présentés. Enfin, l'implémentation de notre outil sera présentée au travers d'un exemple d'application et en parallèle, les étapes nécessaires sans notre solution pour la réalisation du prétraitement seront présentées. Pour terminer, les limites et les perspectives de nos recherches seront discutées.

2 REVUE DE LITTERATURE

2.1 Prise de décision spatiale

Il y a trente ans, les analyses de [Schmidt, 1983] révélaient que les décisions de localisation étaient effectuées rapidement, par des gens sans expérience ni connaissance des problématiques impliquées. Les décisions étaient prises de façon subjective avec peu d'exigences et en envisageant qu'une petite partie des options existantes.
Il y a une quinzaine d'années [MacEachren and Kraak, 2001] remarquent que de nombreux problèmes dans les domaines scientifiques et sociaux avaient un aspect spatial. Il ajoute qu'à ce moment-là, la quantité de données disponibles ayant une composante spatiale ne cessait de croitre. Cependant, à cause du manque de méthodologie adaptée pour les analyser, ces données étaient rarement utilisées pour construire de la connaissance utilisable.
Plus récemment, selon Thompson [Thompson and Walker, 2005], l'intérêt grandissant pour l'analyse des données spatiales est associé à l'augmentation de la compétitivité. [Keenan, 2006] identifie quant à lui un besoin d'information spatiale chez les managers de haut niveau.
De nombreux chercheurs se sont penchés sur ces sujets, de leurs travaux ont résulté de nombreux outils qui permettent de valoriser les informations spatiales.
Les outils développés dans cette optique sont souvent appelés Spatial Decision Support System (SDSS), introduit par [Armstrong et al., 1990] et [Densham, 1991] au début des années quatre-vingt-dix.

2.2 Datamining spatial : particularités et challenges

De nombreux SDSS sont basés sur le datamining spatial qui selon [Miller, 2007] permet de révéler des modèles intéressants sur des éléments ou des évènements distribués spatialement. Ces modèles peuvent être basés sur les propriétés spatiales de ces éléments, ou sur les relations spatiales qui existent entre les éléments (en plus des attributs non spatiaux d'intérêt usuel dans le datamining traditionnel).
Alors que de nombreux projets de SDSS ont été réalisés, il reste toujours plusieurs challenges mentionnés dans la littérature. [Erskine et al., 2013] et [MacEachren and Kraak, 2001] indiquent que le stockage, la manipulation et l'utilisation des données spatiales doivent être étudiés pour simplifier la prise de décisions stratégiques ou organisationnelles.
Dans le même sens [Sugumaran, 2007] signale le besoin de construire des SDSS capables d'utiliser la technologie de manière à faciliter l'interopérabilité entre les données spatiales et les SDSS.
Pour [Keenan, 2004], une approche qui regroupe des interfaces, des techniques de modélisation et des bases de données appropriées doit être développée.
[Pretorius and Matthee, 2006] indique que pour réaliser ses analyses il n'existait pas, à l'époque de ses recherches, de méthodologie uniforme et générique pour le datamining spatial.
[Alatrista Salas et al., 2013] disent que les algorithmes de datamining classiques prennent en entrée des données stockées dans des tables, et qu'ils ne tiennent pas compte de l'information spatiale directement.
Pour [Bogorny et al., 2005], il existe peu d'outils qui gèrent les directement les données spatiales.
Dans la majorité des cas, comme le mentionne [Pretorius and Matthee, 2006] il faut préparer ces données à la main avant de les importer pour pouvoir en tirer parti.

2.3 Phase de prétraitement des données spatiales

2.3.1 Particularités liées aux données spatiales

Un des premiers aspects à prendre en compte pour la préparation des données spatiales est qu'il faut rendre les différents jeux de données compatibles entre eux.
Cette compatibilité est nécessaire pour pouvoir ensuite visualiser ces données, ou étudier les relations entre les jeux de données [Flowerdrew, 1991]. Comme [Mennis and Guo, 2009] le disent, les données spatiales proviennent souvent de sources différentes.
Une autre particularité des données spatiales, comme l'avancent [Ester et al., 1999], est que des éléments voisins spatialement peuvent s'influencer entre eux.
[Shekhar and Chawla, 2003] disent que les données spatiales doivent être transformées en prédicats spatiaux pour pouvoir être traités par les outils d'analyse de données classiques.
[Bogorny et al., 2005] définissent les prédicats spatiaux comme la matérialisation des relations spatiales, qui ne sont pas stockées explicitement dans les bases de données. Ces relations entre des éléments positionnés dans l'espace peuvent être calculées par des opérations spatiales.
[Ester et al., 1999] et [Bogorny et al., 2005] dénombrent trois types de relations spatiales qui peuvent exister entre deux entités géospatiales : les relations topologiques, les relations de distances, et les relations d'orientation (voir Figure 1, extraite et traduite de [Bogorny et al., 2005]).
Dans [Bogorny et al., 2005] les relations topologiques caractérisent le type d'intersection qui existe entre les éléments, par exemple : touche, contient, se chevauche, etc. (voir Figure 1.A). Les relations de distances, montrées dans la Figure 1.B, peuvent être basées sur différentes métriques, par exemple la distance euclidienne. Les relations de direction prennent en compte la position des éléments par rapport aux autres (exemple en Figure 1.C).

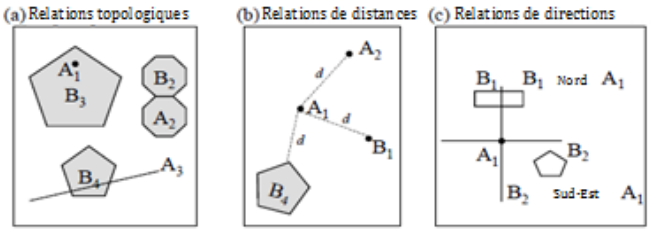


Figure 1 - les relations spatiales

2.3.2 Complexité et durée des opérations spatiales

Les données spatiales sont la plupart du temps prétraitées au travers de Systèmes d'Information Géographique (SIG). Cette préparation des données spatiales doit être effectuée avec précaution, et cela peut prendre beaucoup de temps, en particulier les opérations de calcul des relations spatiales. En effet [Flowerdrew, 1991] avance que l'intégration des données spatiales n'est pas un processus évident et que, bien que les

systèmes d'information géographique puissent paraitre faciles d'utilisation, cela reste du travail soigneux.
Dans de nombreuses recherches sur les méthodologies d'Extraction de Connaissances à partir des Données (ECD, la traduction de KDD), la phase de préparation des données est souvent considérée comme une des plus gourmandes en temps (voir Figure 2 extraite et traduite de [Cios et al., 2007]).

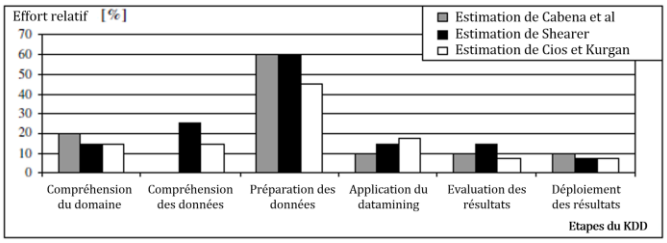


Figure 2 - Effort relatif par étape

Même dénuée du caractère spatial, la phase de prétraitement est déjà considérée de loin comme la plus chronophage. [Pretorius and Matthee, 2006] avancent que la phase la plus impactée par la composante spatiale dans les processus de datamining est celle de la préparation des données.
[Clementini et al., 1993] affirment que stocker les résultats de toutes les relations spatiales est très couteux en espace mémoire. Il en déduit qu'au lieu de stocker toutes les relations spatiales entre éléments, il est plus pratique de les calculer quand elles sont nécessaires. Pour cela, il faut avoir une compréhension complète sur les techniques de calcul des relations spatiales.
Par rapport à l'application de ces techniques [Egenhofer, 1992] mentionne que des concepts fondamentaux pour la réalisation d'analyses spatiales sont à maitriser, tels que ceux relatifs aux relations géométriques entre les éléments spatiaux.
Dans le même sens [West, 2000] dit que les SIG sont complexes à utiliser, et que leur utilisation nécessite la maitrise de notions cartographiques complexes, qui peuvent paraitre inaccessibles pour les non-initiés.
[Vahedi et al., 2016] ajoutent qu'au sein des SIG les fonctions spatiales et leurs paramètres sont souvent difficiles à comprendre et à utiliser. Entre autres parce que les experts en analyse sont souvent non connaisseurs des SIG et qu'ils sont sans formation particulière dans ce domaine. Il ajoute que le manque d'alternative aujourd'hui rend malgré tout obligatoire l'utilisation des SIG pour effectuer certaines opérations spatiales.
[Bogorny et al., 2005] mentionnent aussi que, fréquemment, les personnes qui procèdent aux analyses de données ne sont pas forcément des experts en base de données spatiale.
[Appice et al., 2003] remarquent eux que l'expertise requise pour traiter les données spatiales est souvent un frein à la réalisation d'analyses de données spatiales.
En plus d'être complexe [Bogorny et al., 2005] ajoutent que la préparation pour rendre les données spatiales prêtes pour les algorithmes de datamining est longue et doit être faite à la main.
Pour [Gibert et al., 2008] cette préparation des données spatiales, en plus d'être chronophage, doit être reproduite pour chaque application.
Enfin, comme le mentionnent [Mennis and Guo, 2009] plusieurs choix sont à faire durant cette préparation, par exemple sur les métriques choisies ou sur le type de relations à considérer. [Alatrista Salas et al., 2013] mettent en avant le fait qu'une connaissance du domaine d'application peut être nécessaire pour pratiquer la préparation des données.

2.3.3 Prétraitement dans les cas d'étude d'analyse spatiale

De nombreux articles plus ou moins récents présentent leurs recherches sur des analyses de données réalisées sur des cas d'études.
Dans plusieurs cas, sans donner trop d'explication ni de détails sur l'acquisition, le prétraitement et les types des relations spatiales utilisées [Knezic and Mladineo, 2006], [Previl et al., 2003], [Ghaemi et al., 2009], [Vlachopoulou et al., 2001].
D'autres recherches donnent quelques détails, comme celle de [Evans and Sabel, 2012] qui mentionne prendre plusieurs données spatiales en compte, mais sans explications plus détaillées.
On trouve dans [Wanderer and Herle, 2015] et [Andrienko et al., 2001] quelques relations spatiales calculées, mais sans explication non plus sur les choix de ces relations. Il existe aussi, plus rarement, des recherches comme celle de [Roig-Tierno et al., 2013] qui présentent des calculs de relations spatiales avancés, réalisés avec des SIG.

2.3.4 Solutions d'amélioration du prétraitement des données spatiales

[Bogorny et al., 2005] avaient déjà identifié certains des problèmes induits par la composante spatiale des données. Ils avaient alors proposé un environnement avec des outils qui permettent de réaliser la préparation des données spatiales. L'utilisation de leurs outils requière cependant des connaissances en SIG, et le calcul des relations spatiales n'y est pas automatisé. De plus, l'approche de [Bogorny et al., 2005] ne propose pas d'assistance pour déterminer les relations spatiales à prendre en compte.
À notre connaissance, il n'y a aujourd'hui pas de solution qui permet de préparer efficacement les données spatiales sans connaissance particulière des SIG et des relations spatiales à prendre en compte dans un domaine d'application donné.
Pour pallier ce manque, nos recherches se concentrent sur l'automatisation du processus de pré traitement des données.
Comme la préparation des données spatiales est une notion qui peut être compliquée à appréhender, la prochaine section se concentre sur l'explication des concepts qui y sont liés. Les effets que peuvent avoir les choix impliqués dans cette préparation sur les résultats d'analyses sont aussi présentés.

3 NECESSITE ET COMPLEXITE DU PRETRAITEMENT

Comme mentionné précédemment, dans la majorité des cas, les données spatiales sont inadaptées pour l'application d'algorithmes de datamining classiques.
Comme cette notion est parfois difficile à appréhender, nous allons dans cette partie essayer d'expliquer, dans un premier temps, la nécessité du prétraitement. Dans un second temps, nous essayerons de montrer la complexité des choix qui peuvent être impliqués au cours de ce prétraitement et ce que ces choix peuvent induire sur la suite des analyses.
Pour commencer, il faut savoir que dans les SIG, un ensemble de données relatives à un type d'éléments est appelé une couche. Une couche est composée d'éléments semblables géométriquement : un ensemble de points (par exemple, pour définir des emplacements précis), un ensemble de polygones (par exemple, pour définir des zones), ou un ensemble de lignes (par exemple, des réseaux routiers). Par ailleurs à chaque couche est associée une table de données qui contient des informations relatives à chaque élément de la couche. La Figure 3 donne une représentation de plusieurs couches de

données spatiales. Les zones géographiques (représentées en Figure 3.A), tout comme les lacs (Figure 3.E) sont stockés dans les données comme des polygones. À ces polygones peuvent être associées d'autres données comme le nombre d'habitants pour les zones. Les données associées à des emplacements, comme les points de vente (Figure 3.B) où les stations de ski (Figure 3.D) sont elles stockées sous forme de points, auxquels peuvent être associées d'autres données dans les tables correspondantes.

Comme le mentionnent [Cliquet et al., 2006] et [Dubelaar et al., 2002] dans le cas particulier du secteur de la distribution, la connaissance de l'environnement peut être un atout compétitif majeur pour améliorer les performances.

Nous allons donc, pour illustrer les problématiques liées au prétraitement des données, nous pencher sur le cas d'une entreprise travaillant dans les matériaux de construction de chalets, et distribuant ses produits au travers de détaillants.

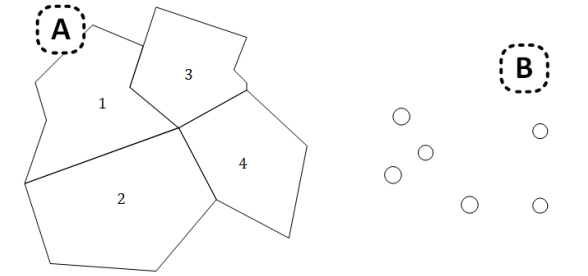
Cette entreprise souhaite mieux comprendre les effets de l'environnent sur ses résultats dans chaque région où elle à des partenariats avec des détaillants.

3.1 La nécessité du prétraitement.

Dans un premier temps, il faut rassembler les données disponibles. L'entreprise dispose des données de ventes (chiffres d'affaires entre autres) de chacun de ses magasins, ainsi que de leurs emplacements, sous forme de coordonnées GPS (Figure 3.B).

Les données des régions correspondent quant à elles aux limites géographiques, associées à des données sociodémographiques (nombre d'habitants par région par exemple) comme on le voit sur la Figure 3.A.

Les commerciaux de l'entreprise suspectent que les ventes sont influencées par la proximité des stations de ski, des terrains de golf, ainsi que des plans d'eau. Les données correspondantes à ces éléments sont alors récupérées (respectivement Figure 3. C, 3. D, et 3.E).



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

Détaillants	Ventes (CA)
1	3 000
2	2 400
...	

Club de golf
1
2
...

Station de ski
1
2
...

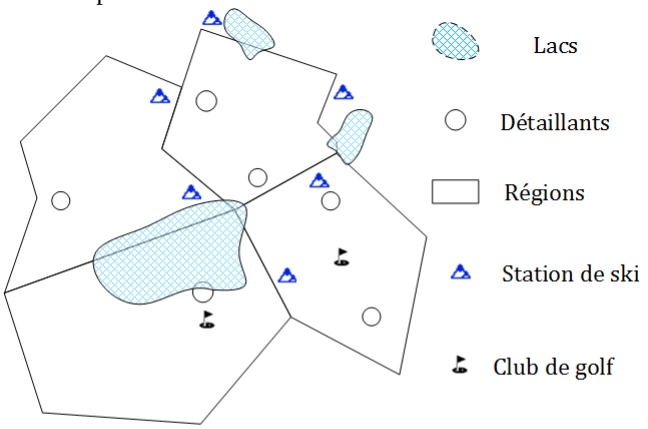
Lac
1
2
...

Figure 3 - ensemble de données

Ces différents jeux de données, encore dissociés, ne permettent pas l'application de la plupart des algorithmes de datamining classique. Il faut « fusionner » les données au sein d'une table dans laquelle les prédicats spatiaux, issus des relations spatiales entre les différents jeux de données seront calculés, comme dans la Figure 4.

Avec la combinatoire entre le nombre de relations spatiales et la quantité de données, il n'est pas envisageable de calculer toutes les relations spatiales existantes. Il faut donc sélectionner les relations spatiales pertinentes avant de les importer dans une table sur laquelle pourront être appliqués des algorithmes de datamining.

La section suivante aspire à mettre en avant les difficultés liées aux biais dans les résultats induits par les choix réalisés pendant le prétraitement des données.



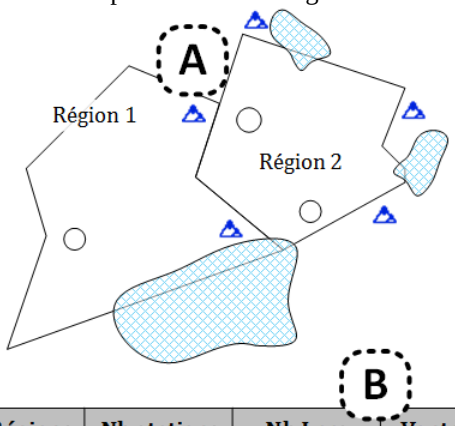
Régions	Surface	Population	Prédicats spatiaux					Ventes totales
1	45	1 200	...	...	...	...	...	12 000
2	18	4 500	...	...	...	...	...	8 000
...	...	...	...	...	...	...	...	...

Figure 4 - intégration des données

3.2 Conséquence des choix effectués lors du prétraitement.

En fonction des prédicats spatiaux choisis et importés dans la table, les résultats d'analyse sur les effets de l'environnement peuvent varier.

Pour illustrer cette influence du choix de la relation par rapport au résultat d'analyse, on se replace dans notre cas particulier, en se concentrant uniquement sur deux régions, avec la situation suivante représentée sur la Figure 5.A.

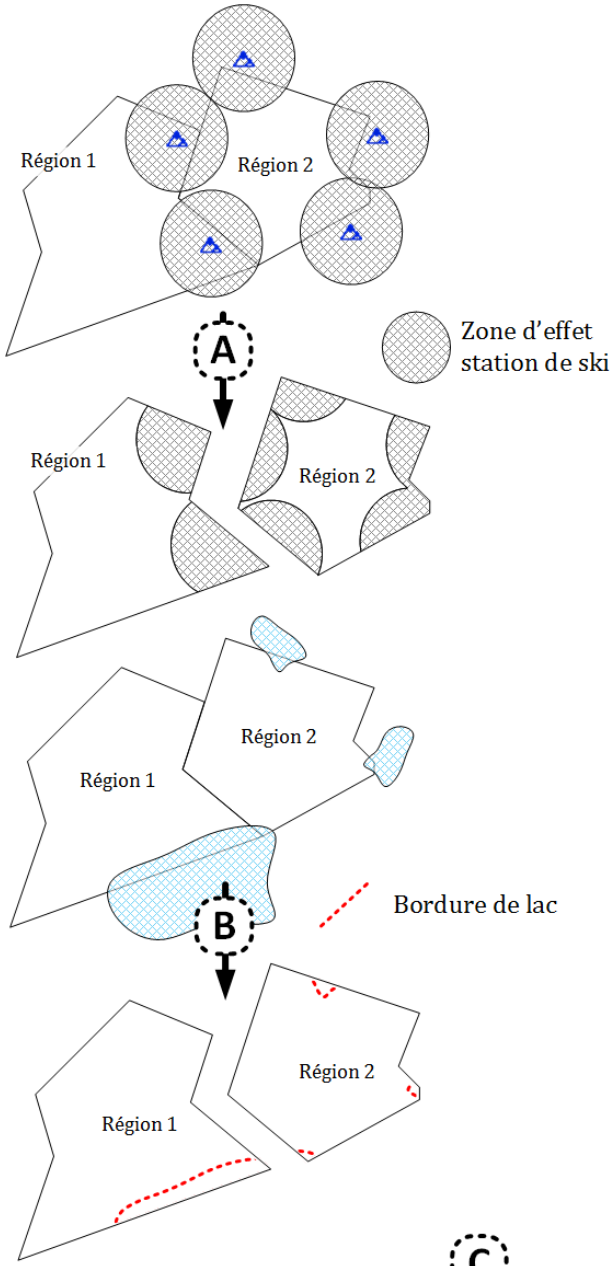


Régions	Nb stations	Nb Lacs	Ventes
1	2	1	12 000
2	0	3	8 000

Figure 5 - calcul du nombre d'éléments

Dans le cas où nous cherchons à étudier l'effet des stations de ski et des lacs sur les ventes, et que nous choisissons de prendre comme relation spatiale le nombre de stations de ski et

le nombre de lacs par régions, nous obtiendrons une table de données comme celle en Figure 5.B.
On peut aussi choisir d'autres relations spatiales, par exemple pour les stations de ski, la couverture des régions par leurs zones d'attraction (relation spatiale illustrée en Figure 6.A).
Pour les lacs, on peut imaginer prendre en compte le nombre de kilomètres de côte dans chaque région (relation spatiale illustrée en figure 6.B).
Dans le cas où ce sont plutôt ces dernières relations qui sont sélectionnées, nous obtiendrions une table de données qui ressemblera à celle en 6.c.



Régions	Couverture chalandise ski (%)	Côtes (km)	Ventes
1	15 %	12 km	12 000
2	40 %	5 km	8 000

Figure 6 - calculs de relations spatiales

Les différences entre les tables de données 5.A et 6.C laissent à penser que les analyses faites sur celles-ci ne donneront par les mêmes interprétations quant aux effets des lacs et des stations de ski sur la performance.

Ainsi la préparation des données introduit un biais lié au choix des relations spatiales choisies, calculées, et importées dans la table qui servira de données entrantes pour les analyses.
Par ailleurs, comme mentionné dans la revue de littérature, le calcul de ces relations est une tâche gourmande en temps qui demande des compétences avancées en SIG. Notre approche, présentée dans la section suivante, aspire à résoudre ces problèmes.

4 ALGORITHME D'EXTRACTION SEMI-AUTOMATIQUE DES PREDICATS SPATIAUX

Afin d'expliquer clairement notre approche, un premier schéma présentant l'enchainement des étapes est proposé. Chaque étape est ensuite présentée avec plus de détails. Les prérequis nécessaires à la mise en place de cette solution sont mentionnés ensuite, puis une architecture possible est proposée. Pour finir, l'implémentation réelle de l'outil est présentée au travers d'un exemple dans lequel on comparera la réalisation des étapes de prétraitement nécessaires avec et sans notre solution.

4.1 Déroulement du prétraitement

4.1.1 Schéma général

Notre approche, qui aspire à automatiser au maximum le processus du prétraitement des données spatiales, part du principe que les données ont été récupérées et stockées dans une base de données. À la suite du processus présenté Figure 7, un ensemble de données sera formaté et mis à disposition pour l'application d'algorithmes de datamining. Dans la Figure 7, les tâches en traits pleins sont celles totalement automatisées, celles en pointillés sont celles à réaliser par l'utilisateur.

4.1.2 Description des étapes

La première étape correspond à la sélection d'une couche cible. La couche cible est celle qui contient une variable que l'on souhaite étudier (par exemple, nous sommes intéressés par ce qui se passe dans les régions). La deuxième étape consiste à sélectionner la variable d'étude de la couche cible. Le choix de la variable d'étude se fait dépendamment des données disponibles et du type d'information recherchée. Dans notre cas d'étude, la variable d'étude sera relative aux ventes, par exemple : le chiffre d'affaires.
L'étape trois permet de sélectionner une couche source pour en étudier l'influence. Une couche source est une couche de données dont on suspecte qu'elle a une influence sur les données de la couche cible. Dans notre cas d'étude, il y a plusieurs couches sources potentielles à prendre en compte, comme celle des stations de ski par exemple.
Lorsque l'utilisateur a sélectionné la couche cible et sa variable d'étude, différentes relations spatiales sont calculées entre les deux couches. Dans notre cas particulier, pour chaque région de la couche cible plusieurs nouvelles données relatives aux relations spatiales avec les stations de ski seront créées. Ainsi plusieurs des prédicats spatiaux seront consultables : comme le nombre de stations de ski par région, la distance à la station la plus proche, ou taux de couverture des régions par les zones d'influences des stations de ski.
Pour chacun des prédicats spatiaux intégrés à l'outil, un score de corrélation avec la variable d'étude est calculé. Ce score de corrélation permet à l'utilisateur d'avoir des indications sur quelles relations spatiales influencent potentiellement la variable d'étude choisie préalablement (à l'étape 2).

L'utilisateur peut alors sélectionner les relations spatiales à conserver. Concrètement, si l'utilisateur choisit de conserver le nombre de stations de ski par régions, l'ensemble de données associé sera stocké en mémoire.
Ces étapes de calculs de relations et d'évaluation de la corrélation peuvent être répétées pour chaque couche de données sources que l'utilisateur souhaite étudier.

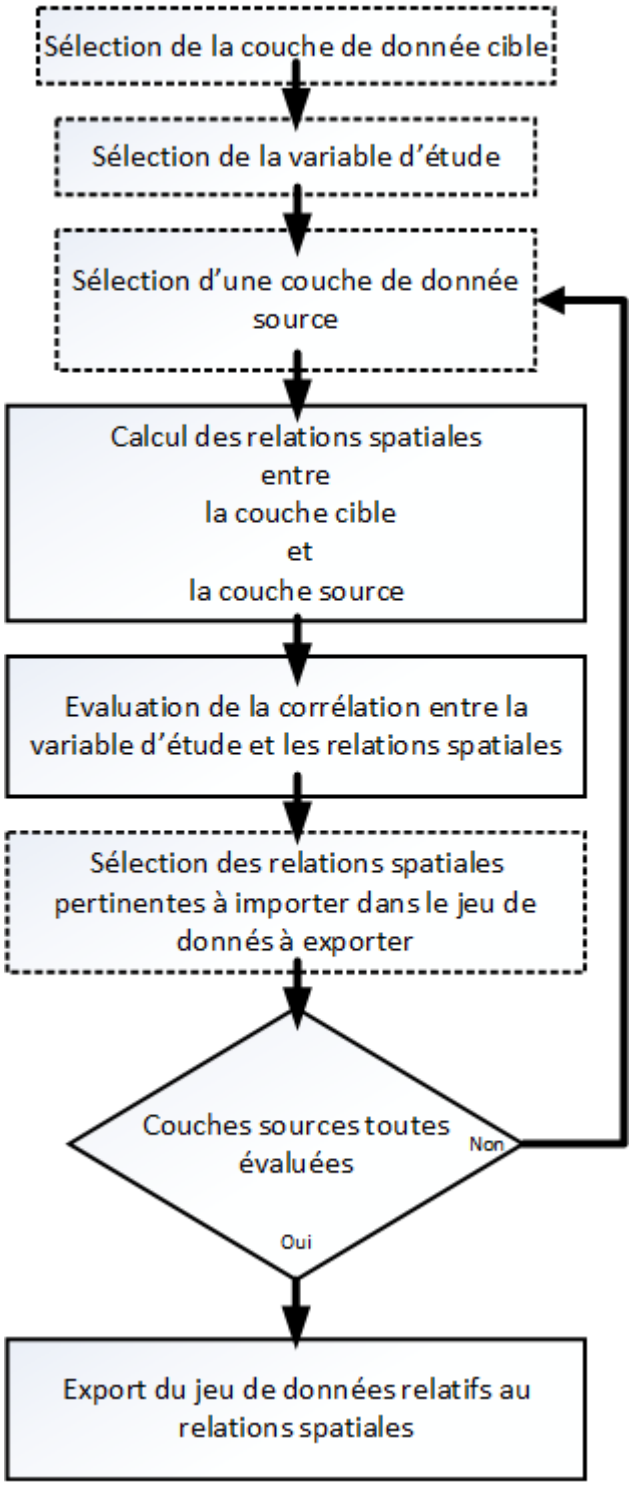


Figure 7 - étapes du processus

Enfin, lorsque l'utilisateur estime avoir fait le tour des couches sources, il peut décider d'exporter l'ensemble des données correspondant aux relations spatiales d'intérêts.
La mise en place de l'outil permettant la réalisation de ce processus est présentée dans la section suivante.

4.2 Mise en place

La mise en place de cette approche n'est réalisable que sous certaines conditions. Il existe des prérequis sur le rassemblement des données et sur la structuration de la base de données.
Par ailleurs, il faut avoir à disposition des outils logiciels qui permettent de réaliser les différentes étapes. Les outils utilisés et l'architecture de notre approche sont présentés dans cette partie.

4.2.1 Prérequis sur la structure des données

Les différentes données rassemblées doivent être mises dans le même système de projection, sinon les calculs de relations spatiales peuvent renvoyer des résultats erronés.
Les couches potentiellement cibles et celles potentiellement sources doivent être stockées dans des endroits prévus à cet effet, de manière à pouvoir les extraire automatiquement par la suite, comme avec la structure de notre base de données en Figure 8. La mise en place d'une telle structure de base de données est un procédé très simple et rapide pour tout utilisateur expérimenté.

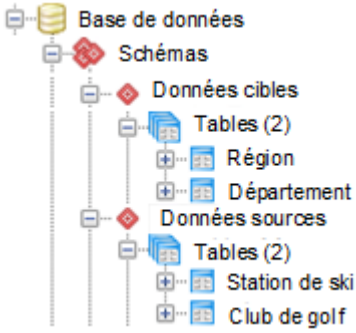


Figure 8 - structure base de données

Enfin, il faut disposer d'une base de données qui permette de faire des requêtes spatiales d'une part, et qui d'autre part, offre des possibilités d'interfaçage avec les outils qui seront utilisés pour la réalisation du système.

4.2.2 Outils utilisés pour l'implémentation

Dans la cadre de précédentes études [Daras et al., 2015], une architecture composée d'outils open sources a été mise en place. Les outils proposés ont été choisis en fonction des éléments suivants:- Ils ont été présentés comme efficaces dans la littérature récente [Zhang et al., 2010, Evans and Sabel, 2012, Agrawal and Gupta, 2014]
- Des interfaces facilitant leurs interactions et leurs intégrations sont disponibles.
Beaucoup de ces outils font parti de la suite Boundless Géo [BoundlessGeo, (accessed 13 July 2016)].
La base de données utilisée est PostgreSQL avec l'extension PostGIS permettant la prise en compte des données spatiales.
Le langage d'analyse utilisé dans notre approche est R, au travers de l'environnement de développement RStudio. L'application de prétraitement des données spatiales a été réalisée à partir du Framework Shiny.
L'application se connecte à la base de données et récupère la liste des couches cibles et des couches sources potentielles. Dans notre implémentation, les outils d'interfaçage entre le langage R et les bases de données PostGreSQL permettent de le faire simplement (voir Figure 9).

```
##Get Available Target an Sources Layers
TargetLayers <- dbGetQuery
    (
    DATABASE_Connection,
    "select tablename from
        pg_tables
    where
        schemaname='analysis_target'"
    )
SourceLayers <- dbGetQuery
    (
    DATABASE_Connection,
    "select tablename from
        pg_tables
    where
        schemaname='analysis_source'"
    )
```

Figure 9 - récupération des couches

Après sélection de la couche cible, de sa variable d'étude, et d'une couche source, les calculs des relations spatiales sont réalisés. Pour l'instant, notre implémentation calcule trois types de relations spatiales pour chaque élément de la couche cible :

-nombre d'éléments de la couche source à l'intérieur,

-distance à l'élément de la couche source le plus proche

-surface couverte par les zones d'influences des éléments de la couche source (exemple en Figure 6.A).

Ces calculs sont effectués au travers de requêtes transmises depuis notre outil à la base de données, comme dans la Figure 10 avec l'exemple du calcul du nombre d'éléments par zone.

```
##Compute cardinal
Spatial_Relation <- dbGetQuery(
    DATABASE_Connection,
    "SELECT Selected_targetLayer,
            Count(Selected_sourceLayer)
    FROM Selected_targetLayer
    LEFT JOIN Selected_sourceLayer,
    ON st_contains( Selected_targetLayer.geom,
                    Selected_sourceLayer.geom)
    GROUP BY Selected_targetLayer
    ORDER BY Selected_targetLayer
    )
```

Figure 10 - requête spatiale pour le cardinal

Ensuite, un score de corrélation est calculé entre la variable d'étude et les relations calculées avec la fonction cor, nativement présente dans le langage R. Le choix des calculs des relations à exporter est alors laissé à l'utilisateur.

4.3 Réalisation du prétraitement

Pour mettre en avant le processus de prétraitement proposé, cette section va montrer les démarches à réaliser avec et sans notre outil. Afin d'éviter les répétitions, nous montrerons dans la suite les étapes nécessaires à l'étude d'une seule relation spatiale (couverture des zones d'influences) pour une seule couche source (dans notre exemple, les clubs de golf).

Dans les deux cas, la récupération des données et l'uniformisation de leur projection sont les premières étapes à réaliser.

4.3.1 Étapes sans notre solution

Sans notre outil, il faut d'abord importer les différentes couches de données dans un SIG pour pouvoir réaliser différentes opérations successives représentées en Figure 11.

Après l'importation des données, il faut créer une nouvelle couche composée des buffers représentants les zones d'influences des clubs de golf (voir Figure 11.B).

Il faut ensuite découper cette couche en fonction des limites des zones géographiques (Figure 11.C). Pour éviter le double comptage des intersections entre zones d'influences, il faut fusionner ces couches (voir Figure 11.D). Enfin, il faut calculer la superficie des zones d'influence sur chacune des zones cibles. Pour finir, il faudra exporter ces données au format qui convient.

L'ensemble de ces étapes peut prendre plusieurs dizaines de minutes à réaliser et ne permet pas d'avoir d'indications quant à la pertinence de l'unique relation spatiale calculée.

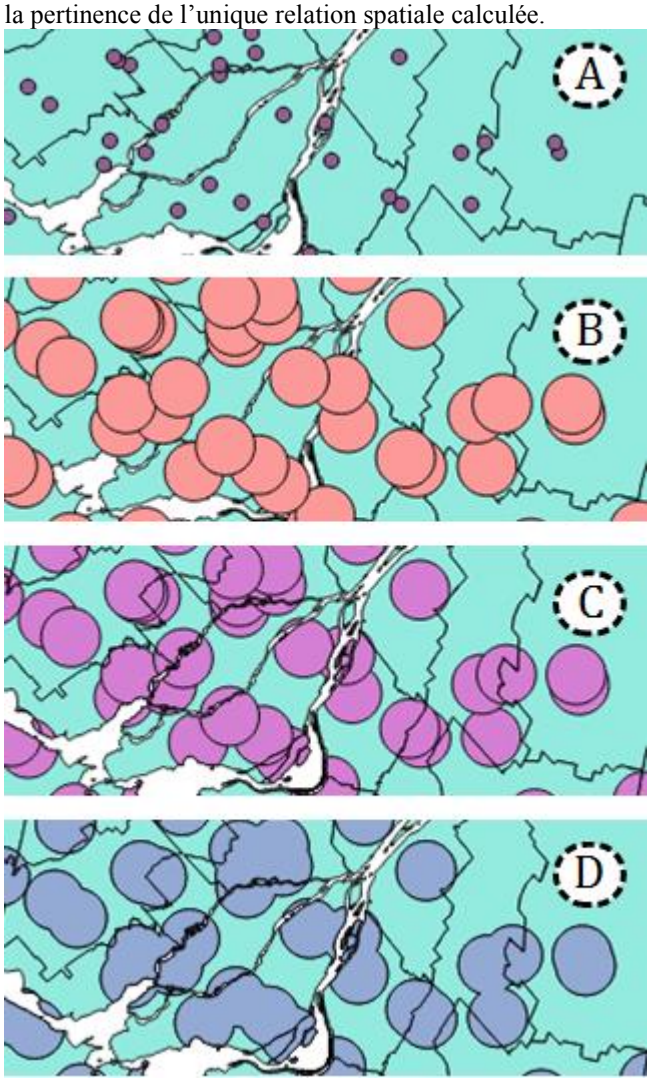


Figure 11 - étapes pour la couverture des zones

4.3.2 Étapes avec notre solution

Les données doivent être disposées dans les dossiers correspondants dans la base de données. L'utilisateur lance l'application, et choisit la couche cible (Figure 12.A), ensuite les différentes variables d'étude possibles sont proposées (Figure 12.B). L'utilisateur a accès à différentes informations statistiques sur la variable en cours de sélection. L'utilisateur peut, lorsqu'il le souhaite, valider son choix en cliquant sur le bouton prévu à cet effet (Figure 12.C).

L'interface bascule ensuite sur la sélection des couches sources (Figure 13) où, après sélection d'une des couches (Figure 13.A), différentes relations s'affichent, avec leurs scores de corrélation avec la variable d'étude (Figure 13.B). L'utilisateur peut ajouter les relations qu'il souhaite à son ensemble de données d'intérêt (figure 13.C). Lorsque l'utilisateur estime avoir terminé son analyse, il clique sur le bouton permettant l'export du jeu de données (figure 13.D).

La réalisation du prétraitement pour les trois relations spatiales intégrées à notre outil prend moins d'une minute et permet à l'utilisateur d'avoir une indication sur leurs pertinences.

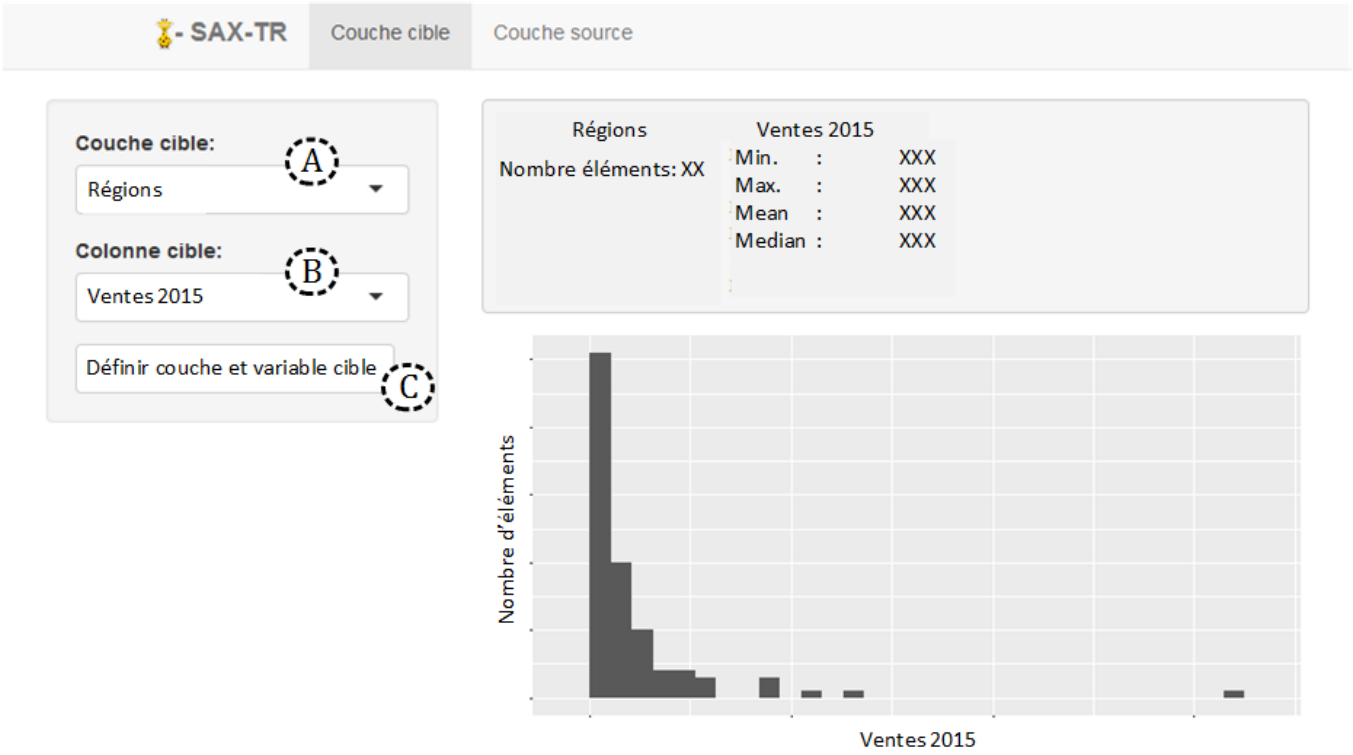


Figure 12 - interface sélection couche cible

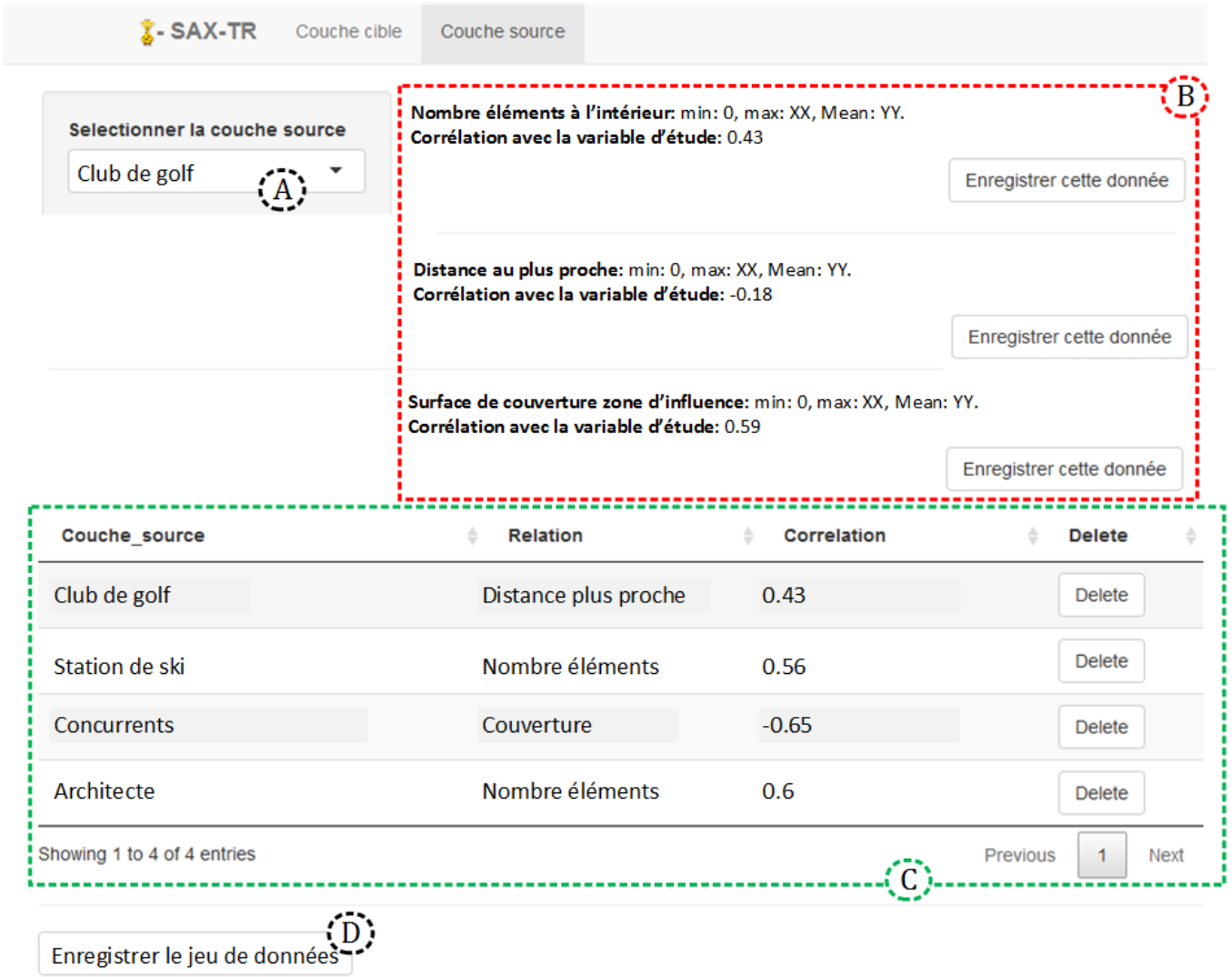


Figure 13 - interface de sélection des couches sources

Les données sont ensuite à disposition de l'utilisateur dans l'environnement de travail RStudio.
Cet exemple montre la réduction du nombre de manipulations que l'utilisateur doit maitriser et effectuer pour réaliser le prétraitement des données spatiales.

5 DISCUSSIONS AND CONCLUSIONS

De nombreuses recherches dénoncent la complexité et le côté fastidieux du prétraitement des données spatiales. Sur ce constat, nos recherches proposent une approche pour automatiser en grande partie ce processus. Nous proposons dans cette recherche une solution logicielle décomposée en plusieurs étapes qui sont décrites ici. Des outils open sources sont proposés pour permettre l'implémentation à coûts réduits de notre solution.
L'approche que l'on propose permet de rendre accessible l'analyse de données spatiales aux utilisateurs ne disposant pas de connaissance en SIG.
En plus de cela, la réalisation des étapes de prétraitement peut être réalisée plus rapidement, et avec des choix guidés quant à la sélection des relations spatiales à prendre en compte.
Différentes perspectives sont envisagées pour améliorer encore ce prétraitement des données spatiales :

- L'ajout d'autres relations spatiales calculées automatiquement ;
- L'optimisation des requêtes spatiales pour les rendre plus rapides ;
- La possibilité de catégoriser ou de pondérer les relations calculées par certains des attributs des éléments des couches cibles ou des couches sources.

REFERENCES

Agrawal, S., Gupta, R. D., 2014. 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 XL-4, 1-5.

Alatrista Salas, H., S. Bringay, F. Flouvat, N. Selmaoui-Folcher et M. Teisseire (2013). A Spatial-based KDD Process to Better Understand the spatiotemporal Phenomena. Conference on Advanced Information Systems Engineering - CAiSE, Valencia, Spain, Jun 2013.

Andrienko, N., G. Andrienko, A. Savinov, H. Voss et D. Wettschereck (2001). Exploratory Analysis of Spatial Data Using Interactive Maps and Data Mining. Cartography and Geographic Information Science **28**(3): 151-166.

Appice, A., M. Ceci, A. Lanza et F. A. Lisi (2003). Discovery of spatial association rules in geo-referenced census data: A relational mining approach. Intelligent Data Analysis **7**(6): 541-566.

Armstrong, M. P., S. De, P. J. Densham, P. Lolonis, G. Rushton et V. K. Tewari (1990). A knowledge-based approach for supporting locational decisionmaking. Environment and Planning B: Planning and Design **17**(3): 341-364.

Bogorny, V., P. Martins Engel et L. O. Alvares (2005). Spatial Data Preparation for Knowledge Discovery. In: I IFIP academy on the state pf software theory and practice - PhD Colloquim. Porto Alegre, Brazil (2005).

BoundlessGeo, (accessed 13 July 2016).
URL http://boundlessgeo.com/

Cios, K. J., W. Pedrycz, R. W. Swiniarski et L. Kurgan (2007). The knowledge discovery process, in: Data Mining A knowledge discovery approach. Springer, pp 9-24.

Clementini, E., P. Felice et P. Oosterom (1993). A small set of formal topological relationships suitable for end-user interaction. In the Third International Symposium on Advances in Spatial Databases - SSD'93. pp 277-295**.**

Cliquet, G., A. Fady et G. Basset (2006). Management de la distribution, 2nd Edition, Dunod.

Daras, G., B. Agard, H. Cambazard et B. Penz (2015). Development of business spatial analysis tools: methodology and framework. 2015 IFAC Symposium on Information Control in Manufacturing – INCOM 2015. Ottawa (Ontario), Canada.

Densham, P. J. (1991). Spatial decisions support systems. In: Geographical Information Systems: Principles and Applications. Wiley, Wiley**:** 403-412.

Dubelaar, C., M. Bhargava et D. Ferrarin (2002). Measuring retail productivity: what really matters? Journal of Business Research **55**(5): 417-426.

Egenhofer, M. J. (1992). Categorizing Binary Topological Relations Between Regions, Lines, and Points in Geographic Databases. Third International Conference on Principles of Knowledge, Representation and Reasoning - KR 1992. pp. 165-176.

Erskine, M., D. Gregg, J. Karimi et J. Scott (2013). Business Decision-Making Using Geospatial Data: A Research Framework and Literature Review. Axioms **3**(1): 10-30.

Ester, M., H.-P. Kriegel et J. Sander (1999). Knowledge Discovery in Spatial Databases. Mustererkennung 1999. Springer Berlin Heidelberg. pp 1-14

Evans, B. et C. E. Sabel (2012). Open-Source web-based Geographical Information System for health exposure assessment. International Journal of Health Geographics **11**: 1-11.

Fayyad, U., G. Piatetsky-Shapiro et P. Smyth (1996). From data mining to knowledge discovery in databases. AI Magazine **17**(3): 37-54.

Flowerdrew, R. (1991). Spatial data integration. Geographical information systems **1**: 375-387.

Ghaemi, P., J. Swift, C. Sister, J. P. Wilson et J. Wolch (2009). Design and implementation of a web-based platform to support interactive environmental planning. Computers, Environment and Urban Systems **33**(6): 482-491.

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

Keenan, P. B. (2006). Spatial Decision Support Systems: A coming of age. Control and Cybernetics **35**: 9-27.

Keenan, P. B. (2004). Using a GIS as a DSS generator. DSSResources.COM, 12/17/2004

Knezic, S. et N. Mladineo (2006). GIS-based DSS for priority setting in humanitarian mine-action. International Journal of Geographical Information Science **20**(5): 565-588.

Koperski, K. et J. W. Han (1995). Discovery of spatial association rules in geographic information databases. Advances in Spatial Databases **951**: 47-66.

MacEachren, A. M. et M.-J. Kraak (2001). Research Challenges in Geovisualization. Cartography and Geographic Information Science **28**(1): 3-12.

Mennis, J. et D. Guo (2009). Spatial data mining and geographic knowledge discovery—An introduction. Computers, Environment and Urban Systems **33**(6): 403-408.

Miller, H. J. (2007). Geographic Data Mining and Knowledge Discovery. Handbook of Geographic Information Science. Blackwell.

PostGIS (accessed 13 July 2016) http://postgis.net/.

PostgreSQL (accessed 13 July 2016) https://www.postgresql.org/.

Pretorius, J. et M. Matthee (2006). The Impact of Spatial Data on the Knowledge Discovery Process. Conference on Information Technology in Tertiary Education. Pretoria, South Africa.

Prévil, C., M. Thériault et J. Rouffignat (2003). Combining Multicriteria Analysis and GIS to help decision making processes in Portneuf County (Québec, Canada), Proceedings of 2nd Annual URISA Public Participation GIS Conference. URISA Summer Conference. Portland Oregon**:** 529-554.

R (accessed 13 July 2016) http://www.r-project.org/.

Roig-Tierno, N., A. Baviera-Puig, J. Buitrago-Vera et F. Mas-Verdu (2013). The retail site location decision process using GIS and the analytical hierarchy process. Applied Geography **40**: 191-198.

RStudio (accessed 24 October 2016) https://www.rstudio.com/.

Schmidt, C. G. (1983). Location decision-making within a retail corporation Regional Science perspectives **13**: 60-71.

Shekhar, S. and S. Chawla (2003). Spatial Databases: A Tour, Prentice Hall.

Shiny (accessed 24 October 2016) http://shiny.rstudio.com/.

Sugumaran, V. (2007). web-based spatial decision support systems (WebSDSS): evolution, architecture, examples and challenges. Communications of the Association for Information Systems **19**: 844-875.

Thompson, A. and J. Walker (2005). Retail network planning Achieving competitive advantage through geographical analysis. Journal of Targeting, Measurement and Analysis for Marketing **13**(3): 250-257.

Vahedi, B., W. Kuhn and A. Ballatore (2016). Question-Based Spatial Computing—A Case Study. In: Sarjakoski, T. and Santos, M.Y. and Sarjakoski, L.T. (eds.) Geospatial Data in a Changing World. Lecture Notes in Geoinformation and Cartography 1. B. Springer**:** 37-50.

Vlachopoulou, M., G. Silleos and V. Manthou (2001). Geographic information systems in warehouse site selection decisions. International Journal of Production Economics **71**(1-3): 205-212.

Wanderer, T. and S. Herle (2015). Creating a spatial multi-criteria decision support system for energy related integrated environmental impact assessment. Environmental Impact Assessment Review **52**: 2-8.

West, L. A. (2000). Designing End-User Geographic Information Systems. Journal of Organizational and End User Computing **12**(3): 14-22.

Zhang, C., Zhao, T., Li, W., 2010. The framework of a geospatial semantic web-based spatial decision support system for digital earth. International Journal of Digital Earth 3 (2), 111-134.