



**HAL**  
open science

# Contribution à la découverte de sous-groupes corrélés : Application à l'analyse des systèmes territoriaux et des réseaux alimentaires

Mohamed Ali Hammal

► **To cite this version:**

Mohamed Ali Hammal. Contribution à la découverte de sous-groupes corrélés : Application à l'analyse des systèmes territoriaux et des réseaux alimentaires. Apprentissage [cs.LG]. Université de Lyon, 2020. Français. NNT : 2020LYSEI024 . tel-03078791

**HAL Id: tel-03078791**

**<https://theses.hal.science/tel-03078791v1>**

Submitted on 16 Dec 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



N°d'ordre NNT : 2020LYSEI024

**THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON**  
Opérée au sein de :  
**L'INSA Lyon**

**École Doctorale N° 512**  
**Mathématiques et Informatique (InfoMaths)**

**SPÉCIALITÉ/DISCIPLINE DE DOCTORAT : INFORMATIQUE**

Soutenue publiquement le 05/06/2020, par :

**MOHAMED ALI HAMMAL**

---

---

**Contribution à la découverte de sous-groupes corrélés :  
Application à l'analyse des systèmes territoriaux et des  
réseaux alimentaires.**

---

---

Devant le jury composé de :

<b>Christine LARGERON</b>	<b>Professeur, Université Jean Monnet</b>	<b>Présidente</b>
<b>Alexandre TERMIER</b>	<b>Professeur, Université de Rennes 1</b>	<b>Rapporteur</b>
<b>Dino IENCO</b>	<b>Chargé de Recherche (HDR), INRAE</b>	<b>Rapporteur</b>
<b>Benjamin NEGREVERGNE</b>	<b>Maître de conférences, Université Paris-Dauphine</b>	<b>Examineur</b>
<b>Céline ROBARDET</b>	<b>Professeur, INSA Lyon</b>	<b>Directrice de thèse</b>
<b>Luc MERCHEZ</b>	<b>Maître de conférences, ENS Lyon</b>	<b>Co-Directeur de thèse</b>
<b>Marc PLANTEVIT</b>	<b>Maître de conférences, Université Claude Bernard Lyon 1</b>	<b>Invité</b>



**Département FEDORA – INSA Lyon - Ecoles Doctorales – Quinquennal 2016-2020**

<b>SIGLE</b>	<b>ECOLE DOCTORALE</b>	<b>NOM ET COORDONNEES DU RESPONSABLE</b>
<b>CHIMIE</b>	<b><u>CHIMIE DE LYON</u></b> <a href="http://www.edchimie-lyon.fr">http://www.edchimie-lyon.fr</a> Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage <a href="mailto:secretariat@edchimie-lyon.fr">secretariat@edchimie-lyon.fr</a> INSA : R. GOURDON	<b>M. Stéphane DANIELE</b> Institut de recherches sur la catalyse et l'environnement de Lyon IRCELYON-UMR 5256 Équipe CDFA 2 Avenue Albert EINSTEIN 69 626 Villeurbanne CEDEX <a href="mailto:directeur@edchimie-lyon.fr">directeur@edchimie-lyon.fr</a>
<b>E.E.A.</b>	<b><u>ÉLECTRONIQUE,</u></b> <b><u>ÉLECTROTECHNIQUE,</u></b> <b><u>AUTOMATIQUE</u></b> <a href="http://edeea.ec-lyon.fr">http://edeea.ec-lyon.fr</a> Sec. : M.C. HAVGOUDOUKIAN <a href="mailto:ecole-doctorale.eea@ec-lyon.fr">ecole-doctorale.eea@ec-lyon.fr</a>	<b>M. Gérard SCORLETTI</b> École Centrale de Lyon 36 Avenue Guy DE COLLONGUE 69 134 Écully Tél : 04.72.18.60.97 Fax 04.78.43.37.17 <a href="mailto:gerard.scorletti@ec-lyon.fr">gerard.scorletti@ec-lyon.fr</a>
<b>E2M2</b>	<b><u>ÉVOLUTION, ÉCOSYSTÈME,</u></b> <b><u>MICROBIOLOGIE, MODÉLISATION</u></b> <a href="http://e2m2.universite-lyon.fr">http://e2m2.universite-lyon.fr</a> Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : H. CHARLES <a href="mailto:secretariat.e2m2@univ-lyon1.fr">secretariat.e2m2@univ-lyon1.fr</a>	<b>M. Philippe NORMAND</b> UMR 5557 Lab. d'Ecologie Microbienne Université Claude Bernard Lyon 1 Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69 622 Villeurbanne CEDEX <a href="mailto:philippe.normand@univ-lyon1.fr">philippe.normand@univ-lyon1.fr</a>
<b>EDISS</b>	<b><u>INTERDISCIPLINAIRE</u></b> <b><u>SCIENCES-SANTÉ</u></b> <a href="http://www.ediss-lyon.fr">http://www.ediss-lyon.fr</a> Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : M. LAGARDE <a href="mailto:secretariat.ediss@univ-lyon1.fr">secretariat.ediss@univ-lyon1.fr</a>	<b>Mme Sylvie RICARD-BLUM</b> Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (ICBMS) - UMR 5246 CNRS - Université Lyon 1 Bâtiment Curien - 3ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tel : +33(0)4 72 44 82 32 <a href="mailto:sylvie.ricard-blum@univ-lyon1.fr">sylvie.ricard-blum@univ-lyon1.fr</a>
<b>INFOMATHS</b>	<b><u>INFORMATIQUE ET</u></b> <b><u>MATHÉMATIQUES</u></b> <a href="http://edinfomaths.universite-lyon.fr">http://edinfomaths.universite-lyon.fr</a> Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 <a href="mailto:infomaths@univ-lyon1.fr">infomaths@univ-lyon1.fr</a>	<b>M. Hamamache KHEDDOUCI</b> Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tel : 04.72.44.83.69 <a href="mailto:hamamache.kheddouci@univ-lyon1.fr">hamamache.kheddouci@univ-lyon1.fr</a>
<b>Matériaux</b>	<b><u>MATÉRIAUX DE LYON</u></b> <a href="http://ed34.universite-lyon.fr">http://ed34.universite-lyon.fr</a> Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bât. Direction <a href="mailto:ed.materiaux@insa-lyon.fr">ed.materiaux@insa-lyon.fr</a>	<b>M. Jean-Yves BUFFIÈRE</b> INSA de Lyon MATEIS - Bât. Saint-Exupéry 7 Avenue Jean CAPELLE 69 621 Villeurbanne CEDEX Tél : 04.72.43.71.70 Fax : 04.72.43.85.28 <a href="mailto:jean-yves.buffiere@insa-lyon.fr">jean-yves.buffiere@insa-lyon.fr</a>
<b>MEGA</b>	<b><u>MÉCANIQUE, ÉNERGÉTIQUE,</u></b> <b><u>GÉNIE CIVIL, ACOUSTIQUE</u></b> <a href="http://edmega.universite-lyon.fr">http://edmega.universite-lyon.fr</a> Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bât. Direction <a href="mailto:mega@insa-lyon.fr">mega@insa-lyon.fr</a>	<b>M. Jocelyn BONJOUR</b> INSA de Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69 621 Villeurbanne CEDEX <a href="mailto:jocelyn.bonjour@insa-lyon.fr">jocelyn.bonjour@insa-lyon.fr</a>
<b>ScSo</b>	<b><u>ScSo*</u></b> <a href="http://ed483.univ-lyon2.fr">http://ed483.univ-lyon2.fr</a> Sec. : Véronique GUICHARD INSA : J.Y. TOUSSAINT Tél : 04.78.69.72.76 <a href="mailto:veronique.cervantes@univ-lyon2.fr">veronique.cervantes@univ-lyon2.fr</a>	<b>M. Christian MONTES</b> Université Lyon 2 86 Rue Pasteur 69 365 Lyon CEDEX 07 <a href="mailto:christian.montes@univ-lyon2.fr">christian.montes@univ-lyon2.fr</a>

\*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie





# Remerciements

Je souhaite remercier en premier lieu ma directrice de thèse, Céline Robardet, professeur à l'INSA Lyon et mon Co-Encadrant Marc Plantevit, Maître de Conférences à l'université de Lyon 1 les deux membres de l'équipe DM2L du laboratoire LIRIS, pour leur confiance, encouragements et surtout leur patience. J'ai pu apprendre à leurs côtés les bases de la fouille de motifs sous contraintes et ce qu'exige un bon travail dans ce domaine. Leurs qualités scientifiques, personnelles et leur dévouement hors commun m'ont toujours aidé à avancer, ils n'ont jamais cessé de croire en moi et en cette thèse. Tous les mots que je pourrai écrire ne pourront exprimer ma reconnaissance envers eux, MILLE mercis.

Un grand merci à mes encadrants de l'École Normale Supérieure (ENS) Lyon, Luc Merchez, Maître de conférences et Hélène Mathian, Ingénieure de recherche CNRS, du laboratoire Environnement Ville et Société (EVS) pour tout ce que j'ai pu apprendre à leurs côtés. Ils étaient toujours attentifs à mes demandes, aux exigences de mon travail de thèse et à mes questions autour du domaine de la géographie. Collaborer avec ces deux personnes était un grand ajout tant sur le plan personnel que sur le plan académique et cela m'a permis de voir l'intérêt qu'apporte la recherche en fouille de données dans l'amélioration de la vie quotidienne des gens.

Je remercie sincèrement, Christine LARGERON, Professeur des universités à l'Université de Jean Monnet, pour avoir accepté d'être présidente de mon jury et Benjamin Nègrevergne, Maître de conférences à l'Université Dauphine - PSL, parce qu'ils ont accepté de participer à mon jury de thèse, et je suis très honoré pour l'intérêt qu'ils ont porté à mon modeste travail, et pour l'échange que nous avons eu durant et après la soutenance de thèse.

Également, je remercie Dino Ienco, Chargé de recherche HDR à l'INRAE et Alexandre Termier, Professeur des universités à l'Université de Rennes 1, d'avoir accepté d'être rapporteurs de ce manuscrit de thèse, aussi pour le travail qu'ils ont consacré à sa lecture et à l'écriture des rapports, malgré les temps serrés.

Durant ces années de thèse, j'avais la chance d'être entouré de plein de personnes formidables de l'équipe DM2L, spécialement Mehdi, Jean-François et Rémi ainsi que mes collègues avec qui j'ai toujours eu des discussions passionnantes, plus ou moins intéressantes mais qui étaient toujours à l'écoute. Je vous remercie du fond du cœur, pour vos conseils, pour votre amitié et pour ces moments de bonheur.

Pour finir, bien sûr toute ma gratitude va à mes parents, qui malgré notre modeste vie ont toujours fait de leur mieux et ont su m'apprendre l'essentiel dans la vie. À mes trois sœurs, qui même étant distants ont su me faire garder le moral durant toute la période de cette thèse. À toute ma famille, mes proches et mes amis Redouane, Souhaib, Mustapha, Mohamed, Abdelhak, Chakib, Sabrina, Youcef et plein d'autres que je ne pourrai citer mais qui se reconnaîtront. Des gens qui ont toujours cru en moi et qui m'ont toujours encouragé pour aller de l'avant avec mes aspirations et mes rêves, ce travail vous est dédié.



# Résumé

Mieux nourrir les villes en quantité et en qualité, notamment les grandes agglomérations, constitue un défi majeur dont la résolution passe par une meilleure compréhension des relations entre les populations urbaines et leur alimentation. A l'échelle des systèmes alimentaires urbains, on a besoin de diagnostics ciblant la disponibilité des ressources alimentaires croisée avec les profils socio-économiques des territoires et l'on manque d'outils et de méthodes pour appréhender de façon systématique les relations entre les bassins de consommation, l'offre et les comportements alimentaires. L'objectif de cette thèse est de contribuer à l'élaboration de nouveaux outils informatiques pour traiter des données temporelles, hétérogènes et multi-sources afin d'identifier et de caractériser des comportements propres à une zone géographique. Pour cela, nous nous appuyons sur l'exploration conjointe de motifs graduels, identifiant des corrélations de rang, et de sous-groupes afin de découvrir des contextes pour lesquels les corrélations décrites par les motifs graduels sont exceptionnellement fortes par rapport au reste des données. Nous proposons un algorithme d'énumération s'appuyant sur des propriétés d'élagage avec des bornes supérieures, ainsi qu'un autre algorithme qui échantillonne les motifs selon la mesure de qualité. Ces approches sont validées non seulement sur des jeux de données de référence, mais aussi à travers une étude empirique de la formation des déserts alimentaires sur l'agglomération lyonnaise.

**Mots-clés :** Découverte de connaissances, fouille de motifs, sous-groupes corrélés, échantillonnage de motifs, analyse de déserts alimentaires.



# Abstract

Better feeding cities in quantity and quality, especially large cities, is a major challenge, whose resolution requires a better understanding of the relationships between urban populations and their food. On the scale of urban food systems, we need to understand the availability of food resources crossed with the socio-economic profiles of the territories. But we lack tools and methods to systematically understand the relationships between consumption basins, supply and eating habits. The objective of this thesis is to contribute to the development of new IT tools to process temporal, heterogeneous and multi-sources data in order to identify and characterize behaviors specific to a geographic area. For this, we rely on the joint exploration of gradual patterns, to discover rank correlations, and subgroups in order to find contexts for which the correlations described by the gradual patterns are exceptionally strong compared to the remaining of the data. We propose an enumeration algorithm based on pruning properties with upper bounds, as well as another algorithm which samples the patterns according to the quality measure. These approaches are validated not only on benchmark datasets, but also through an empirical study of the formation of food deserts in the Lyon urban area.

**Keywords :** Knowledge Discovery, Pattern mining, Correlated subgroups, food deserts.



# Table des matières

<b>Résumé</b>	<b>i</b>
<b>Abstract</b>	<b>iii</b>
<b>Table des matières</b>	<b>v</b>
<b>Introduction</b>	<b>1</b>
<b>1 État de l'art sur la fouille de données sous contraintes</b>	<b>7</b>
1.1 Le cadre générique de la fouille de motifs sous-contraintes . . . .	7
1.1.1 Le langage de motifs . . . . .	8
1.1.2 Des mesures d'intérêt aux contraintes . . . . .	10
1.1.3 Les algorithmes . . . . .	11
1.2 Les verrous . . . . .	12
1.3 La fouille de motifs graduels . . . . .	14
1.4 Découverte de sous-groupes . . . . .	16
1.5 Discussion . . . . .	18
<b>2 Fouille de sous-groupes corrélés</b>	<b>21</b>
2.1 Introduction . . . . .	21
2.2 Sous-groupes corrélés . . . . .	23
2.2.1 Évaluer la corrélation d'un ensemble d'attributs . . . . .	23
2.2.2 Corrélations contextualisés . . . . .	26
2.3 Algorithme . . . . .	28
2.4 Expérimentations . . . . .	32
2.4.1 Jeux de données et objectifs . . . . .	32
2.4.2 Étude expérimentale quantitative . . . . .	34
2.4.3 Étude expérimentale qualitative . . . . .	37
2.5 Conclusion . . . . .	41
<b>3 Échantillonnage de sous-groupes corrélés</b>	<b>45</b>
3.1 Introduction . . . . .	45
3.2 Méthodes d'échantillonnage pour la fouille de motifs . . . . .	46
3.2.1 Méthode par échantillonnage direct . . . . .	46



3.2.2	Méthodes de Monte-Carlo par chaînes de Markov . . . .	47
3.2.3	Échantillonnage biaisé vers les motifs maximaux . . . .	48
3.3	Échantillonnage aléatoire de sous-groupes corrélés maximum . .	48
3.4	Résultats expérimentaux . . . . .	49
3.4.1	Étude quantitative. . . . .	50
3.4.2	Étude qualitative. . . . .	53
3.5	Conclusion . . . . .	53
<b>4</b>	<b>Application sur les données « Commerces »</b>	<b>55</b>
4.1	Contexte . . . . .	55
4.2	Description des données et objectifs . . . . .	56
4.3	Modélisation . . . . .	61
4.4	Découverte de motifs graduels / corrélations globales . . . . .	64
4.5	Découverte de sous groupes corrélés . . . . .	64
4.6	Conclusion . . . . .	66
	<b>Conclusion</b>	<b>71</b>
	<b>Bibliographie</b>	<b>73</b>

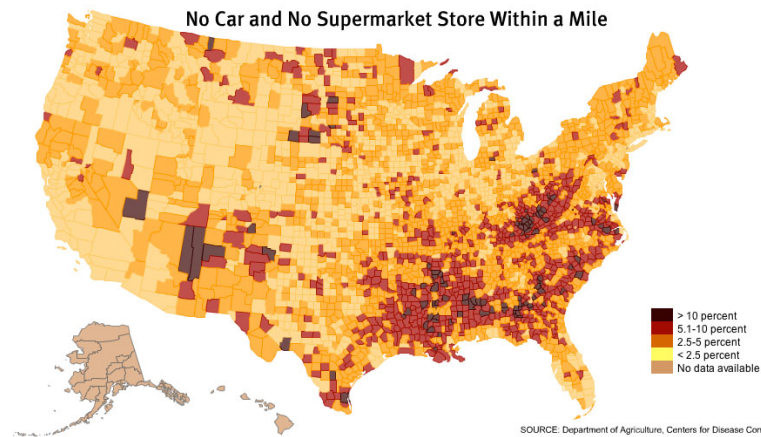
# Introduction

## De nouveaux outils pour l'étude des déserts alimentaires

Mieux nourrir les villes en quantité et en qualité, notamment les grandes agglomérations, constitue un défi majeur dont la résolution passe par une meilleure compréhension des relations entre les populations urbaines et leur alimentation. Dès lors, même si le recours à l'agriculture de proximité s'affiche comme une garantie de qualité face aux crises alimentaires, l'analyse ne peut s'arrêter aux modes et lieux de production. Les villes sont nourries par un système alimentaire urbain (Rastoin and Ghersi, 2012) qui comprend l'ensemble des espaces, processus et acteurs impliqués dans leur alimentation et leur approvisionnement (production, transformation, commercialisation, distribution, consommation des aliments). L'évolution des attentes des consommateurs envers le contenu de leur assiette, particulièrement notable en ce qui concerne la proximité et la qualité des produits consommés, tend à la fois à resserrer ces systèmes alimentaires géographiquement et topologiquement (pour une meilleure traçabilité) et à en multiplier les composantes (pour mieux répondre aux différentes attentes). Cependant, ces transformations touchent de manière inégale les consommateurs. Alors que les initiatives de type circuits courts qui connectent les « *petits producteurs locaux* » font désormais partie du quotidien de certaines populations urbaines aisées et éduquées, d'autres espaces et d'autres populations, plus défavorisés du point de vue social et/ou culturel, apparaissent non seulement déconnectés de leur agriculture proche, mais aussi dépourvus d'un approvisionnement de proximité et de qualité : une situation croissante qui mettrait en évidence l'émergence, suivant la terminologie américaine, de « déserts alimentaires » ou, du moins, de « déserts de circuits-courts » (Nikolli, 2014). Parallèlement, les recherches issues du champ de la nutrition se focalisent sur les symptômes (obésité, insécurité alimentaire) et leurs déterminants sociaux (précarité, isolement, acculturation). Ces réflexions autour de l'alimentation de proximité et de qualité des espaces urbains, et en particulier de ceux qui sont défavorisés, appellent la notion de justice alimentaire. Celle-ci désigne la répartition équitable des ressources alimentaires sur un territoire donné, dans les modes de production et de dis-

tribution (Gottlieb and Joshi, 2010).

Mieux comprendre ces déserts alimentaires, c'est être en capacité de cibler la disponibilité des ressources alimentaires (agricoles et surtout commerciales) croisée avec les profils socio-économiques des territoires. Plusieurs articles (Fyfe, 1994, Lemoy et al., 2010, Morland et al., 2002, Raja et al., 2008, Walker et al., 2010) ont étudié les effets des déserts alimentaires sur la vie des gens, mais aussi les conditions qui peuvent influencer l'apparition de ce genre de phénomènes. Or ces données sont géo-localisées, hétérogènes et évoluent dans le temps. Une telle complexité est rarement considérée simultanément dans les outils existants d'analyse de données et justifie le développement de nouvelles approches.



Carte montrant la formation de déserts alimentaires aux USA : l'échelle de couleurs indique le pourcentage de la population qui se trouve à une distance supérieure à un mile d'un supermarché.

Ce mémoire présente de nouveaux outils informatiques qui permettent de traiter des données temporelles, hétérogènes et multi-sources afin d'identifier et de caractériser des comportements propres à une zone géographique. Pour cela, nous proposons d'identifier des motifs exprimant des corrélations entre des variables quantitatives qui sont exceptionnellement fortes dans certains voisinages géographiques, ou tout autre caractérisation symbolique d'une partie des données. Par exemple nous recherchons des zones géographiques pour lesquelles on observe des corrélations fortes entre des variables quantitatives telles que la durée de vie d'un commerce et le nombre de commerces par type dans son voisinage. Ces nouveaux outils permettent de répondre aux questions relatives à la formation des déserts alimentaires comme nous le montrons dans

notre étude de cas portant sur la répartition de l'offre alimentaire dans la ville de Lyon.

## Contributions

**Fouille de sous-groupes corrélés :** Produire un outil de fouille de données hétérogènes nécessite d'analyser conjointement des attributs numériques et symboliques pour faire émerger les aspects saillants des données. Pour cela, nous utilisons la découverte de sous-groupes (Atzmüller and Puppe, 2006, Klösgen, 1996, Lavrač et al., 2004, Wrobel, 1997), qui est un cadre très flexible pour l'extraction de motifs permettant d'exprimer facilement et de répondre efficacement à des questions liées à l'identification de phénomènes remarquables. Par sous-groupe, nous considérons un sous-ensemble d'observations qui peut être séparé par une description basée sur les attributs décrivant les observations. C'est-à-dire que l'ensemble des observations satisfaisant la description correspond exactement au sous-groupe. L'aspect remarquable d'un sous-groupe est évalué à l'aide d'une mesure qui permet de comparer un type de modèle évalué sur le sous-groupe au même modèle évalué sur l'ensemble des observations.

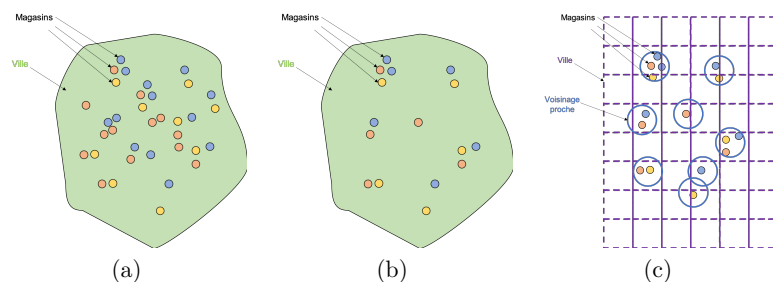
Le modèle que nous observons sur les attributs numériques est la corrélation évaluée par la mesure  $\tau$  de Kendall. Cette mesure a l'avantage de ne pas nécessiter d'hypothèse particulière sur les variables considérées et de permettre la comparaison d'un nombre quelconque (supérieur ou égal à deux) de variables. Ainsi, un groupe de variables numériques sera d'autant plus corrélé que le nombre de paires d'observations, ordonnées de la même façon par les différentes variables, sera grand. Les sous-groupes sont quant à eux identifiés par des descriptions, c'est-à-dire des restrictions sur les variables descriptives qu'elles soient numériques ou symboliques.

Notre contribution recouvre (1) la formalisation de ce nouveau domaine de motifs, (2) la définition d'un nouvel algorithme qui calcule tous les sous-groupes corrélés en exploitant certaines propriétés d'élagage basées sur deux bornes supérieures et une propriété de fermeture sur ce langage, ainsi que (3) l'évaluation empirique sur plusieurs jeux de données de l'efficacité de l'approche proposée.

**Échantillonnage de sous-groupes corrélés :** La méthode développée dans notre première contribution, qui est exacte et optimisée, a néanmoins une complexité importante due à la taille de l'espace des motifs qui est exponentielle. Même avec l'exploitation des contraintes pour réduire l'espace de recherche à son maximum, celui-ci est dans certains cas de taille exponentielle. Cette difficulté peut toutefois être surmontée en utilisant des techniques d'échantillonnage qui permettent de réduire le coût du calcul tout en identifiant les motifs les plus importants. Dans cette seconde contribution, nous

proposons une technique d'échantillonnage permettant de tirer aléatoirement des sous-groupes corrélés. Nous montrons également que cette approche permet d'obtenir des motifs de haute qualité et ce de manière efficace.

**Application sur les données « Commerces » :** Dans notre troisième contribution, nous utilisons les méthodes développées précédemment pour étudier la formation de déserts alimentaires. Pour cela, nous nous basons sur des données constituées de l'ensemble des baux commerciaux sur l'agglomération lyonnaise depuis 1957. Pour chaque bail, nous avons l'emplacement du commerce, sa nature, sa date de création et son éventuelle date de fermeture. L'objectif est de mettre en évidence des configurations de voisinage qui ont un impact (positif ou négatif) sur la durée de vie des commerces, et peuvent ainsi constituer des indicateurs de formation de déserts alimentaires. La figure suivante illustre la façon dont les voisinages sont construits : (a) A partir des données initiales, (b) on procède à une épuration de la base de données en sélectionnant les commerces d'analyse ; (c) puis on calcule dans le voisinage géographique proche (à une distance inférieure à 400 m) le nombre de commerces de chaque type.



Méthodologie appliquée pour l'analyse des données « Commerces ».

L'étude menée permet d'observer des corrélations locales non observables sur l'ensemble des données.

## Structure de la thèse et lien avec les publications

La suite de ce manuscrit est organisé en quatre chapitres et une conclusion. Le chapitre 1 présente le domaine de recherche de la fouille de motifs sous contraintes et les verrous de ce domaine de recherche. Ensuite, ce chapitre se focalise sur l'état de l'art de la fouille de motifs graduels et de la découverte de sous-groupes. Le chapitre 2 expose notre formalisation du problème de fouille de sous-groupes corrélés ainsi que la solution algorithmique que nous lui proposons. Cette contribution a été publiée à la fois dans la conférence nationale

EGC'2019 (Hammal et al., 2019c) et dans la revue *Journal of Intelligent Information Systems* (Hammal et al., 2019b). Le chapitre 3 propose un nouvel algorithme pour répondre au problème de la fouille de sous-groupes corrélés, basé sur l'échantillonnage de motifs. Ce travail a été publié dans la conférence internationale DCAI'2019 (Hammal et al., 2019a). Le chapitre 4 est consacré à l'utilisation de ces techniques pour l'analyse des données « Commerces ». Enfin, le dernier chapitre regroupe nos conclusions et perspectives.



# Chapitre 1

## État de l'art sur la fouille de données sous contraintes

Ce chapitre vise à donner une vue d'ensemble du domaine de la fouille de motifs sous contraintes en introduisant quelques définitions qui seront utiles par la suite. Dans un premier temps, nous allons présenter le cadre générique fondateur, ainsi que les premières méthodes de ce champ de recherche. Nous présenterons ensuite les limites de ces approches et les verrous à surmonter. Une des difficultés majeures est d'extraire des motifs suffisamment expressifs pour permettre à l'utilisateur d'étendre ses connaissances à partir des données. C'est dans cette perspective que s'inscrit l'extraction de motifs graduels dont nous exposerons les principes ainsi que les méthodes de l'état de l'art. Enfin, l'extraction de motifs peut être guidée par une variable ou un ensemble de variables que l'on cherche à caractériser. C'est l'objet de la découverte de sous-groupes ou de la fouille de modèles exceptionnels dont la présentation terminera cet état de l'art.

### 1.1 Le cadre générique de la fouille de motifs sous-contraintes

Alors que l'interrogation classique de base de données permet d'obtenir l'ensemble des transactions qui satisfont des contraintes, la fouille de motifs vise à extraire les éléments d'un langage qui satisfont certaines contraintes dans les données. Ce procédé a été formalisé dans (Mannila and Toivonen, 1997). Si on note  $\mathcal{D}$  les données,  $\mathcal{L}$  le langage de motifs, la fouille de motifs sous la contrainte  $q$  consiste à extraire l'ensemble des éléments  $\varphi$  du langage  $\mathcal{L}$  tels que  $\varphi$  vérifie le prédicat  $q$  dans  $\mathcal{D}$  :

$$(1.1) \quad Th(\mathcal{L}, \mathcal{D}, q) = \{ \varphi \in \mathcal{L} \mid q(\mathcal{D}, \varphi) = \mathbf{vrai} \}$$

Nous allons étudier plus en détail ces différents composants.



### 1.1.1 Le langage de motifs

Le langage de motifs est un ensemble prédéfini de descriptions que l'analyste de données cherche à examiner. Il représente l'espace de recherche des hypothèses dont l'exploration vise à trouver celles qui se vérifient dans les données  $\mathcal{D}$ . Ces hypothèses sont dites locales lorsqu'elles correspondent à un sous-groupe, une partie des données. Cette partie des données peut être sélectionnée, ou séparée du reste, par une description ou motif dans le langage de descriptions considéré. La définition d'un langage est donc intimement liée à la nature des données analysées. Le lien entre un motif et les données est exprimé par la notion d'extension qui associe à une description l'ensemble des instances des données qui la satisfont. Cette notion d'extension est spécifique à chaque langage de motifs. Nous donnons ci-dessous quelques exemples.

**Le langage des Itemsets :** Les itemsets sont définis dans des données transactionnelles, où  $\mathcal{D} = \{t_1, t_2, \dots, t_n\}$  est un ensemble de transactions dont chacune est un sous-ensemble d'items :  $t_i \in \mathcal{I}$ . La Table 1.1 donne un exemple de telles données.

	items			
$t_1$	A	B		D
$t_2$		B	C	D
$t_3$	A	B	C	D
$t_4$		B		D
$t_5$	A		C	D
$t_6$		B	C	

TABLE 1.1 – Base de données transactionnelle.

Le langage des itemsets est l'ensemble de tous les sous-ensembles d'items possibles sur  $\mathcal{I}$ , c'est-à-dire

$$\mathcal{L} = 2^{\mathcal{I}}.$$

Un exemple de motifs de  $\mathcal{L}$  est le motif  $BD$  (notation simplifiée de l'itemset  $\{B, D\}$ ). L'extension d'un itemset  $\varphi \in 2^{\mathcal{I}}$  est l'ensemble des instances de  $\mathcal{D}$  vérifiant le motif  $\varphi$ , plus formellement :

$$ext(\varphi) = \{t \in \mathcal{D} \mid \varphi \subseteq t\}$$

Par exemple,  $ext(BD) = \{t_1, t_2, t_3, t_4\}$ .

**Le langage des règles d'association :** Sur des données transactionnelles, on peut également exprimer des hypothèses sous la forme de règles d'association (Agrawal et al., 1994), qui mettent en correspondance deux itemsets

afin d'exprimer des co-occurrences entre différents sous-ensembles d'items. Le langage est alors

$$\mathcal{L} = \{X \rightarrow Y \mid X \in 2^{\mathcal{I}}, Y \in 2^{\mathcal{I}} \text{ et } X \cap Y = \emptyset\}.$$

Un exemple de motif est  $C \rightarrow AD$ . L'extension d'une règle d'association  $X \rightarrow Y$  est alors l'ensemble des instances de  $\mathcal{D}$  vérifiant la prémisse et la conclusion de la règle :

$$\text{ext}(\varphi \equiv X \rightarrow Y) = \{t \in \mathcal{D} \mid X \subseteq t \text{ et } Y \subseteq t\}$$

Par exemple,  $\text{ext}(C \rightarrow AD) = \{t_3, t_5\}$ .

**Le langage des intervalles pour les données numériques :** Lorsque les données sont numériques, les motifs peuvent exprimer des restrictions sur les valeurs possibles des attributs. Ces données sont composées d'un ensemble d'objets  $\mathcal{O}$ , chacun prenant une valeur numérique sur les attributs  $\{A_1, \dots, A_m\}$ . Les données sont alors définies par des  $m$ -uplets de valeurs numériques prises dans un domaine de valeurs spécifique à chaque attribut ( $\mathbf{Dom}(A_k)$ ) :  $\mathcal{D} = \{o \in \mathcal{O} \mid o = (A_1(o), \dots, A_m(o)), A_k(o) \in \mathbf{Dom}(A_k), \forall k = 1 \dots m\}$ . Le langage de motifs est alors composé du produit cartésien d'intervalles sur chacun des attributs (Kaytoue et al., 2011) :

$$\mathcal{L} = \left\{ \prod_{k=1}^m [a_k, b_k] \mid a_k, b_k \in \mathbf{Dom}(A_k), a_k < b_k \right\}$$

	attributs numériques				
	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$
$o_1$	1	20	75	5.4	300
$o_2$	1	25	83	6.7	100
$o_3$	2	10	30	5.6	500
$o_4$	5	15	52	3.8	760
$o_5$	1	22	44	2.0	850
$o_6$	3	30	37	1.5	410

TABLE 1.2 – Base de données numériques.

La Table 1.2 donne un exemple de données numériques et  $[4, 5] \times [15, 25] \times [52, 44] \times [1.5, 6.7] \times [760, 850]$  est un exemple de motif de  $\mathcal{L}$ . L'extension d'un motif  $\varphi$  de ce langage est alors définie par l'ensemble des instances de  $\mathcal{D}$  dont toutes les valeurs sur les attributs du jeu de données appartiennent aux intervalles constituant le motif. Si l'on note  $A_k(o)$  la valeur prise par l'instance  $o$  sur l'attribut  $A_k$ , on a :

$$\text{ext}(\varphi \equiv \prod_k [a_k, b_k]) = \{o \in \mathcal{D} \mid \forall k, a_k \leq A_k(o) \leq b_k\}$$

Par exemple,  $ext([4, 5] \times [15, 25] \times [52, 44] \times [1.5, 6.7] \times [760, 850]) = \{o_4\}$ .

### 1.1.2 Des mesures d'intérêt aux contraintes

Les contraintes permettent de sélectionner les motifs adaptés aux données et de diriger le processus de fouille pour extraire les motifs potentiellement intéressants par rapport à l'application considérée.

La contrainte la plus classique, que l'on peut utiliser sur les différents langages de motifs, est la fréquence,

$$\mathbf{freq}(\varphi) = \frac{|ext(\varphi)|}{|\mathcal{D}|},$$

qui mesure la proportion des objets de  $\mathcal{D}$  qui vérifient le motif  $\varphi$ .

D'autres mesures, telles que l'aire peuvent être également utilisées :

$$\mathbf{aire}(\varphi) = \frac{|ext(\varphi)|}{|\mathcal{D}|} \times |\varphi|,$$

avec  $|\varphi|$  le nombre d'items d'un itemset  $\varphi$  ou  $|\varphi| = \prod_k \frac{1}{b_k - a_k}$  dans le cas d'un motif intervalles  $\times_k [a_k, b_k]$ .

Pour évaluer l'intérêt d'une règle d'association  $X \rightarrow Y$ , on utilise généralement, en plus de la fréquence, la mesure de confiance qui évalue la proportion de transactions qui satisfont la conclusion  $Y$  parmi celles qui satisfont la prémisse  $X$  :

$$\mathbf{conf}(X \rightarrow Y) = \frac{|ext(X \cup Y)|}{|ext(X)|}$$

On peut également évaluer l'intérêt d'une règle en comparant le nombre de transactions supportant  $X$  et  $Y$  à l'estimation de ce nombre dans l'hypothèse où  $X$  et  $Y$  seraient indépendants.

$$\mathbf{lift}(X \rightarrow Y) = \frac{|ext(X \cup Y)|}{|ext(X)| \times |ext(Y)|}$$

Un « lift » supérieur à 1 traduit une corrélation positive de  $X$  et  $Y$ , et donc le caractère significatif de l'association. Il ne s'agit ici de donner que deux mesures de qualité des règles parmi les dizaines qui ont été proposées.

Ces différentes mesures d'intérêt sont alors exploitées pour extraire les motifs du langage potentiellement les plus pertinents pour l'utilisateur. Pour ce faire, on peut utiliser un seuil, dont la valeur est fixée par l'utilisateur, et extraire les motifs ayant une valeur sur la mesure d'intérêt supérieure à ce seuil. On peut aussi rechercher les top- $k$  motifs (Ke et al., 2009, Wang et al., 2005) ayant les valeurs les plus élevées sur la mesure d'intérêt, ou encore considérer plusieurs mesures et extraire les motifs appartenant au front de Pareto (Ugarte et al., 2017).

### 1.1.3 Les algorithmes

Une contrainte est d'autant plus intéressante qu'elle permet d'évaluer correctement la qualité des motifs et qu'elle possède des propriétés qui, combinées à un processus d'énumération habile, permettent de déduire sa valeur sur de nombreux motifs sans avoir à les évaluer individuellement. Plusieurs propriétés de contraintes utiles ont été identifiées qui permettent la conception d'algorithmes corrects et complets passant à l'échelle. De tels algorithmes implémentent des stratégies d'élagage et des mécanismes de propagation et évitent un grand nombre de calculs par rapport à une énumération exhaustive. Cependant, certains domaines de motifs ne peuvent pas être calculés de manière exacte, et requièrent l'utilisation d'approches heuristiques.

**Les algorithmes de fouille corrects et complets :** Ces algorithmes extraient tous les motifs satisfaisant la contrainte. Or, comme la taille du langage est généralement exponentielle dans l'espace des attributs, et même dans certains cas infinie (par exemple pour les langages de séquences), ces algorithmes reposent sur une exploration partielle de l'espace de recherche, mais avec la garantie que les parties non explorées ne contiennent pas de motifs solution (c'est-à-dire de motifs satisfaisant l'équation 1.1). Pour cela, les motifs sont énumérés selon un ordre qui permet d'exploiter des propriétés des contraintes. C'est principalement les propriétés de monotonie et d'anti-monotonie qui sont les plus efficaces dans ces processus d'élagage, ainsi que l'identification de bornes supérieures sur la contrainte.

**Les algorithmes heuristiques :** Ces algorithmes permettent de réaliser des processus de fouille de motifs dans des contextes difficiles, lorsque les mesures d'intérêt n'ont pas de bonnes propriétés (Bonchi and Lucchese, 2007), ou que l'espace de recherche est trop grand, malgré la possibilité de mécanismes d'élagage. L'optimalité de la solution, jusque là garantie par les approches correctes et complètes, est alors sacrifiée au profit de l'obtention d'une solution approchée. Une stratégie heuristique couramment utilisée est la recherche par faisceau (Gamberger and Lavrac, 2002, Lavrac et al., 2004), où à chaque niveau de l'énumération seulement un sous-ensemble de fils sont considérés. D'autres principes heuristiques ont également été utilisés, tels que les algorithmes évolutifs (del Jesús et al., 2007, Pachón et al., 2011, Rodríguez et al., 2012) ou de manière plus originale la recherche arborescente de type Monte Carlo (Bosc et al., 2018). Des approches heuristiques dédiés à une tâche particulière ont également été élaborés, pour, par exemple, calculer de manière itérative des itemsets maximaux (Moens and Goethals, 2013), ou extraire des sous-graphes attribués (Bendimerad et al., 2016).

On peut également citer un article récent qui propose d'effectuer une recherche *anytime* (Belfodil et al., 2018) : les sous-groupes sont générés de telle sorte que leur qualité et leur diversité s'améliorent au cours du processus

d'énumération. Ainsi, l'algorithme peut être interrompu à tout moment tout en fournissant une borne sur la mesure de qualité des motifs extraits.

**Les algorithmes par échantillonnage :** L'idée consiste à échantillonner les motifs sur la base d'une distribution de probabilité qui favorise les motifs ayant une grande valeur sur la mesure d'intérêt. On peut distinguer deux types de techniques d'échantillonnage :

- Celles basées sur les méthodes MCMC (Markov Chain Monte Carlo) qui nécessitent effectuer un parcours aléatoire sur un graphe de transition représentant la probabilité d'atteindre un motif en fonction d'un motif actuel. Bien qu'un tel processus garantisse que la distribution de la mesure est proportionnelle sur l'ensemble échantillonné à celui de l'ensemble des motifs (Boley et al., 2010), elle est obtenue à un coût qui en empêche son utilisation dans de nombreux contextes.
- Les approches qui renvoient directement des motifs sans qu'il soit nécessaire de simuler une chaîne de Markov. Un exemple de telles méthodes est la procédure d'échantillonnage en deux étapes (two-step sampling procedure de (Boley et al., 2011, 2012). Cette méthode est efficace pour extraire des itemsets. Elle a également été généralisée à d'autres types de motifs (Diop et al., 2018, Giacometti and Soulet, 2018).

## 1.2 Les verrous

Les défis liés à la fouille de motifs sous contraintes sont nombreux. Nous considérons dans cette section les principaux verrous et quelques travaux les ayant abordés.

**L'expressivité des langages :** Comme expliqué dans la section précédente, le langage de motifs est dépendant du type de données considéré. Alors que les premiers travaux (Agrawal et al., 1994) ont porté sur l'analyse de données transactionnelles, des travaux postérieurs ont eu pour ambition de considérer des données plus complexes et de proposer des domaines de motifs (langages et contraintes associées) plus expressifs. Tout un pan de recherche a considéré l'analyse de séquences (Fournier-Viger et al., 2017) qui consiste à découvrir des sous-séquences intéressantes dans un ensemble de séquences, où l'intérêt d'une sous-séquence peut être mesuré en fonction de divers critères tels que sa fréquence d'apparition, sa longueur ou encore son profit. De nombreux travaux ont aussi visé à extraire des motifs sur des graphes, que ce soit un ensemble de graphes (Borgelt and Berthold, 2002), un seul graphe attribué (Bendimerad et al., 2016, Kaytoue et al., 2017) pouvant également être dynamique (Desmier et al., 2013).

**La redondance des motifs :** Une autre difficulté des méthodes de fouille est la redondance dans les collections de motifs extraits. En effet, plusieurs descriptions peuvent avoir la même extension : ces motifs sont dits non fermés par rapport à la classe d'équivalence définie par l'extension. Des algorithmes ont ainsi été proposés pour extraire seulement les motifs fermés (Besson et al., 2005, Zaki and Hsiao, 2002) ou calculer une représentation condensée de collections de motifs (Calders et al., 2004). Cependant, l'extraction de motifs fermés ne règle pas complètement le problème : des descriptions aux extensions similaires, sans toutefois être identiques, ont bien souvent des valeurs proches sur les mesures d'intérêt ce qui conduit à leur extraction. Des post-traitements, visant à limiter l'intersection entre les motifs sont généralement utilisés pour résoudre a posteriori ce problème de redondance tout en cherchant à conserver un maximum de diversité entre les motifs. En effet, la réduction de la redondance des motifs d'une même collection ne doit pas se faire au détriment de la diversité de ceux-ci : il faut veiller à ce que les motifs non redondants reflètent la diversité des motifs de la collection complète.

**La détermination des seuils de contrainte :** Une autre difficulté réside dans le réglage des valeurs des seuils utilisés pour transformer les mesures d'intérêt en contrainte. Il n'existe pas de méthode permettant réellement d'aider l'utilisateur lors de cette étape, celui-ci doit alors procéder par tâtonnement en fonction de ses objectifs d'analyse mais aussi en fonction des temps de calcul de l'algorithme pour un seuil donné. Or, le temps de calcul de l'extraction est difficilement prévisible ce qui constitue une difficulté importante à cette étape. Pour contourner ce problème, des travaux ont proposé d'extraire les  $k$  meilleurs motifs (Ke et al., 2009, Wang et al., 2005) ayant les valeurs les plus élevées sur la mesure d'intérêt, ou les motifs les meilleurs selon plusieurs mesures d'intérêt (Ugarte et al., 2017).

**Prise en compte des connaissances de l'utilisateur :** Enfin, comme dernier verrou, on peut citer la difficulté de retourner des motifs apportant de la connaissance à l'utilisateur. Cela peut être fait en intégrant des feed-backs de l'utilisateur au cours de l'extraction dans les mesures d'intérêt pour guider l'extraction vers les motifs qui l'intéressent (Bendimerad et al., 2019b). On peut aussi intégrer les connaissances antérieures de l'utilisateur, son intérêt subjectif (Bie, 2011), au processus d'extraction pour retourner les motifs apportant de la connaissance par rapport à ses connaissances antérieures, et aussi prendre en compte la connaissance apportée par des motifs déjà extraits au cours de l'extraction (Bendimerad et al., 2019a), ce mécanisme ayant également l'avantage de traiter le problème de redondance des motifs.

Dans notre travail, nous avons abordé principalement trois verrous de la fouille de motifs : (1) l'expressivité des motifs, en travaillant sur un domaine de

motifs peu étudié, les motifs graduels, qui permettent d'extraire des connaissances à partir de données numériques, (2) en veillant à calculer des collections de motifs peu redondantes à l'aide de mécanismes de fermeture, et (3) en nous confrontant au verrou du passage à l'échelle afin de pouvoir utiliser nos algorithmes sur des jeux de données réels. Dans la suite, nous présentons les travaux de l'état de l'art sur la fouille de motifs graduels.

### 1.3 La fouille de motifs graduels

L'étude des corrélations dans des données numériques est un vecteur de connaissance essentiel permettant d'identifier des quantités qui varient conjointement. Ainsi, des domaines de motifs adaptés aux données numériques et basés sur l'étude de corrélations entre attributs ont été développés. L'étude des co-variations entre attributs dans des données numériques ou ordinales a été étudiée sous plusieurs vocables : les itemsets corrélés sur les rangs (Calders et al., 2006), les dépendances graduelles (Hüllermeier, 2002), les itemsets graduels (Do et al., 2010, 2015) ou les motifs de co-variation (Prado et al., 2013).

Pour mesurer l'intérêt des itemsets graduels, Do et al. (2010, 2015) utilisent une mesure de support basée sur la longueur du plus long chemin entre deux objets, un chemin étant un ordonnancement d'objets selon leurs valeurs sur les items de l'itemset.

**Definition 1** (Itemset graduel). Soit  $\mathcal{D}$  une base de données numériques sur  $\mathcal{A}$  définie par  $\mathcal{D} = \{o \in \mathcal{O} \mid o = (A_1(o), \dots, A_m(o)), A_k(o) \in \mathbf{Dom}(A_k), \forall k = 1 \dots m\}$ . Le langage des itemsets graduels utilise l'ensemble  $S = \{+, -\}$  pour désigner le sens des variations de chacun des attributs du motif : + signifie une variation positive et - une variation négative. Un motif de co-variation  $\varphi$  est un sous-ensemble d'attributs de  $\mathcal{A}$  où chaque attribut est signé :

$$\mathcal{L} = \{\varphi \mid \varphi \subseteq \mathcal{A} \times S\}.$$

Ainsi, chaque élément de  $\varphi$  est de la forme  $p^+$  ou  $p^-$  avec  $p \in \mathcal{A}$ .

L'extension d'un itemset graduel est définie par la plus longue liste d'objets ordonnés selon l'itemset graduel :

$$ext(\varphi) = \langle o_{i_1}, \dots, o_{i_k} \rangle$$

tel que  $\forall j = 1 \dots k, o_{i_j} \in \mathcal{O}$  et  $\forall j = 1 \dots k - 1, \forall p^s \in \varphi, p(o_{i_j}) <_s p(o_{i_{j+1}})$  avec  $<_+ \equiv <$  et  $<_- \equiv >$ .

**Exemple 1.** L'extension de  $A^+B^-$  sur l'exemple de la Table 1.3 est  $\langle o_3, o_2, o_4 \rangle$ .

Cette mesure présente certains inconvénients que ce soit du point de vue sémantique ou du point de vue du calcul. En effet, du point de vue du calcul,

l'extension d'un itemset graduel nécessite de calculer le plus long ordre partiel cohérent avec l'itemset. Cela revient à calculer le plus long chemin dans le graphe représentant l'ordre induit sur les objets par l'itemset graduel. En général, ce problème est NP-difficile, mais ici, comme le graphe représente un ordre partiel, il s'agit par définition d'un DAG (graphe acyclique orienté), de sorte que le chemin le plus long peut être trouvé en temps linéaire à l'aide d'un algorithme de tri topologique. Du point de vue de la sémantique, la séquence supportant le motif est difficilement interprétable, n'ayant pas par exemple de fondement statistique.

Calders et al. (2006) propose d'utiliser certaines mesures statistiques bien établies pour évaluer les corrélations comme mesure d'intérêt pour la fouille de motifs. Trois types de corrélations statistiques sont considérées : la corrélation de Pearson, le  $\tau$  de Kendall et la corrélation de rang de Spearman. La corrélation de Pearson est la mesure la plus utilisée, mais elle nécessite des attributs numériques de type continu (ou encore à échelle d'intervalle). De plus, elle est basée sur des hypothèses fortes (les deux attributs doivent être distribués selon une loi Normale, avec une relation linéaire et homoscedastique) qui ne sont généralement pas satisfaites dans la pratique. Les mesures de corrélation de rang (par exemple, la corrélation de rang  $\tau$  de Kendall ou celle de Spearman) sont mieux adaptées car elles ne sont pas basées sur les hypothèses mentionnées ci-dessus. La principale différence entre ces deux mesures est que le  $\tau$  de Kendall est un test non-paramétrique basé sur le nombre de paires triées (par exemple le premier élément de la paire est plus petit que le second) de la même manière par les deux attributs, tandis que la formule de corrélation de Spearman est basée sur la différences des rangs d'un objet dans l'ordre global des objets engendré par les variables numériques. Contrairement au coefficient de Spearman, la mesure  $\tau$  de Kendall est facile à interpréter. Cette approche a été généralisée par Prado et al. (2013) à un nombre quelconque d'attributs étant corrélés positivement ou négativement. Regardons formellement comment ce langage est défini.

**Definition 2** (Langage des motifs de co-variation). *Soit  $\mathcal{D}$  une base de données numériques,  $\mathcal{D} = \{o \in \mathcal{O} \mid o = (A_1(o), \dots, A_m(o)), A_k(o) \in \mathbf{Dom}(A_k), \forall k = 1 \dots m\}$ . Le langage des motifs de co-variation utilise l'ensemble  $S = \{+, -\}$  pour désigner le sens des variations de chacun des attributs du motif : + signifie une variation positive et - une variation négative. Un motif de co-variation  $\varphi$  est un sous-ensemble d'attributs de  $\mathcal{A}$  où chaque attribut est signé :*

$$\mathcal{L} = \{\varphi \mid \varphi \subseteq \mathcal{A} \times S\}.$$

*Ainsi, chaque élément de  $\varphi$  est de la forme  $p^+$  ou  $p^-$  avec  $p \in \mathcal{A}$ .*

*L'extension d'un motif de co-variation est l'ensemble des couples d'objets de  $\mathcal{D}$  ordonnés selon  $\varphi$ , c'est-à-dire :*

$$\text{ext}(\varphi) = \{(o_i, o_j) \in \mathcal{O}^2, o_i \neq o_j \mid \forall p^s \in \varphi, p(o_i) <_s p(o_j)\}$$



avec  $\langle_+ \equiv \langle$  et  $\langle_- \equiv \rangle$ .

Par exemple, un motif de co-variation sur les données de la Table 1.3 est  $A^+B^-$ . L'extension de ce motif est  $ext(A^+B^-) = \{(o_1, o_4), (o_2, o_4), (o_3, o_2), (o_3, o_4), (o_5, o_4)\}$ .

	A	B	C
$o_1$	1	2	10
$o_2$	3	3	30
$o_3$	2	4	11
$o_4$	4	-1	20
$o_5$	1	0	15

TABLE 1.3 – Exemple de base de données numériques.

**Definition 3** (Mesure d'intérêt pour le langage des motifs de co-variation). *Comme mesure d'intérêt, nous considérons le  $\tau$  de Kendall qui est le nombre de paires d'objets concordantes selon le motif  $\varphi$  moins le nombre de paires discordantes, divisé par le nombre total de paires. En observant que le nombre de paires total n'est autre que  $ext(\emptyset)$ , on a :*

$$\tau(\varphi) = \frac{|ext(\varphi)| - (|ext(\emptyset)| - |ext(\varphi)|)}{|ext(\emptyset)|} = \frac{2 \times |ext(\varphi)| - |ext(\emptyset)|}{|ext(\emptyset)|}$$

**Exemple 2.** Dans l'exemple de la Table 1.3, si l'on considère le motif de co-variation  $\varphi = \{A^+, B^-\}$ , on a  $|ext(\emptyset)| = 10$  et  $\tau(\varphi) = \frac{2 \times 5 - 10}{10} = 0$ .

Rechercher de fortes corrélations peut être d'autant plus intéressant que celles-ci apparaissent localement dans les données : par exemple, les variables âge et revenu peuvent ne pas être corrélées globalement dans les données, mais le devenir si on considère une sous catégorie de la population, comme par exemple les cadres. Pour identifier des motifs qui sont corrélés localement dans les données, nous considérons la découverte de sous-groupes, dont les principaux résultats sont présentés ci-dessous.

## 1.4 Découverte de sous-groupes

Les motifs définis par des contraintes peuvent également être utilisés dans un contexte supervisé où l'on cherche à caractériser un phénomène représenté par une variable. Cette tâche a été introduite dans (Piatetsky-Shapiro et al., 1996) et formalisée sous le terme de découverte de sous-groupes dans (Wrobel, 1997).

Cette tâche vise à rechercher des motifs tels que la distribution d'une variable cible soit différente sur le motif, de celle observée sur l'ensemble des

données. La capacité d'un sous-groupe à discriminer les modalités de la variable cible est évaluée par une mesure de qualité. Il existe une grande variété de mesures dans la littérature (le F-Score, le coefficient de Jaccard, la divergence de Kullback Leibler pondérée – *WKL*) (Belfodil, 2019). Notamment, l'intérêt d'un sous-groupe peut-être évalué par la mesure *WRAcc* (Weighted Relative Accuracy) (Lavrac et al., 1999),

**Definition 4** (*WRAcc*). *Soit  $\varphi$  un sous-groupe d'extension  $ext(\varphi)$  et Cible une variable cible à valeur booléenne sur le même domaine que  $ext$ . On peut alors considérer la matrice de contingence suivante :*

		Cible		$e$
		$T$	$F$	
sous-groupe	$ext$	$TT$	$TF$	$C$
	$-ext$	$FT$	$FF$	

La mesure de *WRAcc* est alors

$$WRAcc(\varphi, Cible) = \frac{e}{E} \left( \frac{TT}{e} - \frac{C}{E} \right)$$

Le principe est que la précision du sous-groupe (la valeur  $\frac{TT}{e}$ ) doit être considérée relativement à la précision obtenue si l'on choisit de manière indépendante la classe  $T$  de la variable cible (la valeur  $\frac{C}{E}$ ), et l'on pondère cette valeur par la taille du sous-groupe ( $\frac{e}{E}$ ) pour s'affranchir des petits sous-groupes qui ont plus facilement une forte précision.

L'exploration de modèle exceptionnel (Duivesteijn et al., 2016, Leman et al., 2008) généralise la tâche de découverte de sous-groupes en considérant des concepts cibles plus complexes que la distribution des modalités de la variable cible sur les sous-groupes. Dans ce paradigme, on utilise des modèles plus sophistiqués sur la variable cible que la simple fréquence des modalités. On peut même considérer plusieurs variables cibles simultanément. EMM vise à trouver des sous-groupes d'objets pour lesquels le modèle induit par rapport aux variables cibles s'écarte considérablement du modèle induit sur l'ensemble du jeu de données. De nombreux modèles EMM définis sur un ensemble de variables cibles ont été proposés : un modèle de régression (Duivesteijn et al., 2012), un modèle de classification (Duivesteijn and Thaele, 2014), un modèle de préférences parmi un ensemble de cibles (de Sá et al., 2018), un réseau bayésien (Duivesteijn et al., 2010), et aussi un modèle de corrélation (Duivesteijn et al., 2016) entre deux variables cibles. Dans ce dernier travail, l'interaction entre les deux attributs cibles numériques est modélisée par plusieurs mesures de corrélation (le coefficient de corrélation de Pearson et le  $\tau$  de Kendall). La principale limite est que les deux attributs numériques cibles sont

spécifiés a priori et que l'approche ne fonctionne pas avec un ensemble arbitraire d'attributs numériques.

Un exemple concret d'EMM est donné dans (Duivesteijn et al., 2012) où l'objectif est d'étudier la relation entre le prix et la demande d'un produit. La loi économique de la demande stipule que (toutes choses égales par ailleurs), si le prix d'un produit augmente, la demande pour le produit diminuera. Dans un modèle de régression, cela entraînerait une pente négative. Toutefois, dans certaines conditions, on peut observer une tendance à acheter davantage d'un produit lorsque son prix augmente (Jensen and Miller, 2008). Par conséquent, dans ces cas exceptionnels, la pente de la régression est positive. Ces cas exceptionnels peuvent être identifiés en utilisant le cadre EMM et ainsi rechercher des restrictions d'attributs identifiant des sous-groupes possédant un modèle de régression de variables cible très différent du modèle de régression global des cibles sur l'ensemble du jeu de données, comme illustré par la Figure 1.1.

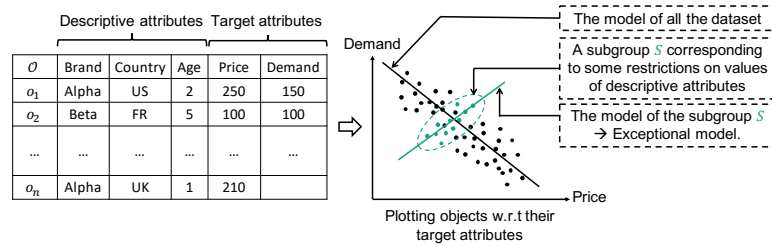


FIGURE 1.1 – Exemple d'EMM avec un modèle de régression sur deux attributs cibles (figure issue de (Bendimerad, 2019)).

## 1.5 Discussion

La découverte de sous-groupes (**SD**) et sa généralisation **EMM** sont deux domaines en plein essor et de nombreux algorithmes ont été développés ces dernières années. En effet, la définition d'un modèle exceptionnel de données permet d'accroître l'interprétation des motifs obtenus par les algorithmes de fouille de données, en les contrastant à une ou plusieurs variables cibles qui représentent les connaissances a priori de l'utilisateur. Dans cette thèse notre contribution est la définition d'une méthode d'EMM autour de la fouille de motifs de co-variation. Cela permet d'avoir une méthode d'extraction de connaissances dans des données hétérogènes où l'on recherche des corrélations entre des attributs numériques de telle sorte que ces corrélations soient exceptionnelles sur un sous-groupe des données décrit par des attributs de type numérique ou nominal. Nous proposons un algorithme correct et complet, mais

aussi une approche par échantillonnage pour réaliser des extractions dans des contextes difficiles.



## Chapitre 2

# Fouille de sous-groupes corrélés

### 2.1 Introduction

L'analyse exploratoire de données (AED) est considérée par les scientifiques comme un domaine de recherche important depuis son introduction (Tukey, 1977). Parmi les différentes techniques d'AED qui visent à comprendre des ensembles de données et à découvrir leur structure, la découverte de sous-groupes (Klosgen, 1996, Wrobel, 1997) est une tâche générique visant à trouver des régions dans les données qui se détachent par rapport à une variable cible. Beaucoup d'autres tâches de fouille ont des objectifs similaires, comme le calcul des motifs émergents (Dong and Li, 1999), des règles significatives (Terada et al., 2013), des ensembles de contraste (Bay and Pazzani, 2001) ou de règles d'association de classification (Liu et al., 1998). Cependant, parmi ces différentes tâches, la découverte de sous-groupes est considérée comme l'approche la plus générique. En particulier, elle est agnostique des données et du domaine de motifs. Par exemple, les sous-groupes peuvent être définis par la conjonction de conditions sur des attributs symboliques (Lavrač et al., 2004) ou numériques (Atzmüller and Puppe, 2006, Grosskreutz and Rüping, 2009) ainsi que des séquences (Grosskreutz et al., 2013). En outre, la cible unique peut être discrète ou numérique (Lemmerich et al., 2016). La fouille de modèles exceptionnels (Leman et al., 2008) étend la découverte de sous-groupes en offrant la possibilité de gérer des cibles complexes, par exemple plusieurs attributs discrets (Bosc et al., 2016, Duivesteijn et al., 2010, 2016, van Leeuwen and Knobbe, 2012), des graphes (Bendimerad et al., 2016, 2017, Kaytoue et al., 2017), des comparaisons par paires (Belfodil et al., 2017), des préférences (de Sá et al., 2016), ou deux cibles numériques (Downar and Duivesteijn, 2017).

Dans ce chapitre, nous présentons notre première contribution qui consiste à rechercher des sous-groupes ayant une très forte corrélation sur un sous-

ensemble (identifié inductivement) d'attributs cibles. Les attributs cibles sont des attributs à valeur numérique. Les sous-groupes sont identifiés par des restrictions sur des attributs descriptifs, à valeur symbolique ou numérique. Nous voulons donc trouver des sous-groupes pour lesquels on observe une forte corrélation sur certains attributs cibles. La découverte de sous-groupes dans le cas où l'on a deux cibles numériques a déjà été définie (Downar and Duijvesteijn, 2017). Cette approche permet de découvrir des sous-ensembles d'objets du jeu de données dont la corrélation sur les deux attributs numériques diffère grandement de celle observée sur l'ensemble des observations. Cependant, cette approche ne permet pas une exploration non supervisée, puisque les deux cibles numériques sont définies avant la recherche de sous-groupes. C'est une limitation lorsque les données contiennent plusieurs attributs numériques dont les corrélations peuvent apporter des informations cruciales à l'analyste.

Pour permettre une exploration non supervisée, nous proposons de rechercher des sous-ensembles arbitraires de cibles numériques dont la corrélation est exceptionnelle pour un sous-groupe trouvé automatiquement. Ainsi, nous introduisons le nouveau problème de *découverte de sous-groupes à forte corrélation de rang sur un sous-ensemble de cibles numériques*. Le sous-groupe est identifié par  $D$ , un ensemble de conditions ou restrictions sur les attributs descriptifs des objets. Ces attributs descriptifs peuvent être symboliques ou numériques. L'ensemble  $C$  désigne le sous-ensemble d'attributs cibles sur lequel les corrélations de rang (positives ou négatives) sont évaluées par une généralisation du  $\tau$  de Kendall. Un sous-groupe  $(C, D)$  est d'autant plus intéressant que la corrélation des attributs de  $C$  est plus forte sur les objets satisfaisant les propriétés de  $D$ .

L'intérêt de ce modèle est illustré sur le jeu de données de la Table 2.1. Les objets  $o_i$  représentent des baux de magasins, c'est-à-dire des contrats de location. Les attributs décrivant ces contrats sont les dates de début et de fin de contrat, les coordonnées GPS  $(x, y)$  du magasin et le type de commerce (boulangerie, boucherie, pharmacie, etc.). Les attributs cibles décrivent les changements dans l'environnement géographique de chaque magasin entre le début et la fin du bail. Ces attributs désignent la variation du nombre de magasins d'un type donné dans un rayon de 200 m du magasin. Un attribut supplémentaire indique la durée du bail de magasin (durée de vie). Sur cet ensemble de données, le motif  $LifeTime^+, Bakery^-, Boucherie^+$  a une plus forte corrélation sur le sous-groupe décrit par  $x \in [1, 3], y \in [1, 3], cat \in \{Bakery\}$  ( $\tau(C, \sigma(D)) = 6/6$ ) que sur l'ensemble du jeu de données ( $\tau(C, \mathcal{O}) = 6/21$ ) et indique que les boulangeries dans la région  $[1, 3] \times [1, 3]$  ont une longévité qui est corrélée à l'augmentation du nombre de boucheries dans leur voisinage et sont anti-corrélées avec leur nombre de concurrents (augmentation du nombre de boulangeries à proximité).

Pour identifier de tels modèles, nous commençons par formaliser le problème de découverte de sous-groupes corrélés. Ensuite, nous définissons un algorithme qui exploite certaines propriétés d'élagage basées sur deux bornes

Id	Attributs descriptifs					Attributs cibles			
	Date début	Date fin	x	y	Type	Durée de vie	Pharmacie	Boulangerie	Boucherie
o1	1991	2000	1	3	Boulangerie	10	5	7	1
o2	2000	2013	3	3	Boulangerie	13	7	5	3
o3	1975	1992	2	1	Boulangerie	18	3	2	7
o4	1986	2005	2	3	Boulangerie	20	9	1	9
o5	1999	2008	2	3	Pharmacie	10	7	2	2
o6	1995	2014	5	3	Boucherie	20	8	3	1
o7	1980	1999	4	4	Boulangerie	20	6	3	1

TABLE 2.1 – Exemple de jeu de données.

supérieures et une propriété de fermeture. Enfin, une étude empirique sur plusieurs jeux de données montre l'efficacité de l'approche proposée.

## 2.2 Sous-groupes corrélés

Un sous-groupe corrélé est un sous-ensemble d'attributs cibles fortement corrélés pour un sous-groupe de données. Cet ensemble d'objets est identifié par une description, c'est-à-dire par une conjonction de conditions sur les attributs descriptifs. Plus formellement, un tel motif est composé de deux ensembles : un ensemble d'attributs corrélés positivement ou négativement, et une conjonction de restrictions sur certains attributs descriptifs. Les objets qui satisfont la description constituent le sous-groupe de données sur lequel les corrélations sont évaluées. Nous définissons d'abord comment évaluer une corrélation positive et négative sur plusieurs (supérieur ou égal à 2) attributs numériques. Ensuite, nous introduisons le langage de motifs de sous-groupes corrélés sur les rangs.

Nous avons besoin de définir quelques notations. L'ensemble de données est noté  $\mathbb{D} = (\mathcal{O}, \mathcal{C}, \mathcal{R})$  où  $\mathcal{O}$  est un ensemble d'objets de taille  $n$ ,  $\mathcal{C}$  un ensemble d'attributs numériques qui associe à chaque objet une valeur numérique ( $\forall c \in \mathcal{C}, c : \mathcal{O} \rightarrow \mathbb{R}$ ), appelés attributs cibles sur lesquelles nous allons rechercher des corrélations.  $\mathcal{R}$  est l'ensemble des attributs descriptifs qui sont soit numériques soit symboliques, et dont les restrictions de leurs domaines de valeur identifient les sous-groupes.

### 2.2.1 Évaluer la corrélation d'un ensemble d'attributs

Les mesures de corrélation évaluent la force de l'association entre deux attributs ainsi que la direction de la relation. La valeur absolue du coefficient de corrélation, qui varie entre 0 et 1, évalue l'intensité, tandis que la direction de la relation est indiquée par le signe du coefficient. Une valeur positive décrit une corrélation positive – par exemple, plus le nombre d'années de travail est élevé, plus le salaire est élevé – alors qu'une valeur négative indique une relation inversée – par exemple, plus le nombre d'absences est élevé, plus les



notes baissent. Trois types de corrélations statistiques existent : la corrélation de Pearson, le  $\tau$  de Kendall et les corrélations de rang de Spearman.

La corrélation de Pearson est la mesure la plus utilisée, mais elle nécessite des attributs numériques continus. De plus, elle est basée sur des hypothèses fortes (les deux attributs doivent être distribués selon une loi Normale, avec une relation linéaire et homoscédastique) qui ne sont généralement pas satisfaites dans la pratique, en particulier lorsque des sous-groupes de données sont considérés.

Les mesures de corrélation de rang (par exemple, corrélation de rang  $\tau$  de Kendall, Corrélation de rang de Spearman) sont mieux adaptées car elles ne sont pas basées sur les hypothèses mentionnées ci-dessus. La principale différence entre ces deux mesures est que le  $\tau$  de Kendall est un test non-paramétrique basé sur le nombre de paires triées de la même manière par les deux attributs, tandis que la formule de corrélation de Spearman est basée sur la différences des rangs. Contrairement au coefficient de Spearman, la mesure  $\tau$  de Kendall est facile à interpréter et peut être étendue facilement (Calders et al., 2006). La corrélation  $\tau$  de Kendall sur deux attributs numériques  $a, b \in \mathcal{C}$  est définie par

$$\tau(ab) = \frac{|\eta(ab)| - |\overline{\eta(ab)}|}{\frac{1}{2}n(n-1)}$$

avec  $\eta(ab) = \{(o_i, o_j) \in \mathcal{O}^2 \mid ((a(o_i) < a(o_j)) \text{ et } (b(o_i) < b(o_j))) \text{ ou } ((a(o_i) > a(o_j)) \text{ et } (b(o_i) > b(o_j)))\}$ , c'est-à-dire le nombre de paires concordantes d'objets.  $\overline{\eta(ab)} = \mathcal{O}^2 \setminus \eta(ab)$  est l'ensemble des paires discordantes. Ainsi,  $|\eta(ab)| + |\overline{\eta(ab)}| = \frac{1}{2}n(n-1) = \binom{n}{2}$ . Ainsi, le dénominateur normalise la mesure. La mesure peut être réécrite comme :

$$\tau = \frac{|\eta(ab)| - \left( \binom{n}{2} - |\eta(ab)| \right)}{\binom{n}{2}} = \frac{2|\eta(ab)|}{\binom{n}{2}} - 1$$

Nous pouvons généraliser cette mesure pour plus de 2 attributs. On dit qu'une paire d'objets  $(o_i, o_j)$  est concordante sur les attributs  $a, b$  et  $c \in \mathcal{C}$  si et seulement si :

$$\begin{aligned} & ((a(o_i) < a(o_j)) \text{ et } (b(o_i) < b(o_j)) \text{ et } (c(o_i) < c(o_j))) \\ \text{or } & ((a(o_i) > a(o_j)) \text{ et } (b(o_i) > b(o_j)) \text{ et } (c(o_i) > c(o_j))) \end{aligned}$$

Les paires discordantes, c'est-à-dire  $\overline{\eta(abc)}$ , sont de 3 types différents définis par différentes combinaisons où deux attributs sont ordonnés dans un sens et le troisième dans l'autre sens. Ces différents cas sont représentés dans la Figure 2.1.

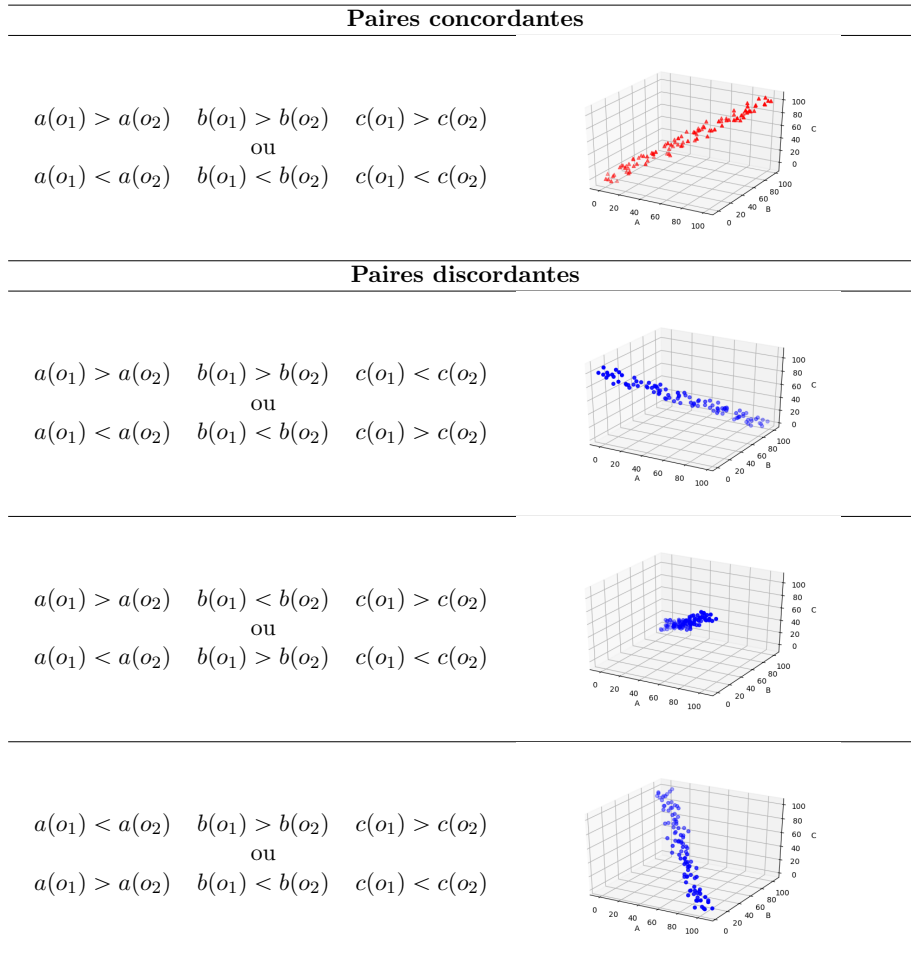


FIGURE 2.1 – Paires concordantes et les 3 types de paires discordantes pour  $abc$ .

Ainsi, ces paires décrivent plusieurs phénomènes et un ensemble  $\overline{\eta(abc)}$  dont le cardinal est grand n'est pas équivalent à une forte corrélation dans le sens opposé. Une corrélation négative se produit seulement si l'un des différents types de paires discordantes est sur-représenté. Pour faire face à cette situation, la généralisation de la corrélation basée sur les rangs  $\tau$  de Kendall à plus de deux attributs nécessite la distinction entre ces différents cas de paires

discordantes. Ceci est fait par l'ajout d'un signe ( $\{+, -\}$ ) à chaque attribut. Par convention, le signe du premier attribut est défini comme étant  $+$  ce qui permet de s'affranchir des motifs équivalents par symétrie. En effet, les motifs  $a^+b^+c^+$  et  $a^-b^-c^-$  représentent la même information.

**Definition 5** (Motifs corrélés sur les rangs). *Un motif corrélié sur les rangs  $C$  est un ensemble d'au moins deux attributs signés de  $\mathcal{C}$ . Il est défini par  $C = \{(a, s) \mid a \in \mathcal{C} \text{ et } s \in \{-, +\}\}$  avec le signe  $+$  pour le premier attribut dans l'ordre canonique par convention. Etant donné un ensemble d'objets  $O \subseteq \mathcal{O}$ , l'ensemble des paires d'objets de  $O$  concordantes avec un motif  $C$  est défini par  $\eta(C, O) = \{(o_i, o_j) \in O \times O \mid \nu_C(o_i, o_j)\}$  avec*

$$\nu_C(o_i, o_j) \equiv \bigwedge_{(a,s) \in C} (a(o_i) <_s a(o_j))$$

et  $<_s$  est la relation binaire classique sur  $\mathbb{R} <$  quand  $s = +$ , et  $>$  quand  $s = -$ . La mesure de corrélation  $\tau$  de Kendall généralisée à un nombre quelconque d'attributs est alors :

$$(2.1) \quad \tau(C, O) = \frac{|\eta(C, O)|}{N(O)} \text{ avec } N(O) = \binom{|O|}{2}.$$

### 2.2.2 Corrélations contextualisés

La tâche de fouille considérée consiste à trouver des sous-groupes d'objets  $O$  pour lesquels la corrélation  $\tau$  des attributs de  $C$  est plus forte que sur l'ensemble des objets  $\mathcal{O}$ . Ces sous-groupes sont définis au moyen d'une conjonction de valeurs que les attributs de  $\mathcal{R}$  peuvent prendre. Rappelons que ces attributs descriptifs, ordonnés dans un ordre canonique  $\mathcal{R} = \langle d_1, \dots, d_{|\mathcal{R}|} \rangle$ , peuvent être numériques ou symboliques, et les conditions sont des restrictions sur leurs domaines de valeur. Par exemple, considérons trois attributs : AGE avec  $\mathbf{Dom}(AGE) = [25, 60]$ , SEXE avec  $\mathbf{Dom}(SEXE) = \{femme, homme\}$  et ODEUR avec  $\mathbf{Dom}(ODEUR) = \{piquant, fleur, menthe\}$ . Un exemple de sous-groupe est défini par  $AGE \geq 30 \wedge SEXE = femme$  qui est équivalent à  $AGE \in [30, 60]$  et  $SEXE \in \{femme\}$  et  $ODEUR \in \{piquant, fleur, menthe\}$ .

**Definition 6** (Sous-groupe et support). *Un sous-groupe d'objets est défini par la description  $D = \langle f_1, \dots, f_{|\mathcal{R}|} \rangle$  et chaque  $f_\ell$  est une restriction sur le domaine d'attributs  $d_\ell \in \mathcal{R}$ . En fonction du type de  $d_\ell$ , la restriction se fait comme suit :*

- Si  $d_\ell$  est symbolique,  $f_\ell = \{v\}$  avec  $v \in \mathbf{Dom}(d_\ell)$ , ou  $f_\ell = \mathbf{Dom}(d_\ell)$
- Si  $d_\ell$  est numérique,  $f_\ell = [v, w]$  avec  $v, w \in \mathbf{Dom}(d_\ell)$  et  $v < w$ .

*L'ensemble des objets  $\mathcal{O}$  qui satisfait  $D$  est le support de la description  $D = \langle d_1, \dots, d_{|\mathcal{R}|} \rangle$  :*

$$\sigma(D) = \{o_i \in \mathcal{O} \mid d_\ell(o_i) \in f_\ell, \forall \ell = 1 \dots |\mathcal{R}|\}$$

Nous avons maintenant tous les ingrédients pour définir les sous-groupes corrélés.

**Definition 7** (Sous-groupes corrélés). *Un sous-groupe corrélé est une paire  $(C, D)$  avec  $C$  un ensemble d'attributs cibles de  $\mathcal{C}$  et  $D$  une description sur les attributs de  $\mathcal{R}$  qui définit un sous-groupe. La corrélation de  $C$  sur  $\sigma(D)$  est mesurée par  $\tau(C, \sigma(D))$ .*

Certains sous-groupes corrélés peuvent être considérés comme équivalents car ils partagent le même support. Ces motifs de la même classe d'équivalence peuvent être filtrés par des opérateurs de fermeture.

**Definition 8** (Opérateurs de fermeture). *Soient  $H$  et  $M$  deux fonctions qui associent un sous-groupe corrélé à son ensemble support de paires d'objets et réciproquement :*

1.  $H(C, D) = \eta(C, \sigma(D))$  comme défini ci-dessus.
2. Étant donné un ensemble de paires d'objets  $X \subseteq \mathcal{O} \times \mathcal{O}$ ,  $M(X)$  est le sous-groupe corrélé  $(C', D')$  défini par
  - $C' = \{(a, s) \in \mathcal{C} \times \{+, -\} \mid \forall (x, y) \in X, a(x) <_s a(y)\}$
  - $D' = \langle f'_1, \dots, f'_{|\mathcal{R}|} \rangle$  avec
    - $f'_\ell = \begin{cases} v, & \text{si } \forall (o_j, o_k) \in X, d_\ell(o_j) = d_\ell(o_k) = v \\ \mathbf{Dom}(d_\ell) & \text{sinon} \end{cases}$  et  $d_\ell$  est symbolique,
    - $f'_\ell = [v, w]$  avec  $\begin{cases} v = \min_{(x,y) \in X \cup (y,x) \in X} d_\ell(x) \\ w = \max_{(x,y) \in X \cup (y,x) \in X} d_\ell(x) \end{cases}$  si  $d_\ell$  est numérique.

L'opérateur de fermeture est  $\mathbf{Closure}(C, D) = M(H(D))$  et le sous-groupe corrélé est fermé ssi  $\mathbf{Closure}(C, D) = (C, D)$ .

Les sous-groupes corrélés et fermés qui capturent le mieux les corrélations locales dans les données doivent avoir une valeur de corrélation élevée par rapport à ce qui est observé sur l'ensemble des données. Une mesure appropriée de ce type de phénomène est la WRAcc (Weighted Relative Accuracy) (Lavraç et al., 1999). Cette mesure prend en compte l'augmentation de la précision par rapport à la corrélation par défaut (c'est-à-dire la corrélation sur l'ensemble des données). En outre, cette mesure réalise un compromis entre la précision relative et la généralité du sous-groupe (c'est-à-dire le nombre de paires couvertes).

**Definition 9** (WRAcc). *L'exceptionnalité d'un sous-groupe corrélé  $(C, D)$  est évaluée en utilisant la mesure Wracc, définie par :*

$$(2.2) \quad \mathbf{WRAcc}(C, D) = \frac{N(\sigma(D))}{N(\mathcal{O})} (\tau(C, \sigma(D)) - \tau(C, \mathcal{O}))$$

Notre tâche de fouille peut maintenant être entièrement exprimée comme le problème suivant :

**Problème 1** (Fouille de sous-groupes corrélés et fermés). *Pour les seuils  $\alpha$  et  $\beta$  fixés, soit  $\mathcal{S}$  la collection de sous-groupes corrélés et fermés définie comme :*

$$\begin{aligned} \forall (C, D) \in \mathcal{S} \\ \mathbf{Closure}(C, D) = (C, D) \\ \sigma(D) \geq \alpha \\ \tau(C, \sigma(D)) \geq \beta \\ \mathbf{WRAcc}(C, D) \geq 0 \end{aligned}$$

Nous voulons également un ensemble concis de motifs inattendus qui maximisent la mesure  $\mathbf{WRAcc}$ . Cependant, il est bien connu (Xin et al., 2006) qu'en général, ces motifs sont très redondants, certains étant une simple variation des autres. Par conséquent, le deuxième problème que nous considérons est défini ci-dessous :

**Problème 2** (Fouille des top- $k$  sous-groupes fermés, corrélés, exceptionnels et diversifiés). *Soit  $\mathcal{K}$  un sous ensemble de  $\mathcal{S}$  contenant les top- $k$  sous-groupes fermés et corrélés par rapport à la mesure  $\mathbf{WRAcc}$  et qui sont aussi diversifiés. La diversité entre deux motifs est évaluée par la mesure de Jaccard, définie dans notre contexte par :*

$$(2.3) \quad \mathbf{Jaccard}(C, D), (C', D') = \frac{|\eta(C, \sigma(D)) \cap \eta(C', \sigma(D'))|}{|\eta(C, \sigma(D)) \cup \eta(C', \sigma(D'))|}$$

Compte tenu d'un seuil  $\delta$ , l'ensemble  $\mathcal{K}$  des  $k$  sous-groupes les plus diversifiés, est défini par :

1.  $\forall (C, D), (C', D') \in \mathcal{K}^2, \mathbf{Jaccard}((C, D), (C', D')) \leq \delta$
2.  $\forall (C, D) \in \mathcal{K}$  et  $\forall (C', D') \in \mathcal{S}$  tel que  $\mathbf{Jaccard}((C, D), (C', D')) > \delta$ , on a  $\mathbf{WRAcc}(C, D) \geq \mathbf{WRAcc}(C', D')$ .
3.  $|\mathcal{K}| = k$

Le point 2 est très difficile à assurer incrémentalement. Cela est dû à la non-transitivité de la mesure de similarité. En effet, si un sous-groupe corrélé  $(C, D)$  est exclu de  $\mathcal{K}$  par un modèle similaire  $(C', D')$  de meilleure qualité, il n'y a aucune garantie que d'autres sous-groupes corrélés exclus par similitude avec  $(C, D)$  sont également similaires à  $(C', D')$ . Ainsi, cette condition est relaxée :

$$\begin{aligned} \forall (C, D) \in \mathcal{K}, \exists (C', D') \in \mathcal{S} \text{ tel que } \mathbf{Jaccard}((C, D), (C', D')) > \delta \\ \text{et } \mathbf{WRAcc}(C, D) \geq \mathbf{WRAcc}(C', D') \end{aligned}$$

### 2.3 Algorithme

Nous énumérons récursivement les sous-groupes corrélés à l'aide d'une recherche en profondeur. L'algorithme LoCoM est présenté dans Algorithme 1.

Étant donné le motif  $(C, D)$  actuellement exploré, LoCoM retourne toutes les spécialisations de ce motif qui sont des sous-groupes corrélés exceptionnels. Lors du premier appel de la fonction, le motif  $(C, D)$  est initialisé à  $(\emptyset, M(\mathcal{O}))$ . L'ordre de spécialisation considéré est  $\leq$ , l'ordre partiel défini par  $(C, D) \leq (C', D')$  si et seulement si  $C \subseteq C'$  et, pour  $D = \langle f_1, \dots, f_{|\mathcal{R}|} \rangle$  et  $D' = \langle f'_1, \dots, f'_{|\mathcal{R}|} \rangle$ ,  $f'_\ell \subseteq f_\ell, \forall \ell = 1 \dots |\mathcal{R}|$ . De plus, pour éviter de générer le même motif plusieurs fois, nous utilisons des ordres arbitraires,  $\ll_C$  sur  $\mathcal{C}$  et  $\ll_{\mathcal{R}}$  sur  $\mathcal{R}$ . L'ordre canonique entre les motifs est donc défini par  $\ll$  avec :

$$(C, D) \ll (C', D') \Leftrightarrow \begin{aligned} & (C, D) \leq (C', D') \\ & \text{et } \forall a \in C, \forall a' \in C' \setminus C, \quad a \ll_C a' \\ & \text{et } : \operatorname{argmax}_\ell f_\ell \neq \mathbf{Dom}(d_\ell) \ll_{\mathcal{R}} \operatorname{argmin}_\ell f'_\ell \neq \mathbf{Dom}(d'_\ell) \neq f'_\ell \end{aligned}$$

Le deuxième paramètre de LoCoM,  $X$ , contient les attributs signés qui peuvent être ajoutés à  $C$  et les valeurs d'attributs qui peuvent être enlevées à  $D$ . Au début,  $X$  est égal à  $\mathcal{C} \times \{+, -\} \cup \{d \in \mathcal{R}_s \times \mathbf{Dom}(d)\} \cup \{(d, [x, y]) \mid d \in \mathcal{R}_n, x = \min_{o_i \in \mathcal{O}} d(o_i), y = \max_{o_i \in \mathcal{O}} d(o_i)\}$ , avec  $\mathcal{R}_s$  (resp.  $\mathcal{R}_n$ ) les attributs symboliques (resp. numériques) de  $\mathcal{R}$ .

Si  $X$  n'est pas vide (Lignes 4 à 37), le sous-groupe corrélé courant est spécialisé soit en ajoutant un attribut signé à  $C$ , soit en réduisant la valeur du domaine d'un attribut de  $\mathcal{R}$ . Si cet attribut est catégorique, son domaine est limité à une seule valeur (Ligne 17). S'il est numérique, deux sous-intervalles peuvent être générés : l'un réduit d'une seule valeur sur la gauche (Ligne 25) et l'autre à droite (Ligne 32). Pour éviter de générer deux fois le même intervalle, la réduction sur la droite n'est autorisée que lorsqu'aucune réduction sur la gauche ne s'est produite précédemment (Kaytoue et al., 2011).

La fonction de fermeture est utilisée pour faire *des sauts* dans le processus d'énumération, et obtenir directement le motif le plus spécifique pour un même ensemble de paires d'objets.

**La fonction Propager**  $(X, (C_c, D_c))$  : elle est utilisée pour éliminer rapidement les candidats de  $X$  qui ne pourront pas satisfaire les seuils. Chaque valeur d'attribut signé  $(a, s)$  dans  $C_c$ , mais aussi  $(a, \bar{s})$ , l'attribut  $a$  associé au signe opposé de  $s$ , sont supprimés de  $X$ . Pour  $(a, v)$  une paire de valeurs d'attribut de  $\mathcal{R}_s \cap D_c$ , toutes les paires  $(a, w)$ , avec  $w \in \mathbf{Dom}(a)$  seront supprimées de  $X$ . Enfin, pour les paires  $(a, [x, y]) \in \mathcal{R}_n \cap D_c$ , l'intervalle de valeurs possibles de  $a$  dans  $X$  est mis à jour à  $[x, y]$ , et la paire est supprimée si  $x = y$ .

Trois techniques d'élagage sont utilisées pour arrêter le processus d'énumération (Ligne 6) : l'anti-monotonie de la mesure support  $\sigma(D)$ , et les deux bornes supérieures, une sur le  $\tau$  de Kendall, et l'autre sur la mesure WRAcc.

**Algorithm 1:** LoCoM( $(C, D)$ ,  $X$ , left)

---

**Input:**  $(C, D)$  le motif en construction,  $X$  l'ensemble des attributs-valeurs de  $\mathcal{C} \cup \mathcal{R}$  à énumérer. *left* : Un tableau de  $|\mathcal{R}_n|$  valeurs booléennes indiquant si les intervalles de l'attribut numérique correspondant ont été réduits sur le côté gauche.

Il y a aussi des variables globales :

- $\beta$ ,  $\alpha$ ,  $\delta$  : les seuils utilisés pour les contraintes
- $\text{minWRAcc}$  : la valeur  $\text{WRAcc}$  minimale des  $k$  motifs actuels

**Output:**  $\mathcal{K}$ , Liste des top- $k$  motifs diversifiés actuels.

```

1 if  $X = \emptyset$  then
2   if  $\tau(C, \sigma(D)) \geq \beta$  et  $\text{WRAcc}(C, D) \geq \text{minWRAcc}$  then
3     |  $\text{minWRAcc} \leftarrow \text{TOPKDIV}(\mathcal{K}, (C, D))$ 
4 else
5   if  $(UB_\tau(C, D) \geq \beta)$  and  $(\sigma(S_D) \geq \alpha)$ 
6     et  $(UB_{\text{WRAcc}}(C, D) \geq \text{minWRAcc})$  then
7     Sois  $(a, v) \in X$ 
8     if  $a \in \mathcal{C}$  then
9       |  $C' \leftarrow C \cup \{(a, v)\}$ 
10      |  $(C_c, D_c) \leftarrow \text{Closure}(C', D)$ 
11      | if  $(C', D) \ll (C_c, D_c)$  then
12        | | LoCoM( $(C_c, D_c)$ , Propager( $X, (C_c, D_c)$ ), left)
13        | | LoCoM( $(C, D)$ ,  $X \setminus \{(a, v)\}$ , left)
14      else
15        //  $a \in \mathcal{R}$ 
16        if ( $a$  est le  $i$ ème attribut symbolique de  $\mathcal{R}$ ) then
17          |  $D' = \langle f_1, \dots, f_{i-1}, \{v\}, f_{i+1}, \dots, f_{|\mathcal{R}|} \rangle$ 
18          |  $(C_c, D_c) \leftarrow \text{Closure}(C, D')$ 
19          | if  $(C, D') \ll (C_c, D_c)$  then
20            | | LoCoM( $(C_c, D_c)$ , Propager( $X, (C_c, D_c)$ ), left)
21            | | LoCoM( $(C, D)$ ,  $X \setminus \{(a, v)\}$ , left)
22          else
23            //  $a$  est le  $i$ ème attribut numérique de  $\mathcal{R}$ 
24            |  $[x, y] \leftarrow v$ 
25            |  $D' = \langle f_1, \dots, f_{i-1}, [x+1, y], f_{i+1}, \dots, f_{|\mathcal{R}|} \rangle$ 
26            |  $(C_c, D_c) \leftarrow \text{Closure}(C, D')$ 
27            | if  $(C, D') \ll (C_c, D_c)$  then
28              | |  $\text{left}[i]' \leftarrow \text{true}$ 
29              | | LoCoM( $(C_c, D_c)$ , Propager( $X, (C_c, D_c)$ ), left')
30            | | LoCoM( $(C, D)$ ,  $X \setminus \{(a, v)\}$ , left)
31            | if  $\text{left}[i] = \text{false}$  then
32              | |  $D' = \langle f_1, \dots, f_{i-1}, [x, y-1], f_{i+1}, \dots, f_{|\mathcal{R}|} \rangle$ 
33              | |  $(C_c, D_c) \leftarrow \text{Closure}(C, D')$ 
34              | | if  $(C, D') \ll (C_c, D_c)$  then
35                | | |  $\text{left}[i] \leftarrow \text{false}$ 
36                | | | LoCoM( $(C_c, D_c)$ , Propager( $X, (C_c, D_c)$ ), left')
37                | | | LoCoM( $(C, D)$ ,  $X \setminus \{(a, [x, y])\}$ , left)

```

---

**La fonction  $UB_\tau(C, D)$**  : la mesure de corrélation de rang de Kendall est monotone par morceau (Cerf et al., 2009) par rapport à  $\leq$  et nous pouvons en déduire une borne supérieure comme suit. Soit  $\mathcal{E}(C, D) = \{(C', D') \mid (C, D) \leq (C', D'), \text{ pour tout } (C', D') \in \mathcal{E}(C, D)\}$ , nous avons :

$$\tau(C', \sigma(D')) = \frac{|\eta(C', \sigma(D'))|}{N(\sigma(D'))} \leq \frac{\max_{(X, Y) \in \mathcal{E}(C, D)} |\eta(X, \sigma(Y))|}{\min_{(X, Y) \in \mathcal{E}(C, D)} N(\sigma(Y))} = \frac{|\eta(C, \sigma(D))|}{N(\alpha)}$$

En effet,  $\eta$  est décroissant par rapport à  $\leq$  et  $\sigma$  est également décroissant, tout en étant contraint par le paramètre  $\alpha$ . Ce qui donne la borne supérieure définie par  $UB_\tau(C, D) = \frac{2 \times |\eta(C, \sigma(D))|}{\alpha \times (\alpha - 1)}$ .

**La fonction  $UB_{\text{WRAcc}}(C, D)$**  : elle est une borne supérieure de la mesure  $\text{WRAcc}$ , définie par :

$$UB_{\text{WRAcc}}(C, D) = \frac{|\eta(C, \sigma(D))|}{N(\mathcal{O})} (1 - \tau(C, \mathcal{O}))$$

**Propriété 1.** Pour tout  $(C', D') \in \mathcal{E}(C, D)$ ,

$$\text{WRAcc}(C', D') \leq UB_{\text{WRAcc}}(C, D)$$

**Preuve 1.** Notons  $y = |\eta(C, \sigma(D))|$ ,  $x = N(\sigma(D))$ ,  $\alpha = N(\mathcal{O})$  et  $\beta = |\eta(C, \mathcal{O})|$ . Puisque  $\alpha$  et  $\beta$  sont indépendants de  $D$ , les valeurs de  $x$  et  $y$  déterminent uniquement  $\text{WRAcc}(D, P)$  et nous avons  $\text{WRAcc}(x, y) = \frac{x}{\alpha} \left( \frac{y}{x} - \frac{\beta}{\alpha} \right)$ . Comme indiqué dans (Kaytoue et al., 2017),  $\text{WRAcc}(x, y)$  est une fonction convexe. Suite aux résultats de (Morishita and Sese, 2000), cette fonction prend sa valeur maximale en  $(y, y)$ , c'est-à-dire :

$$\text{WRAcc}(C', D') \leq \frac{y}{\alpha} \left( 1 - \frac{\beta}{\alpha} \right) = \frac{|\eta(C, \sigma(D))|}{N(\mathcal{O})} (1 - \tau(C, \mathcal{O}))$$

Ainsi, lorsque  $UB_\tau(C, D)$  est inférieure à  $\beta$ , nous sommes sûrs qu'aucun des motifs de  $\mathcal{E}(C, D)$  peut satisfaire la contrainte et le processus de dénombrement peut être arrêté tout en garantissant l'exhaustivité de l'extraction. La limite supérieure  $UB_{\text{WRAcc}}(C, D)$  est exploitée conjointement avec  $\text{minWRAcc}$ , la valeur  $\text{WRAcc}$  la plus faible des motifs  $k$  dans  $\mathcal{K}$ . Lorsque  $|\mathcal{K}| \leq k$ ,  $\text{minWRAcc} = 0$ , sinon  $\text{minWRAcc} = \min_{(C, D) \in \mathcal{K}} \text{WRAcc}(C, D)$ .  $\text{minWRAcc}$  est mis à jour par la fonction  $\text{TOPKDiv}$  présentée ci-dessous.

**La fonction  $\text{TopKDiv}$**  calcule les top- $k$  motifs corrélés sur rang et diversifiés définis dans le problème 2. Le pseudo-code est présenté dans l'Algorithme 2. L'algorithme boucle sur la liste ordonnée  $\mathcal{K}$  des motifs corrélés par rang diversifiés déjà trouvés (Lignes 3 à 10) et stocke ceux qui sont similaires



au motif  $(C, D)$  (variable *similarPatterns*). Si un motif similaire a une valeur **WRAcc** plus élevée que  $(C, D)$ , la boucle s'arrête puisque le motif  $(C, D)$  ne sera pas inséré dans *mathcal{K}* (Lignes 8 et 9). S'il n'y a pas de motifs similaires dans  $\mathcal{K}$ , soit (1)  $(C, D)$  remplacera un mauvais motif, si la taille de  $\mathcal{K}$  est  $k$  (Lignes 13 à 17), ou sinon (2) il sera simplement ajouté à  $\mathcal{K}$  (Ligne 19) sans que *minWRAcc* soit mis à jour pour ne pas diminuer sa valeur. Si des motifs similaires existent et si  $(C, D)$  a une valeur **WRAcc** plus élevée que ceux-ci, alors tous les motifs similaires sont supprimés et  $(C, D)$  est inséré dans  $\mathcal{K}$ . *minWRAcc* n'est mis à jour que si la liste  $\mathcal{K}$  contient  $k$  motifs.

## 2.4 Expérimentations

Nous présentons nos résultats expérimentaux. Nous commençons par décrire les jeux de données que nous avons utilisés, ainsi que les questions auxquelles nous voulons répondre. Ensuite, nous fournissons une étude approfondie des performances, y compris une comparaison avec une méthode de base basée sur un algorithme de l'état de l'art. Enfin, nous donnons quelques résultats qualitatifs.

### 2.4.1 Jeux de données et objectifs

Nous considérons 4 jeux de données dont les caractéristiques sont données dans la Table 2.2. Le premier ensemble de données<sup>1</sup>, nommé **SA-heart**, rassemble des données provenant d'un échantillon d'hommes provenant d'une région à haut risque de maladie cardiaque du Cap-Occidental d'Afrique du Sud. Les trois autres jeux de données proviennent de l'UC Irvine Machine Learning repository<sup>2</sup>. Ils couvrent d'autres types de domaines d'application : la description d'ormeaux (**Abalone**), les secousses sismiques (**Seismic-bumps**), le risque de crédit (**German-credit**) et contiennent un plus grand nombre d'objets.

Dans cette étude empirique, nous visons à examiner le comportement de LOCOM concernant les questions suivantes :

- Quelle est l'efficacité de LOCOM en ce qui concerne les données caractéristiques qui peuvent influencer sur son temps d'exécution ?
- Comment se comporte LOCOM en fonction de ses paramètres ?
- Quelle est l'efficacité des propriétés d'élagage de LOCOM ?
- Est-ce que les sous-groupes corrélés fournissent des connaissances pertinentes et inconnues auparavant ?

Dans le cadre de cette évaluation, nous considérons un algorithme de référence (baseline), appelé PAIRMINING qui est dérivé de (Calders et al., 2006). PAIRMINING génère un motif en énumérant les descriptions à partir

1. <https://web.stanford.edu/~hastie/ElemStatLearn/datasets/>

2. <https://archive.ics.uci.edu/ml/index.php>

**Algorithm 2:** TOPKDIV( $\mathcal{K}$ ,  $(C, D)$ )

**Input:**  $\mathcal{K}$  est une liste ordonnée par rapport à la mesure **WRAcc** jusqu'au  $k$ ème sous-groupe corrélé diversifié,  $(C, D)$  un sous-groupe corrélé.

**Output:**  $\mathcal{K}$ , La liste triée mise à jour et la valeur de mise à jour de  $\text{minWRAcc}$ .

```

1 similarPatterns  $\leftarrow \emptyset$ 
2  $i \leftarrow 0$ 
3 isBetter  $\leftarrow true$ 
4 while isBetter do
5    $(C_K, D_K) \leftarrow \mathcal{K}[i]$ 
6   if Jaccard $((C_K, D_K), (C, D)) \geq \delta$  then
7      $insert((C_K, D_K), similarPatterns)$ 
8     if WRAcc $(C_K, D_K) > \mathbf{WRAcc}(C, D)$  then
9        $isBetter \leftarrow false$ 
10     $i \leftarrow i + 1$ 
11 if  $(|similarPatterns| = 0)$  then
12   if  $(\mathbf{WRAcc}(C, D) \geq minWRAcc)$  then
13     if  $(|\mathcal{K}| = k)$  then
14       //ici nous avons la garantie que
15        $(\mathbf{WRAcc}(C, D) \geq \mathbf{WRAcc}(\mathcal{K}[0]))$ 
16        $remove(\mathcal{K}[0], \mathcal{K})$ 
17        $insert((C, D), \mathcal{K})$ 
18        $minWRAcc \leftarrow \mathbf{WRAcc}(\mathcal{K}[0])$ 
19     else
20        $insert((C, D), \mathcal{K})$ 
21 else
22   if isBetter then
23     forall  $(C_S, D_S) \in similarPatterns$  do
24        $remove((C_S, D_S), \mathcal{K})$ 
25        $insert((C, D), \mathcal{K})$ 
26       if  $(|\mathcal{K}| = k)$  then
27          $minWRAcc \leftarrow \mathbf{WRAcc}(\mathcal{K}[0])$ 

```

Datasets	$ \mathcal{O} $	$ \mathcal{C} $	$ \mathcal{R} $	Description
SA-heart	462	5	5 (4+1)	La variable de classe à prédire est la maladie cardiaque avec différents attributs décrivant la santé et le mode de vie des personnes.
Abalone	4 177	4	4 (4+2)	Mesures physiques d'ormeaux.
Seismic-bumps	2 584	12	7 (5+2)	Les données décrivent des signes avant-coureurs d'épisodes sismiques à haute énergie dans une mine de charbon.
German-credit	1 000	10	10 (10+2)	Des personnes sont décrites par un ensemble d'attributs et la variable de classe indique le niveau de risque du crédit.

TABLE 2.2 – Principales caractéristiques des jeux de données.

des attributs de  $\mathcal{R}$  puis appelle Topominer (Prado et al., 2013) pour identifier des sous-groupes corrélés. Dans la suite, nous montrons que LOCOM surpasse cet algorithme de référence et est capable de découvrir des sous-groupes corrélés dans les contextes où PAIRMINING échoue. Nous fournissons également une analyse plus poussée de LOCOM en faisant varier la valeur de ses paramètres ainsi que les caractéristiques des jeux de données. En particulier, nous étudions l'efficacité de la propriété d'élagage. Nous étudions la diversité des motifs produits par LOCOM, c'est-à-dire à quel point les sous-groupes corrélés sont divers, et nous produisons également quelques résultats qualitatifs.

### 2.4.2 Étude expérimentale quantitative

La Figure 2.2 montre le temps d'exécution de LOCOM et PAIRMINING en fonction des paramètres  $\alpha$  et  $\beta$  pour chaque jeu de données. Notez que pour une comparaison équitable, LOCOM renvoie tous les sous-groupes corrélés satisfaisant les contraintes du Problème 1. Nous pouvons observer que LOCOM surpasse PAIRMINING dans toutes les configurations, sauf lorsque  $\sigma(D)$  est

suffisamment élevé pour qu'il n'y ait presque aucun motif. En outre, PAIRMINING est capable de réaliser la tâche de découverte de sous-groupes corrélés uniquement pour des valeurs très élevées de  $\alpha$ . PAIRMINING échoue pour les autres configurations. Notez qu'en pratique, les sous-groupes qui couvrent plus de 80% du jeu de données ne sont pas intéressants (ils sont généralement déjà connus). Les sous-groupes vraiment intéressants couvrent de plus petites parties des données. Cela démontre le besoin d'un algorithme dédié à la découverte de tels motifs.

Examinons maintenant plus en détail le comportement de LOCOM, en particulier son efficacité en fonction des caractéristiques du jeu de données, des paramètres de l'algorithme et de l'efficacité de ses propriétés d'élagage. Pour étudier l'efficacité de  $UB_{\mathbf{WRAcc}}$ , la borne supérieure de la valeur  $\mathbf{WRAcc}$  d'un motif défini dans la Section 2.3, nous comparons le temps de calcul de LOCOM avec et sans cette optimisation. La Figure 2.3 montre ces temps d'exécution, le nombre de motifs explorés pour LOCOM implémentant  $UB_{\mathbf{WRAcc}}$  et LOCOM sans  $UB_{\mathbf{WRAcc}}$  en fonction de  $k$ . La distribution des valeurs de  $\mathbf{WRAcc}$  des motifs renvoyés est également affichée sur cette figure. Dans tous les jeux de données, l'optimisation basée sur  $UB_{\mathbf{WRAcc}}$  permet d'accélérer la découverte des top- $k$  sous-groupes corrélés. Ce gain peut atteindre un facteur de 2, comme pour les jeux de données German-crédit et Abalone. En effet, cette optimisation permet d'élaguer une grande partie de l'espace de recherche (voir le nombre de motifs explorés). Plus la valeur  $k$  est faible, plus la valeur  $\mathbf{WRAcc}$  du  $k^{ième}$  motif est grande et plus l'élagage des candidats peu prometteurs est efficace. Notez que, pour la plupart des jeux de données, il y a beaucoup moins de 1000 motifs qui vérifient les contraintes, ce qui explique les temps d'exécution similaires pour certaines valeurs de  $k$ .

Nous étudions également le comportement de notre algorithme en fonction du facteur de diversité  $\delta$ . La Figure 2.4 rend compte du temps d'exécution de LOCOM, du nombre de motifs retournés et de la valeur  $\mathbf{WRAcc}$  la plus basse des motifs découverts en fonction de  $\delta$ . Les résultats obtenus confirment les résultats précédents : le temps d'exécution augmente lorsque les seuils deviennent moins restrictifs. En effet, plus le seuil  $\delta$  est bas, plus la contrainte de diversité est stricte et plus le nombre de motifs découverts est petit. Par conséquent, la plus petite valeur de  $\mathbf{WRAcc}$  parmi les motifs découverts augmente lorsque  $\delta$  diminue et les capacités d'élagage sont alors plus efficaces.

Enfin, la Figure 2.5 montre le temps d'exécution de LOCOM selon les différentes dimensions des jeux de données, c'est-à-dire le nombre d'attributs de description, le nombre d'attributs de corrélation et le nombre d'objets. Évidemment, le temps d'exécution de LOCOM augmente avec la taille du jeu de données. En particulier, le nombre d'attributs (attributs de description et de corrélation) est la dimension qui a le plus d'impact sur les performances de LOCOM. Ceci est bien connu en fouille de motifs car la taille de l'espace de recherche dépend de ces dimensions et non directement du nombre d'objets.

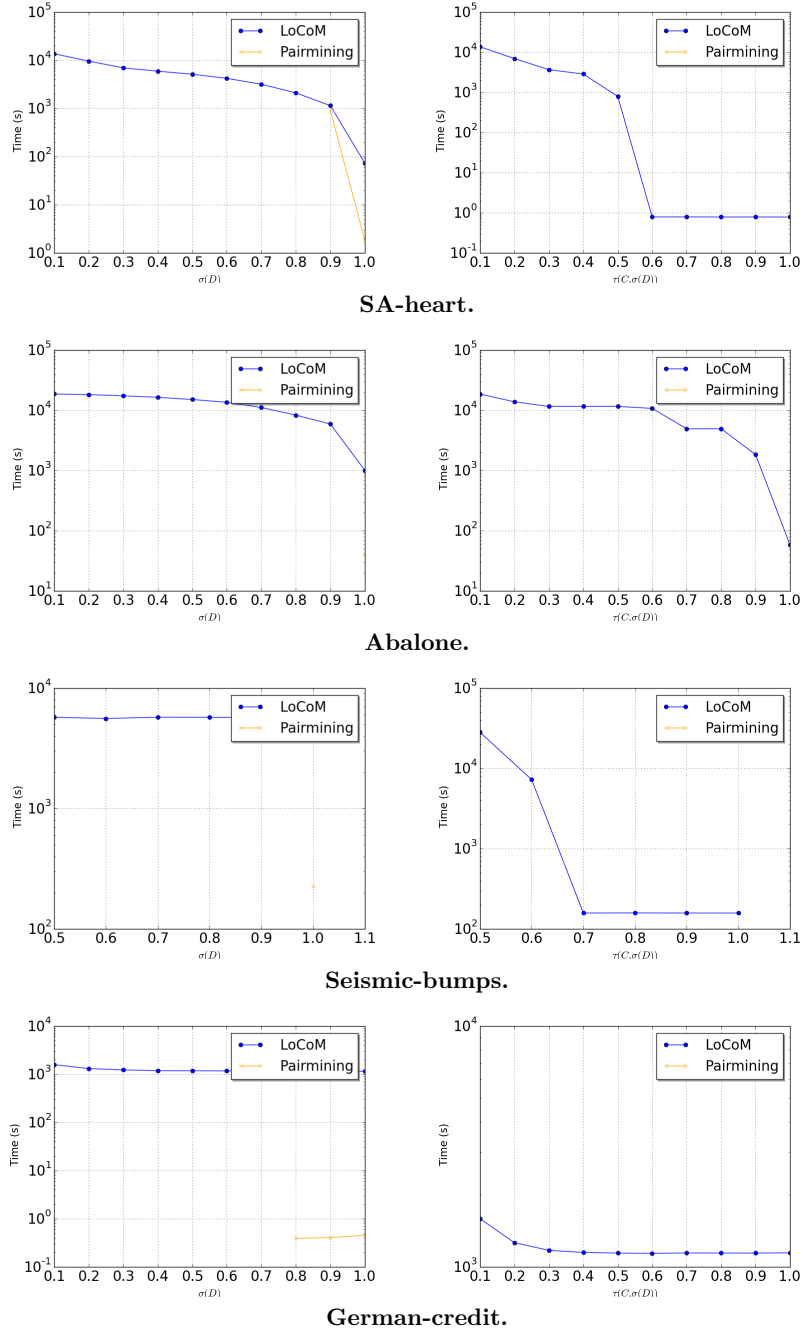


FIGURE 2.2 – Temps d'exécution de LoCoM et PAIRMINING en fonction de  $\alpha$  (gauche) et  $\beta$  (droite) pour les quatre jeux de données (valeurs par défaut :  $\alpha = 0.1$  et  $\beta = 0.1$ ).

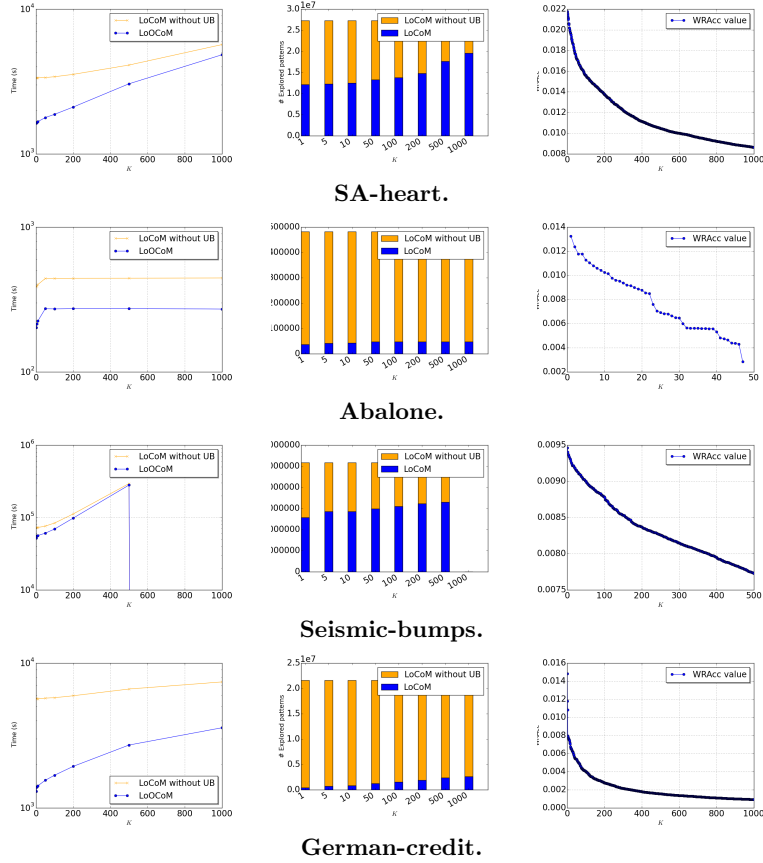


FIGURE 2.3 – Temps d’exécution (à gauche), nombre de candidats explorés (au centre) et distribution des top- $k$  motifs (à droite) pour LoCoM avec et sans optimisation  $UB_{WRAcc}$  selon  $k$  (valeurs par défaut :  $\alpha = 0.05$  et  $\beta = 0.4$ ).

### 2.4.3 Étude expérimentale qualitative

Nous montrons maintenant les motifs les mieux classés trouvés dans chaque jeu de données. Les meilleurs motifs par rapport à la mesure WRAcc sont rapportés dans le Tableau 2.3.

- Les motifs obtenus sur le jeu de données SA-heart sont cohérents avec les connaissances communes : la maladie coronarienne (chd) est positivement corrélée avec l’âge et l’obésité des sujets. Plus intéressant,

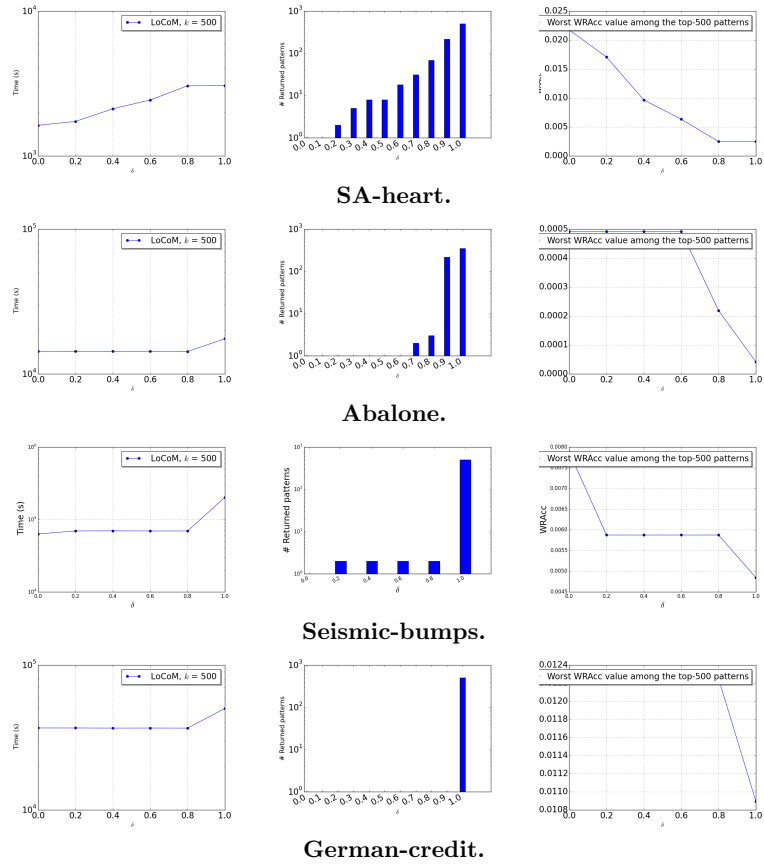


FIGURE 2.4 – Temps d'exécution (gauche), nombre de motifs retournés (milieu) et valeur **WRAcc** du dernier motif parmi les top-100 motifs (droite) (valeurs par défaut :  $\alpha = 0.05$  et  $\beta = 0.4$ ).

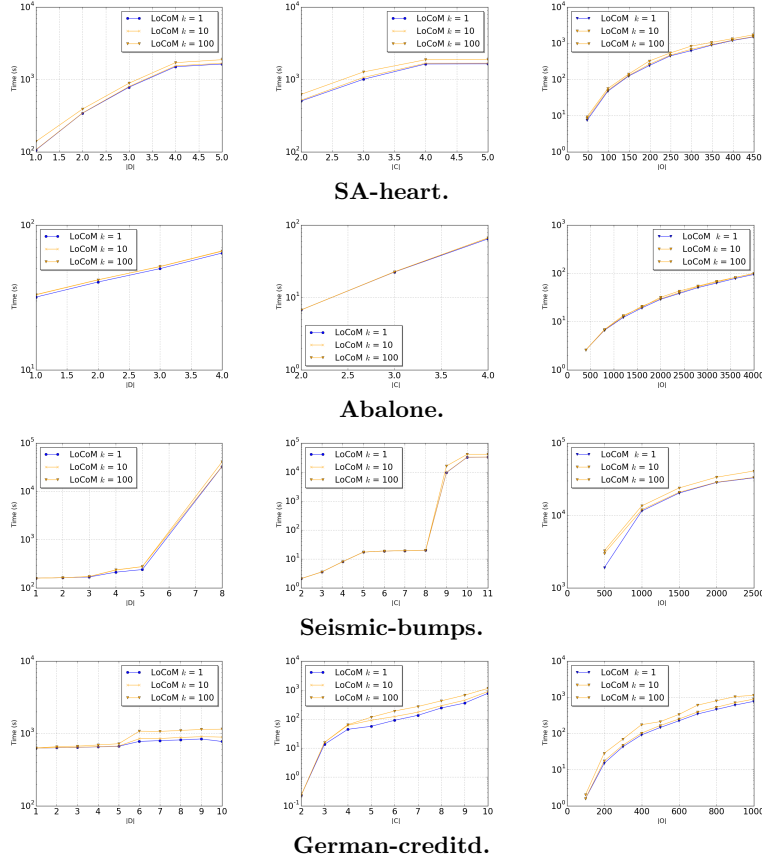


FIGURE 2.5 – Temps d'exécution de LoCoM en fonction du nombre d'attributs de description ( $|\mathcal{R}|$ ) (à gauche), du nombre d'attributs de corrélation ( $|\mathcal{C}|$ ) (au milieu) et du nombre d'objets ( $|\mathcal{O}|$ ) (à droite) (valeurs par défaut :  $\alpha = 0.05$  et  $\beta = 0.4$ ).

l'âge et la consommation d'alcool sont anti-corrélés pour les valeurs de tension artérielle systolique élevée.

- Les motifs décrivant les ormeaux (jeu de données Abalone) mettent en évidence les corrélations entre les mesures de poids et les dimensions des coquillages sur l'ensemble des données ( $\sigma(D) \geq 0,99$ ). Ces mesures sont également corrélées avec l'âge de l'ormeau (Anneaux) quand ils sont plutôt petits.



$C$	$D$	$\tau(C, O)$	$\sigma(D)$	$\tau(C, \sigma(D))$	<b>WRAcc</b>
Motifs SA-heart ( $\alpha = 0.05$ , $\beta = 0.5$ and $\delta = 0.2$ ).					
Alcool <sup>+</sup> , Age <sup>-</sup>	Sdb = [121, 194]	0.512	0.803	0.536	0.015
Age <sup>+</sup> , Chd <sup>+</sup>	Sdb = [101, 138]	0.591	0.628	0.617	0.010
Obésité <sup>+</sup> , Chd <sup>+</sup>	Sdb = [117, 148]	0.510	0.665	0.523	0.005
Motifs Abalone ( $\alpha = 0.05$ , $\beta = 0.5$ and $\delta = 0.7$ ).					
Poids décoquillé <sup>+</sup> , Poids de la coque <sup>+</sup>	Diam. =[0.055, 0.630], Hauteur = [0.01, 0.230]	0.879	0.996	0.880	0.0005
Poids décoquillé <sup>+</sup> , Poids de la coque <sup>+</sup> , anneaux <sup>+</sup>	Diam. =[0.055, 0.545], Hauteur = [0.01, 0.130]	0.874	0.406	0.876	0.0004
Poids total <sup>+</sup> , Poids décoquillé <sup>+</sup>	Diam. =[0.055, 0.650], Hauteur = [0.01, 0.250]	0.699	0.999	0.699	0.0002
Motifs Seismic-bumps ( $\alpha = 0.05$ , $\beta = 0.5$ and $\delta = 0.8$ ).					
Nombre-de- secousses5 <sup>+</sup> , Nombre-de- secousses6 <sup>+</sup>	période = coal_getting	0.661	0.643	0.757	0.039
motifs German-credit ( $\alpha = 0.05$ , $\beta = 0.3$ and $\delta = 0.8$ ).					
Age <sup>+</sup> , Nombre de crédits <sup>-</sup>	état civil = homme_célibataire, garants = aucun, Type.d'apartment = own	0.447	0.369	0.503	0.007

TABLE 2.3 – Les meilleurs motifs par rapport à la mesure **WRAcc**.

- Dans le jeu de données Seismic-bumps, le nombre de secousses sismiques dont l'énergie varie entre  $[10^6, 10^7)$  et  $[10^7, 10^8)$  est corrélé pendant la période d'extraction du charbon.
  - Dans les données de German-credit, l'âge est anti-corrélé avec le nombre de crédits pour les hommes célibataires qui possèdent leur appartement.
- Étonnamment, les meilleurs motifs selon la mesure **WRAcc** ont une valeur

élevée de  $\sigma(D)$  même si le seuil est bas ( $\alpha = 0.05$ ). Ceci s'explique par le fait que la mesure WRAcc est corrigée par le support de la description. Par conséquent, les motifs avec un support de description élevé ont tendance à être favorisés par cette mesure. La Figure 2.6 (à gauche) confirme cette affirmation. Pour évaluer seulement le gain de corrélation avec la description, nous considérons le taux de croissance (Dong and Li, 1999) qui est défini comme le rapport de la corrélation locale à la corrélation globale (c'est-à-dire  $Gr = \frac{\tau(C,\sigma(D))}{\tau(C,\sigma(O))}$ ). La Figure 2.6 montre la distribution des motifs en fonction du support de description et du taux de croissance (au milieu) et de la mesure WRAcc et du taux de croissance (à droite). Le taux de croissance est optimisé avec une description de support faible. Ainsi, les meilleurs motifs par rapport au taux de croissance sont, dans la plupart des cas, différents des meilleurs selon la WRAcc comme indiqué dans le Tableau 2.4.

## 2.5 Conclusion

Dans ce chapitre, nous avons présenté le nouveau problème de la découverte de sous-groupes corrélés pour un nombre arbitraire de cibles numériques (supérieur ou égal à 2). Ce problème permet de découvrir des sous-groupes d'objets – identifiés par des conditions sur des attributs numériques et/ou nominaux – pour lesquels la corrélation de rang entre un sous-groupe d'attributs cibles est exceptionnellement élevée par rapport à l'ensemble de données. Les motifs de corrélation de rang que nous considérons sont basés sur une généralisation du  $\tau$  de Kendall qui permet de représenter un sous-ensemble d'attributs numériques qui covarient d'une manière positive ou négative. Nous avons proposé LOCOM, un algorithme qui exploite certaines propriétés d'élagage basées sur deux bornes supérieures et sur des propriétés de fermeture. Une étude empirique sur plusieurs jeux de données démontre l'efficacité de LOCOM.

Nous croyons que cela ouvre de nouvelles perspectives de recherche. Par exemple, d'autres mesures et paradigmes peuvent être étudiés pour évaluer l'intérêt des sous-groupes, en particulier l'intérêt subjectif qui permet de prendre en compte les connaissances a priori de l'utilisateur (Bie, 2011). Une autre direction de recherche intéressante consiste à concevoir des méthodes d'exploration instantanée en supprimant la complétude et en échantillonnant directement l'espace des motifs (Boley et al., 2011) (Boley et al., 2012, Chaoji et al., 2008). Cette voie est explorée dans le chapitre suivant.

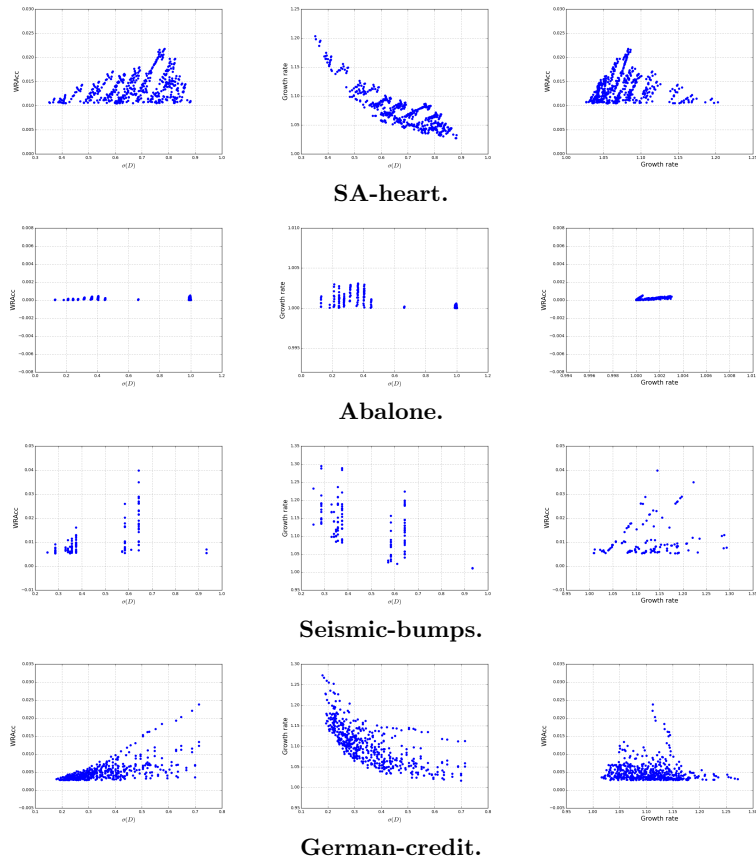


FIGURE 2.6 – Répartition des motifs selon le support de la description et de la WRAcc (gauche), le support de la description et le taux de croissance (growth rate au milieu) et le taux de croissance et le WRAcc (droite) ( $\alpha = 0.05$ ,  $\beta = 0.4$  et  $\delta = 1, 0$ ).

$C$	$D$	$\tau(C, O)$	$\sigma(D)$	$\tau(C, \sigma(D))$	<b>WRAcc</b>	Gr
Top 3 motifs de SA-heart par rapport au taux de croissance ( $\alpha = 0.05$ , $\beta = 0.4$ et $\delta = 1$ ).						
Obésité <sup>+</sup> , Age <sup>-</sup>	sdb = [140 , 198]	0.424	0.352	0.511	0.011	1.203
Obésité <sup>+</sup> , Chd <sup>-</sup>	sdb = [134 , 170]	0.423	0.470	0.480	0.011	1.115
Obésité <sup>+</sup> , Al- cool <sup>+</sup>	sdb = [134 , 178]	0.460	0.498	0.507	0.012	1.101
Top 3 des motifs de Abalone par rapport au taux de croissance ( $\alpha = 0.05$ , $\beta = 0.4$ et $\delta = 1$ ).						
Poids de la coque <sup>+</sup> , Anneaux <sup>+</sup>	diam. = [0.055, 0.515], height = [0, 0.125]	0.939	0.366	0.942	0.0004	1.003
Poids décoquillé <sup>+</sup> , Poids de la coque <sup>+</sup>	diam. = [0.055, 0.515], height = [0.01, 0.105]	0.879	0.212	0.882	0.0001	1.003
Poids décoquillé <sup>+</sup> , Poids de la coque <sup>+</sup> , Anneaux <sup>+</sup>	diam. = [0.055, 0.545], height = [0.01, 0.130]	0.874	0.406	0.876	0.0004	1.003
Top 3 motifs de Seismic-bumps par rapport au Taux de croissance ( $\alpha = 0.05$ , $\beta = 0.4$ et $\delta = 1$ ).						
nbumps6 <sup>+</sup> , nbumps89 <sup>+</sup>	sismique = faible danger, période = coal-getting	0.315	0.285	0.407	0.007	1.294
nbumps5 <sup>+</sup> , nbumps89 <sup>-</sup>	sismique = faible danger, période = coal-getting	0.313	0.285	0.403	0.007	1.287
nbumps2 <sup>+</sup> , nbumps4 <sup>+</sup>	sismique = faible danger, période = coal-getting	0.378	0.357	0.467	0.011	1.235
Meilleurs motifs de German-credit par rapport au Taux de croissance ( $\alpha = 0.05$ , $\beta = 0.4$ et $\delta = 1$ ).						
Versement <sup>+</sup> Age <sup>+</sup>	Objet de valeur = immobilier , crédit-concret = aucun , Tel = au- cun	0.342	0.18	0.435	0.007	1.294
DuratCurr <sup>+</sup> No_of_dependents <sup>-</sup>	Objectif = ra- dio/television Typeaparte- ment = détenu, Etranger = oui	0.340	0.222	0.419	0.003	1.230
Tau- versement <sup>+</sup> Nb_crédits <sup>+</sup>	Objet de valeur = Voiture ou autre	0.350	0.23	0.424	0.00388382	1.209

TABLE 2.4 – Meilleurs motifs par rapport au taux de croissance.



## Chapitre 3

# Échantillonnage de sous-groupes corrélés

### 3.1 Introduction

Dans le chapitre précédent, nous avons proposé d'identifier des motifs permettant de caractériser des observations décrites par des attributs numériques et symboliques. Cette méthode consiste à rechercher des corrélations entre plusieurs attributs numériques qui sont bien plus importantes dans un sous-groupe des données que dans l'ensemble du jeu de données. Le sous-groupe et les attributs fortement corrélés sont identifiés automatiquement par l'approche. Ces corrélations sont d'autant plus intéressantes qu'elles se produisent localement dans le jeu de données : par exemple, alors que les variables âge et revenu ne sont pas corrélées globalement, celles-ci le deviennent lorsqu'on limite les observations aux personnes cadres. L'approche de fouille de motifs proposée identifie les motifs qui sont corrélés localement dans les données : les ensembles d'attributs numériques (par exemple, âge et revenu) qui co-varient fortement dans un sous-ensemble des données, identifié par des restrictions sur la valeurs de certains autres attributs (par exemple, catégorie d'employé = cadre). Les ensembles d'attributs numériques et le sous-ensemble de données sont automatiquement (inductivement) identifiés par la méthode et les motifs, avec une valeur élevée sur une mesure d'intérêt qui évalue le niveau de corrélation locale, sont calculés.

Alors que l'espace des motifs à explorer est potentiellement exponentiel, la complexité du problème peut être surmontée en utilisant des techniques d'échantillonnage permettant de réduire le coût du calcul tout en identifiant les motifs les plus importants. Dans ce qui suit, nous présentons les principes des algorithmes d'échantillonnage conçus pour les problèmes d'exploration de motifs. Ensuite, nous présentons la technique d'échantillonnage que nous utilisons et montrons qu'elle permet d'obtenir des motifs de haute qualité de manière efficace. Cette méthode d'échantillonnage calcule chaque motif

indépendamment des autres et permet ainsi de traiter de grandes bases de données en répartissant les calculs sur plusieurs machines.

## 3.2 Méthodes d'échantillonnage pour la fouille de motifs

Les méthodes exactes d'extraction de motifs peuvent être trop consommatrices de ressources informatiques pour être utilisées sur de grands jeux de données. Cette complexité peut néanmoins être surmontée en utilisant des techniques d'échantillonnage qui présentent plusieurs avantages : elles réduisent le coût de calcul des motifs, permettent l'identification des motifs les plus importants, et peuvent être exécutées sur plusieurs machines en parallèles afin de distribuer le calcul.

Il existe deux familles de techniques d'échantillonnage de motifs locaux. La première famille est basée sur des méthodes d'échantillonnage direct Boley et al. (2011, 2012) qui tire un échantillon sans matérialiser de graphe de transition entre un motif aléatoire et un autre (voir Sous-section 3.2.1). La seconde famille est basée sur les méthodes de Monte-Carlo par chaînes de Markov (MCMC) qui effectuent une marche aléatoire sur un graphe de transition représentant la probabilité d'atteindre un motif en fonction du motif actuel. Cela peut être fait avec la garantie que la distribution de la mesure de qualité considérée est proportionnelle sur l'ensemble échantillonné à celle de l'ensemble des motifs dans le jeu de données de départ (Boley et al., 2010). Mais le coût de calcul d'une telle approche est très élevé, car le graphe de transition représentant la probabilité d'atteindre un motif donné étant donné le motif actuel, doit être matérialisé dans les deux sens (voir la Sous-section 3.2.2). Une autre approche (Chaoji et al., 2008, Moens and Goethals, 2013) assouplissant cette contrainte est présentée dans la Sous-section 3.2.3, une approche heuristique qui calcule les motifs maximaux (Moens and Goethals, 2013). Cette approche pragmatique donne de très bons résultats dans la pratique, comme nous le verrons dans la suite.

### 3.2.1 Méthode par échantillonnage direct

Pour tirer un échantillon de motifs avec une probabilité proportionnelle à une mesure, Boley et al. (2011) proposent une procédure aléatoire en deux étapes qui échantillonne des motifs sans simuler de processus stochastiques. Ils étudient l'échantillonnage d'itemsets selon une distribution proportionnelle à la fréquence. Dans un premier temps, ils sélectionnent de manière aléatoire une transaction (un objet) en fonction d'une première distribution  $w_1$ . Ensuite, dans un deuxième temps, ils tirent un sous-ensemble d'items de cette transaction (un itemset) à l'aide d'une deuxième distribution  $w_2$ . La combinaison des deux étapes suit la distribution souhaitée.

Dans notre cas, nous voulons échantillonner des sous-ensemble d'attributs corrélés selon la mesure  $\tau$ . Ceci nécessite une étape de calcul de la pondération  $w_1$  en prétraitement, un calcul sur l'ensemble des observations, qui est inutilisable dans notre cas où l'on considère des paires d'observations.

---

**Algorithm 3: ECHANTILLONNAGE DIRECT**


---

**Input:**  $\mathbb{D} = (\mathcal{O}, \mathcal{C}, \mathcal{R})$

**Output:** Un échantillon  $C$ .

- 1 Calcul des poids  $w_1$  pour chaque paire d'objets  $P = (o_i, o_j)$ ,  $w_1$  étant proportionnel au nombre d'attributs ordonnés suivant la paire  $|\{A \mid A(o_i) < A(o_j)\}|$ .
  - 2 **Tirer**  $P \sim w_1$
  - 3 **Tirer**  $C \sim w_2(P)$
  - 4 **Retourner**  $C$
- 

### 3.2.2 Méthodes de Monte-Carlo par chaînes de Markov

Boley et al. (2010) utilisent le cadre des méthodes de Monte-Carlo par chaînes de Markov (MCMC) pour échantillonner des motifs fermés selon une distribution proportionnelle à une mesure de qualité. Ils définissent une chaîne de Markov sur l'espace d'état des motifs fermés. Les transitions entre deux états, qui correspondent à deux motifs fermés  $F = ((C, D), \eta(C, \sigma(D)))$  et  $F' = ((C', D'), \eta(C', \sigma(D')))$ , existe s'il existe au moins un élément générateur  $x$  permettant de passer du premier motif fermé au second : **Closure** $\left(\left((C, D) \cup x, \eta(C, \sigma(D))\right)\right) = F'$  ou **Closure** $\left(\left((C, D), \eta(C, \sigma(D))\right) \cup x\right) = F'$ . Pour appliquer le cadre MCMC et assurer la convergence de la chaîne de Markov vers sa distribution stationnaire, la chaîne doit être ergodique, c'est-à-dire que tout état peut être atteint à partir de n'importe quel autre. Ceci est obtenu par un mélange aléatoire de deux chaînes de Markov sur le même espace d'état, dont les probabilités de transition sont proportionnelles au nombre d'éléments générateurs.

La première chaîne va de  $F$  vers ses successeurs  $F'$  (tels qu' $\exists x$ , **Closure** $\left(\left((C, D) \cup x, \eta(C, \sigma(D))\right)\right) = F'$ ) en ajoutant un attribut signé à  $C$  ou une restriction à  $D$  et la probabilité de transition est alors

$$P_s(F, F') \propto |\{x \mid H(C, D) = (C', D')\}|$$

Réciproquement, la seconde va de  $F$  à ses précédentes  $F'$  en ajoutant une paire d'objets à  $F$  avec pour probabilité de transition :

$$P_s(F, F') \propto |\{x \mid M(\eta(C, \sigma(D)) \cup x) = F'\}|$$



Cet algorithme peut avoir un problème de passage à l'échelle car le nombre d'étapes de simulation requises doit être nettement supérieure à la taille de l'espace des états pour assurer la convergence de la chaîne de Markov vers sa distribution stationnaire. De plus, chaque étape a une forte complexité car elle nécessite le calcul des fermetures pour estimer chaque probabilité de transition.

### 3.2.3 Échantillonnage biaisé vers les motifs maximaux

Cette méthode, présentée dans (Chaoji et al., 2008, Moens and Goethals, 2013) échantillonne aléatoirement les motifs maximaux mais sans garantir l'uniformité de l'échantillonnage. En revanche, elle est très efficace lorsque la taille de l'ensemble à échantillonner est grande. C'est l'approche que nous avons choisi de suivre.

## 3.3 Échantillonnage aléatoire de sous-groupes corrélés maximum

Dans le chapitre précédent, nous avons proposé un algorithme, appelé LO-COM, qui calcule la collection complète de sous-groupes corrélés sur les rangs. Pour énumérer un sous-groupe  $(C, D)$  corrélé sur les rangs, il génère d'abord la partie  $C$ , en utilisant une borne supérieure sur la mesure  $\tau$  pour éliminer les motifs peu prometteurs, puis recherche le sous-groupe  $D$  pour lequel la corrélation est très élevée. La première étape est la plus chronophage. Dans ce qui suit, nous proposons de la remplacer par une approche par échantillonnage.

Nous suivons l'approche proposée dans (Chaoji et al., 2008), pour l'échantillonnage aléatoire de graphes, également utilisée dans (Moens and Goethals, 2013) pour générer des ensembles d'éléments maximaux. Cette méthode ne garantit pas l'uniformité de l'échantillonnage mais est très efficace lorsque la taille de l'ensemble à échantillonner est grande comme c'est le cas ici. Par conséquent, nous voulons échantillonner la partie corrélée du motif, c'est-à-dire les ensembles  $C \in \mathcal{C}$ . Si nous notons par  $m$  la taille de  $\mathcal{C}$ , il y a  $\frac{3^m-1}{2}$  motifs corrélés possibles<sup>1</sup>.

La méthode d'échantillonnage, voir Algorithme 4, consiste à tirer aléatoirement un attribut signé positivement et à ajouter progressivement des attributs signés suivant une probabilité  $p$  proportionnelle à la valeur de qualité du motif obtenu : pour un motif  $C$ ,  $p(C) \sim \frac{\tau(C)}{Z}$  avec  $Z$  un facteur de normalisation. Le processus s'arrête lorsqu'aucun attribut signé ne peut être ajouté sans que la valeur de  $\tau$  soit inférieure au seuil.

1. Pour chaque motif corrélé de taille  $s$ , il y a  $2^{s-1}$  motifs signés. Comme il existe  $\binom{m}{s}$  tels motifs, nous avons  $\frac{1}{2} \sum_{s \geq 0} \binom{m}{s} 2^m - 1 = \frac{3^m - 1}{2}$ .

**Algorithm 4:** LOMAX

---

**Input:** Les données :  $\mathcal{D} = \mathcal{O} \times \mathcal{C}$  et  $\leq$  un ordre canonique sur  $\mathcal{C} \times \{+, -\}$ .  
**Output:** Un sous-groupe échantillonné  $C$ .

```

1 Tire  $a \sim u([0, |\mathcal{C}|])$ 
2  $C \leftarrow \{a^+\}$ 
3  $Candidates \leftarrow \{x^s \in \mathcal{C} \times \{+, -\} \setminus C : \tau(x^s) \geq \beta\}$ 
4 while  $|Candidates| \geq 0$  do
5    $sum \leftarrow 0$ 
6   for  $x^s \in Candidates$  ordonné par  $\leq$  do
7      $sum \leftarrow sum + p(C \cup x^s)$ 
8      $N[x^s] \leftarrow sum$ 
9   Tire  $v \sim u([0, 1])$ 
10  Trouve le maximum  $x^s$  dans  $N$  tel que  $v < N[x^s]$ 
11   $C \leftarrow \{C \cup x^s\}$ 
12   $Candidates \leftarrow \{x^s \in Candidates \setminus C : \tau(C \cup x^s) \geq \beta\}$ 

```

---

Une fois le motif de corrélation de rang  $C$  obtenu, ses descriptions  $D$ , qui maximisent la mesure WRAcc, sont obtenues à l'aide de l'algorithme d'énumération du Chapitre 2.

### 3.4 Résultats expérimentaux

Pour les expériences, nous avons également utilisé les quatre jeux de données de l'UCI suivants : (1) SA-heart qui regroupe la description de la santé et du mode de vie ( $\mathcal{C} = 6$ ,  $\mathcal{R} = 1$ ) de 462 patients au sujet de maladies cardiaques ; (2) Abalone, qui contient des mesures physiques de 4177 ormeaux ( $\mathcal{C} = 6$ ,  $\mathcal{R} = 2$ ) ; (3) Seismic-bumps qui décrit des secousses sismiques à haute énergie dans une mine de charbon (2584 observations,  $\mathcal{C} = 14$ ,  $\mathcal{R} = 5$ ) ; et (4) German-credit, un ensemble de données où les clients sont décrits par différentes variables et une variable de classe indique le risque de crédit de chaque client (1000 observations,  $\mathcal{C} = 15$ ,  $\mathcal{R} = 6$ ). Notez également que LOCOM et LOMAX sont implémentés en C et que les expériences sont exécutées sur une machine équipée de 8 processeurs Intel (R) Xeon (R) W-2125 avec des cœurs de 4.00 GHz, 126 Go de mémoire, sous Debian GNU / Linux.

Pour évaluer la qualité de notre approche d'échantillonnage, nous comparons LOMAX avec LOCOM, l'approche complète. Nous cherchons à répondre aux questions suivantes : Comment l'échantillon aléatoire se rapproche-il de l'ensemble complet des sous-groupes corrélés par rang ? Les motifs de qualité supérieure sont-ils plus susceptibles d'être tirés aléatoirement ? Les motifs extraits couvrent-ils bien toutes les données ? Le temps de calcul des motifs échantillonnés est-il inférieur au temps de calcul complet ?

### 3.4.1 Étude quantitative.

La Figure 3.1 présente la proportion de motifs échantillonnés par rapport à la mesure  $\tau$  pour différentes tailles d'échantillon (en fonction du nombre total de motifs de corrélation de rang dans l'ensemble de données). Nous pouvons observer que LOMAX se rapproche bien de la distribution globale des motifs le long de la mesure  $\tau$ . Nous pouvons également constater que les petits échantillons contiennent des motifs de haute valeur  $\tau$  et que, par conséquent, de petits échantillons peuvent être suffisants pour obtenir des motifs de haute qualité. Notez que nous pouvons observer deux distributions mixtes dans le jeu de données Abalone. Cela est dû au fait que certains attributs contiennent beaucoup plus de valeurs égales que d'autres, ce qui conduit à des motifs de corrélations plus faibles.

Pour étudier l'efficacité de la méthode d'échantillonnage pour trouver des motifs avec des valeurs de  $\tau$  élevées, nous avons échantillonné de manière aléatoire un nombre de motifs beaucoup plus grand que celui obtenu avec l'algorithme et complet (le nombre calculé par LOCOM). Par conséquent, le même motif peut être tiré plusieurs fois. La Figure 3.2 indique, pour chaque unique motif le nombre de fois où il a été tiré et nous considérons ce nombre par rapport à la valeur  $\tau$  du motif. Nous pouvons observer que plus la valeur  $\tau$  est élevée, plus les motifs sont tirés. Cela montre que LOMAX favorise les motifs avec des valeurs  $\tau$  élevées, ce qui est exactement ce que nous recherchons.

Dans la Figure 3.3, nous évaluons si les motifs extraits couvrent l'ensemble du jeu de données. À cette fin, nous calculons la mesure de Jaccard entre les paires d'observations décrites par les motifs de l'approche complète et celles décrites par des motifs de l'échantillon aléatoire. Plus formellement, supposons que  $\mathcal{M}_1$  soit l'ensemble des motifs de corrélation de rang maximaux extraits par LOMAX et  $\mathcal{M}_2$  celui extrait par LOCOM. Soit  $N_i = \bigcup_{C \in \mathcal{M}_i} \eta(C, \mathcal{O})$ . La mesure de Jaccard est donc :  $Jaccard(\mathcal{M}_1, \mathcal{M}_2) = \frac{N_1 \cap N_2}{N_1 \cup N_2}$  : plus le nombre de paires couvertes est élevé, plus le nombre de paires couvertes est élevé. La Figure 3.3 (en haut) montre qu'un échantillon aléatoire de taille un tiers environ de l'ensemble des motifs suffit à couvrir une très grande proportion des paires couvertes par LOCOM. Ainsi, LOMAX génère des motifs très divers. La Figure 3.3 (au centre) montre que le temps d'exécution de LOMAX est inférieur à celui de LOCOM, de sorte que le calcul d'un échantillon de motifs est plus rapide tout en offrant un échantillon de motifs de très bonne qualité. De plus, comme chaque tirage aléatoire est indépendant des autres, LOMAX peut être exécuté en parallèle sur plusieurs machines. Nous étudions également la qualité **WRAcc** des sous-groupes fournis par LOCOM. Sur la Figure 3.3 (en bas), on peut observer la distribution de la mesure pour le jeu de données Seismic-bump. On voit que les échantillons approximent bien la collection complète de sous-groupes corrélés.

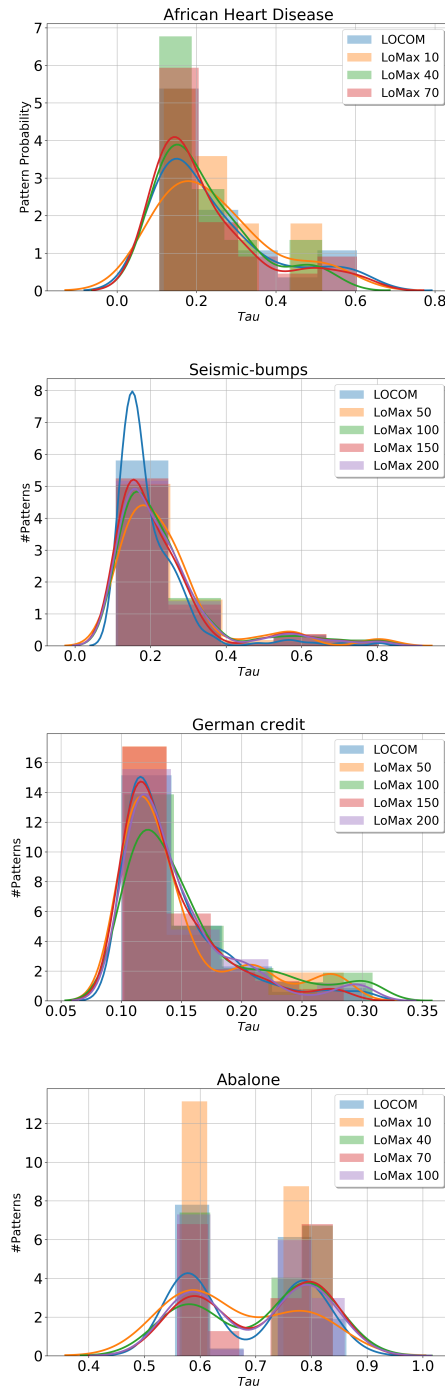


FIGURE 3.1 – Distribution des motifs (collection complète et échantillonnée de différentes tailles) par rapport à la valeur  $\tau$ .



FIGURE 3.2 – Nuage de points du nombre de motifs identiques par rapport à  $\tau$ . La taille de l'échantillon est 10 fois supérieure au nombre de motifs obtenus par LOCOM.

### 3.4.2 Étude qualitative.

LOMAX permet de découvrir des sous-groupes corrélés maximaux, en particulier de mettre en évidence simultanément un sous-espace de données et un modèle de corrélation de rang dont la corrélation est supérieure à celle observée dans l'ensemble de données. Pour le jeu de données Abalone, le motif aléatoire ayant la valeur de  $\tau$  et de **WRAcc** la plus élevée est :  $\langle \{Weight^+, Shweight^+, length^+, diameter^+, height^+, \{sex = infant\}\} \rangle$  ( $supp(D) = 0.321$ ,  $\tau(C, O) = 0.78$ ,  $\tau(C, \sigma(D)) = 0.8$ , et  $WRAcc(C, D) = 0.0025$ ). Ce motif signifie que, pour les ormeaux en bas âge, la corrélation entre le poids total, le poids écaillé, la longueur de la coquille la plus longue, le diamètre et la hauteur sont plus puissants que sur l'ensemble du jeu de données. De même, pour le jeu de données German-credit, nous obtenons  $\langle \{Age^+, Duration^+, PrevCred^+, \{Housing = own\}\} \rangle$  ( $supp(D) = 0.71$ ,  $\tau(C, O) = 0.15$ ,  $\tau(C, \sigma(D)) = 0.18$ , et  $WRAcc(C, D) = 0.012$ ). Cette tendance montre que la corrélation entre les attributs Âge, Durée et le nombre de crédits existants est faible mais plus forte pour les personnes qui sont propriétaires de leur maison.

## 3.5 Conclusion

Dans ce chapitre, nous avons étudié le problème de la découverte de sous-groupes corrélés à l'aide d'un algorithme d'échantillonnage. Cet algorithme permet de calculer ce domaine de motifs de manière stochastique. Sur quatre jeux de données, nous avons pu observer que la méthode d'échantillonnage réduit le coût de calcul, tout en identifiant les motifs les plus importants. Dans le chapitre suivant, nous utilisons cette approche pour analyser un jeu de données réel contenant la description de tous les magasins de Lyon et sa région pendant les 50 dernières années.

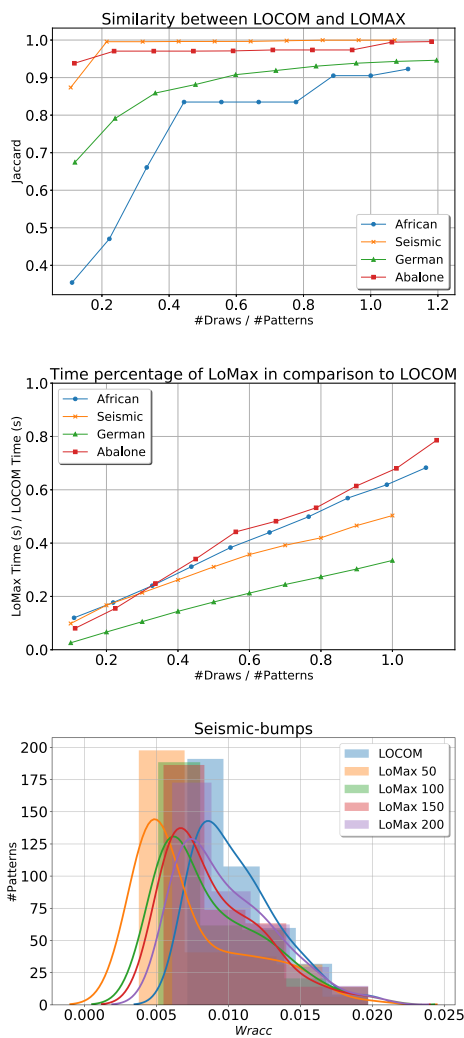


FIGURE 3.3 – (Haut) Valeur de Jaccard entre les paires d’observations décrites par les motifs de l’approche complète et celles décrites par des motifs échantillonnés. L’axe des  $x$  est le nombre de tirages divisé par le nombre de motifs retournés par LOCOM. (Centre) Temps de calcul de LOMAX divisé par celui de LOCOM par rapport au nombre de tirages divisé par le nombre de motifs retournés par LOCOM. (Bas) Valeur de **WRacc** pour différentes taille d’échantillons sur le jeu de données Seismic-bumps.

## Chapitre 4

# Application sur les données

## « Commerces »

Dans ce chapitre, nous montrons comment les algorithmes présentés dans les chapitres précédents peuvent permettre de mettre en évidence des configurations de voisinage qui ont un impact (positif ou négatif) sur la durée de vie des commerces. Dans le cadre du projet RESALI<sup>1</sup>, nous nous intéressons particulièrement aux commerces de type alimentaire afin d'étudier les disparités à l'échelle de l'agglomération lyonnaise.

### 4.1 Contexte

Mieux nourrir les villes en quantité et en qualité, et en particulier les grandes agglomérations, constitue un défi pour les mondes urbains futurs, pensés notamment en termes de durabilité et de justice alimentaire. A l'échelle des systèmes alimentaires urbains, on a besoin de diagnostics et on manque d'outils pour appréhender de façon systématique les relations entre les bassins de consommation, l'offre et les comportements alimentaires. Le projet RESALI propose donc de tester des outils et des méthodes quantitatives pour analyser finement l'organisation des systèmes alimentaires urbains et saisir de façon systématique les connexions / déconnexions entre l'offre alimentaire et la demande ou entre les ressources alimentaires et certains bassins de consommation, même les plus relégués et les moins informés.

Dans la littérature, des travaux proposent déjà des méthodes computationnelles pour mettre en évidence des disparités. Grauwin et al. (2012) proposent un modèle dynamique de la ségrégation résidentielle. Ils présentent une solution analytique pour un modèle de ségrégation non voulue (ségrégation de Schelling). Dujardin and Goffette-Nagot (2010) utilisent des modèles statistiques pour évaluer les effets du voisinage sur le chômage. Concernant l'étude

---

1. RESALI - REseaux et Système ALimentaire Systèmes d'information innovants et exploratoires pour plus de justice alimentaire dans les métropoles (2015).



des commerces, de nombreux travaux cherchent à examiner le rôle collaboratif ou compétitif des types de commerces, par exemple, l'impact d'une nouvelle librairie sur les librairies déjà existantes. Plus récemment, Cambe et al. (2019) étudient l'impact d'un nouveau commerce sur l'écosystème local en considérant un certain nombre de facteurs qui agissent de manière synchrone. Ils définissent un cadre de modélisation qui examine le rôle des nouveaux commerces dans leurs régions respectives. Via une modélisation sous forme de réseaux, ils mettent en évidence différents types d'interactions (i.e., coopération, concurrence) entre les différents commerces. Ils construisent également un modèle pour prédire l'impact d'un nouveau commerce sur son écosystème (i.e., la vente au détail). Cependant, la finalité prédictive de ce travail ne permet pas de mettre en évidence les différents facteurs qui ont un impact sur les commerces.

Dans ce chapitre, nous allons appliquer nos algorithmes afin d'étudier l'impact du voisinage sur la durée de vie des commerces. La section suivante présente les données sur lesquelles notre étude se basera et les questions auxquelles nous souhaitons apporter des éléments de réponse dans le cadre du projet RESALI.

## 4.2 Description des données et objectifs

Nous disposons de données d'ouvertures et de fermetures de baux commerciaux entre 1918 et 2016, fournies par la Chambre de Commerce et d'Industrie (CCI) de Lyon Métropole<sup>2</sup>.

En tout, 225245 baux sont décrits à l'aide de 9 attributs :

- **ID** : Un code unique sur huit caractères identifiant chaque bail.
- **Date début** : Date de début du bail (l'installation du commerce). Sur les données fournies par la CCI, les dates d'ouverture vont de 1918 à 2015.
- **Date fin** : Date de fin du bail. La date de fin la plus ancienne est 1945.
- **Code RIVOLI**<sup>3</sup> : Identifie un lieu (i.e., une commune, une voie, ou dans la majorité des cas un lieu-dit). Ces informations sont extraites du fichier FANTOIR<sup>4</sup> accessible via l'application MAJIC (Mise A Jour des Informations Cadastreales), des services de la DGFIP<sup>5</sup> opérant des missions cadastrales.
- **Code NAF (Nomenclature d'Activités Française)** : Nomenclature des activités économiques productives, principalement élaborée pour faciliter l'organisation de l'information économique et sociale. Afin de faciliter les comparaisons internationales, elle a la même structure

2. <https://www.lyon-metropole.cci.fr>

3. RIVOLI : Répertoire Informatisé des VOies et Lieux-dits

4. FANTOIR : Fichier ANnuaire TOpographique Initialisé Réduit

5. <https://portail.dgfp.finances.gouv.fr/portail/accueilIAM.pl>

que la nomenclature d'activités européenne NACE, elle-même dérivée de la nomenclature internationale CITI. De ce NAF, on peut extraire le code APE (pour Activité Principale Exercée) qui est un code de cinq caractères (quatre chiffres et une lettre) attribué par l'INSEE à toute entreprise et à chacun de ses établissements lors de son inscription au répertoire SIRENE (Système national d'identification et du répertoire des entreprises et de leurs établissements), qui est le répertoire français géré par l'INSEE qui attribue un numéro SIREN aux entreprises, aux organismes et aux associations et un SIRET aux établissements de ces mêmes entreprises, organismes et associations.

- **code ZIP** : Code postal du bail.
- **Longitude** : Longitude correspondant à l'adresse du bail.
- **Latitude** : Latitude correspondant à l'adresse du bail.
- **Grande famille** : Correspond à une agrégation des codes APE par rapport au code hiérarchique, pour décrire la grande famille dans laquelle appartient un commerce. Des exemples de correspondances entre code APE et Grande Famille sont illustrés dans le Tableau 4.1.

APE	Libellé APE	Grande Famille
1071C	Boulangerie et Boulangerie-pâtisserie	Boulangerie
1071B	Cuisson de produits de boulangerie	Boulangerie
1071D	Pâtisserie	Boulangerie
4724Z	Comm. détail pain pâtisserie & confiserie	Boulangerie
47222	Comm. détail viandes & produits à base de viandes	Boucherie
1013B	Charcuterie	Boucherie
4711A	Commerce de détail de produits surgelés	Produits surgelés

TABLE 4.1 – Exemples de correspondances entre code APE et grandes familles.

Notons que les attributs Code ZIP, longitude, latitude et Grande Famille sont dérivés des autres attributs.

La Figure 4.1 décrit le nombre d'ouvertures et de fermetures de baux par année depuis 1900. Même si les premiers baux apparaissent avant les années 1920, le nombre d'ouvertures commence à être important à partir de 1950. De manière similaire, le nombre de fermetures de commerces commence à être important à partir des années 1980. Cela aura une influence sur le choix des données à considérer pour nos analyses. Nous détaillons cela dans la suite de ce chapitre.

La Figure 4.2 représente la distribution par année des différents types de baux pour chaque type de grande famille. On observe qu'en moyenne les commerces ouvrent et ferment autour des années 2000 et que la distribution des ouvertures et des fermetures se concentre entre les années 1990 et 2010.

L'ensemble des baux est cartographié dans les Figures 4.3, 4.4, 4.5 et

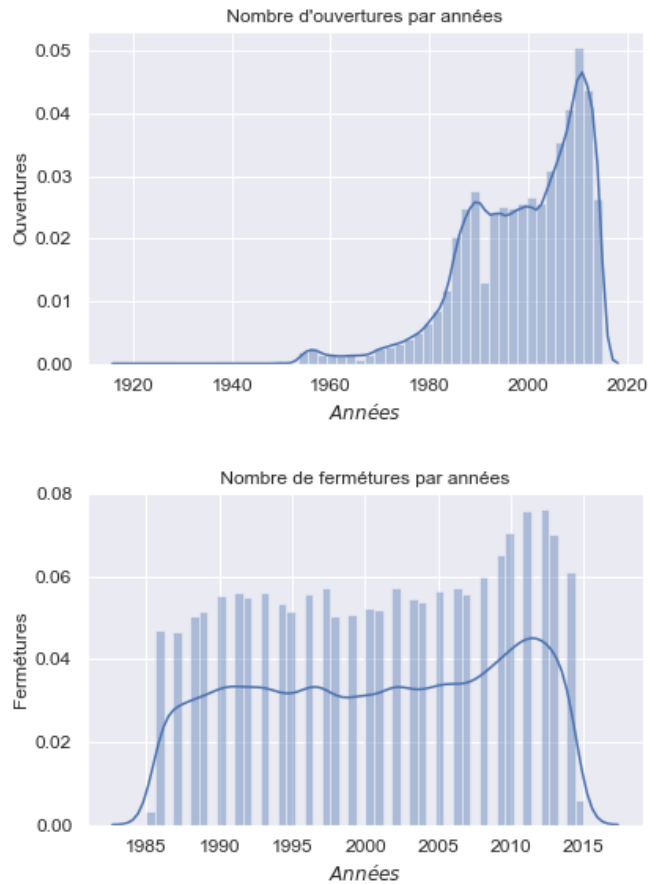


FIGURE 4.1 – Distribution des ouvertures/nombre d'ouvertures totales (resp. fermetures /nombre d'ouvertures totales) des magasins par année.

4.6 en utilisant le système d'information géographique QGIS (Las Palmas 2.18), pour suivre le nombre d'ouvertures et de fermetures de baux à travers différentes périodes temporelles. Ces différentes visualisations permettent de mettre en évidence des zones géographiques plus dynamiques que d'autres. Toutefois, elles s'avèrent insuffisantes pour répondre à des questionnements plus complexes.

L'analyse de ces données à l'aide des algorithmes présentés dans les chapitres précédents peut nous permettre de mettre en évidence des corrélations

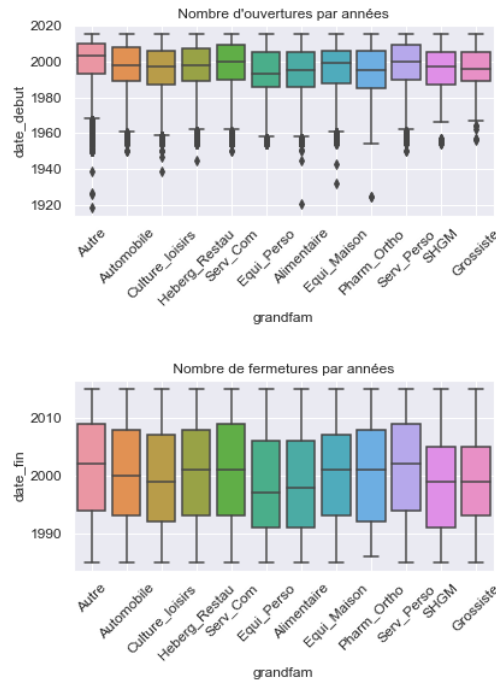


FIGURE 4.2 – Distribution en boxplots des ouvertures/fermetures des magasins par grande famille de commerce à travers le temps.

locales entre la durée de vie d'un commerce et des éléments de son voisinage. Plus particulièrement, l'utilisation des algorithmes développés dans cette thèse peut apporter des éléments de réponses aux questions suivantes :

- Quelles sont les configurations spatiales propices à l'installation d'un commerce ?
- Est-ce que la structure du voisinage a un impact sur la durée de vie d'un commerce ? Si oui, quelles sont les conditions qui garantissent une durée de vie plus importante ?
- Y a-t-il des conditions sur le voisinage qui favorisent la disparition des magasins ?

Dans la suite de ce chapitre, nous allons étudier l'application des algorithmes proposés dans le contexte de l'étude de l'évolution des baux, avec une attention particulière aux commerces de type alimentaire. Nous allons donc essayer d'apporter des éclairages aux questions précédentes à l'aide des motifs

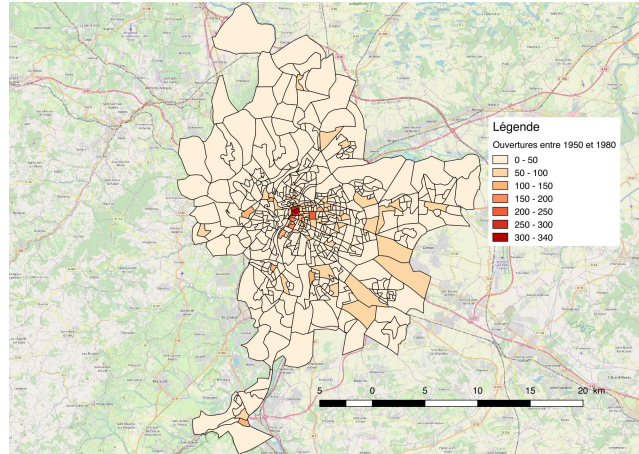


FIGURE 4.3 – Nombres d’ouvertures de commerces de 1950 à 1980.

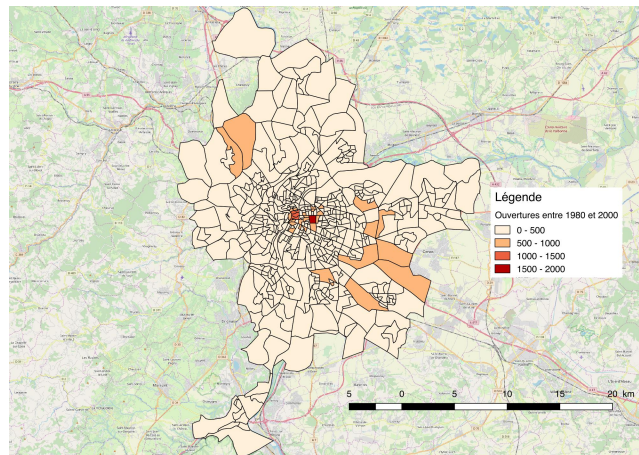


FIGURE 4.4 – Nombres d’ouvertures de commerces de 1980 à 2000.

découverts. Tout d’abord, il est nécessaire de présenter comment nous passons des données de baux commerciaux brutes à une modélisation nous permettant d’appliquer nos algorithmes. L’ensemble des résultats, des scripts et des algorithmes utilisés peut être trouvé sur le lien suivant <https://gitlab.liris.cnrs.fr/mahammal/these>. Les données, quant à elles, ne sont pas disponibles librement.

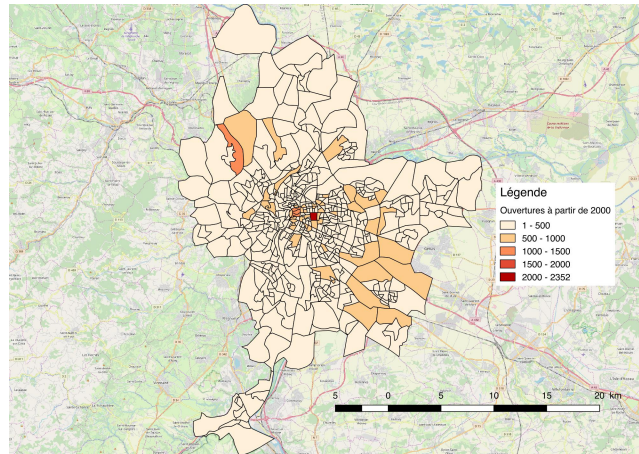


FIGURE 4.5 – Nombres d’ouvertures de commerces depuis 2000.

### 4.3 Modélisation

Nous expliquons comment nous prétraitons les données présentées dans la section précédente afin d’exploiter nos algorithmes pour mettre en évidence des corrélations entre des caractéristiques des baux et celles de leur voisinage. Pour cela, nous adoptons une démarche classique au travers du processus complet d’extraction de connaissances à partir des données (i.e., sélection, prétraitement, transformation, fouille, et interprétation).

Les différentes représentations graphiques et cartographiques présentées dans la section précédente ont permis d’identifier avec les experts géographes les périodes et les zones d’étude pertinentes.

Nous définissons la terminologie fixée avec les géographes :

**Commerces d’analyse :** Ce sont les commerces que l’on cible dans notre étude, c’est-à-dire les commerces alimentaires. Ils sont sélectionnés par le code NAF. En tout, les commerces de type alimentaires sont regroupés au sein de 15 catégories décrites dans la Table 4.2.

**Commerces contextuels :** Des commerces non alimentaires dont nous ne cherchons pas à caractériser l’évolution. Toutefois, ces commerces vont être utilisés pour caractériser les commerces d’analyse. Plus précisément, nous allons définir le voisinage d’un commerce d’analyse en fonction de l’ensemble des commerces (analyse ou contextuels). C’est l’ensemble des commerces y compris ceux qui ne sont pas alimentaires et comportent 70 catégories APE regroupés dans de grandes familles de commerces en récapitulatif dans la Table 4.3.



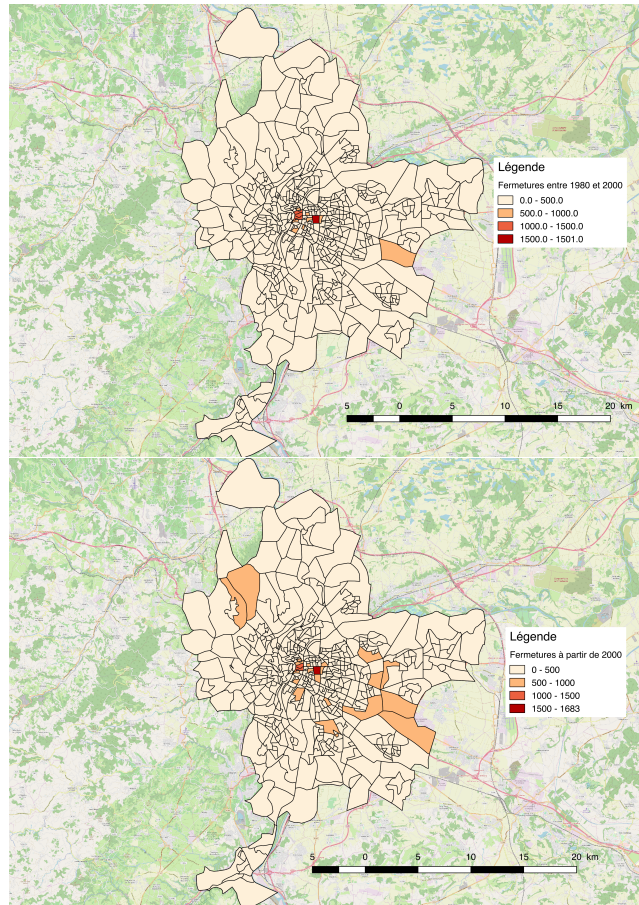


FIGURE 4.6 – Nombres de fermetures de commerces de 1980 à 2000 (haut), et depuis 2000 (bas).

Pour chaque commerce d'analyse, on peut caractériser son voisinage lors de son apparition et lors de sa disparition. Pour calculer un voisinage, nous nous appuyons sur la notion de rayon. Nous considérons tous les commerces dans un rayon de 400 mètres autour du commerce considéré. Cette distance de 400 mètres a été choisie par les géographes. Nous utilisons la formule de Haversine qui permet de prendre en compte le rayon de courbure de la Terre lors du calcul de distance, même si à cette échelle son impact est négligeable.

Nous transformons ainsi le jeu de données initial en un nouveau jeu de données contenant deux types d'attributs :

APE_RESALI	LabelAPE2008	APE_RESALI2
4711B	Commerce d'alimentation générale	EPICERIE
4711C	Supérettes	EPICERIE
4729Z	Autres commerces de détail alimentaires en magasin spécialisé	AUTDETAIL
1013B	Charcuterie	BOUCHAR
4722Z	Commerce de détail de viandes et de produits à base de viande en magasin spécialisé	BOUCHAR
1071B	Cuisson de produits de boulangerie	BOULANG
1071C	Boulangerie et boulangerie-pâtisserie	BOULANG
1071D	Pâtisserie	BOULANG
4724Z	Commerce de détail de pain, pâtisserie et confiserie	BOULANG
4721Z	Commerce de détail de pain, pâtisserie et confiserie en magasin spécialisé	FRUITLEG
4711F	Hypermarchés	HYPERM
4723Z	Commerce de détail de poissons, crustacés et mollusques en magasin spécialisé	POISSON
4711D	Supermarchés	SUPERM
4711E	Magasins multi-commerces	SUPERM
4711A	Commerce de détail de produits surgelés	SURGELE

TABLE 4.2 – Descriptif des principaux commerces alimentaires.

- Les *attributs cibles* sur lesquels nous cherchons à découvrir des corrélations.
- Les *attributs contextuels* permettant d'extraire des sous-groupes pour lesquels des attributs cibles sont davantage corrélés que sur l'ensemble des données, permettant ainsi de comprendre les conditions favorisant ces corrélations.

Au final, ces traitements nous permettent de pouvoir générer plusieurs jeux de données en fonction de ce que nous voulons étudier précisément (e.g., durée de vie, conditions d'ouverture, conditions de fermetures).



GroupeVoisin	Nb CodeAPE
ALIM	11
EQUIP	27
PERSONNE	7
RESTAU	11
SERVICE0	14
Total	70

TABLE 4.3 – Les grandes familles utilisées pour décrire le voisinage des commerces d’analyse.

#### 4.4 Découverte de motifs graduels / corrélations globales

La recherche de corrélations globales ne permet que de mettre en évidence des observations évidentes. Par exemple, le motif (Apparitions+, ALIMD+) décrit le fait que le nombre de commerces apparaissant dans le voisinage d’un commerce est corrélé ( $\mathcal{T}(C, O) = 0.72$ ) au nombre de commerces de types alimentaires en début de bail. On observe aussi que plus la durée de vie d’un commerce est importante, plus le nombre d’apparitions dans son voisinage est importante ( $\mathcal{T}(C, O) = 0.6$ ). Ces corrélations globales ne permettent pas de répondre aux questions que l’on se pose et qui nécessitent la prise en considération de corrélations locales par l’intermédiaire de sous-groupes corrélés.

#### 4.5 Découverte de sous groupes corrélés

Nous présentons maintenant quelque résultats obtenus à l’aide de l’algorithme LOMAX sur les données des baux commerciaux. Plus particulièrement, par rapport aux questionnements exposés précédemment, nous nous intéressons à l’étude de certains facteurs - le nombre d’apparitions de commerces (APP), le nombre de disparitions de commerces (DISP) et la durée de vie de ces commerces (DDV) -, à comment ils se trouvent corrélés avec certaines variables contextuelles et à découvrir les conditions (locales) favorisant certaines corrélations. Pour les variables contextuelles, nous allons considérer séparément le type de commerces dans le voisinage, le voisinage temporel (e.g., est-ce qu’il y a une période temporelle plus propice pour une corrélation), et la spatialité de la corrélation (e.g., est-ce qu’on a des corrélations observées plus fortement dans un sous-espace de la zone d’étude?).

La Table 4.4 décrit des sous-groupes corrélés incluant le nombre d'apparitions de commerces dans les variables corrélées. Le premier motif montre une corrélation globale entre le nombre d'apparitions de commerces et le nombre de commerces services0. Cette corrélation est positive et améliorée quand on s'intéresse à des commerces spécifiques (autdetail). Le deuxième motif montre aussi une corrélation positive entre l'apparition de magasins et des magasins de restauration et de services0, sur le sous-groupe qui définit la période temporelle pour laquelle la corrélation augmente de 20%. Ceci est en accord avec la distribution observée dans la Figure 4.2. Enfin les motifs suivants mettent en évidence les zones géographiques où l'on observe des corrélations bien plus importantes que sur l'ensemble des données. Ainsi, alors que le nombre d'apparitions de commerces et le nombre de commerces de type équipement et alimentaire à l'installation sont corrélés, pour une certaine région cette corrélation augmente significativement de 13%, également pour le nombre d'apparitions de commerces et des magasins de type restauration et SERVICE0 (les commerces de services), l'augmentation est visible et concerne la zone géographique représentée sur la Figure 4.7.

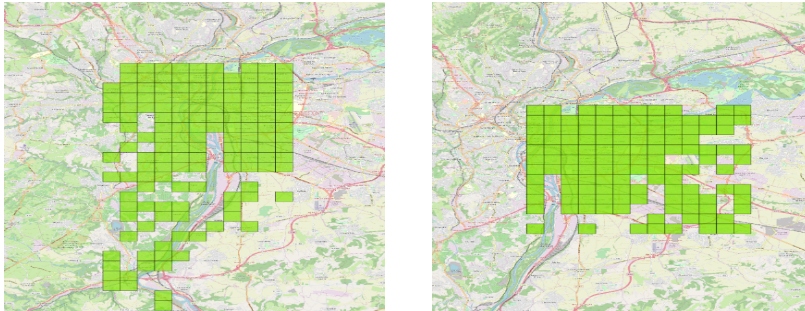


FIGURE 4.7 - : Visualisation des sous-groupes corrélés :  $\langle C=(APP^+, ALIMD^+, EQUIPD^+), D=(\mathbf{x} = [8373540, 847540], \mathbf{y} = [6498135, 6522135]) \rangle$  (gauche) et  $\langle C=(APP^+, RESTAUD^+, SERVICE0D^+), D=(\mathbf{x} = [842540, 854540], \mathbf{y} = [6510135, 6522135]) \rangle$  (droite)

De la même façon, la Table 4.5 illustre des sous-groupes corrélés incluant le nombre de disparitions de commerces dans les attributs cibles. Encore une fois, un contexte sur le type de commerces n'améliore que faiblement la corrélation globale observée. Le second motif est conforme à ce qu'on observe dans la Figure 4.2. Enfin, les contextes mettant en évidence une augmentation de corrélation pour les deux sous-groupes de corrélations pour le contexte géographique, représentés dans la Figure 4.8. Pour la première figure de gauche, plus la disparité -écart type- sur le nombre de disparitions de commerces

est importante plus la disparité sur le nombre de fermetures de commerces alimentaires et de services à la personne, le deuxième motif on retrouve une augmentation de corrélation de 8% entre le nombre de disparitions de commerces dans un voisinage contenant des magasins de services à la personne et de services, pour le sous-groupes représenté sur la figure de gauche (Figure 4.2).

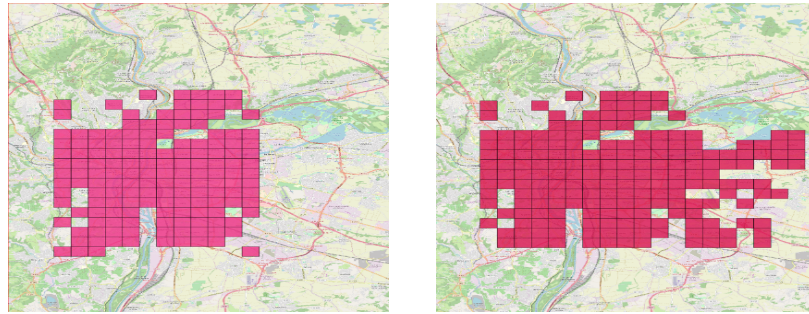


FIGURE 4.8 – : Visualisation des sous-groupes corrélés :  $\langle C=(std(DISP)^+, std(ALIMF)^+, std(PERSONNEF)^+), D=(x=[8373540, 847540], y=[6498135, 6522135]) \rangle$  (gauche) et  $\langle C=(DISP^+, PERSONNEF^+, SERVICE0F^+), D=(x=[842540, 854540], y=[6510135, 6522135]) \rangle$  (droite)

Enfin, la Table 4.6 représente les sous-groupes corrélés incluant la durée de vie des commerces comme attribut cible. Pour le premier sous-groupe on remarque que la durée de vie des magasins est anti-corrélée avec l'apparition de magasins alimentaires, cette corrélation est faiblement améliorée pour si les magasins sont de types épiceries. Dans le deuxième sous-groupe on a une corrélation positive plus la disparité sur la durée de vie est importante plus est la disparité des disparitions de magasins d'équipements, de restaurations et de services dans le voisinage, cette corrélation est un peu améliorée pour des magasins qui durent longtemps sur presque toute la période de l'étude et qui sont de type commerces de détail. Au final, le contexte géographique montre une amélioration de la corrélation sur certaines régions de la ville, pour le premier sous-groupe on a une corrélation positive sur la durée de vie et la disparition de magasins de types restaurants et de services et sur le deuxième une corrélation avec la disparition de magasins de types alimentaires et restaurations avec une amélioration de la corrélation de 14% sur ce dernier.

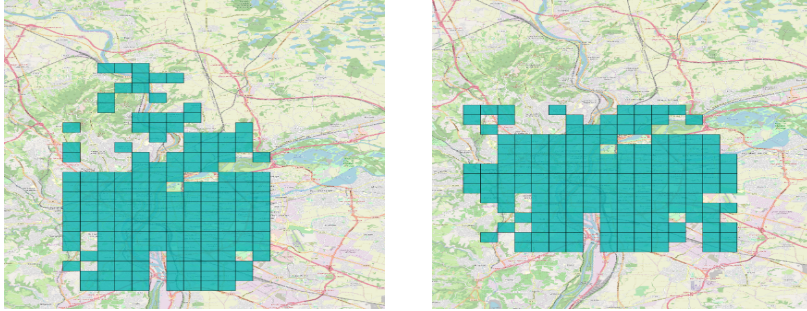


FIGURE 4.9 - : Visualisation des sous-groupes corrélés :  $\langle C=(DDV^+, RESTAUF^+, SERVICE0F^+), D=(x=[836540, 852540], y=[6502135, 6522135]) \rangle$  (gauche) et  $\langle C=(DDV^+, ALIMF^+, RESTAUF^+), D=(x=[842540, 852540], y=[6506135, 6522135]) \rangle$  (droite)

## 4.6 Conclusion

Dans ce chapitre, nous avons étudié l'application des algorithmes définis précédemment pour l'étude des commerces de type alimentaire (et leur durée de vie) afin de mettre en évidence des disparités à l'échelle de l'agglomération lyonnaise. Ces résultats sont encourageants dans la mesure où l'on peut observer des corrélations locales non observables sur l'ensemble des données. La prise en compte de la position spatiale des commerces est l'élément principal qui nous permet de découvrir de tels sous-groupes.

Les perspectives d'amélioration sont nombreuses. Tout d'abord, il faut mieux intégrer les experts dans l'analyse des résultats, et donc définir des processus pleinement interactifs. Ensuite, d'autres axes d'amélioration sur la mesure considérée sont possibles. Tout d'abord, la mesure que l'on considère (WRAcc) n'est pas intuitive pour un utilisateur, il faut penser une mesure qui véhicule une information plus directement assimilable pour un utilisateur non expert des processus de découverte de connaissances dans les données et/ou de découverte de sous-groupes. Enfin, nous utilisons le  $\tau$  de Kendall pour capturer les corrélations observées assez faibles. Des travaux s'attaquent à ce problème, notamment dans la communauté de la logique floue (Marsala et al., 2018). L'intégration de telles mesures dans nos algorithmes pourrait nous permettre de capturer d'autres sous-groupes corrélés intéressants.

$C$	$D$	$\tau(C, O)$	$\sigma(D)$	$\tau(C, \sigma(D))$	WRAcc
	Catégorie				
APP <sup>+</sup> , SERVICE0D <sup>+</sup>	APERESALI = AUTDE/TAIL	0.665	0.11	0.711	0.00063
	Contexte temporel				
APP <sup>+</sup> , RESTAUD <sup>+</sup> , SERVICE0D <sup>+</sup>	Début = [1980, 2012], Fin = [1985, 2015]	0.583	0.72	0.701	0.0661
	Contexte Spatial				
APP <sup>+</sup> , ALIMD <sup>+</sup> , EQUIPD <sup>+</sup>	$x = [8373540, 847540], y = [6498135, 6522135]$	0.558	0.541	0.634	0.0223
APP <sup>+</sup> , RESTAUD <sup>+</sup> , SERVICE0D <sup>+</sup>	$x = [842540, 854540], y = [6510135, 6522135]$	0.625	0.499	0.696	0.0174

TABLE 4.4 – Étude des conditions d'apparition des commerces de type alimentaire.

$C$	$D$	$\pi(C, O)$	$\sigma(D)$	$\pi(C, \sigma(D))$	WRAcc
	Catégorie				
DISP <sup>+</sup> , ALIMF <sup>+</sup> , EQUIPF <sup>+</sup>	APERESALI = AUTDETAIL	0.618	0.11	0.628	0.000131
	Contexte temporel				
DISP <sup>+</sup> , RESTAUF <sup>+</sup> , SERVICE0F <sup>+</sup>	Début = [1982, 2012], Fin = [1986, 2013]	0.608	0.714	0.677	0.035
	Contexte Spatial				
$std(DISP)^+$ , $std(ALIMF)^+$ , $std(PERSONNEF)^+$	$x = [8373540, 847540], y = [6498135, 6522135]$	0.585	0.634	0.629	0.028
DISP <sup>+</sup> , PERSONNEF <sup>+</sup> , SERVICE0F <sup>+</sup>	$x = [842540, 854540], y = [6510135, 6522135]$	0.617	0.763	0.671	0.0316

TABLE 4.5 – Étude des conditions de disparition des commerces de type alimentaire.

$C$	$D$	$\tau(C, O)$	$\sigma(D)$	$\tau(C, \sigma(D))$	$WR_{Acc}$
<b>Catégorie</b>					
$DDV^+, ALIMD^-$	$APERESALI = EPICERIE$	0.492	0.443	0.513	0.004
<b>Contexte temporel</b>					
$std((DDV)^+, std(EQUIPF)^+, std(SERVICEOF)^+)$	Début = [1954, 2014], Fin = [1958, 2015]	0.3363	0.149	0.400	0.0008
<b>Contexte Spatial</b>					
$DDV^+, RESTAUF^+, SERVICEOF^+$	$x = [836540, 852540], y = [6502135, 6522135]$	0.397	0.700	0.422	0.0125
$DDV^+, ALIMF^+, RESTAUF^+$	$x = [842540, 852540], y = [6506135, 6522135]$	0.356	0.490	0.406	0.0120

TABLE 4.6 – Étude des conditions de durabilité des commerces de type alimentaire.

# Conclusion

De nombreuses problématiques géographiques, et notamment celles liées à l'alimentation, peuvent être étudiées à partir de données massives, qu'elles proviennent de VGI (Volunteered Geographic Information) – des informations géographiques participatives et citoyennes issues de réseaux sociaux – ou de collectes exhaustives de données issues de la numérisation d'activités spécifiques, comme les baux commerciaux enregistrés par la Chambre de Commerce et d'Industrie (CCI). Dans les deux cas, les données à analyser partagent des caractéristiques communes : elles sont massives, hétérogènes (alliant attributs symboliques et numériques), intègrent une composante géospatiale, ainsi qu'une dimension temporelle. Une telle complexité est rarement prise en compte dans les outils standard d'analyse de données, or l'analyse de leurs contenus peut offrir une vision incomparable des structures socio-culturelles et de leurs dynamiques, tout en étant aisément cartographiables.

Au cours de notre travail de thèse, nous avons alors cherché à répondre à cet enjeu en proposant de nouvelles méthodes pour identifier des phénomènes exceptionnels propres à des zones géographiques. Dans les approches que nous avons développées, les phénomènes recherchés, des corrélations entre attributs numériques, et les zones géographiques sont obtenus de manière inductive, c'est-à-dire à partir des données sans que l'une ou l'autre ne soit donnée par l'analyste.

Notre première contribution a consisté à formaliser le nouveau problème de la découverte de sous-groupes corrélés sur le rang avec un nombre arbitraire de cibles numériques (supérieur ou égal à 2). Chaque sous-groupe d'objets est identifié par des conditions sur des attributs numériques et/ou nominaux et a pour particularité d'avoir une corrélation de rang exceptionnellement grande par rapport à celle observée sur la totalité des données. Nous avons défini un nouvel algorithme, LOCOM, qui exploite certaines propriétés d'élagage basées sur deux bornes supérieures et sur des propriétés de fermeture. Une étude empirique sur plusieurs ensembles de données de référence a démontré l'efficacité de LOCOM.

Dans une deuxième contribution, nous avons proposé un algorithme par échantillonnage, LOMAX, pour calculer le même domaine de motifs, mais de manière heuristique avec pour objectif de réduire le temps de calcul des motifs. Sur quatre jeux de données de référence, nous avons pu observer que



la méthode d'échantillonnage réduit le coût de calcul, tout en identifiant les motifs les plus importants.

Notre troisième contribution a consisté à utiliser ces algorithmes pour analyser l'implantation et la durée de vie des commerces de type alimentaire. Cette étude a permis de mettre en évidence des disparités à l'échelle de l'agglomération lyonnaise.

## Perspectives

Ce travail ouvre de nouvelles perspectives de recherche.

**Améliorer l'algorithme LOMAX** L'approche par échantillonnage que nous avons proposée consiste à partir d'un attribut signé positivement et à ajouter progressivement d'autres attributs signés selon une probabilité proportionnelle à la valeur de qualité du motif obtenu. Le processus s'arrête quand aucun attribut signé ne peut être ajouté sans que la valeur de corrélation ne soit inférieure au seuil. Une fois le motif de corrélation obtenu, les sous-groupes maximisant la mesure WRAcc sont obtenus en utilisant l'algorithme d'énumération Locom. Une extension de ce travail pourrait être d'utiliser une approche par échantillonnage pour identifier le motif corrélé mais aussi le sous-groupe afin de rendre l'algorithme encore plus rapide.

**Utiliser une autre mesure de corrélation** La mesure  $\tau$  de Kendall permet d'évaluer le degré de corrélation de rangs entre plusieurs attributs numériques. Cependant, cette mesure est très sensible aux égalités sur un même attribut alors que ce phénomène se produit fréquemment dans les données réelles, ce qui engendre des corrélations observées assez faibles numériquement. Des travaux s'attaquent à ce problème, notamment dans la communauté de la logique floue (Marsala et al., 2018). L'intégration de telles mesures dans nos algorithmes pourrait nous permettre de capturer d'autres sous-groupes corrélés intéressants.

**Prendre en compte les connaissances a priori de l'utilisateur** Afin de produire des motifs plus adaptés aux besoins de l'utilisateur, d'autres mesures et paradigmes peuvent être étudiés pour évaluer l'intérêt des sous-groupes, en particulier l'intérêt subjectif qui permet de prendre en compte les connaissances a priori de l'utilisateur (Bie, 2011). Cette méthode vise à intégrer les connaissances de l'utilisateur et prend en compte le coût d'assimilation d'un motif par celui-ci. Ainsi, on recherche les motifs à la fois inattendus par rapport à un modèle de connaissances, et faciles à assimiler par l'utilisateur. Cela pourrait permettre de pallier les difficultés d'interprétation de la mesure WRAcc.

# Bibliographie

- Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.
- Martin Atzmüller and Frank Puppe. Sd-map - A fast algorithm for exhaustive subgroup discovery. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *Knowledge Discovery in Databases : PKDD 2006, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, Germany, September 18-22, 2006, Proceedings*, volume 4213 of *Lecture Notes in Computer Science*, pages 6–17. Springer, 2006.
- Stephen D Bay and Michael J Pazzani. Detecting group differences : Mining contrast sets. *Data mining and knowledge discovery*, 5(3) :213–246, 2001.
- Adnene Belfodil, Sylvie Cazalens, Philippe Lamarre, and Marc Plantevit. Flash points : Discovering exceptional pairwise behaviors in vote or rating data. In *ECML PKDD*, pages 442–458, 2017.
- Aimene Belfodil. *An Order Theoretic Point-of-view on Subgroup Discovery*. Theses, Université de Lyon, September 2019. URL <https://hal.archives-ouvertes.fr/tel-02355529>.
- Aimene Belfodil, Adnene Belfodil, and Mehdi Kaytoue. Anytime subgroup discovery in numerical domains with guarantees. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part II*, pages 500–516, 2018.
- Ahmed Anes Bendimerad, Marc Plantevit, and Céline Robardet. Unsupervised exceptional attributed sub-graph mining in urban data. In Francesco Bonchi, Josep Domingo-Ferrer, Ricardo Baeza-Yates, Zhi-Hua Zhou, and Xindong Wu, editors, *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, pages 21–30. IEEE Computer Society, 2016.

- Ahmed Anes Bendimerad, Rémy Cazabet, Marc Plantevit, and Céline Robardet. Contextual subgraph discovery with mobility models. In *COMPLEX NETWORKS*, pages 477–489, 2017.
- Anes Bendimerad. *Mining Useful Patterns in Attributed Graphs*. Theses, Université de Lyon, September 2019. URL <https://hal.archives-ouvertes.fr/tel-02284436>.
- Anes Bendimerad, Jefrey Lijffijt, Marc Plantevit, Céline Robardet, and Tijl De Bie. Contrastive antichains in hierarchies. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis, editors, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019.*, pages 294–304. ACM, 2019a.
- Anes Bendimerad, Marc Plantevit, Céline Robardet, and Sihem Amer-Yahia. User-driven geolocated event detection in social media. *IEEE Transactions on Knowledge and Data Engineering*, 2019b.
- Jérémy Besson, Céline Robardet, Jean-François Boulicaut, and Sophie Rome. Constraint-based concept mining and its application to microarray data analysis. *Intelligent Data Analysis*, 9(1) :59–82, 2005.
- Tijl De Bie. Maximum entropy models and subjective interestingness : an application to tiles in binary databases. *Data Mining and Knowledge Discovery*, 23(3) :407–446, 2011.
- Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Zelezný, editors. *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part I*, volume 8188 of *Lecture Notes in Computer Science*, 2013. Springer. ISBN 978-3-642-40987-5. doi : 10.1007/978-3-642-40988-2. URL <https://doi.org/10.1007/978-3-642-40988-2>.
- Mario Boley, Thomas Gärtner, and Henrik Grosskreutz. Formal concept sampling for counting and threshold-free local pattern mining. In *SDM*, pages 177–188, 2010.
- Mario Boley, Claudio Lucchese, Daniel Paurat, and Thomas Gärtner. Direct local pattern sampling by efficient two-step random procedures. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pages 582–590, 2011.

- Mario Boley, Sandy Moens, and Thomas Gärtner. Linear space direct pattern sampling using coupling from the past. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, pages 69–77, 2012.
- Francesco Bonchi and Claudio Lucchese. Extending the state-of-the-art of constraint-based pattern discovery. *Data & Knowledge Engineering*, 60(2) : 377–399, 2007.
- Christian Borgelt and Michael R. Berthold. Mining molecular fragments : Finding relevant substructures of molecules. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*, pages 51–58. IEEE Computer Society, 2002.
- Guillaume Bosc, Jérôme Golebiowski, Moustafa Bensafi, Céline Robardet, Marc Plantevit, Jean-François Boulicaut, and Mehdi Kaytoue. Local subgroup discovery for eliciting and understanding new structure-odor relationships. In Calders et al. (2016), pages 19–34.
- Guillaume Bosc, Jean-François Boulicaut, Chedy Raïssi, and Mehdi Kaytoue. Anytime discovery of a diverse set of patterns with monte carlo tree search. *Data Mining and Knowledge Discovery*, 32(3) :604–650, 2018.
- Toon Calders, Christophe Rigotti, and Jean-François Boulicaut. A survey on condensed representations for frequent sets. In Jean-François Boulicaut, Luc De Raedt, and Heikki Mannila, editors, *Constraint-Based Mining and Inductive Databases, European Workshop on Inductive Databases and Constraint Based Mining, Hinterzarten, Germany, March 11-13, 2004, Revised Selected Papers*, volume 3848 of *Lecture Notes in Computer Science*, pages 64–80. Springer, 2004.
- Toon Calders, Bart Goethals, and Szymon Jaroszewicz. Mining rank-correlated sets of numerical attributes. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 96–105. ACM, 2006.
- Toon Calders, Michelangelo Ceci, and Donato Malerba, editors. *Discovery Science - 19th International Conference, DS 2016, Bari, Italy, October 19-21, 2016, Proceedings*, volume 9956 of *Lecture Notes in Computer Science*, 2016. ISBN 978-3-319-46306-3. doi : 10.1007/978-3-319-46307-0. URL <https://doi.org/10.1007/978-3-319-46307-0>.
- Jordan Cambe, Krittika D'Silva, Anastasios Noulas, Cecilia Mascolo, and Adam Waksman. Modelling cooperation and competition in urban retail ecosystems with complex network metrics. 2019.

- Loïc Cerf, Jérémy Besson, Céline Robardet, and Jean-François Boulicaut. Closed patterns meet  $n$ -ary relations. *ACM Transactions on Knowledge Discovery from Data*, 3(1) :3 :1–3 :36, 2009.
- Vineet Chaoji, Mohammad Al Hasan, Saeed Salem, Jérémy Besson, and Mohammed J. Zaki. ORIGAMI. *Statistical Analysis and Data Mining*, 1(2) : 67–84, 2008.
- Cláudio Rebelo de Sá, Wouter Duivesteijn, Carlos Soares, and Arno J. Knobbe. Exceptional preferences mining. In Calders et al. (2016), pages 3–18.
- Cláudio Rebelo de Sá, Wouter Duivesteijn, Paulo J. Azevedo, Alípio Mário Jorge, Carlos Soares, and Arno J. Knobbe. Discovering a taste for the unusual : exceptional models for preference mining. *Machine Learning*, 107 (11) :1775–1807, 2018.
- María José del Jesús, Pedro González, Francisco Herrera, and Mikel Mesonero. Evolutionary fuzzy rule induction process for subgroup discovery : A case study in marketing. *IEEE Trans. Fuzzy Systems*, 15(4) :578–592, 2007.
- Elise Desmier, Marc Plantevit, Céline Robardet, and Jean-François Boulicaut. Trend mining in dynamic attributed graphs. In Blockeel et al. (2013), pages 654–669.
- Lamine Diop, Cheikh Talibouya Diop, Arnaud Giacometti, Dominique Li, and Arnaud Soulet. Sequential pattern sampling with norm constraints. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*, pages 89–98. IEEE Computer Society, 2018.
- Trong Dinh Thac Do, Anne Laurent, and Alexandre Termier. PGLCM : efficient parallel mining of closed frequent gradual itemsets. In Geoffrey I. Webb, Bing Liu, Chengqi Zhang, Dimitrios Gunopulos, and Xindong Wu, editors, *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, pages 138–147. IEEE Computer Society, 2010.
- Trong Dinh Thac Do, Alexandre Termier, Anne Laurent, Benjamin Negre-Verigne, Behrooz Omidvar-Tehrani, and Sihem Amer-Yahia. Pglcm : efficient parallel mining of closed frequent gradual itemsets. *Knowledge and Information Systems*, 43(3) :497–527, 2015.
- Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns : Discovering trends and differences. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 43–52. ACM, 1999.

- Lennart Downar and Wouter Duivesteijn. Exceptionally monotone models—the rank correlation model class for exceptional model mining. *Knowledge and Information Systems*, 51(2) :369–394, May 2017.
- Wouter Duivesteijn and Julia Thaele. Understanding where your classifier does (not) work - the scape model class for EMM. In *2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, December 14-17, 2014*, pages 809–814, 2014.
- Wouter Duivesteijn, Arno J. Knobbe, Ad Feelders, and Matthijs van Leeuwen. Subgroup discovery meets bayesian networks – an exceptional model mining approach. In *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, pages 158–167, 2010.
- Wouter Duivesteijn, Ad Feelders, and Arno J. Knobbe. Different slopes for different folks : mining for exceptional regression models with cook’s distance. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’12, Beijing, China, August 12-16, 2012*, pages 868–876, 2012.
- Wouter Duivesteijn, Ad Feelders, and Arno J. Knobbe. Exceptional model mining - supervised descriptive local pattern mining with complex target concepts. *Data Mining and Knowledge Discovery*, 30(1) :47–98, 2016.
- Claire Dujardin and Florence Goffette-Nagot. Neighborhood effects on unemployment ? : A test a la altonji. *Regional Science and Urban Economics*, 40 (6) :380 – 396, 2010.
- Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage Uday Kiran, and Yun Sing Koh. A survey of sequential pattern mining. *Data Science and Pattern Recognition*, 1(1) :54–77, 2017.
- Gillian Fyfe. *Poor and paying for it : The price of living on a low income*. HM Stationery Office, 1994.
- Dragan Gamberger and Nada Lavrac. Expert-guided subgroup discovery : Methodology and application. *Journal of Artificial Intelligence Research (JAIR)*, 17 :501–527, 2002.
- Arnaud Giacometti and Arnaud Soulet. Dense neighborhood pattern sampling in numerical data. In Martin Ester and Dino Pedreschi, editors, *Proceedings of the 2018 SIAM International Conference on Data Mining, SDM 2018, May 3-5, 2018, San Diego Marriott Mission Valley, San Diego, CA, USA*, pages 756–764. SIAM, 2018.
- Robert Gottlieb and Anupama Joshi. Food justice. *Hardback : MIT Press*, 2010.

- Sébastien Grauwin, Florence Goffette-Nagot, and Pablo Jensen. Dynamic models of residential segregation : An analytical solution. *Journal of Public Economics*, 96(1) :124 – 141, 2012.
- Henrik Grosskreutz and Stefan Rüping. On subgroup discovery in numerical domains. *Data Mining and Knowledge Discovery*, 19(2) :210–226, 2009.
- Henrik Grosskreutz, Bastian Lang, and Daniel Trabold. A relevance criterion for sequential patterns. In Blockeel et al. (2013), pages 369–384.
- Mohamed-Ali Hammal, Bernardo Abreu, Marc Plantevit, and Céline Robardet. Sampling rank correlated subgroups. In Francisco Herrera, Kenji Matsui, and Sara Rodríguez-González, editors, *Distributed Computing and Artificial Intelligence, 16th International Conference, DCAI 2019, Avila, Spain, 26-28 June, 2019*, volume 1003 of *Advances in Intelligent Systems and Computing*, pages 217–225. Springer, 2019a.
- Mohamed-Ali Hammal, Hélène Mathian, Luc Merchez, Marc Plantevit, and Céline Robardet. Rank correlated subgroup discovery. *J. Intell. Inf. Syst.*, 53(2) :305–328, 2019b.
- Mohamed-Ali Hammal, Céline Robardet, and Marc Plantevit. Quand les sous-groupes rencontrent les graduels : découverte de sous-groupes identifiant des corrélations exceptionnelles. In Marie-Christine Rousset and Lydia Boudjeloud-Assala, editors, *Extraction et Gestion des connaissances, EGC 2019, Metz, France, January 21-25, 2019*, volume E-35 of *RNTI*, pages 201–212. Éditions RNTI, 2019c.
- Eyke Hüllermeier. Association rules for expressing gradual dependencies. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Principles of Data Mining and Knowledge Discovery, 6th European Conference, PKDD 2002, Helsinki, Finland, August 19-23, 2002, Proceedings*, volume 2431 of *Lecture Notes in Computer Science*, pages 200–211. Springer, 2002.
- Robert T Jensen and Nolan H Miller. Giffen behavior and subsistence consumption. *American Economic Review*, 98(4) :1553–77, 2008.
- Mehdi Kaytoue, Sergei O. Kuznetsov, and Amedeo Napoli. Revisiting numerical pattern mining with formal concept analysis. In Toby Walsh, editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 1342–1347. IJCAI/AAAI, 2011.
- Mehdi Kaytoue, Marc Plantevit, Albrecht Zimmermann, Anes Bendimerad, and Céline Robardet. Exceptional contextual subgraph mining. *Machine Learning*, 2017.

- Yiping Ke, James Cheng, and Jeffrey Xu Yu. Top-k correlative graph mining. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA*, pages 1038–1049. SIAM, 2009.
- Willi Klösgen. Knowledge discovery in databases and data mining. In Zbigniew W. Raś and Maciek Michalewicz, editors, *Foundations of Intelligent Systems*, pages 623–632, Berlin, Heidelberg, 1996. Springer Berlin Heidelberg.
- Willi Klogsen. Explora : A multipattern and multistrategy discovery assistant. *Advances in knowledge discovery and data mining*, 1996.
- Nada Lavrac, Peter A. Flach, and Blaz Zupan. Rule evaluation measures : A unifying view. In *Inductive Logic Programming, 9th International Workshop, ILP-99, Bled, Slovenia, June 24-27, 1999, Proceedings*, pages 174–185, 1999.
- Nada Lavrac, Bojan Cestnik, Dragan Gamberger, and Peter A. Flach. Decision support through subgroup discovery : Three case studies and the lessons learned. *Machine Learning*, 57(1-2) :115–143, 2004.
- Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupčo Todorovski. Subgroup discovery with cn2-sd. *Journal of Machine Learning Research*, 5(Feb) : 153–188, 2004.
- Dennis Leman, Ad Feelders, and Arno J. Knobbe. Exceptional model mining. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML/PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II*, pages 1–16, 2008.
- Florian Lemmerich, Martin Atzmueller, and Frank Puppe. Fast exhaustive subgroup discovery with numerical target concepts. *Data Mining and Knowledge Discovery*, 30(3) :711–762, 2016.
- Rémi Lemoy, Charles Raux, and Pablo Jensen. An agent-based model of residential patterns and social structure in urban areas. *Cybergeo : European Journal of Geography*, 2010.
- Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. In *KDD*, pages 80–86, 1998.
- Heikki Mannila and Hannu Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data mining and knowledge discovery*, 1(3) :241–258, 1997.



- Christophe Marsala, Anne Laurent, Marie-Jeanne Lesot, Maria Rifqi, and Arnaud Castelltort. Discovering ordinal attributes through gradual patterns, morphological filters and rank discrimination measures. In *Scalable Uncertainty Management - 12th International Conference, SUM 2018, Milan, Italy, October 3-5, 2018, Proceedings*, pages 152–163, 2018.
- Sandy Moens and Bart Goethals. Randomly sampling maximal itemsets. In Duen Horng Chau, Jilles Vreeken, Matthijs van Leeuwen, and Christos Faloutsos, editors, *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics, IDEA@KDD 2013, Chicago, Illinois, USA, August 11, 2013*, pages 79–86. ACM, 2013.
- Shinichi Morishita and Jun Sese. Transversing itemset lattices with statistical metric pruning. In *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '00*, pages 226–236, New York, NY, USA, 2000. ACM.
- Kimberly Morland, Steve Wing, Ana Diez Roux, and Charles Poole. Neighborhood characteristics associated with the location of food stores and food service places. *American journal of preventive medicine*, 22(1) :23–29, 2002.
- Emirgenâ Nikolli. Economic growth and unemployment rate. case of albania. *European Journal of Social Sciences Education and Research Articles*, 1 :1, 2014.
- Victoria Pachón, Jacinto Mata Vázquez, Juan Luis Domínguez, and Manuel J. Maña López. Multi-objective evolutionary approach for subgroup discovery. In *Hybrid Artificial Intelligent Systems - 6th International Conference, HAIS 2011, Wroclaw, Poland, May 23-25, 2011, Proceedings, Part II*, pages 271–278, 2011.
- Gregory Piatetsky-Shapiro, Ronald J. Brachman, Tom Khabaza, Willi Klösgen, and Evangelos Simoudis. An overview of issues in developing industrial data mining and knowledge discovery applications. In Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, pages 89–95. AAAI Press, 1996.
- Adriana Prado, Marc Plantevit, Céline Robardet, and Jean-Francois Boulicaut. Mining graph topological patterns : Finding covariations among vertex descriptors. *IEEE Transactions on Knowledge and Data Engineering*, 25(9) :2090–2104, 2013.
- Samina Raja, Changxing Ma, and Pavan Yadav. Beyond food deserts : measuring and mapping racial disparities in neighborhood food environments. *Journal of Planning Education and Research*, 27(4) :469–482, 2008.

- Jean-Louis Rastoin and Gérard Gherzi. Le système alimentaire mondial : concepts et méthodes, analyses et dynamiques. *Economie rurale*, 329 :98–99, 2012.
- Daniel Rodríguez, Roberto Ruiz, José C. Riquelme, and Jesús S. Aguilar-Ruiz. Searching for rules to detect defective modules : A subgroup discovery approach. *Information Sciences*, 191 :14–30, 2012.
- Aika Terada, Mariko Okada-Hatakeyama, Koji Tsuda, and Jun Sese. Statistical significance of combinatorial regulations. *Proceedings of the National Academy of Sciences*, 110(32) :12996–13001, 2013.
- John W Tukey. *Exploratory data analysis*. Reading, Mass., 1977.
- Willy Ugarte, Patrice Boizumault, Bruno Crémilleux, Alban Lepailleur, Samir Loudni, Marc Plantevit, Chedy Raïssi, and Arnaud Soulet. Skypattern mining : From pattern condensed representations to dynamic constraint satisfaction problems. *Artificial Intelligence*, 244 :48–69, 2017.
- Matthijs van Leeuwen and Arno J. Knobbe. Diverse subgroup set discovery. *Data Mining and Knowledge Discovery*, 25(2) :208–242, 2012.
- Renee E Walker, Christopher R Keane, and Jessica G Burke. Disparities and access to healthy food in the united states : A review of food deserts literature. *Health & place*, 16(5) :876–884, 2010.
- Jianyong Wang, Jiawei Han, Ying Lu, and Petre Tzvetkov. TFP : an efficient algorithm for mining top-k frequent closed itemsets. *IEEE Transactions on Knowledge Data Engineering*, 17(5) :652–664, 2005.
- Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In Jan Komorowski and Jan Zytkow, editors, *Principles of Data Mining and Knowledge Discovery*, pages 78–87, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg.
- Dong Xin, Hong Cheng, Xifeng Yan, and Jiawei Han. Extracting redundancy-aware top-k patterns. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 444–453. ACM, 2006.
- Mohammed J Zaki and Ching-Jui Hsiao. Charm : An efficient algorithm for closed itemset mining. In *Proceedings of the 2002 SIAM international conference on data mining*, pages 457–473. SIAM, 2002.





## FOLIO ADMINISTRATIF

### THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : HAMMAL

DATE de SOUTENANCE : 05/06/2020

(avec précision du nom de jeune fille, le cas échéant)

Prénom : Mohamed Ali

**TITRE : Contribution à la découverte de sous-groupes corrélés : Application à l'analyse des systèmes territoriaux et des réseaux alimentaires.**

NATURE : Doctorat

Numéro d'ordre : 2020LYSEI024

Ecole doctorale : InfoMaths (ED 512)

Spécialité : Informatique

#### RESUME :

Mieux nourrir les villes en quantité et en qualité, notamment les grandes agglomérations, constitue un défi majeur dont la résolution passe par une meilleure compréhension des relations entre les populations urbaines et leur alimentation. A l'échelle des systèmes alimentaires urbains, on a besoin de diagnostics ciblant la disponibilité des ressources alimentaires croisée avec les profils socio-économiques des territoires et l'on manque d'outils et de méthodes pour appréhender de façon systématique les relations entre les bassins de consommation, l'offre et les comportements alimentaires. L'objectif de cette thèse est de contribuer à l'élaboration de nouveaux outils informatiques pour traiter des données temporelles, hétérogènes et multi-sources afin d'identifier et de caractériser des comportements propres à une zone géographique. Pour cela, nous nous appuyons sur l'exploration conjointe de motifs graduels, identifiant des corrélations de rang, et de sous-groupes afin de découvrir des contextes pour lesquels les corrélations décrites par les motifs graduels sont exceptionnellement fortes par rapport au reste des données. Nous proposons un algorithme d'énumération s'appuyant sur des propriétés d'élagage avec des bornes supérieures, ainsi qu'un autre algorithme qui échantillonne les motifs selon la mesure de qualité. Ces approches sont validées non seulement sur des jeux de données de référence, mais aussi à travers une étude empirique de la formation des déserts alimentaires sur l'agglomération lyonnaise.

**MOTS-CLÉS :** Découverte de connaissances, fouille de motifs, sous-groupes corrélés, échantillonnage de motifs, analyse des déserts alimentaires.

Laboratoire (s) de recherche : **L**aboratoire d'**I**nfo**R**matique en **I**mage et **S**ystèmes d'**I**nformation (**LIRIS**)

#### Directeur de thèse :

Céline ROBARDET (Professeure des Universités, INSA Lyon, Directrice de thèse)

Luc MERCHEZ (Maitre de Conférences, ENS Lyon, Co-Directeur)

#### Président de jury :

Christine LARGERON (professeure à l'université Jean Monnet, Présidente)

#### Composition du jury :

Dino IENCO (Chargé de recherche HDR à l'INRAE, rapporteur)

Alexandre TERMIER (professeur en informatique à l'université Rennes 1, rapporteur)

Christine LARGERON (professeure à l'université Jean Monnet, examinatrice)

Benjamin NEGREVERGNE (maître de conférences à l'université Paris-Dauphine, examinateur)

Marc PLANTEVIT (Maitre de conférences HDR UCBL, Invité)

