Méthodes d'Extraction de Connaissances à partir de Données (ECD) appliquées aux Systèmes d'Information Géographiques (SIG)

Christophe Candillier

UNIVERSITÉ DE NANTES
FACULTÉ DES SCIENCES

ÉCOLE DOCTORALE SCIENCES ET TECHNOLOGIES
DE L'INFORMATION ET DES MATÉRIAUX

Année 2006

N° attribué par la bibliothèque

Méthodes d'Extraction de Connaissances à partir de Données (ECD) appliquées aux Systèmes d'Information Géographiques (SIG)

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE NANTES

Discipline : Informatique

présentée et soutenue publiquement par

Christophe CANDILLIER

Le 21 septembre 2006

à l'UFR Sciences et Techniques, Université de Nantes

devant le jury ci-dessous

Président	:	Michel SCHOLL, Professeur	CNAM Paris
Rapporteurs	:	Anne DOUCET, Professeur	Université Paris 6
		Mohand-Saïd HACID, Professeur	Université Claude-Bernard Lyon 1
Examinateurs	:	Pierre COURONNÉ, PDG	Société GÉOBS
		Marc GELGON, Maître de Conférence	Université de Nantes
Directeur de thèse	:	Noureddine MOUADDIB, Professeur	Université de Nantes
Directeur scientifique	:	Mohamed QUAFAFOU, Professeur	Université de la Méditerranée

Laboratoire d'Informatique de Nantes Atlantique N° ED 366-262

COLLABORATION DE RECHERCHE ET PROJET DE R&D GÉOBS

Collaboration de Recherche

Le travail de thèse s'est déroulé principalement dans les locaux de la société GÉOBS. Il a fait l'objet d'un contrat de collaboration de recherche dans le cadre d'une Convention Industrielle de Formation par la Recherche (CIFRE). Ce contrat a associé l'entreprise GÉOBS et l'Université de Nantes agissant en présence de l'École Polytechnique de l'Université de Nantes devenue Polytech'Nantes et de l'Institut de Recherche en Informatique de Nantes (IRIN) devenu Laboratoire Informatique de Nantes Atlantique (LINA).

Entreprise GÉOBS SA 8 avenue des Thébaudières 44800 SAINT-HERBLAIN	Université de Nantes 1 quai de Tourville 44035 NANTES cedex 1	Polytech'Nantes rue Christian Pauc 44306 Nantes Cedex 3	LINA 2 rue de la Houssinière 44322 Nantes cedex 3
http://www.geobs.com/	http://www.univ-nantes.fr/	http://www.polytech-nantes.fr/	http://www.sciences.univ-nantes.fr/lina

Projet de R&D Géobs

Lauréat, le 21 juin 1999, du concours du « Ministère de l'Éducation Nationale, de la Recherche et de la Technologie - Direction de la Technologie - Concours de création d'entreprises innovantes » pour son **projet de R&D innovant de couplage Géomatique[1] &**

[1] «Terme canadien qui désigne la contraction des mots géographie et informatique, la Géomatique est une discipline ayant pour objet la gestion des données à référence spatiale et qui fait appel aux sciences et aux technologies reliées à leur acquisition (satellites, GPS…), leur stockage (Systèmes de Gestion de Bases de Données (SGBD), outils SIG bureautiques, …), leur traitement et leur analyse (SGBD, outils SIG bureautiques, Statistiques, Data Mining Spatial, …)

Data Mining Spatial, la société GÉOBS a été créée en 1999. L'objet de ce projet, dont le Maître d'Ouvrage et le Maître d'OEuvre sont la société GÉOBS, est la « **Constitution d'un outil de Data Mining Spatial dédié à l'observation économique et au marketing du territoire** ».

Le domaine d'exploitation prévu concerne tous les domaines d'activité de GÉOBS liés aux problématiques de l'Aménagement du Territoire dans toutes ses dynamiques applicatives, et, en particulier l'observation économique et le marketing territorial. Ces domaines d'application sont les suivants :

- le Développement Économique,
- le Géomarketing,
- l'Analyse de Risque
- l'Environnement,
- la Santé,
- l'Urbanisme,
- …

Ce projet R&D, démarré le 21 juin 1999 et prévu en clôture courant juin 2007, est organisé en trois phases qui sont les suivantes :

- Phase 1 : Etude et Analyse du Besoin, du 21 juin 1999 jusqu'à fin avril 2001,
- Phase 2 : Recherche, de janvier 2001 à octobre 2005,
- Phase 3 : Prototype (Développement logiciel), de avril 2005 à juin 2007.

Les travaux des Phase 1 et Phase 3 ont été menés exclusivement par GÉOBS avec ses propres ressources et ses propres compétences. Certains travaux relatifs à la Phase 2 de ce projet R&D ont fait l'objet de collaborations de recherche avec trois laboratoires externes à l'entreprise. Chacune de ces collaborations s'est appuyée sur une convention CIFRE. C'est dans ce cadre que cette thèse a été réalisée.

Depuis le début du projet et jusqu'à ce jour, l'équipe du projet R&D GÉOBS a été composée de la manière suivante :

- un Responsable R&D depuis le 21 juin 1999 jusqu'au début 2005 : le PDG de GÉOBS expert en géomatique et tout particulièrement en problématiques de «Société et Aménagement du Territoire». Il a assuré l'expertise géomatique pour tous les travaux de recherche ;
- un Conseiller Scientifique et Technique du projet R&D depuis le 21 juin 99 : il est chercheur expert international en KDD (Knowledge Discovery in Databases) et en Data Mining ;
- un Chargé d'Études en Géomatique de GÉOBS, expert en géomarketing, du 21 juin 1999 jusqu'au début 2005 : il a participé activement aux travaux d'étude et d'analyse du besoin et a assuré l'expertise cartographique et géomarketing pour tous les travaux de recherche ;
- un Ingénieur R&D de GÉOBS, profil Informatique Décisionnelle, de février 2002 à octobre 2003 : il a intégré des méthodes de traitement et d'analyse des données spatiales sous la forme de composants logiciels. Ces méthodes sont issues de différents domaines : l'analyse spatiale, la statistique multidimensionnelle, le Data Mining Spatial ;

---

et leur diffusion / restitution (cartographie, …)». Source Marcel Bergeron, Vocabulaire de la Géomatique Québec (http://geomatique.ifrance.com)

- un Ingénieur R&D de GÉOBS, profil Informatique Décisionnelle, de février 2003 à juin 2003 : il a participé à la conception des IHM des outils d'analyse de données géographiques ;
- un Statisticien Analyste de Données de GÉOBS, de septembre 2003 à juin 2004 : il a effectué l'analyse de cas d'étude concrets pour vérifier les méthodes statistiques. Il a aussi constitué des jeux de données nécessaires à l'avancement des travaux de recherche ;
- un Géographe doctorant CIFRE à GÉOBS («CNRS, Université de Nantes» - GÉOBS), dans la discipline Géographie, de avril 2001 à mars 2004 : son sujet était «Les dynamiques de localisation des activités économiques et les motivations des chefs d'entreprises dans leur choix d'implantation». **Il a soutenu sa thèse le 19 janvier 2005** ;
- un Ingénieur Informatique, d'abord Ingénieur R&D à GÉOBS et ensuite doctorant, CIFRE à GÉOBS («LIPN-CNRS, Université Paris 13, Institut Galilée» - d'abord GÉOBS puis Numsight), dans la discipline Informatique, de mars 2002 à octobre 2004 : il a d'abord effectué son mémoire de master de recherche dont le sujet était «Intégration des Réseaux de Neurones aux SIG». Il a ensuite réalisé sa thèse pendant environ 2 ans à GÉOBS et il a continué à Numsight en région parisienne. Le sujet de la thèse est «Approches Connexionnistes pour l'Exploration et l'Extraction de Connaissance à partir de Données Spatio-Temporelles, Application aux données Géographiques». **La soutenance de la thèse est prévue pour la fin de l'année 2006**.
- un Ingénieur Informatique, depuis février 2002, d'abord Ingénieur R&D GÉOBS et ensuite doctorant CIFRE («Université de Nantes, Ecole Polytechnique de l'Université de Nantes, IRIN devenu LINA » - GÉOBS), dans la discipline Informatique : le sujet de la thèse est «Méthodes d'Extraction de Connaissances à partir de Données (ECD) appliquées aux Systèmes d'Information Géographiques (SIG)». La thèse CIFRE s'est déroulée d'octobre 2002 à octobre 2005. **La soutenance de la thèse est prévue en septembre 2006**.
- Un Directeur R&D – Industrialisation à GÉOBS, depuis début 2005, Ingénieur Informatique Senior expert en édition de logiciels, en ingéniérie des systèmes d'information, en bases de données et en analyse de données. Il est le pilote de la finalisation des travaux de recherche, le concepteur et le maître d'œuvre du projet «Valorisation de la R&D» qui comprend en premier lieu la troisième phase du projet R&D (phase Prototype).

RÉSUMÉ / ABSTRACT

Résumé : Le travail effectué durant cette thèse concerne l'étude des méthodes d'Extraction de Connaissances à partir de Données (ECD) dans le cadre des Systèmes d'Information Géographiques (SIG). Nous avons non seulement mis en œuvre et amélioré des méthodes d'ECD classique (*Classification de Données, Visualisation de Classifications)* mais aussi des méthodes d'ECD spatiales liées à des méthodes d'analyse spatiale (*Lissage Spatial, Détermination de Pôles*, *Sectorisation*). Nous avons effectué notre travail de recherche au sein de la société GÉOBS spécialisée dans l'analyse des données géographiques (spatiales), et nous avons donc expérimenté, appliqué et vérifié ces méthodes sur des jeux de données fournis par GÉOBS et liés à des problématiques de Développement Économique, de Géomarketing, d'Analyse de Risque, d'Environnement, de Santé, etc. Ce mémoire offre une vision globale concernant un ensemble de problématiques et de méthodes d'analyse. Il met ainsi en avant la complémentarité des méthodes utilisées qui sont souvent connectées entre elles soit du point de vue technique soit du point de vue de leur utilisation. Finalement, ce fut un travail très enrichissant car il a touché à de nombreuses problématiques et à d'aussi nombreuses méthodes d'extraction de connaissances.

Mots-clés : Fouille de données, Extraction de Connaissances à partir de Données (ECD), Systèmes d'Information Géographiques (SIG), Classification de Données, Visualisation de Classifications, Arbres de Décision, Lissage Spatial, Sectorisation, Autocorrélation Spatiale, Modélisation des Flux

Abstract: During this PhD thesis, we have studied methods for Knowledge Discovery in Databases (KDD) applied to Geographic Information Systems (GIS). We have improved both classical KDD methods (*Data Clustering, Cluster Visualization*) and spatial KDD methods linked with spatial analysis methods (*Spatial Smoothing, Hot Spot Extraction*, *Spatial Partitionning*). We have worked in GÉOBS, a company expert in spatial data analysis. So our KDD methods have been implemented and tested with data sets provided by GÉOBS in relation with Economic Development, Geomarketing, Risk Analysis, Environnement, Health, etc. This report gives a wide point of view on a range of analysis methods and their related problems. It points up the complementarity between theses methods which can be connected either in a technical way or in a user way. Eventually, this work was very enriching because it has concerned many problems and as many KDD tools.

Key-words: Data Mining, Knowledge Discovery in Databases (KDD), Geographic Information Systems (GIS), Data Clustering, Cluster Visualization, Decision Tree, Spatial Smoothing, Spatial Partitionning, Spatial Autocorrelation, Flow Modeling

REMERCIEMENTS

À mon épouse Sanae et à ma fille Myriam.

À mes parents Claire et Guy et à ma sœur Murielle.

Merci au Professeur Mohamed QUAFAFOU, mon co-directeur de thèse, pour avoir trouvé un sujet de thèse aussi intéressant, pour son implication dans le développement des recherches effectuées et pour ses conseils.

Je remercie Pierre COURONNÉ, PDG de GÉOBS, qui m'a fait confiance pour effectuer mes travaux de recherche au sein de GÉOBS et qui a participé à mon travail de thèse en me conseillant et en m'apportant les compétences géomatiques nécessaires à l'avancement de mes travaux de recherche.

Un grand merci à Florent CHARRON, doctorant CIFRE en thèse de géographie à GÉOBS, d'avril 2001 à mars 2004, pour m'avoir expliqué de façon pédagogique, dans le cadre du proget R&D GÉOBS, les méthodes d'analyse spatiale retenues par GÉOBS pour être étudiées dans le cadre de ma thèse.

Merci à Aziz ZBITOU, Directeur R&D Industrialisation à GÉOBS, pour son implication dans la finalisation de mes travaux de thèse et pour ses conseils.

Je remercie Noureddine MOUADDIB, mon directeur de thèse pour m'avoir conseillé et soutenu tout au long de ma thèse.

Je remercie Marc GELGON, maître de conférences, pour ses remarques sur le rapport de thèse et pour sa participation au jury.

Merci aux professeurs Anne DOUCET et Mohand-Saïd HACID pour avoir lu ma thèse et en avoir fait le rapport. Merci pour leurs commentaires et leurs remarques intéressantes.

Je remercie le professeur Michel SCHOLL d'avoir accepté de présider le jury de la soutenance de thèse.

Merci à tous ceux qui, de prêt ou de loin, ont contribué à l'élaboration de cette thèse et à ma recherche.

SOMMAIRE

INTRODUCTION

L'analyse de données permet d'obtenir des informations synthétiques à partir de données recueillies. Il existe de nombreuses problématiques et autant de méthodes permettant ou essayant d'y répondre. Par ailleurs, depuis l'explosion des capacités de stockage informatique au moins à partir du début des années 1990, la question de l'analyse de grands volumes de données s'est imposée. L'Extraction de Connaissances à partir de Données (ECD) ou en anglais, *Knowledge Discovery in Databases (KDD)* est le domaine de recherche tentant de répondre à cette question en mettant au point des outils de fouille de données (outils de *DataMining* en anglais*)* capables d'analyser les grands volumes de données. Par ailleurs, les Systèmes d'Information Géographiques (SIG) sont des systèmes d'information spécialement conçus pour la manipulation des données géographiques. Ce sont des données particulières car elles disposent d'une composante spatiale et de ce fait, certaines problématiques portant sur la nature spatiale des données sont spécifiques aux SIG. Toutefois, beaucoup de méthodes d'analyse utilisées en Géographie sont issues du domaine de la statistique traditionnelle [*Sand89*, *Char89*, *Chad97*] : histogramme, courbe de concentration, régression linéaire, analyse des corrélations, Analyse en Composantes Principales (ACP), Analyse Factorielle des Correspondances (AFC), Classifications Ascendante Hiérarchique (CAH), … D'un autre côté, certaines méthodes de DataMining ont été adaptées pour le traitement et l'analyse de données spatiales ou étaient déjà directement utilisables (règles d'association, classifications) [*Zeit99*, *AYK00*]. Ainsi notre démarche de recherche dans le cadre de l'ECD et des SIG porte donc à la fois sur des méthodes d'ECD classique et aussi sur des méthodes d'ECD propres au SIG (comme la méthode d'ECD relative au *Lissage Spatial*).

Nous avons effectué notre recherche sur les *Méthodes d'Extraction de Connaissances à partir de Données (ECD) appliquées aux Systèmes d'Information Géographiques (SIG)* au sein de la société GÉOBS. Cette entreprise est spécialisée dans l'analyse de données géographiques (étude, traitement, analyse et restitution de l'information géographique), l'ingénierie des SIG et l'édition de logiciels géodécisionnels. Elle intervient sur des problématiques liées au Développement Économique, au Géomarketing, à l'Analyse de Risque, à l'Environnement, à la Santé, à l'Aménagement, etc. Ainsi, nous avons pu identifier les problématiques rencontrées par les professionnels utilisant les SIG. Nous avons ensuite réutilisé et aussi amélioré les méthodes d'analyse adéquates pour résoudre ces problématiques. Ce faisant, nous avons expérimenté, appliqué et validé ces méthodes sur des jeux de données réelles fournis par GÉOBS.

Nous allons maintenant présenter les quatre problématiques étudiées :

- La *Classification de Données* permet de regrouper (ou de classer) des données dans un petit nombre de groupes (ou de classes) typiques. Le but est de faciliter l'analyse en regroupant ensemble les objets ayant des caractéristiques similaires.
- La *Visualisation de Classifications* permet la présentation des informations d'une classification sous une forme claire et facilement compréhensible. Les visualisations

sont variées et peuvent être des profils de classes (histogramme, boite et moustache,…), un tableau ou encore une hiérarchie.

- le *Lissage Spatial* permet la mise en lumière des tendances se produisant à des grandes échelles spatiales et invisibles aux petites échelles.
- La *Sectorisation* découpe un territoire en plusieurs secteurs. Les méthodes de sectorisation répondent à des besoins d'organisation optimale, de gestion et de management comme la construction de secteurs commerciaux à prospecter (par exemple, des besoins de logistique centrés sur de grandes villes dans le cadre de la *Sectorisation à partir de Centres*).

Cette thèse est découpée en trois grands chapitres : le premier concerne l'état de l'art relativement à ces quatre problématiques, le second expose nos contributions et le dernier traite des expérimentations et applications réalisées.

Dans le **premier chapitre**, nous nous intéresserons à l'**état de l'art** concernant ces quatre sujets. En ce qui concerne la *Classification de Données*, nous nous intéresserons aux algorithmes de classification incrémentaux [*Berk02*] et aux mesures de ressemblance entre les individus dans le cas des variables mixtes [*MC02*]. Nous constaterons que les algorithmes incrémentaux utilisés comportent souvent de nombreux paramètres à choisir judicieusement et qu'ils ont un fonctionnement complexe mettant en jeu de nombreuses opérations. Dans le cas de la *Visualisation de Classifications*, nous traiterons des techniques existantes concernant la visualisation sous forme de profils [*Siir00*], la visualisation sous forme de tableau [*EJBB98*] et l'optimisation de l'ordre des variables et des individus [*BDGHJS02*] dans les visualisations. Cependant, les graphiques de résumés sont souvent trop chargés en « chiffres » et il n'y a pas de visualisation intéressante pour les classifications hiérarchiques. Ensuite, nous aborderons les méthodes de *Lissage Spatial* en nous intéressant principalement à une famille de méthodes robustes et faciles à paramétrer : les méthodes à noyaux de densité (*kernel density*) [*BG95*]. Mais les résultats produits ne sont pas synthétiques et il n'est pas aisé de comparer simplement les différentes cartes de lissage obtenues pour plusieurs échelles. Nous terminerons ce tour d'horizon par les méthodes de *Sectorisation*. Nous verrons que certaines problématiques de *Sectorisation* sont largement traitées en théorie des graphes, ainsi la problématique du partitionnement de graphes correspond à la problématique de la *Sectorisation Équilibrée* et la problématique du rééquilibrage de partitionnements de graphes correspond au *Rééquilibrage de Sectorisations*. Nous constaterons qu'il existe de bons algorithmes [*SKK03*] en théorie des graphes et que nous pourrons les utiliser par la suite pour réaliser des sectorisations. Cependant, ils ne permettent pas de traiter toutes les problématiques de sectorisation.

Le **deuxième chapitre** expose **nos contributions**. La première partie concerne notre méthode de *Classification de Données pour de grands volumes de données mixtes*. Nous présenterons d'abord notre mesure de dissimilarité qui fonctionne avec des données mixtes, puis nous exposerons notre algorithme de classification incrémental : la Classification Ascendante Approximative (CAA) et ses différentes variations (version hiérarchique et version parallèle). Notre algorithme est très simple à paramétrer et à mettre en œuvre. Il est de plus très rapide car sa complexité est linéaire. Nous continuerons ensuite avec les *Visualisations de Classifications* que nous avons mises au point : la *Hiérarchie Évoluée des profils de classes* et le *Tableau Évolué des profils de classes*. Leur lecture est très simple et très rapide, ce qui rend aisé leur analyse. De plus, nous montrerons l'approche originale que nous avons développée pour l'*optimisation de l'ordre des variables et des classes* utilisée dans les visualisations. La troisième partie de ce chapitre concerne le travail effectué pour la *Détermination et la Hiérarchisation de Pôles*. Nous expliquerons notre méthode permettant de résumer les cartes de lissage par des pôles dont nous établirons ensuite la hiérarchie. Notre méthode permet d'expliquer les pôles des échelles spatiales les plus grandes à l'aide de ceux des échelles spatiales les plus courtes. La dernière partie de ce chapitre traite

des problématiques relatives à la *Sectorisation :* la *Sectorisation Équilibrée*, la *Sectorisation à partir de Centres* et le *Rééquilibrage de Sectorisations.* Nous présenterons d'abord notre indice évaluant la qualité d'une sectorisation. Ensuite nous exposerons notre l'algorithme itératif de *Sectorisation Équilibrée* qui s'appuie sur un algorithme de partitionnement de graphe éprouvé. Puis nous décrirons nos algorithmes de *Sectorisation à partir de Centres* : un algorithme progressif basique et un algorithme itératif encapsulant ce premier algorithme. Nous nous intéresserons ensuite au *Rééquilibrage de Sectorisations* en détaillant notre algorithme progressif de mise en œuvre des transferts. Il permet de converger progressivement vers une solution de rééquilibrage.

Le **troisième chapitre** est dédié aux **expérimentations et applications** de nos méthodes. Nous aborderons d'abord la *Classification de Données & de la Visualisation de Classifications*. Nous montrerons que notre algorithme de classification incrémental produit de bons résultats en nous appuyant sur la mesure de la perte d'inertie et en le comparant avec un algorithme de référence, les *k-moyennes* (ou *nuées dynamiques*). Ensuite, nous présenterons la *Classification de Variables* qui est une extension permettant de réaliser une analyse des variables complémentaires de l'analyse des données. Nous montrerons ensuite l'intérêt de la *Classification de Données* et la *Classification de Variables* en procédant à l'analyse des données socioprofessionnelles de Paris et sa Petite Couronne. Puis, nous terminerons cette partie par une autre extension : la *Classification Hiérarchique Spatiale* qui découpe de façon hiérarchique un territoire en secteurs situés autour de zones de fortes valeurs. Dans la deuxième partie de ce chapitre, nous verrons les applications du *Lissage Spatial & de la Détermination des Pôles.* Tout d'abord, nous constaterons que la *Détermination et la Hiérarchisation des Pôles* permet d'analyser la répartition de la population française pour plusieurs échelles simultanément : l'échelle locale (50 km) et l'échelle nationale (130 km). Ensuite, nous montrerons que le lissage spatial des données en prétraitement de la *Classification de Données* permet d'analyser les données pour des échelles spatiales plus grandes et d'obtenir aussi une carte des classes plus facile à analyser. La dernière partie de ce chapitre concernera les applications des méthodes de *Sectorisation* : la *Sectorisation Équilibrée*, la *Sectorisation à partir de Centres* et le *Rééquilibrage de Sectorisations.* Il s'agira tout d'abord du partage du territoire français en 22 secteurs de population égale à l'aide de la *Sectorisation Équilibrée*. Puis nous constaterons que notre méthode de *Sectorisation à partir de Centres* donne de bons résultats en terme de contiguïté et d'objectif de taille, mais nécessite une amélioration de la compacité des secteurs. Nous poursuivrons alors avec notre algorithme de *Rééquilibrage de Sectorisations* et nous montrerons qu'il permet d'atteindre les objectifs de taille sans trop altérer la compacité.

Finalement, nous conclurons sur l'impact positif et enrichissant de notre recherche sur le thème de l'Extraction de Connaissances à partir de Données (ECD) dans le cadre des Systèmes d'Information Géographiques (SIG). Cela nous a donné une vision globale des problématiques et des méthodes d'analyse. En conséquence, nous avons pu mettre au point des méthodes innovantes s'appuyant sur d'autres méthodes (par exemple, la *Détermination des Pôles* se base sur les résultats du *Lissage Spatial*), et en retour, nous avons pu décliner une même méthode pour plusieurs utilisations (par exemple, la *Classification Hiérarchique Spatiale* est une variation de la *Classification de Données*). Cela n'aurait pas été possible en se focalisant uniquement sur une seule problématique particulière.

En annexe, nous verrons des méthodes non centrales dans notre travail de recherche, elles concernent l'*Arbre de Décision*, l'*Autocorrélation Spatiale* et la *Modélisation des Flux*. Dans la première partie de cette annexe, nous nous intéresserons à la méthode de l'*Arbre de Décision* [*Murt98*] en considérant deux problèmes : d'une part, la recherche du meilleur partitionnement binaire concernant les modalités d'une variable qualitative [*MAR96*] et d'autre part, le choix préférentiel de règles fiables (c'est-à-dire toujours vraies). Nous montrerons tout d'abord que notre algorithme de la CAHA (déjà exposé dans la *Classification de Données*) permet de trouver

très rapidement des bonnes solutions au premier problème. Puis, concernant le deuxième problème, nous exposerons notre coefficient favorisant les règles fiables et illustrerons son intérêt par un exemple. La partie suivante de cette annexe traite de l'*Autocorrélation Spatiale* [*PS97*] qui mesure le degré de ressemblance entre les objets proches. Après avoir montré que les différents coefficients existants ont un « biais » car ils pénalisent les objets situés en bordure, nous proposerons nos propres coefficients d'autocorrélation et nous démontrerons qu'il corrige effectivement ce biais. Enfin, dans la dernière partie de cette annexe, nous aborderons la *Modélisation des Flux* de G. Dorigo et W. Tobler [*DT83*] dont les résultats souffrent d'un défaut : les flux résultats peuvent être négatifs, ce qui est impossible dans la réalité. Nous montrerons que notre amélioration effectue des corrections minimes et ne perturbent pas la modélisation des flux.

Chapitre 1

1. ÉTAT DE L'ART

Introduction

Ce premier chapitre concerne l'état de l'art des différents sujets abordés : la *Classification de Données*, la *Visualisation de Classifications*, le *Lissage Spatial* et la *Sectorisation*.

Dans la partie consacrée à la *Classification de Données*, nous verrons les méthodes existantes utilisées pour la classification de données. Nous examinerons d'abord les différentes familles d'algorithmes de classification en axant notre démarche sur la façon de gérer les grands volumes de données (il s'agit de méthodes de complexité linéaire en temps de calcul et en espace-mémoire). Puis nous étudierons ensuite un autre aspect commun à la plupart de ces méthodes : la mesure de la ressemblance entre les individus (objets) en nous attachant au cas des données ayant des variables mixtes.

La deuxième partie est consacrée à la *Visualisation de Classifications*. Elle traite trois aspects : la *visualisation sous forme de résumés*, la *visualisation sous forme de tableau* et l'*optimisation de l'ordre des variables et des individus*. Nous verrons tout d'abord les techniques de visualisation de résumés, issues pour la plupart des graphiques utilisés en statistiques. La visualisation sous forme de tableau permet, quant à elle, de facilement visualiser les informations des données mixtes. Nous soulignerons aussi l'intérêt d'un ordre optimal pour les variables et des individus dans une visualisation et nous présenterons les différentes techniques d'optimisation utilisables.

Dans la troisième partie, nous verrons le *Lissage Spatial*, largement utilisé pour mettre en évidence les tendances apparaissant aux grandes échelles spatiales. Nous verrons les différentes techniques de lissage et nous nous focaliserons sur la méthode des noyaux de densité (*kernel density*) qui est puissante et facile à paramétrer. Celle-ci se base sur des fonctions d'interaction spatiale (ou de voisinage) qui permettent de définir une « gradation » de la notion de voisinage. À partir de ces fonctions, nous verrons les fonctions de lissage spatial servant à calculer la carte de lissage spatial.

Nous clôturerons ce chapitre par l'étude des techniques de *Sectorisation* existantes : la *Sectorisation Équilibrée* et le *Rééquilibrage de Sectorisations*. La *Sectorisation Équilibrée* consiste en la création de secteurs de « taille » égale (La taille est dans ce cas une quantité à partager : surface,

population, clients, …). La localisation des secteurs n'est pas définie à l'avance. Nous intéressons au partitionnement de graphe (*Graph Partitioning*) qui est l'équivalent de ce problème en théorie des graphes. Comme c'est un problème très étudié, nous nous focaliserons donc sur les méthodes les plus intéressantes, c'est-à-dire les méthodes récursives et multi-échelles. Nous verrons ensuite le *Rééquilibrage de Sectorisations* qui permet d'optimiser une sectorisation existante. En effet, il est parfois plus intéressant d'améliorer une sectorisation existante que d'en calculer une nouvelle. Le rééquilibrage des secteurs vise donc à « améliorer » les quantités de chaque secteur en transférant des objets géographiques entre les secteurs. La technique de rééquilibrage étudiée se décompose en deux étapes : le calcul de tous les transferts à effectuer entre les secteurs et la phase opérationnelle pendant laquelle les transferts sont effectués.

1 Classification de Données

Nous allons d'abord examiner les différentes techniques utilisées par les algorithmes de classification pour gérer les grands volumes de données. Puis nous étudierons les différentes mesures de ressemblance (similarités, dissimilarités et distances) entre les individus (objets) en nous attachant au cas des individus ayant des variables mixtes.

1.1 Méthodes de classification sur les grands volumes de données

Dans un premier temps, nous allons donner les différentes familles d'algorithmes de classification telles qu'elles sont traditionnellement représentées [*JD88, JMF99, Berk02*]: c'est-à-dire en familles d'algorithmes regroupés selon la façon dont ils classent les individus (objets).

Les deux principales familles de méthodes sont les suivantes :

- Les méthodes par partitionnement génèrent des classes autour de noyaux choisis initialement au hasard puis elles améliorent itérativement ces partitions initiales en se basant sur une fonction de coût à minimiser. Les principaux algorithmes sont les k-moyennes (Nuées dynamiques) et k-medoïds (Médoïdes), cette dernière famille comprend CLARA (*Clustering LARge Applications*), PAM (*Partitionning Around Medoids*) [*KR90*], CLARANS (*Clustering LARge Applications based upon RANdomized Search*) [*NH94*]. Les algorithmes de mélange tels que EM [*DLR97*] sont une variante intéressante car les noyaux utilisés sont des lois de probabilité (gaussienne généralement).
- Les méthodes hiérarchiques :
    - Les méthodes par agglomération créent une hiérarchie ascendante en procédant par regroupements successifs. Il s'agit de AGNES (*AGglomerative NESting*) [*KR90*], CURE (*Clustering Using REpresentatives*) [*GRS98*], CHAMELEON [*KHK99*].

- Les méthodes par division créent une hiérarchie descendante en procédant par divisions successives comme DIANA (*DIvisive ANAlysis*) [*KR90*].
- Les méthodes incrémentales descendantes qui construisent globalement une hiérarchie descendante en ajoutant les individus un par un. On trouve COBWEB [*Fish87*] et BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*) [*ZRL96*].

À ces deux familles, ils convient d'ajouter d'autres familles de méthodes généralement plus récentes et qui parfois se recoupent avec les anciennes familles:

- Les méthodes par densité qui groupent les individus situés dans des zones de forte densité. Elles comprennent DBSCAN (*Density Based Spatial Clustering of Applications with Noise*) [*EKSX96*], OPTICS (*Ordering Points To Identify Clustering Structure*) [*ABKS99*], DENCLUE (*DENsity-based CLUstEring*) [*HK98*].
- Les méthodes par grilles qui appliquent une grille multi-niveaux (ou un maillage multi-niveaux) dans l'espace des données et groupent ainsi les données situées dans une même cellule. Il s'agit de STING (*STatistical INformation Grid*) [*WYM97*], WAVECLUSTER [*SCZ98*], CLIQUE (*CLustering In QUEst*) [*AGGR98*].
- Les méthodes par modèle qui définissent des modèles de classes et y affectent les données correspondant au modèle.

Nous allons maintenant étudier les différentes stratégies mises en œuvre par ces algorithmes pour fonctionner sur des grands volumes de données. Nous ferons aussi référence aux algorithmes déjà cités précédemment :

- ***L'échantillonnage*** est la méthode la plus utilisée pour faire fonctionner n'importe quel algorithme de classification sur les grands volumes de données. Le principe est simple, à partir de l'échantillon, on utilise la méthode de classification de son choix, cela fait partie de la phase de prétraitement. Une fois les classes déterminées sur l'échantillon, le reste des données est lue en une seule fois, chaque individu lu étant affecté à la classe la plus proche, c'est la phase de traitement de la base de données. Cette méthode à l'avantage de permettre la réutilisation de toutes les méthodes de classification. Mais elle a l'inconvénient de dépendre entièrement de l'échantillon qui peut très bien ne pas être « représentatif » dans le cas où des individus atypiques seraient absents de l'échantillon. La conséquence est qu'il n'y a pas de classes adaptées pour ces individus et qu'ils sont donc obligatoirement « mal classés », c'est-à-dire associés à des classes les représentants assez mal. Cette méthode est celle utilisée par les algorithmes CLARA [*KR90*], CLARANS [*NH94*], CURE [*GRS98*] afin qu'ils soient fonctionnels sur les grandes bases de données.
- ***La création descendante d'un arbre hiérarchique de classes*** construit progressivement un arbre de manière descendante. Chaque nouvel individu descend de la classe racine jusqu'à la classe la plus basse dans laquelle il sera incorporé. Lors de la descente, divers opérateurs de création, de division et de fusion de classes peuvent être appliqués. C'est la méthode utilisée par COBWEB [*Fish87*].
- ***La création de résumés*** permet de transformer un grand volume de données en un volume beaucoup plus petit sur lequel seront appliqués les algorithmes classiques ou leur variations proches. On trouve deux méthodes souvent utilisées :
  - ***La création descendante d'un arbre hiérarchique de résumés (Clustering Feature Tree)*** se fait suivant des paramètres d'effectif maximal et de diamètre maximal par résumé. La construction des résumés se fait de manière descendante car un nouvel individu descend de la racine jusqu'au résumé le

plus bas dans lequel il sera incorporé. C'est la méthode utilisée par BIRCH [*ZRL96*] qui utilise ensuite l'algorithme CLARANS pour réaliser la classification finale à partir des résumés. On peut aussi citer SaintEtiQ [*RM02*] qui fonctionne sur le même pricipe et permet de résumer des données floues.

- ***La création de bulles de données (Data Bubbles)*** [*BKKS01*] est une méthode alternative car contrairement à la création descendante d'un arbre hiérarchique de résumés, les bulles de données ne sont pas structurées. Les bulles de données sont des résumés de données qui sont soit définis par les résumés issus de BIRCH soit initialisés par un échantillon et « remplis » par le reste des données en les agrégeant aux bulles les plus proches, durant cette étape a lieu la lecture des données. À partir de là, sont calculées les caractéristiques nécessaires à l'utilisation de ces bulles de données par l'algorithme OPTICS [*ABKS99*].

- ***L'utilisation d'un graphe de voisinage*** permet à des outils tel que CHAMELEON [*KHK99*] d'être très rapides. Cependant, le principal obstacle vient de la construction du graphe de voisinage qui est très coûteuse en temps et en mémoire dès lors que le nombre de dimensions (ou variables) est élevé (plus de 2 ou 3 dimensions). Il est possible d'utiliser des méthodes heuristiques telles la création de cellules afin de créer un graphe de voisinage approché.
- ***L'utilisation d'un index spatial (R-tree)*** permet l'indexation d'objets spatiaux et la recherche rapide d'objets compris dans un certain rayon. Le nombre de dimensions des données est ainsi limité à deux. C'est cet index spatial qu'utilisent DBSCAN [*EKSX96*], OPTICS.
- ***La création d'intervalles*** se fonde sur le découpage de chacune des dimensions (ou variables) en intervalles. Chaque individu est ensuite affecté aux intervalles auxquels il appartient, c'est la phase de traitement de la base de données. Les algorithmes de classification travaillent alors ces intervalles. Cette méthode est utilisée par CLIQUE [*AGGR98*] et WAVECLUSTER [*SCZ98*].
- ***La création de cellules*** part d'un découpage de chacune des dimensions en intervalles. Ce découpage créé de fait un « quadrillage » de l'espace par des cellules. Chaque individu est ensuite affecté dans la cellule à laquelle il appartient, c'est la phase de traitement de la base de données. Les cellules voisines sont connectées et les cellules vides sont supprimées. Les algorithmes de classification travaillent alors sur ces cellules. Cette méthode est utilisée par DENCLUE [*HK98*]. Il faut cependant noter que même en prenant un nombre constant d'intervalles par dimension, le nombre de cellules croît exponentiellement avec le nombre de dimensions. C'est une limite importante à son utilisation sur des données comportant un grand nombre de dimensions.

Transversalement à l'analyse précédente, il existe une catégorie de méthodes particulièrement bien adaptées aux grands volumes de données incrémentaux tels les entrepôts de données (Data Warehouses) : il s'agit des ***méthodes incrémentales***. Techniquement, un algorithme de classification incrémental reçoit un flux de données qu'il doit « classifier » et il ne doit (peut) conserver en mémoire qu'une petite partie des données reçues. On peut résumer ce principe par le respect des contraintes suivantes :

- la lecture des données se fait en une seule passe
- les individus sont pris en compte un par un
- le processus peut à tout instant être stoppé et redémarré

- un résultat temporaire est disponible à tout moment.

On constate ainsi que ces méthodes sont bien adaptées aux entrepôts de données car la mise à jour de la classification s'effectue par l'ajout des nouvelles données, sans avoir à recommencer le traitement depuis le début. Parmi les algorithmes incrémentaux, on peut citer BIRCH et COBWEB, toutefois ces algorithmes n'ont pas un fonctionnement simple car ils mettent en œuvre beaucoup d'opérations différentes (comme l'inclusion d'un objet dans une classe, l'éclatement de classes, la fusion de classes, …) et nécessitent plusieurs paramètres (le paramètre de taille maximale d'une classe, le paramètre de taille minimale d'une classe, le nombre maximum de classes, …). Nous verrons donc ultérieurement notre choix pour la mise au point d'un algorithme simple, robuste et efficace.

1.2 Similarités, dissimilarités et distances

Toutes les méthodes de classification décrites précédemment utilisent les notions de similarité ou de dissimilarité pour qualifier par une grandeur numérique la ressemblance entre deux individus. Elles peuvent aussi utiliser la distance qui est une notion plus contraignante de la dissimilarité.

Nous allons voir en premier les propriétés des notions de similarité, de dissimilarité et de distance. Puis nous verrons les dissimilarités pour les variables quantitatives puis pour les variables binaires et qualitatives. Nous finirons par le cas qui nous intéresse particulièrement : les dissimilarités pour les distances mixtes fonctionnant avec des variables quantitatives et qualitatives.

1.2.1 Propriété des notions de similarité, dissimilarité et distance

Toutes ces notions sont des mesures de dissemblances (ou de ressemblances) entre deux individus. Les propriétés que vérifient la similarité (*sim*) ou la dissimilarité (*dis*) sont :

- la positivité :

$$\forall a,b \qquad dis(a,b) \geq 0$$
$$sim(a,b) \geq 0$$

- la symétrie :

$$\forall a,b \qquad dis(a,b) = dis(b,a)$$
$$sim(a,b) = sim(b,a)$$

La similarité vérifie que la similarité maximum K est atteinte seulement pour la similarité entre un individu et lui-même :

$$\forall a,b,a \neq b \qquad K = sim(a,a) = sim(b,b)$$
$$K > sim(a,b)$$

La dissimilarité vérifie que la dissimilarité minimum est atteinte seulement pour la dissimilarité entre un individu et lui-même et qu'elle vaut 0:

$$\forall a,b,a \neq b \qquad 0 = dis(a,a) = dis(b,b)$$
$$0 < dis(a,b)$$

La distance (*d*) est un indice de dissimilarité vérifiant également l'inégalité triangulaire :

$$\forall a,b,c \qquad d(a,c) \leq d(a,b) + d(b,c)$$

Il existe de nombreux indices de similarité, de dissimilarité et de nombreuses distances.

Nous allons d'abord examiner les distances pour les variables (attributs) quantitatives (numériques).

1.2.2 Distances pour les variables quantitatives

La distance la plus utilisée pour les données quantitatives est la distance euclidienne :

$$\forall o_i, o_j \; d(o_i,o_j) = \left(\sum_{k=1}^{m} \left| V_k(o_i) - V_k(o_j) \right|^2 \right)^{\frac{1}{2}} = \sqrt{\sum_{k=1}^{m} \left(V_k(o_i) - V_k(o_j) \right)^2}$$

Avec *m* variables quantitatives V_k

et $V_k(o_i)$, la valeur de la variable V_k pour l'objet o_i

Cependant, elle est très souvent appliquée sur les variables centrées-réduites plutôt que directement sur les variables afin que la distance soit significative.

1.2.3 Dissimilarités pour les variables binaires

La première méthode consiste à considérer les variables binaires comme des variables quantitatives et donc d'utiliser les distances décrites précédemment. Une autre méthode est l'utilisation d'indices concernant spécifiquement les variables binaires. Elle nécessite le calcul des correspondances pour chaque variable.

Le coefficient d'appariement suivant donne le pourcentage de valeurs différentes :

$$\forall o_i, o_j \; ca\_b(o_i,o_j) = \frac{t_{10}(o_i,o_j) + t_{01}(o_i,o_j)}{t_{00}(o_i,o_j) + t_{11}(o_i,o_j) + t_{10}(o_i,o_j) + t_{01}(o_i,o_j)}$$
$$= \frac{t_{10} + t_{01}}{t_{00} + t_{11} + t_{10} + t_{01}}(o_i,o_j)$$

Avec $t_{xy}(o_i,o_j)$ la fonction qui compte le nombre de variables binaires pour lesquelles on a $V_k(o_i) = x$ et $V_k(o_j) = y$, *x* et *y* étant égale à 1 ou à 0. Cette fonction peut se définir de la manière suivante :

$$t_{xy}(o_i,o_j) = \sum_{k=1}^{m} match(V_k(o_i) = x \text{ et } V_k(o_j) = y) \text{ avec x, y égale à 1 ou 0}$$

Avec $match(p)$ = 1 si la proposition *p* est vrai
= 0 sinon

On remarque toutefois qu'il existe une écriture plus simple du coefficient d'appariement :

$$\forall o_i, o_j \; ca\_b(o_i, o_j) = \frac{1}{m}\sum_{k=1}^{m} match(V_k(o_i) \neq V_k(o_j))$$

Très utilisé aussi, le coefficient de dissimilarité de Jaccard [*vRij79*] est presque identique au coefficient d'appariement mais ne prend pas en compte les valeurs nulles pour les deux objets :

$$\forall o_i, o_j \; cj(o_i, o_j) = \frac{t_{10} + t_{01}}{t_{11} + t_{10} + t_{01}}(o_i, o_j)$$

L'exemple suivant montre les différences entre le coefficient d'appariement binaire (*ca_b*) et le coeeficient de Jaccard (*cj*). Soit $o_1 = (1,1,0,1,0)$ et $o_2 = (0,1,0,0,0)$, on a :

	V_1	V_2	V_3	V_4	V_5
o_1	1	1	0	1	0
o_2	0	1	0	0	0
Pris en compte par	t_{10}	t_{11}	t_{00}	t_{10}	t_{00}

Fig. 1 – Exemple de variables binaires

D'où : $t_{00} = 2$ (cela concerne V3 et V5), $t_{10} = 2$ (V1 et V4), $t_{01} = 0$ (aucune variable), $t_{11} = 1$ (V2)

Nous avons donc : $ca\_b(o_1, o_2) == \frac{2+0}{1+2+2+0} = \frac{2}{5}$ et $cj(o_1, o_2) == \frac{2+0}{2+2+0} = \frac{2}{4}$

Nous allons maintenant voir les distances pour les variables qualitatives ayant plusieurs modalités.

1.2.4 Dissimilarités pour les variables qualitatives à plusieurs modalités

Une première méthode consiste à créer un codage disjonctif créant autant de variables binaires qu'il y a de modalités. Ensuite, il suffit d'utiliser les distances pour les variables binaires ou quantitatives.

Par exemple, supposons que nous avons les individus $o_1 = (Rouge, Oui)$ et $o_2 = (Vert, Oui)$, avec la première variable qui est la *Couleur* pouvant prendre les valeurs « Rouge », « Verte » ou « Bleu » et que la seconde variable est la *Présence* pouvant prendre les valeurs « Oui » ou « Non ».

	Couleur	Présence
o_1	Rouge	Oui
o_2	Vert	Oui

Fig. 2 – Exemple de variables qualitatives

Le codage disjonctif crée donc cinq nouvelles variables binaires qui sont « Couleur=Rouge », « Couleur=Vert », « Couleur=Bleu », « Présent=Oui » et « Présent=Non ».

Cela donne $disj(o_1) = (\{1,0,0\},\{1,0\})$ et $disj(o_2) = (\{0,1,0\},\{1,0\})$. On peut ainsi calculer le coefficient d'appariement binaire :

	Couleur = Rouge	Couleur = Vert	Couleur = Bleu	Présence = Oui	Présence = Non
o_1	1	0	0	1	0
o_2	0	1	0	1	0
Pris en compte par	t_{10}	t_{01}	t_{00}	t_{11}	t_{00}

Fig. 3 – Exemple dedisjonction de variables qualitatives

D'où : $t_{00} = 2$, $t_{11} = 1, t_{10} = 1$, $t_{01} = 1$

Nous avons ainsi : $ca\_b(disj(o_1), disj(o_2)) = \frac{1+1}{5} = \frac{2}{5}$

Une autre méthode consiste à utiliser des distances spécifiques aux variables à modalités. La plus utilisée est le coefficient d'appariement qui donne le pourcentage de valeurs différentes, elle s'exprime de la même façon que pour les variables binaires :

$$\forall o_i, o_j \; ca\_m(o_i, o_j) = \frac{1}{m}\sum_{k=1}^{m} match(V_k(o_i) \neq V_k(o_j))$$

À partir de l'exemple, nous obtenons :

$$ca\_m(o_1, o_2) =$$

$$\frac{1}{2}(match(Couleur(o_1) \neq Couleur(o_2)) + match(Présence(o_1) \neq Présence(o_2)))$$

$$ca\_m(o_1, o_2) = \frac{1}{2}(1+0) = 0{,}5$$

En effet, les deux objets ont la même valeur « présence » mais pas la même « couleur ».

Nous remarquons que le coefficient d'appariement directement calculé ne donne pas les mêmes résultats que le coefficient d'appariement binaire calculé sur les variables disjonctives. En effet, le total des variables est $m_{dis} = \sum_{k=1}^{m} m_k$ au lieu de *m* et de plus, pour chaque groupe de variables binaires issues d'une même variable à modalités, les différences sont comptées en double. Ce « double comptage » vient du fait que lorsqu'il y a une différence pour la variable de départ, on a pour le groupe de variables binaires issues de cette disjonction une fois la différence entre 0 et 1 et une autre fois la différence entre 1 et 0, le reste étant des 0 qui ne changent pas. Ainsi, dans l'exemple $Rouge \neq Vert$ devient $\{1,0,0\} \neq \{0,1,0\}$ dans le codage

disjonctif et le nombre de variables est $m_1 = 3$ au lieu de 1 et le nombre de différences constatées est 2 au lieu de 1.

Ainsi, pour avoir une équivalence entre $ca\_b(disj(o_1), disj(o_2))$ et $ca\_m(o_1, o_2)$, il convient de pondérer le résultat par $p = \frac{\sum_{k=1}^{m} m_k}{2 \times m}$. Nous avons ainsi :

$$ca\_m(o_1, o_2) = p \times ca\_b(disj(o_1), disj(o_2))$$

Dans l'exemple, $p = \frac{1}{2 \times 2}(2+3) = \frac{5}{4}$

Après avoir vu les différences entre les mesures de dissimilarités pour les variables qualitatives, nous allons maintenant étudier les dissimilarités pour les variables mixtes.

1.2.5 Dissimilarités pour les variables mixtes

Dans ce cas, les variables sont de natures différentes et les données sont représentées par des variables quantitatives et des variables qualitatives. Le but est ici de trouver une distance s'appliquant sur ces données. Il existe de nombreuses techniques permettant de définir une dissimilarité pour des données mixtes [*MC02*].

La première méthode consiste à transformer les variables quantitatives en variables qualitatives en les discrétisant. Le principal inconvénient est la perte d'information liée à la discrétisation.

Une autre méthode, inverse de la précédente si l'on peut dire, est la transformation des variables qualitatives en variables quantitatives en créant une disjonction. Le principal inconvénient est l'augmentation du nombre de variables qui peut être très important pour les variables possédant beaucoup de modalités.

Enfin, il est aussi possible d'utiliser des distances s'appliquant aux variables mixtes.

L'une des plus utilisées est :

$$\forall o_i, o_j \; dm(o_i, o_j) = \frac{1}{m} \sum_{k=1}^{m} contribution(V_k(o_i), V_k(o_j))$$

Avec $contribution(V_k(o_i), V_k(o_j))$

= 0 si V_k est une variable qualitative et $V_k(o_i) = V_k(o_j)$

= 0 Si V_k est une variable quantitative normalisée et $|V_k(o_i) - V_k(o_j)| \leq s$ avec s un seuil fixé

= 1 sinon

Cette mesure ressemble beaucoup au coefficient d'appariement et l'étend aux variables quantitatives normalisées en utilisant un seuil pour décider si les données sont semblables ou dissemblables, cette mesure est binaire car le résultat est soit 0 soit 1.

Une variante de cette méthode est la suivante :

$contribution2(V_k(o_i), V_k(o_j))$ = 0 si V_k est une variable qualitative et $V_k(o_i) = V_k(o_j)$

= 1 si V_k est une variable qualitative et si $V_k(o_i) \neq V_k(o_j)$

$= |V_k(o_i) - V_k(o_j)|$ si V_k est une variable quantitative normalisée

Dans ce cas, la contribution d'une variable quantitative normalisée est directement donnée par la différence entre les valeurs des deux objets. On peut cependant faire la remarque que les valeurs de cette mesure pour les variables qualitatives sont soit 0 soit 1, tandis que pour les variables quantitatives, la valeur peut prendre n'importe quel valeur positive, y compris des valeurs au dela de 1.

Nous verrons ultérieurement la distance que nous utiliserons pour la classification. Notre choix sera guider par un souci d'homogénéisation afin de traiter le plus identiquement possible les variables quantitatives et qualitatives.

2 Visualisation de Classifications

Nous allons tout d'abord voir les techniques existantes de visualisation de résumés qui permettent de visualiser les principales caractéristiques d'une classe. Puis nous verrons, les différents types de visualisation de classifications sous forme de tableau. Nous finirons par les algorithmes d'optimisation de l'ordre des variables et des classes qui permettent une lecture aisée des visualisations (aussi bien pour les résumés ou que pour les tableaux).

2.1 Visualisation de classes sous forme de résumés

La visualisation des résumés (ou des classes) s'apparente à la visualisation d'informations synthétiques concernant un groupe d'individus. Il existe de nombreux outils graphiques en statistique permettant de faire cela. Les plus simples permettent de visualiser les informations pour une seule variable d'un ensemble d'individus. Nous verrons ensuite les outils plus évolués permettant de visualiser simultanément les informations pour plusieurs variables.

2.1.1 Visualisation des caractéristiques d'une seule variable.

Ces graphiques sont très communs en statistiques. Nous allons étudier les points forts et les points faibles des graphiques les plus intéressants :

- L'histogramme des fréquences
- Les graphique de type « Feuille et branche »
- La Boite à moustaches

2.1.1.1 L'histogramme des fréquences

C'est l'un des graphiques les plus utilisés pour décrire une « variable » pour un ensemble d'individus. C'est un graphique constitué par des rectangles de même base placés les uns à côté des autres, et dont la hauteur est proportionnelle à la fréquence d'apparition de différentes classes de valeurs. Dans l'exemple suivant, on observe la fréquence (nombre de zones IRIS de Paris et sa petite couronne) des classes de taux d'ouvriers pour différents intervalles de valeurs.

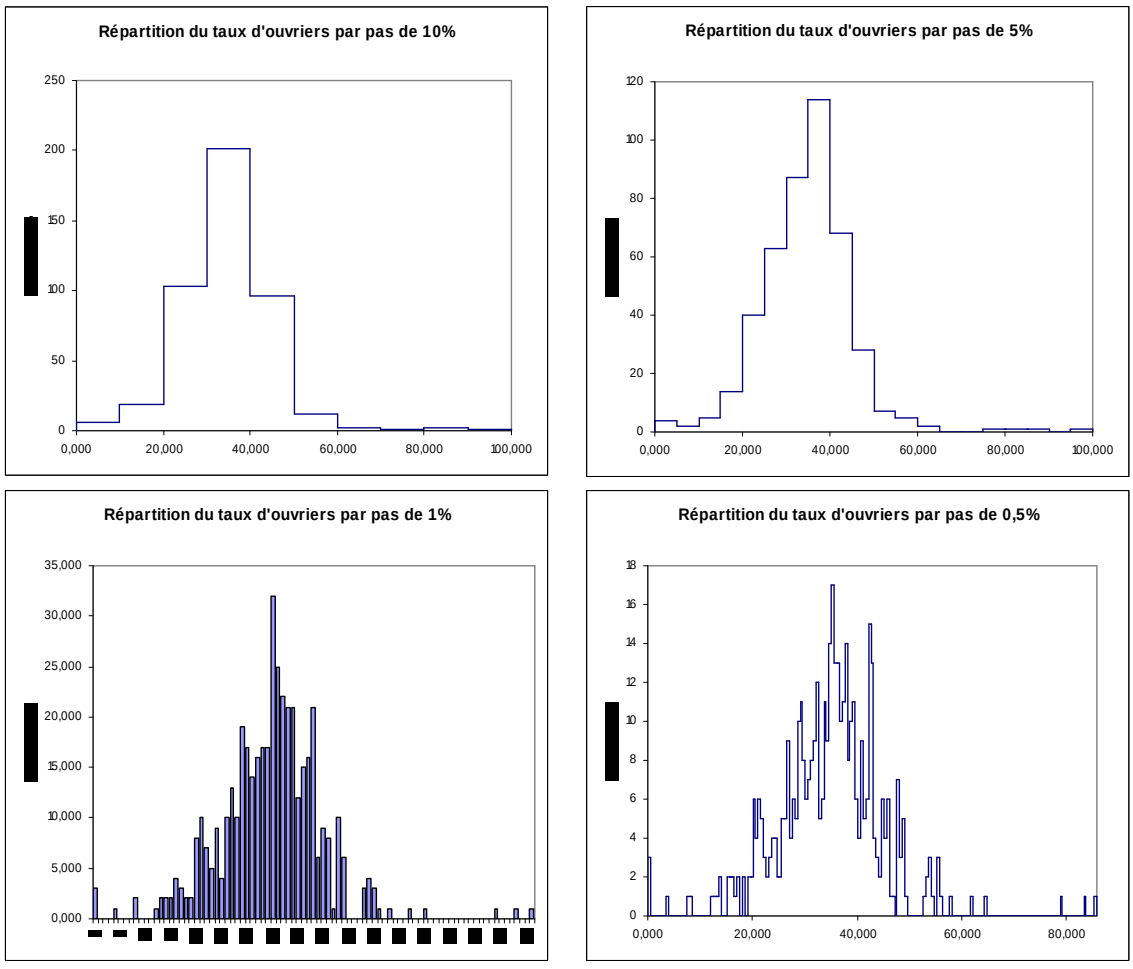


Fig. 4 –Les différents histogrammes obtenus selon la taille de l'intervalle

Le choix de l'intervalle produit des formes assez différentes. Ainsi un problème courant de la construction de l'histogramme est le choix de la taille de l'intervalle des classes car la forme de l'histogramme dépend grandement de ce choix arbitraire. Il existe des règles empiriques pour déterminer un nombre d'intervalles adéquat à partir du nombre d'individus n [*Scot92*]:

Règle de Sturge : nb_intervalles = 1 + (3,3 log n)

Règle de Yule : nb_ intervalles = $2{,}5\times\sqrt[4]{n}$

Dans le cas présent, *n*=442, cela donne 9,72 pour la règle de Sturge et 11,46 pour la règle de Yule. Cela correspond à environ 10 intervalles ayant une grandeur de 10%.

2.1.1.2 Feuille et branche (Stem and leaf plots)

Le "stem and leaf" (feuille et branche) est un histogramme des fréquences sous forme textuelle qui crée une ligne par valeur entière possible de la variable (ou une puissance de dix de la variable) et qui remplit les lignes avec les premières décimales. L'exemple suivant reprend les taux d'ouvrier par zone et les classes sont de 10 en 10.

0	000378
1	2233355566667788999
2	00000000001111111111122222…99999999999999999999
3	000000000000001111111111…444455555555555555555555…99999999999999999
4	00000000000001111111111122…7788888888889
5	233333455558
6	14
7	8
8	35

Fig. 5 –Exemple d'histogramme

La première ligne est 0|000378 ce qui signifie qu'elle contient les valeurs 00, 00, 00, 03, 07 et 08. La dernière ligne est 8|35 ce qui signifie qu'elle contient les valeurs 83 et 85. L'avantage de ce graphique est la possibilité de voir comment se répartissent exactement les valeurs à l'intérieur d'un même intervalle. Les inconvénients sont la nécessité d'utiliser des intervalles de puissance de 10 et l'impossibilité de tracer le graphique en entier quand il a trop de valeurs.

2.1.1.3 Boite à moustaches (*box-and-whisker plot*)

La « boite à moustaches » a été définie par J. Tukey [*Tuke77*] et elle permet de donner sous forme d'un graphique horizontal ou vertical les informations les plus importantes de la distribution :

- La valeur médiane (50% des valeurs sont plus faibles et 50 % sont plus fortes)
- La valeur minimale et la valeur maximale.
- Le premier quartile (25% des valeurs sont plus faibles et 75 % sont plus fortes)
- Le troisième quartile (75% des valeurs sont plus faibles et 25 % sont plus fortes)

En outre, les moustaches séparent les valeurs ordinaires des valeurs exceptionnelles. La moustache « gauche » est située à plus de 1,5 fois l'écart interquartile au dessous du 1er quartile et la moustache « droite » au dessus du 3ème quartile. En effet, si la distribution de la variable suit une distribution normale (gaussienne), plus de 99% des individus sont situés entre les moustaches.

Dans l'exemple suivant, on observe la distribution du taux d'ouvriers dans les IRIS (zones) de Paris et sa petite couronne.

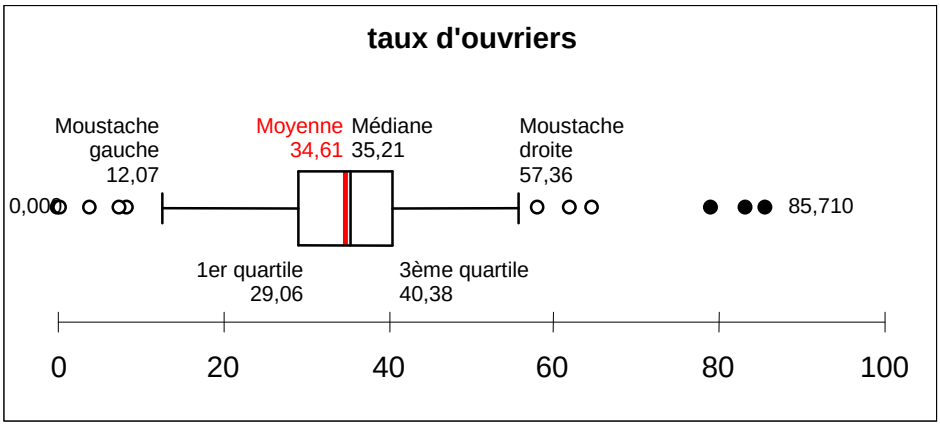


Fig. 6 –Exemple de boite à moustache décrivant la variable « taux d'ouvrier »

L'avantage de ce graphique est que le nombre d'informations affichées est indépendant du nombre d'individus et que les informations sont très synthétiques.

2.1.2 Visualisation des caractéristiques de plusieurs variables

Les graphiques permettant de visualiser les caractéristiques de plusieurs variables peuvent être directement utilisés pour visualiser les caractéristiques des résumés. Les graphiques les plus intéressants sont :

- La visualisation en coordonnées parallèles
- Le diagramme en étoile
- La visualisation de « Boites à moustaches » en parallèles

Mais nous allons tout d'abord parler de la standardisation des variables qui est nécessaire dans la plupart des cas pour obtenir des graphiques lisibles.

2.1.2.1 La standardisation

Le fait de visualiser plusieurs variables simultanément sur un même graphique oblige dans la plupart des cas à adopter une échelle commune pour toutes les variables. Cette étape s'appelle la « standardisation ». En effet, sans standardisation, l'échelle commune est par nécessité la plus grande et de plus, ce choix n'a aucune signification lorsque les variables ne sont pas de même nature. L'exemple ci-dessous montre la représentation commune du taux de cadre (un pourcentage) et le revenu mensuel par ménage (en francs). Cela aboutit à la mauvaise représentation du taux de cadre.

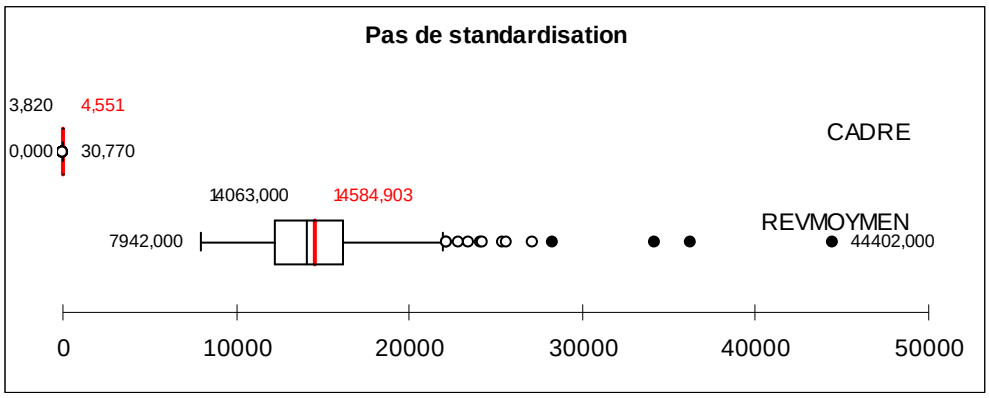


Fig. 7 –Visualisation sur une même échelle des variables « Cadre » (en %) et « Revenu » (en francs)

La standardisation peut se faire de plusieurs manières différentes :

- **Centrer et réduire** : pour chaque valeur de la variable, on retire la moyenne et on divise par l'écart-type. Les variables sont dites « centrées-réduites ». L'échelle dépend alors des valeurs minimum et maximum des variables centrées-réduites. On peut aussi fixer arbitrairement l'échelle en considérant des hypothèses statistiques : pour un répartition des individus suivant une loi normale, 95 % des valeurs sont comprises entre -2 et +2. On risque cependant de ne pas voir les valeurs exceptionnelles qui seraient hors limites.
- **Réduire l'intervalle des valeurs à un intervalle standard (0 à 100)** : on applique une règle de trois pour que la nouvelle valeur minimum soit 0 et la nouvelle valeur maximum soit 100 pour chacune des variables.

Avec la standardisation, les deux variables sont lisibles. Pour chaque variable, on a réduit l'intervalle des valeurs à l'intervalle 0 à 100.

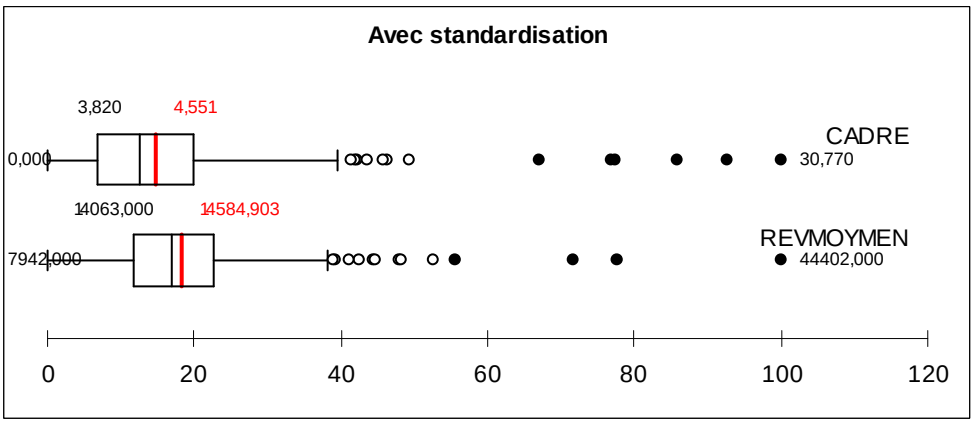


Fig. 8 –Visualisation sur une même échelle des variables « Cadre » et « Revenu » standardisées selon leur valeur minimum et leur valeur maximum

Pour la suite, sauf précision contraire, on considèrera que les variables ont été standardisées d'une manière où d'une autre.

2.1.2.2 Visualisation en coordonnées parallèles

Le graphique en coordonnées parallèles [*ID90, Siir00*] est constitué d'axes verticaux représentant chacun une variable. Ces axes sont placés en parallèle et les coordonnées de chaque individu sont représentées sur ces axes. Une ligne reliant les coordonnées de proche en proche constitue la représentation de l'individu dans le graphique. L'exemple ci-dessous montre le profil des zones de Paris et de sa Petite Couronne en fonction du revenu moyen et des taux de cadres, d'intermédiaires, d'employés et d'ouvriers. L'individu moyen est représenté par la ligne bleue. Un exemple d'individu réel est surligné par une ligne noire.

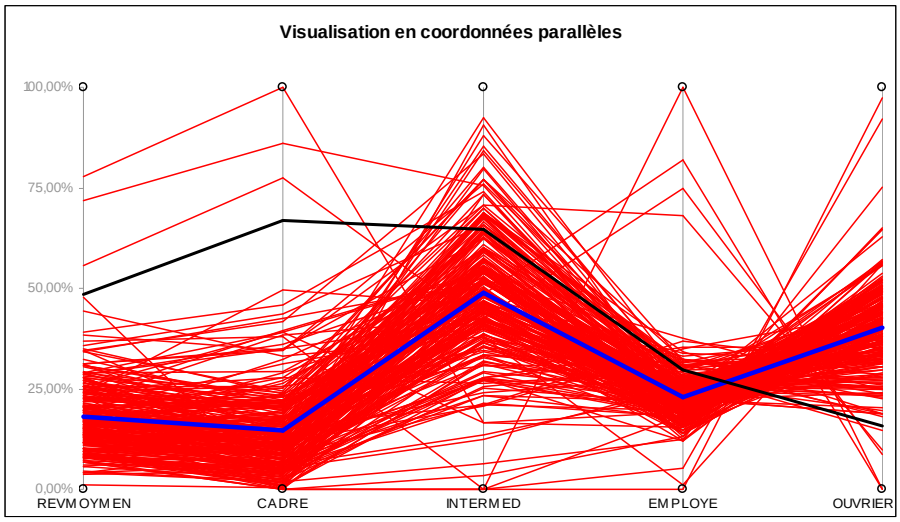


Fig. 9 –Visualisation en coordonnées parallèles de plusieurs variables

L'inconvénient de ce type de graphique est son aspect « embrouillé » lorsque le nombre d'individus devient important comme c'est le cas dans l'exemple précédent.

On peut signaler une amélioration [*YMR03*] permettant de visualiser les différentes classes d'une hiérarchie dans un seul graphique.

2.1.2.3 Diagramme en étoile

Il s'agit d'une variante de la visualisation en coordonnées parallèles. Les axes sont simplement organisés suivant les rayons d'un cercle. Il n'est intéressant que pour des variables périodiques ayant donc un « ordre circulaire » telles que les jours de la semaine ou les mois.

L'exemple suivant montre l'évolution du nombre de mariage suivant les mois (ils constituent les variables) pour quatre années (elles constituent les individus).

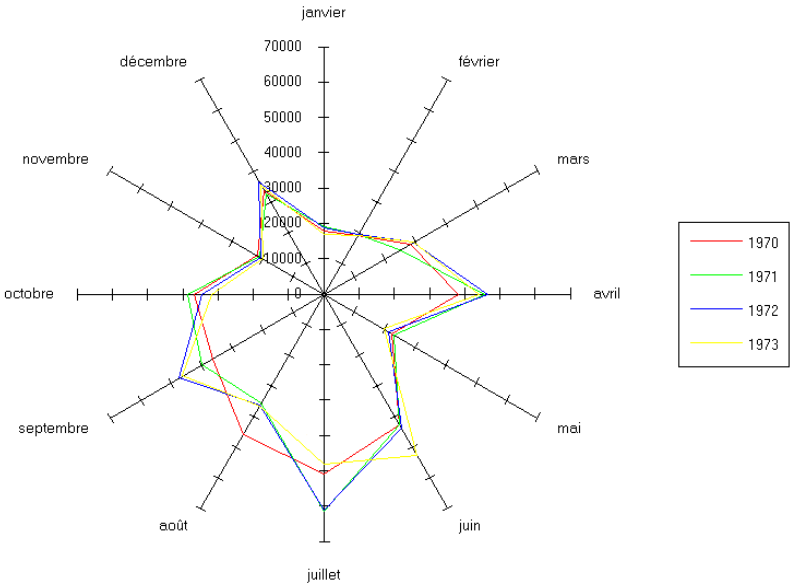


Fig. 10 – Exemple de diagramme en étoiles de plusieurs variables

Ce graphique met en évidence le caractère périodique du nombre de mariages car il y a très peu de variations d'une année à l'autre pour un même mois.

2.1.2.4 Visualisation de boites à moustaches en parallèles

Contrairement à l'histogramme, la boite à moustaches est « unidimensionnel » et se prête bien à une représentation commune sur plusieurs variables, comme l'illustre l'exemple ci-dessous.

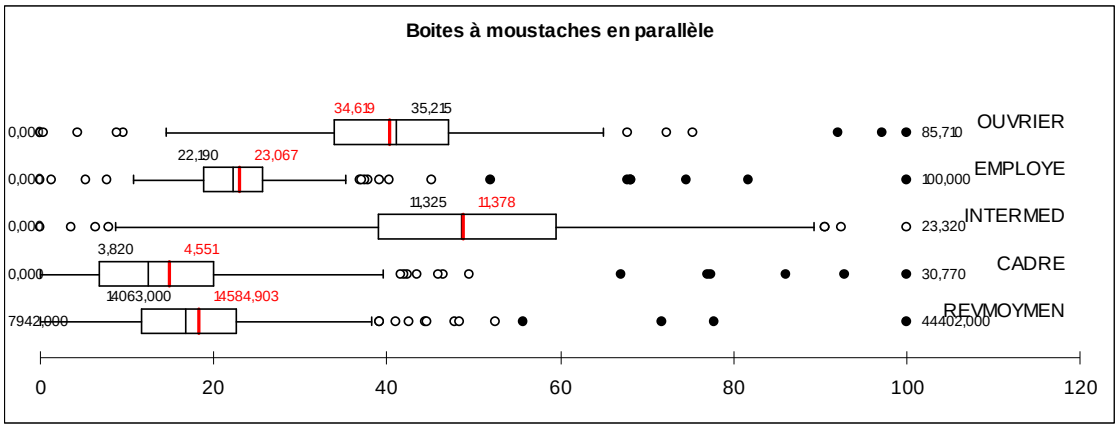


Fig. 11 – Exemple de visualisation de plusieurs variables en utilisant des boîtes à moustaches en parallèle

Son principal avantage est sa clarté. Son principal inconvénient est l'absence de ligne matérialisant les individus dans le graphique comme c'est le cas dans la visualisation en coordonnées parallèles.

2.2 Visualisation de classes sous forme de tableau

Nous allons tout d'abord voir les pseudo-tableaux qui sont des profils de classes juxtaposés avant de voir les différentes techniques de visualisation sous forme de tableaux classes/variables. Nous aborderons aussi le problème de la visualisation des données mixtes.

2.2.1 Pseudo-tableau : Profils de classes juxtaposés

Le graphique décrit ici n'est pas un « vrai » tableau dans la mesure où il s'agit d'une juxtaposition horizontale de graphiques.

Les profils de classes juxtaposés sont la représentation la plus utilisée pour les représentations graphiques. Chaque classe d'individus est visualisée par l'un des graphiques précédents permettant de décrire les individus de la classe en fonction de plusieurs variables. On appellera ce graphique un profil de classe. La visualisation simultanée de plusieurs profils de classes nécessite seulement que les profils aient la même échelle pour une variable donnée et le même ordre des variables.

L'exemple suivant montre le graphique obtenu par la juxtaposition des boites et moustaches de trois classes.

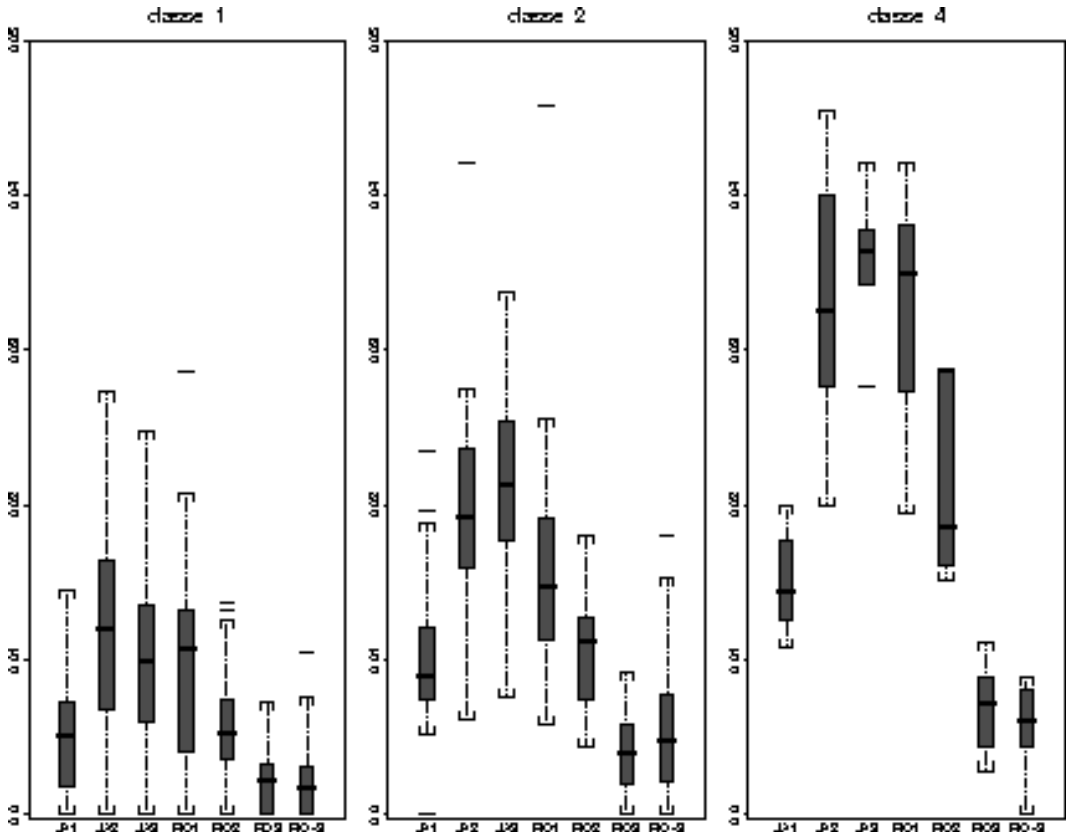


Fig. 12 – Exemple de visualisation parallèle des profils de trois classes, les profils étant des boîtes à moustaches

L'avantage de ce type de graphique est de permettre la réutilisation immédiate de tous les types de graphiques utilisés pour l'analyse de plusieurs variables. Son inconvénient est de ne pas permettre une comparaison aisée des valeurs d'une même variable dans les différents profils.

2.2.2 Tableau des profils de classes

Le tableau des profils de classes est la représentation la plus utilisée pour la visualisation des classes sous forme textuelle [*Jamb99*]. Pour chaque classe et pour chaque variable, la case correspondante indique la valeur moyenne de la classe. Cela permet une lecture dans les deux sens : une lecture horizontale des profils des classes et une lecture verticale des profils des variables.

	Ouvriers	Employés	Intermédiaires	Cadres	Revenu moyen
Classe 1	34.6 %	23.0 %	11.3 %	4.55 %	14500 €
Classe 2	0 %	0 %	0 %	0 %	5930 €
Classe 3	20.2 %	16.2 %	18.8 %	15.4 %	22300 €
Classe 4	8.13 %	10.1 %	15.8 %	33.7 %	33600 €
Classe 5	4.82 %	8.28 %	8.48 %	36.9 %	65800 €

Fig. 13 – Tableau des profils de classes donnant la valeur moyenne

Ainsi, on peut lire de manière horizontale le profil d'une classe et de plus, on peut aussi lire de manière verticale comment se répartissent les différentes valeurs d'une variable dans les classes. Il est important de remarquer que ce dernier mode de lecture n'était pas possible avec les profils de classes juxtaposés.

Il existe des variantes graphiques de ce tableau. La plus simple est la coloration des cases en fonction de la valeur de la variable [*ESBB98*]. L'échelle de couleur est la suivante et est propre à chaque variable :

- Valeurs très faibles en bleu foncé
- Valeurs faibles en bleu clair
- Valeurs moyenne en blanc
- Valeurs faibles en jaune
- Valeurs faibles en rouge foncé

Généralement, la détermination des différentes catégories de valeurs se fait par rapport aux valeurs standardisées.

Cela donne le résultat suivant :

	Ouvriers	Employés	Intermédiaires	Cadres	Revenu moyen
Classe 1	34.6 %	23.0 %	11.3 %	4.55 %	14500 €
Classe 2	0 %	0 %	0 %	0 %	5930 €
Classe 3	20.2 %	16.2 %	18.8 %	15.4 %	22300 €
Classe 4	8.13 %	10.1 %	15.8 %	33.7 %	33600 €
Classe 5	4.82 %	8.28 %	8.48 %	36.9 %	65800 €

Fig. 14 – Tableau des profils de classes donnant la valeur moyenne et sa position par rapport à la moyenne générale

Il est aussi possible d'ajouter d'autres informations telles que les valeurs minimum et maximum, les quartiles, la médiane, l'écart-type, … pour chaque variable dans chaque classe.

L'exemple suivant montre l'ajout des valeurs minimum et maximum.

	Ouvriers	Employés	Intermédiaires	Cadres	Revenu moyen
Classe 1	34.6 % 0 %→85,7 %	23.0 % 0 %→99,9 %	11.3 % 0 %→17,1 %	4.55 % 0 %→34,9 %	14500 € 7940 €→44400 €
Classe 2	0 % 0 %→0 %	0 % 0 %→0 %	0 % 0 %→0 %	0 % 0 %→0 %	5930 € 0 €→44300 €
Classe 3	20.2 % 0 %→46,3 %	16.2 % 0 %→47,6 %	18.8 % 0 %→99,9 %	15.4 % 0 %→50 %	22300 € 8720 €→61400 €
Classe 4	8.13 % 0 %→42,8 %	10.1 % 0 %→33,3 %	15.8 % 0 %→39,3 %	33.7 % 0 %→100 %	33600 € 9230 €→116000 €
Classe 5	4.82 % 0 %→14,2 %	8.28 % 0 %→28,5 %	8.48 % 0 %→21 %	36.9 % 0 %→78.5 %	65800 € 29600 €→131000 €

Fig. 15 – Tableau des profils de classes donnant en plus l'intervalle des valeurs

Une variante moins courante utilise dans chaque case un graphique donnant une ou plusieurs informations sur la variable dans la classe (histogramme, boite à moustaches, …). Son

principal inconvénient est que les graphiques sont beaucoup plus volumineux que les informations textuelles.

2.2.3 Tableau des profils de classes pour des données mixtes

Il s'agit d'une évolution du tableau des profils de classes afin de permettre la représentation de classes d'« objets symboliques », c'est-à-dire mixtes [*BD00*]. Chaque cellule du tableau peut présenter les informations selon deux modes différents en fonctions de la nature de la variable :

- Pour les variables quantitatives, est représenté l'intervalle des valeurs de la classe (c'est-à-dire la valeur minimale et la valeur maximale)
- Pour les variables qualitatives, est indiquée la liste des modalités et de leur fréquence respectives dans la classe.

L'exemple suivant montre la représentation des intervalles pour les variables quantitatives *Ouvriers* et *Revenu Moyen* et la représentation des fréquences des modalités pour la variable qualitative *Nb de voitures*.

	Ouvriers	Revenu moyen	Nb de voitures	…
Classe 3	[0 %→46,3 %]	[8720 €→61400 €]	Aucune : 21% Une : 48 % Deux et + : 31 %	…
Classe 4	[0 %→42,8 %]	[9230 €→116000 €]	Aucune : 16% Une : 47 % Deux et + : 37 %	…
…	…	…	…	…

Fig. 16 – Tableau des profils de classes pour des données mixtes

2.3 Optimisation de l'ordre des variables et des individus

Un problème commun à toutes les méthodes de visualisation est le choix de l'ordre des variables et le choix de l'ordre des individus. En effet, un mauvais ordre peut rendre le graphique illisible car trop complexe en apparence, alors qu'avec un ordre bien choisi, un graphique peut devenir beaucoup plus lisible grâce à l'apparition de zones de valeurs similaires.

Le tableau ci-dessous est difficilement lisible car l'ordre des variables (Ouvrier, Cadres, …) et l'ordre des individus (Secteur 1, Secteur 2, …) sont mauvais.

	Ouvriers	Cadres	Employés	Revenu moyen	Intermédiaires
Secteur 4	8.13 %	33.7 %	10.1 %	33600 €	15.8 %
Secteur 2	0 %	0 %	0 %	5930 €	0 %
Secteur 3	20.2 %	15.4 %	16.2 %	22300 €	18.8 %
Secteur 5	4.82 %	36.9 %	8.28 %	65800 €	8.48 %
Secteur 1	34.6 %	4.55 %	23.0 %	14500 €	11.3 %

Fig. 17 – Tableau dont l'ordre des variables et l'ordre des individus n'ont pas été optimisés.

Au contraire, le même tableau ci-dessous est plus facilement lisible et interprétable car l'ordre des variables et l'ordre des individus ont été optimisés. Ainsi, hormis *Secteur 2* qui est plutôt atypique, on observe une transition graduelle entre les secteurs 1, 3, 4 et 5 avec *Secteur 3* en pivot et une autre transition graduelle entre « Ouvrier », « Employés », « Intermédiaires », « Cadres » et « Revenu moyen » avec la variable « Intermédiaires » en pivot.

	Ouvriers	Employés	Intermédiaires	Cadres	Revenu moyen
Secteur 1	34.6 %	23.0 %	11.3 %	4.55 %	14500 €
Secteur 2	0 %	0 %	0 %	0 %	5930 €
Secteur 3	20.2 %	16.2 %	18.8 %	15.4 %	22300 €
Secteur 4	8.13 %	10.1 %	15.8 %	33.7 %	33600 €
Secteur 5	4.82 %	8.28 %	8.48 %	36.9 %	65800 €

Fig. 18 – Découpage horizontal et vertical du tableau en morceaux de valeur similaires (zones de valeurs fortes, de valeurs moyennes et de valeurs faibles)

Il faut ainsi trouver un ordre optimal pour les variables et pour les individus afin d'avoir une bonne lecture du tableau.

L'optimisation des variables et l'optimisation des individus sont des problèmes similaires. La différence d'approche tient principalement au fait que les variables sont en petit nombre tandis que les individus peuvent être en nombre bien plus grand. Dans la suite, nous traiterons ce problème en utilisant le terme « objet » pour désigner les individus ou les variables selon le problème d'optimisation concerné. Différentes heuristiques plus ou moins efficaces permettent de trouver un ordre assez bon. Nous verrons les méthodes suivantes ordonnées par ordre d'efficacité croissante :

- La courbe remplissante.
- L'ordre selon la moyenne.
- L'ordre selon l'Analyse en Composantes Principales (ACP).
- L'ordre partiel selon la Classification Hiérarchique.
- L'ordre optimal dans une Classification Hiérarchique.

Les exemples de tri sont fait à partir des données socioprofessionelles des Régions française : 9 variables et 21 individus (les régions). Le graphique ci-dessous est un exemple de tableau synthétique: les données ont été centrées-réduites et les valeurs supérieures à la moyenne sont en rouge, celles inférieures en vert et celles proches de la moyenne en blanc.

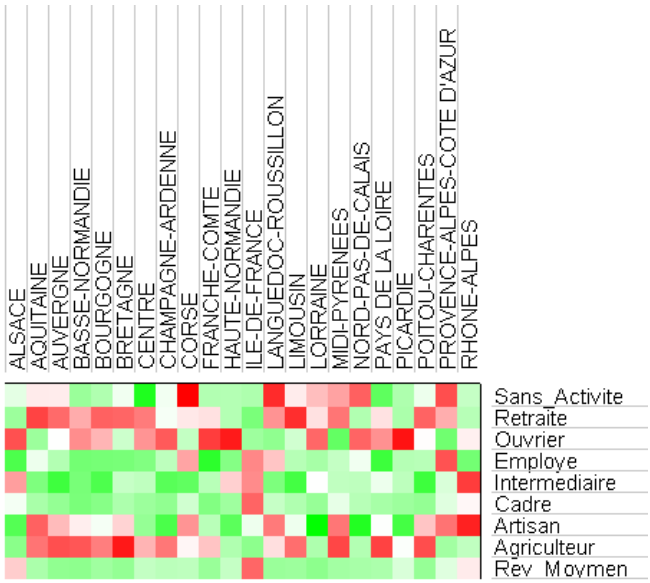


Fig. 19 – Tableau synthétique des Régions pour les variables socioprofessionelles

2.3.1.1 Courbe remplissante pour trier les individus

Cette méthode ne fonctionne qu'avec un petit nombre de variables. Elle n'est donc applicable que pour trier les individus et n'est pas utilisables pour trier les variables.

Une courbe remplissante parcourt tout l'espace des variables avec une précision spécifiable. De cette précision dépend l'ordre et le nombre de secteurs parcouru par la courbe. Il est ainsi possible d'ordonner les données en les affectant au secteur qui les contient. L'exemple suivant montre la courbe remplissante de Peano pour un espace à n=2 dimensions (deux variables). L'espace peut être découpé en $2^{k\times n}$ secteurs avec k un entier quelconque. L'exemple suivant montre la courbe de Morton partant du point ayant les coordonnées les plus basses (1er secteur en bas à droite) jusqu'au point ayant les coordonnées les plus hautes (dernier secteur en haut à gauche) pour 4, 16 et 64 secteurs.

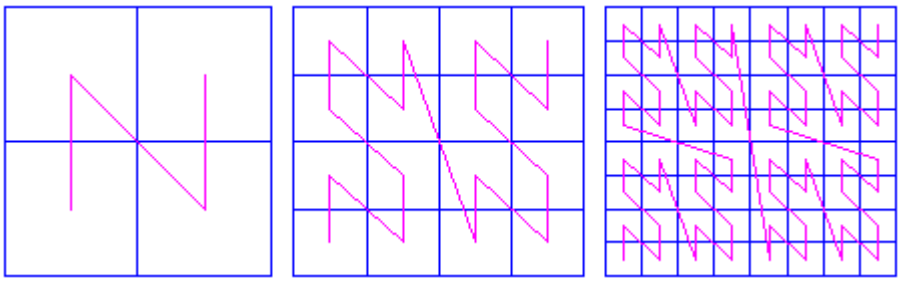


Fig. 20 –Exemple de courbe remplissante d'un espace à deux dimensions (ou deux variables)

L'avantage de cette courbe est d'être aussi précise que l'on souhaite. Si il y a plusieurs individus dans un même secteur, il suffit de découper ce secteur en quatre nouveaux secteurs

plus petits pour les réordonner. L'inconvénient de cette méthode est que la courbe produit obligatoirement quelques « grands sauts » tels que la « zébrure » centrale qui saute de la valeur maximale pour la variable verticale à sa valeur minimale. L'ordre trouvé n'est ainsi pas très pertinent.

2.3.1.2 Tri des individus selon la moyenne de leur valeurs

Cette méthode n'a de sens que pour trier les individus, et elle ne s'applique donc pas au tri des variables [*ESSB98*].

On calcule donc pour chaque individu la moyenne de ses valeurs. Ainsi, il est nécessaire que les variables soient toutes de même nature ou bien standardisées (centrées-réduites par exemple). Ensuite, les individus sont réordonnés selon leur moyenne.

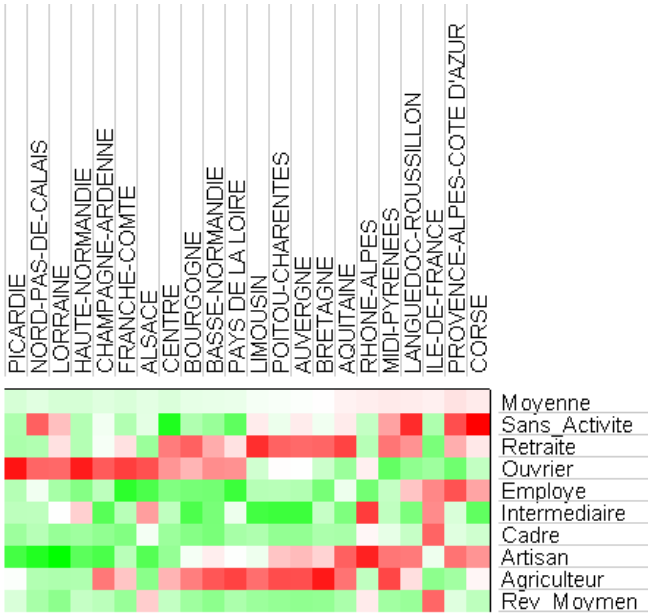


Fig. 21 – Tableau synthétique des Régions triées en fonction de la moyenne des valeurs pour chaque individu (région)

Cette méthode très simple n'est efficace que lorsque toutes les variables sont corrélées positivement. Dans notre exemple, on voit que les Régions sont bien triées pour certaines variables (*Artisan, Ouvrier, …*) mais pas pour d'autres (*Intermédiaire, SansActivite, Retraite,…*). Cette méthode n'est donc pas idéale.

2.3.1.3 Réorganisation via l'Analyse en Composantes Principales (ACP)

L'analyse en composante principale (ACP) permet de créer des variables synthétiques qui s'expriment en fonction des variables naturelles. De plus, ces variables synthétiques peuvent

être ordonnées en fonction de leur facteur d'explication. Ainsi, pour résumer les variables naturelles par une unique variable synthétique, il suffit de prendre la variable synthétique ayant le plus grand facteur d'explication, aussi appelé axe principal. Brièvement, le principe de l'ACP est le suivant : les variables synthétiques sont calculées à partir de la matrice des corrélations des variables. D'un point de vue technique, cela consiste à trouver les valeurs propres de la matrice et les vecteurs associés. Le vecteur ayant la plus grande valeur propre est la variable synthétique ayant la plus grande explication.

L'exemple suivant donne les coefficients de la variable synthétique ayant la plus grande explication, il s'agit du « axe_1 ».

Variables	**axe_1**
Retraite	-0,422408
Agriculteur	-0,361895
Sans_Activite	-0,255668
Artisan	-0,253074
Employe	-0,0120422
Ouvrier	0,161504
Cadre	0,375021
Rev_Moymen	0,439789
Intermediaire	0,448228

Fig. 22 –Coordonnées des variables sur l'axe principal

Ces coefficients définissent la formule linéaire permettant de calculer pour chaque individu X (les régions ici) sa projection sur l'axe principal (appelé ici *axe_1*). Nous avons dans le cas présent :

$$\begin{aligned}\text{axe\_1}(X) = \; & -0.42\times\text{Retraite}(X) - 0.36\times\text{Agriculteur}(X) - 0.25\times\text{Sans\_Activite}(X) \\ & -0.25\times\text{Artisan}(X) - 0.01\times\text{Employe}(X) + 0.16\times\text{Ouvrier}(X) \\ & +0.37\times\text{Cadre}(X) + 0.43\times\text{Rev\_Moyen}(X) + 0.44\times\text{Intermediaire}(X)\end{aligned}$$

La figure suivante montre la matrice des individus (les Régions) ordonnées selon leur valeur sur l'axe principal (*axe_1*) qui est lui aussi indiqué pour information.

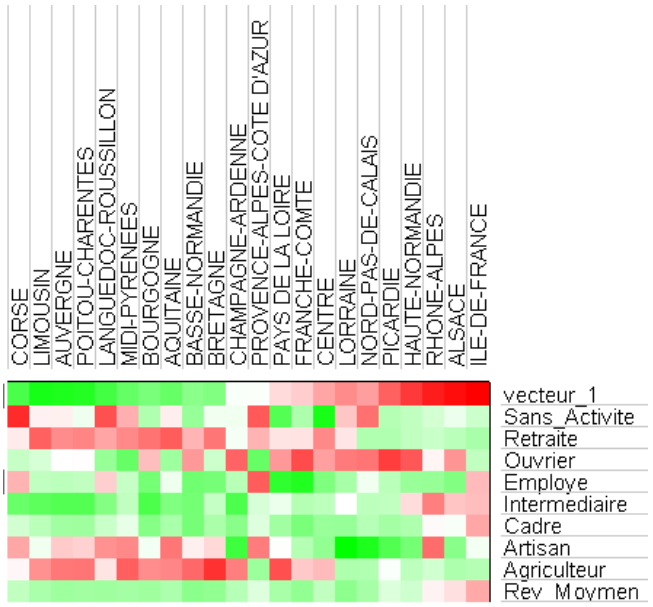


Fig. 23 – Tableau synthétique des Régions triées en fonction de leur projection sur l'axe principal (vecteur_1) de l'ACP

On observe que les régions sont plutôt bien triées pour certaines variables (*Rev_Moyen, Intermédiaire, …*) et beaucoup moins pour d'autres (*Ouvrier, Sans_Activité, …*).

Par ailleurs, les coefficients traduisent assez bien les relations entre les variables : pour des coefficients proches, les variables sont a priori semblables, tandis que pour des coefficients distincts voir opposés, les variables sont dissemblables. Nous pouvons donc aussi trier les variables selon leur coefficient. Le résultat concernant le tri des variables seules est le suivant :

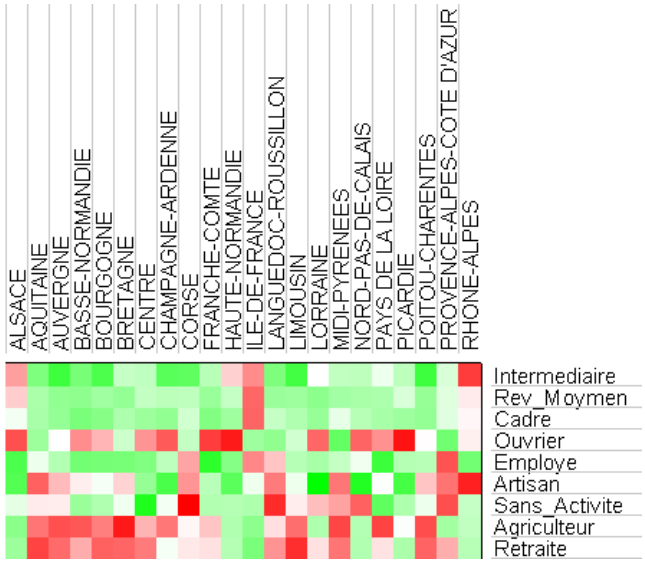


Fig. 24 – Tableau synthétique des Régions dont les variables ont été triées en fonction de leur coefficient sur l'axe principal de l'ACP

On observe, que certaines variables sont vraiment très proches : *Intermédiaire*, *Rev_Moyen* et *Cadre* ou encore *Agriculteur* et *Retraite*. De plus ces deux groupes semblent s'opposer.

Finalement, le tri des individus et le tri des variables combinés donne le résultat suivant :

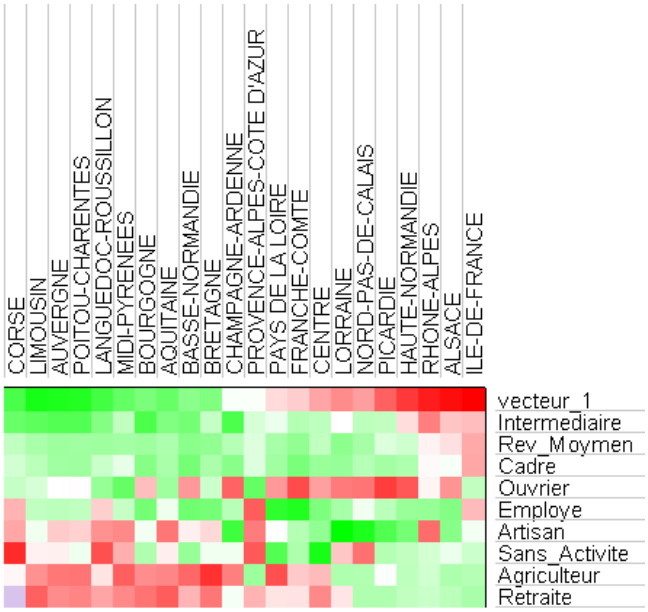


Fig. 25 – Tableau synthétique des Régions dont les variables et les individus ont été triés selon les deux méthodes précédentes

Il apparaît clairement des zones de valeurs extrêmes (en rouge ou en vert) placées dans chaque coin du tableau suivant les deux diagonales : La « diagonale » des valeurs fortes part du coin droit en bas et va au coin gauche en haut tandis que la « diagonale » des valeurs faibles part du coin gauche en haut et va au coin droit en bas. Cependant, la zone centrale du tableau est peu ou pas triée et est difficilement interprétable. Il s'agit du principal inconvénient de l'ACP car elle ne trie bien que les « extrêmes ». Pour cette raison, on obtient de bons résultats que dans le cas où toutes les variables sont fortement corrélées entre elles, ce qui n'était pas le cas ici.

2.3.1.4 Réorganisation via la classification hiérarchique

Il s'agit de créer une classification hiérarchique sur les individus (ou sur les variables) pour les trier. Le principal problème de cette méthode est que l'ordre trouvé n'est que partiel car, pour chaque nœud, ses deux branches peuvent être permutées. L'exemple suivant montre deux hiérarchies équivalentes mais de qualité différente. Les individus sont caractérisés par deux variables et les regroupements (les nœuds) sont représentés par la valeur moyenne :

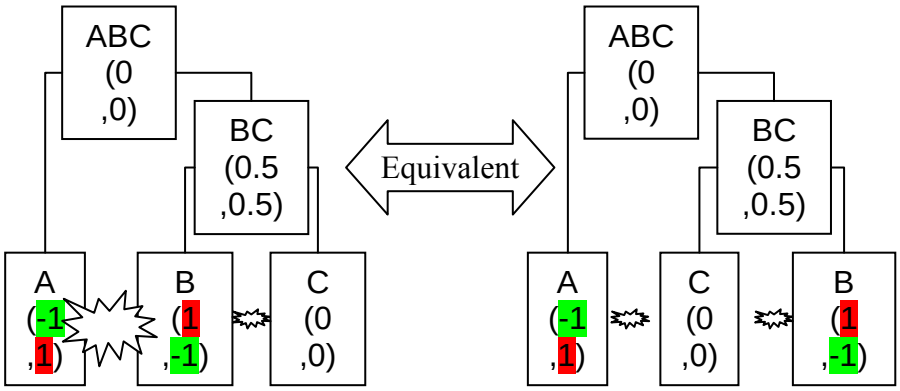


Fig. 26 – Exemple de deux hiérarchies équivalente, mais celle à gauche « trie » moins bien car les individus A et B bien que très différents sont quand même cote à cote.

On observe dans cet exemple que les deux hiérarchies sont équivalentes, la seule différence étant la permutation de B et C. Pourtant, les résultats sont de qualités différentes. Dans la hiérarchie ABC, les individus A et B bien que très différents, sont placés l'un à côté de l'autre, ce qui n'est pas souhaitable. Au contraire, dans la hiérarchie ACB, l'individu C est idéalement placé entre A et B. Nous allons voir par la suite comment améliorer la classification partielle obtenue avec une hiérarchie.

2.3.1.5 Recherche de l'ordre optimal dans une classification existante

Une autre façon d'aborder le problème est de le définir de la façon suivante : parmi tous les ordres possibles, l'ordre optimal est celui qui minimise la somme des distances entre les individus consécutifs. Ce problème d'ordonnancement est connu sous le nom de Sequential Ordering Problem (SOP) ou Linear Ordering Problem (LOP). C'est une variation du problème du voyageur de commerce, Traveling Salesman Problem (TSP), qui à la différence que le chemin (l'ordre des individus) n'est pas une boucle (on ne prend pas en compte la distance entre le dernier individu et le premier individu qui ferme la boucle). Soit P une permutation de $\{1,...,n\}$, l'ordre correspondant des individus est $O_P = \{o_{P(1)},...,o_{P(n)}\}$et son coût est $C_P = \sum_{i=1}^{n-1} d(o_{P(i)}, o_{P(i+1)})$ avec d la distance entre deux individus. L'ordre optimal est O_{Opt} correspondant à $C_{Opt} = \min_{P \in permutation}(C_P)$. Le coût peut être décrit comme la longueur du « chemin » correspondant à l'ordre

En reprenant l'exemple précédent et en calculant son coût en prenant la distance euclidienne au carré, on s'aperçoit que la hiérarchie de droite est bien meilleure. Pour la hiérarchie ABC, nous avons :

$$\begin{aligned} C_{ABC} &= d(A,B) + d(B,C) \\ &= \left((-1-1)^2 + (1-(-1))^2\right) + \left((1-0)^2 + (-1-0)^2\right) \\ &= \left((-2)^2 + 2^2\right) + \left(1^2 + 1^2\right) \\ &= 8 + 2 \end{aligned}$$

$$
\begin{aligned}
&= 10 \\
C_{ACB} &= d(A,C) + d(C,B) \\
&= \left((-1-0)^2 + (1-0)^2\right) + \left((1-0)^2 + (0-(-1))^2\right) \\
&= \left((-1)^2 + 1^2\right) + \left(1^2 + 1^2\right) \\
&= 2 + 2 \\
&= 4
\end{aligned}
$$

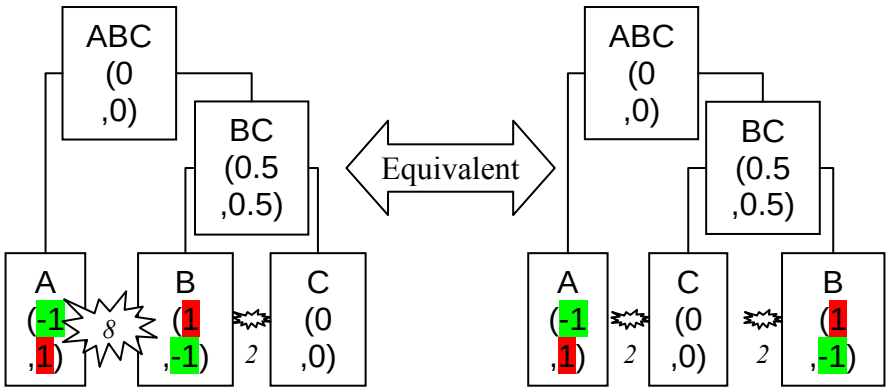


Fig. 27 – Calcul du coût pour les deux hiérarchies équivalente, Le coût élevé de la hiérarchie de gauche est dû à la distance élevée entre les individus A et B.

La recherche exhaustive n'est pas réalisable car pour *n* individus, il y a *n!* ordres différents à tester (ce problème est NP-complet). On utilise donc les heuristiques comme celle indiquée par la suite.

Ce problème peut être simplifié par la création d'une hiérarchie binaire sur les individus. Dans ce cas, le nombre de permutations possibles est limité à 2^{n-1}. Toutefois, une recherche exhaustive n'est toujours pas faisable en un temps raisonnable même pour un nombre d'individus de l'ordre de la centaine. Cependant, Z. Bar-Joseph et d'autres [*BDGHJS02*] ont proposé un algorithme capable de résoudre de manière exact ce problème avec une complexité de $o(n^3)$ en temps et $o(n^2)$ en espace. Cette faible complexité permet le tri rapide pour des individus (ou des variables) de l'ordre du millier.

L'exemple suivant montre le tri des individus réalisé à partir d'une classification hiérarchique de ces individus :

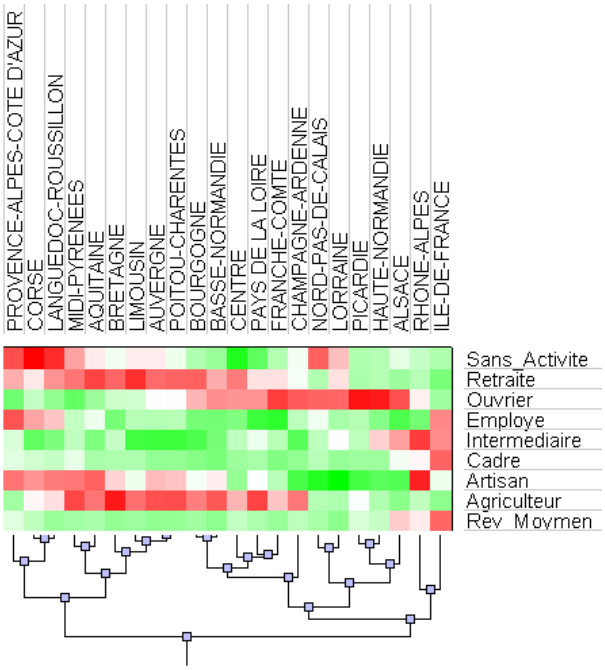


Fig. 28 – Tableau synthétique des Régions triées selon l'ordre optimal dans la hiérarchie

Les résultats sont meilleurs qu'avec les autres méthodes. En effet, pour toutes les variables, les régions sont bien triées à part pour quelques exceptions, l'*Ile de France* et les *Rhone-Alpes* principalement.

Nous effectuons aussi le tri des variables selon cette même méthode :

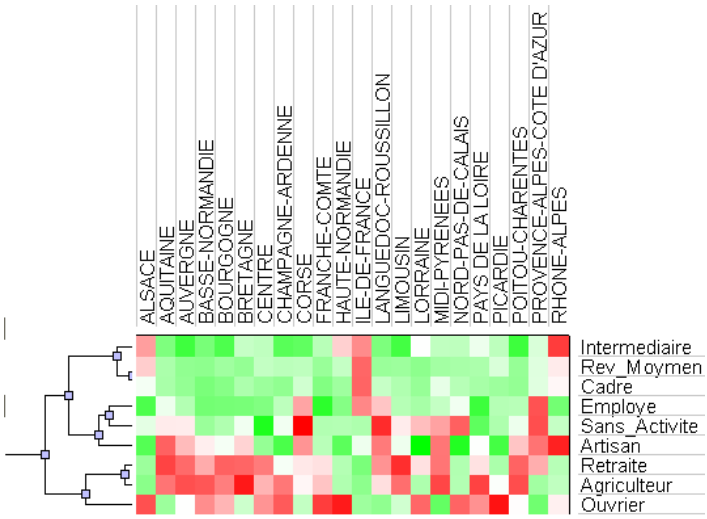


Fig. 29 – Tableau synthétique des Régions dont les variables ont été triées selon l'ordre optimal dans une classification hiérarchique

L'ordre obtenu est sensiblement le même que celui obtenu avec le tri selon l'ACP. La principale différence tient au fait que la variable *Ouvrier*, qui est très différentes des autres, a été placée dans ce cas au bord. En effet, ainsi placée, elle est mise à l'écart. Au contraire, l'ACP a tendance à placer au centre les variables difficiles à trier.

Finalement, en combinant les deux tris nous obtenons le graphique suivant :

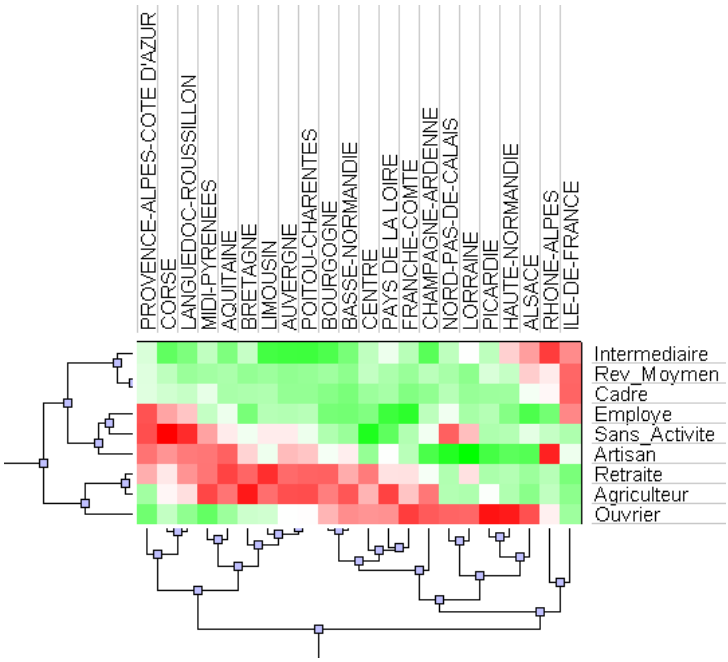


Fig. 30 – Tableau synthétique des Régions dont les variables et les individus ont été triés selon les deux méthodes précédentes

Le résultat est bien meilleur que celui obtenu avec l'ACP. En effet les zones sont clairement séparées : il s'agit principalement de deux diagonales superposées et allant de la gauche en haut vers le bas à droite. À cela s'ajoute une zone de fortes valeurs en haut à droite et le traitement particulier de la variable *Ouvrier*.

Nous pouvons dès lors faire une mini-analyse :

- Tout d'abord, il y a l'opposition Rural/Urbain avec l'*Ile de France* (et dans une moindre mesure le *Rhone-Alpes* et l'*Alsace*) opposé aux autres régions.
- Nous retrouvons ensuite une opposition Nord/Sud avec au Sud plutôt des *Sans_Activités, de l'Artisanat* et des *Retraités* et au Nord beaucoup d'*Ouvriers*.

On remarquera aussi quelques exceptions comme :

- Le fort taux de *Sans_Activité* en Lorraine dans le *Nord-pas-de-Calais*, historiquement lié à la fermeture des aciéries et des mines.
- Le fort taux d'*Artisanat* en *Rhône-Alpes* qui est caractéristique des zones montagneuses et/ou touristiques (sport d'hiver dans les Alpes notamment).

3 Lissage spatial

Le lissage spatial est une technique largement utilisée pour mettre en évidence les tendances à grande échelle spatiale. Nous verrons dans un premier temps les distinctions entre les notions de « lissage spatial » et d' « interpolation spatiale ». Nous étudierons ensuite les composants de

la méthode de noyaux de densité (*kernel density*) qui est puissante et facile à paramétrer. En premier lieu, il s'agira des fonctions d'interaction spatiale qui permettent de définir une « gradation » de la notion de voisinage. Puis nous verrons les fonctions de lissage servant à calculer la carte de lissage spatial. Ce faisant, nous illustrerons l'intérêt du lissage spatial pour la description de la répartition des supermarchés et hypermarchés en Loire-Atlantique.

3.1 Méthodes de lissage

Les méthodes de lissages spatiales sont souvent considérées comme des méthodes d'interpolation [*FN91*, *Wats92*] bien que le but soit totalement différent. Cela est principalement dû au fait que certaines méthodes sont communes à l'interpolation et au lissage. Pourtant, le but du lissage spatial est de mettre en évidence des phénomènes sous-jacents qui n'apparaissent qu'aux grandes échelles et invisibles à l'échelle locale, tandis que l'interpolation a pour objectif la prédiction des valeurs manquantes à partir de quelques valeurs mesurées en quelques points. Toutefois, dans la pratique, certaines méthodes sont communes au lissage et à l'interpolation, ce qui peut ajouter à la confusion. Parmi les méthodes d'interpolation, se trouvent les méthodes dites « exactes » qui prédisent la même valeur que celle mesurée pour chaque point de mesure. Il s'agit par exemple du Kriging [*Krig51*], de l'interpolation par l'algorithme IDW (*Inverse Distance Weighting*) [*Shep68*], de l'interpolation par triangulation ou par voisins naturels [*Sibs81*] ou encore de l'interpolation multiquadratique [*Hard71*]. D'autre part, au contraire des méthodes d'interpolation exactes les méthodes d'interpolation non exactes peuvent être assimilées à des méthodes de lissage. Il s'agit de l'algorithme NDW (*Normal Distance Weighting*), des surfaces de tendances (*trend surfaces*) [*Math76*], et des décompositions en séries (de Fourrier ou autres) [*Davi86*]. Le NDW est en fait une modification du IDW par l'ajout d'un paramètre de lissage dans la formule. L'analyse en surface de tendance est une sorte de régression polynomiale, le paramètre de lissage étant le degré du polynôme, plus celui-ci est bas, plus le résultat est lisse. Enfin, la décomposition en séries donne un résultat d'autant plus lisse que le nombre de séries retenues est faible. Les inconvénients de ces méthodes d'interpolation non exactes sont, d'une part, la difficulté pour manipuler le paramètre de lissage et d'autre part, la qualité insuffisante du lissage qui n'est pas forcément celle recherchée.

La méthode de lissage la plus employée est le lissage basé sur les noyaux de densité (*kernel density*) [*Silv86*, *BG95*]. Le point fort de cette méthode de lissage est qu'elle est très proche de la notion de densité. En effet, c'est une méthode qui calcule localement la moyenne (ou la somme) des valeurs les plus proches. De plus, contrairement aux méthodes précédemment citées, le paramètre de lissage est clair et s'exprime par une distance. L'intensité de la relation de voisinage est une fonction décroissante, généralement la fonction gaussienne. Une variante du lissage par les noyaux de densité est le lissage adaptatif des noyaux de densité [*Sain94*]. Cette méthode calcule dynamiquement le paramètre de lissage. Cependant, les résultats sont plutôt difficiles à interpréter car, le paramètre de lissage pouvant grandement varier (de manière automatique), plusieurs échelles très différentes sont présentes simultanément. Par ailleurs, la visualisation de plusieurs échelles sur une même carte grâce à une hiérarchie est inutilisable dans ce cas. Ainsi, nous utiliserons plutôt une méthode de lissage issue des noyaux de densité avec un paramètre de lissage à fixer.

Nous allons maintenant voir les fonctions d'interaction spatiale utilisées par les noyaux de densité (*kernel density*).

3.2 Fonctions d'interaction spatiale

La fonction d'interaction spatiale permet de définir une « gradation » de la notion de voisin. En effet, la définition traditionnelle du voisinage est trop restrictive :

- Si deux objets sont voisins, ils sont en interaction, elle vaut 1.
- Sinon, ils ne sont pas voisins et leur interaction est nulle, elle vaut 0.

Au contraire, une fonction d'interaction spatiale permet de définir des interactions comprises entre 1 et 0 afin d'éviter un brusque « saut » entre les objets qui sont voisins et ceux qui ne le sont pas. Pour cela, la fonction d'interaction spatiale s'appuie sur la distance séparant les objets. Afin de simplifier l'utilisation de l'interaction spatiale, nous considérons que les objets sont représentés par une position géographique ponctuelle, permettant ainsi de calculer les distances entre ces objets. Nous allons voir successivement les trois fonctions d'interaction spatiale les plus communes, par ordre d'intérêt croissant.

3.2.1 La fonction d'interaction spatiale en plateau

Elle traduit la notion de voisinage habituel : jusqu'à une certaine distance R (le rayon), l'interaction est totale (les objets sont voisins) et elle est nulle au-delà (les objets ne sont plus voisins). Nous avons :

$$f_{R\_plateau}(dist) = 1 \text{ si } dist \le R \text{ et}$$
$$= 0 \text{ sinon.}$$

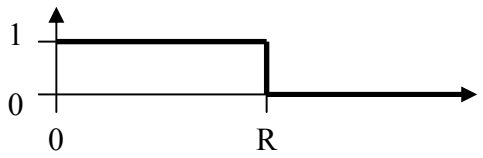


Fig. 31 –Fonction d'interaction spatiale en " plateau "en fonction de la distance

Ainsi pour deux objets X_1 *et* Y_1 tels que $distance(X_1, Y_1) \le R$, $f_{R\_plateau}(distance(X_1, Y_1)) = 1$, ce qui signifie que les deux objets ont un degré de voisinage maximal, ils sont pleinement voisins. Au contraire, pour deux objets X_2 *et* Y_2 tels que $distance(X_2, Y_2) > R$, $f_{R\_plateau}(distance(X_2, Y_2)) = 0$, ce qui signifie que les deux objets ont un degré de voisinage nul, ils ne sont pas voisins.

3.2.2 La fonction d'interaction spatiale triangulaire

Cette fonction permet une décroissance régulière du degré de voisinage. En effet, elle décroît linéairement jusqu'à s'annuler à la distance *R* et reste nulle au-delà. Elle évite le " saut " brusque de la fonction d'interaction spatiale en " plateau ". Elle est définie par :

$$f_{R\_triang}(dist) = \frac{R - dist}{R} \text{ si } dist \le R$$
$$= 0 \text{ sinon.}$$

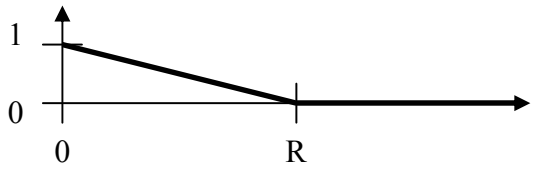


Fig. 32 –Fonction d'interaction spatiale triangulaire en fonction de la distance

3.2.3 La fonction d'interaction spatiale gaussienne

Comme son nom l'indique, elle évolue selon une courbe de Gauss. Bien qu'elle soit assez peu différente de la fonction d'interaction spatiale triangulaire, elle évite les cassures (non dérivabilité) à l'origine et à la distance R. Par ailleurs, bien qu'elle ne soit jamais nulle, on considère qu'elle l'est pour les distances supérieures à $1{,}5 \times R$. Elle est définie par :

$$f_{R\_gauss}(dist) = \exp\left(-\frac{1}{2} \times \left(\frac{dist}{Rg}\right)^2\right)$$

Avec $Rg = \frac{R}{\sqrt{8 \times \ln(2)}} \approx 0.42 \times R$ de manière à avoir $f_{R\_gauss}\left(\frac{R}{2}\right) = f_{R\_triang}\left(\frac{R}{2}\right) = \frac{1}{2}$

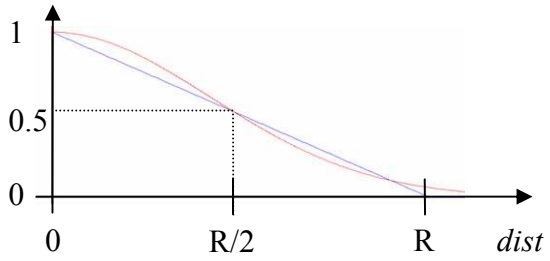


Fig. 33 –Fonction d'interaction spatiale gaussienne en fonction de la distance

La fonction d'interaction spatiale triangulaire et la fonction d'interaction spatiale gaussienne permettent, chacune à leur manière, de quantifier la notion de voisinage de manière à ce que la valeur de voisinage diminue de façon continue par rapport à la distance. Cette propriété est très importante afin d'obtenir un « bon lissage spatial », c'est-à-dire ayant un aspect lisse, sans cassure. Ainsi, la recherche des pôles que nous verrons ultérieurement donne de bons résultats, ce qui n'est pas le cas avec la fonction d'interaction spatiale en plateau.

3.3 Fonctions de lissage spatial

À partir de la fonction d'interaction spatiale f_R, nous pouvons déterminer la valeur lissée en n'importe quel lieu *X*, à partir des caractéristiques des objets O_j . On note respectivement $Lieu(O_j)$ la position et $Val(O_j)$ la valeur de l'objet O_j. Nous pouvons alors déterminer la contribution d'un objet O_j à la valeur lissée du lieu *X*, il s'agit de :

$$Ctrb_{f,R}(X,O_j) = Val(O_j) \times f_R(dist(X, Lieu(O_j)))$$

La valeur lissée qui peut être la Somme Lissée ou la Moyenne Lissée, se définit alors comme suit:

$$SommLiss_{f,R}(X) = \sum_{j=1}^{n} Ctrb_{f,R}(X,O_j)$$

$$MoyLiss_{f,R}(X) = \frac{SommLiss_{f,R}(X)}{\sum_{j=1}^{n} f_R(dist(X, Lieu(O_j)))} = \frac{\sum_{j=1}^{n} (Val(O_j) \times f_R(dist(X, Lieu(O_j))))}{\sum_{j=1}^{n} f_R(dist(X, Lieu(O_j)))}$$

On remarquera que la contribution s'exprime $Ctrb_{f,R}(X,O_j)$ dans la même unité que celle de la valeur $Val(O_j)$.

Nous définissons aussi une version pondérée de ces formules de lissage, en utilisant $Poids(O_j)$ le poids d'un objet O_j, cela donne :

$$CtrbPdr_{f,R}(X,O_j) = Val(O_j) \times Poids(O_j) \times f_R(dist(X, Lieu(O_j)))$$

$$SommPdrLiss_{f,R}(X) = \sum_{j=1}^{n} CtrbPdr_{f,R}(X,O_j)$$

$$MoyPdrLiss_{f,R}(X) = \frac{SommPdrLiss_{f,R}(X)}{\sum_{j=1}^{n} Poids(O_j) \times f_R(dist(X, Lieu(O_j)))}$$

$$= \frac{\sum_{j=1}^{n} (Val(O_j) \times Poids(O_j) \times f_R(dist(X, Lieu(O_j))))}{\sum_{j=1}^{n} Poids(O_j) \times f_R(dist(X, Lieu(O_j)))}$$

La *Somme Lissée* est intéressante pour des variables pouvant être sommées (la population par exemple). Au contraire, la *Moyenne Lissée* est utile pour les autres types de variables pouvant être moyennées, comme les taux ou les pourcentages. Cependant, dans ce cas, il est préférable d'utiliser la *Moyenne Pondérée Lissée* qui reflète mieux la réalité en utilisant la pondération requise (par exemple, la population pour pondérer des variables exprimées en pourcentage de la population). La *Somme Pondérée Lissée* n'a pas beaucoup d'intérêt. L'utilisation du *Maximum Lissé* cherche à traduire l'accessibilité. Les fonctions de lissage possèdent ainsi deux paramètres, le choix de la fonction d'interaction spatiale *f* et le choix du rayon *R*.

La *Somme Lissée* induit une sorte de biais mis en évidence dans l'exemple suivant. L'information sur la surface d'une zone peut donner lieu à deux lissages à peu près équivalent : soit on effectue un lissage par le calcul de la *Somme Lissée* des surfaces, soit par le calcul de la *Moyenne Lissée* des densités des surfaces. Comme les valeurs de densité des surfaces valent 1 (c'est-à-dire 1 km² par km²) pour chaque îlot, ce dernier traitement a pour résultat une valeur lissée valant aussi 1 pour n'importe quel zone. Il en va tout autrement pour la sommation. La carte suivante montre la sommation des surfaces des îlots réalisée avec un rayon de 10 km en utilisant la fonction d'interaction spatiale triangulaire.

D'autre part, la surface lissée théorique pour un lieu situé au milieu (non situé au bord) obtenue avec un maillage infiniment fin et pour une fonction d'interaction spatiale *f* est égal à :

$$SommeThéoMilieu(f_R) = \int_{\theta=0}^{2\pi} \int_{r=0}^{R} f_R \times r \times dr \times d\theta$$

Ainsi, pour un lieu situé sur le bord, ce bord étant en ligne droite, nous n'avons que la moitié des voisins d'un lieu situé au milieu :

$$SommeThéoBord(f_R) = \frac{1}{2} \times SommeThéoMilieu(f_R)$$

Le calcul suivant donne la surface lissée théorique d'un lieu situé au milieu pour la fonction d'interaction spatiale triangulaire avec un rayon de 10 km :

$$SommeThéoMilieu(f_{R\_triang}) = \int_{\theta=0}^{2\pi} \int_{r=0}^{R} \frac{R-r}{R} \times r \times dr \times d\theta = \int_{\theta=0}^{2\pi} \int_{r=0}^{R} \left[\frac{r^2}{2} - \frac{r^3}{3R} \right]_0^R d\theta$$

$$= \int_{\theta=0}^{2\pi} \left(\frac{R^2}{2} - \frac{R^2}{3} \right) d\theta = \int_{\theta=0}^{2\pi} \frac{R^2}{6} d\theta$$

$$= \frac{2 \times \pi \times R^2}{6} = \frac{\pi \times R^2}{3}$$

Avec *R*=10 km, la somme lissée théorique au milieu est donc $\frac{\pi \times 100}{3} \approx 105$ km² et au bord de 52 km². Nous constatons que nous retrouvons bien ces valeurs dans la carte suivante.

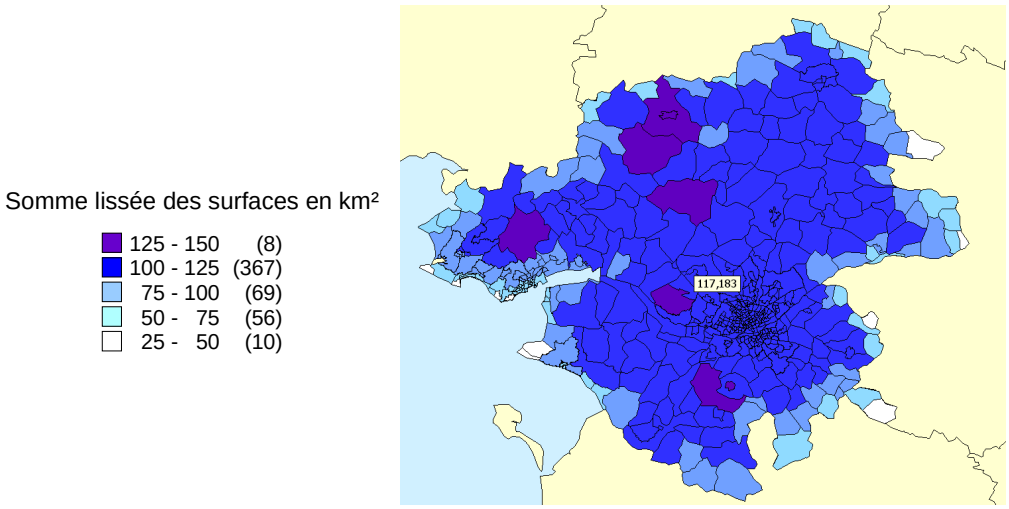


Fig. 34 –Carte de la surface lissée en km²

Les valeurs lissées non situées au bord fluctuent donc autour de la valeur de 105 km² en raison du maillage assez imparfait des îlots et on constate que les lieux situés à la bordure ont effectivement une valeur lissée inférieure, comprise entre un tiers et la moitié de la valeur lissée des lieux non situés au bord.

L'utilisation de la sommation est toutefois justifiée dans la mesure où le bord est réellement une frontière infranchissable ou s'il n'y a aucun objet géographique à prendre en compte de l'autre côté du bord. L'utilisation de la moyenne lissée se justifie dans des conditions contraires, c'est-à-dire lorsque le bord est une frontière artificielle derrière laquelle se situent des objets géographiques ayant des valeurs similaires à celles des objets géographiques pris en compte. Pour ces raisons, pour le bord maritime du département de la Loire-Atlantique il est préférable d'utiliser la *Somme Lissée* et pour le bord administratif marquant la limite entre les départements, il est préférable d'utiliser la *Moyenne Lissée*. D'une manière générale, dans le doute, il vaut mieux considérer les limites comme artificielles et donc d'utiliser la *Moyenne Lissée*.

Nous allons voir maintenant un exemple de lissage permettant d'analyser la localisation des supermarchés et hypermarchés dans le département de Loire Atlantique. Une analyse plus complète est présentée en annexe. Les données brutes sont la densité d'hypermarché et de supermarchés par km², calculé pour chaque zone (îlot) en fonction du nombre de magasins localisés dans la zone et de sa superficie.

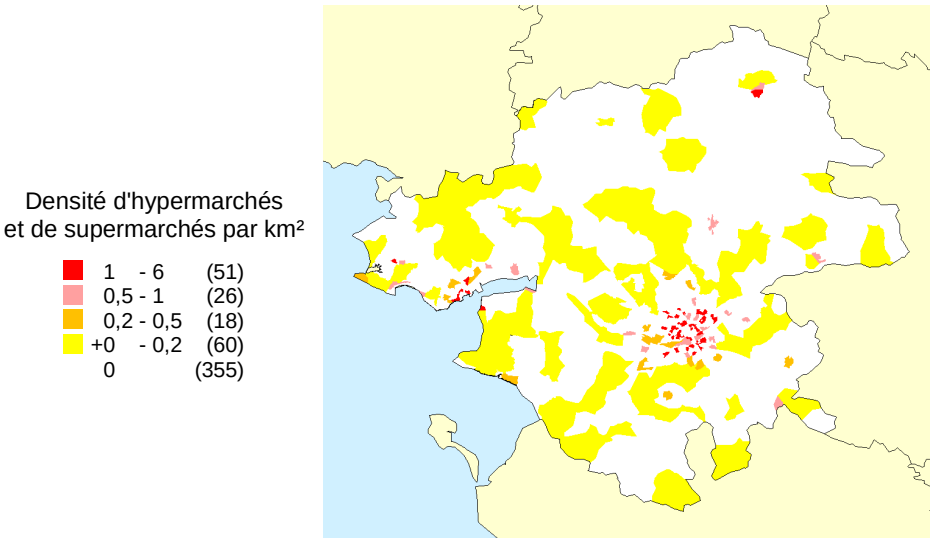


Fig. 35 –Carte du nombre d'hypermarchés et de supermarchés par km² par îlot

On constate que cette carte n'est pas très intéressante à analyser en terme d'accessibilité car les clients d'un magasin ne sont pas « attachés » à leur zone de résidence et peuvent se déplacer. Pour cette raison, nous allons faire un lissage avec un rayon de 10 km, correspondant à l'accessibilité des magasins. Ainsi pour schématiser, un magasin situé à plus de 10 km ne sera pas accessible et 1 magasin situé à 5 km comptera pour une moitié de magasin (à cause de la fonction d'interaction spatiale). Nous utilisons la Moyenne Pondérée Lissée avec la fonction d'interaction spatiale gaussienne avec la surface comme variable de pondération. Nous obtenons le lissage suivant :

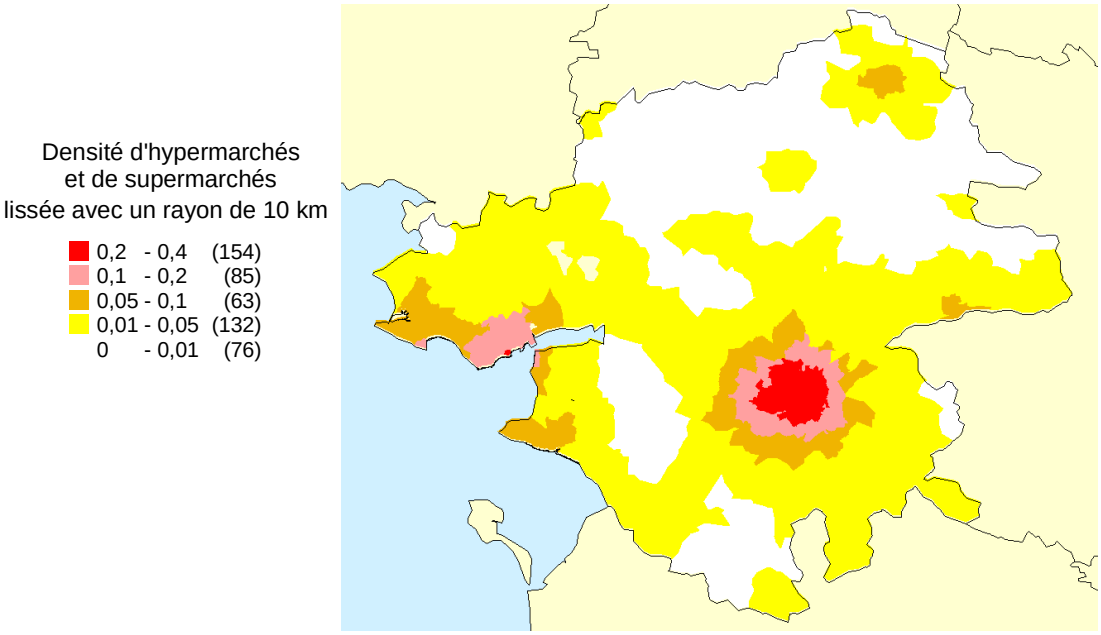


Fig. 36 –Carte du nombre d'hypermarchés et de supermarchés par km² lissé par la Moyenne Pondérée

Cette carte lissée est beaucoup plus lisible et on constate qu'avec une accessibilité de 10 km, les centres urbains (Nantes et St-Nazaire) sont fortement pourvus en supermarchés et

hypermarchés tandis que le reste est pratiquement désert. On constate aussi que cette carte est très semblable à la carte de la densité de population, ce qui est logique car la population est aussi la clientèle.

Le lissage spatial est ainsi une méthode simple et très puissante pour prendre en compte la dimension spatiale. De plus, la cartographie directe des valeurs lissées permet déjà d'obtenir des résultats intéressant à analyser.

4 Sectorisation

Nous étudierons en premier les techniques existantes permettant de réaliser une sectorisation équilibrée. Comme cette problématique est équivalente à la problématique du partitionnement de graphe (sujet largement exploré en théorie des graphes), nous nous focaliserons donc sur les méthodes les plus intéressantes, c'est-à-dire les méthodes récursives et multi-échelles car elles offrent de très bons résultats.

Nous verrons ensuite les méthodes de rééquilibrage qui permettent d'optimiser une sectorisation existante. En effet, il est parfois plus intéressant d'améliorer une sectorisation existante que d'en calculer une nouvelle. Nous verrons en détails les techniques de rééquilibrage qui se décomposent en deux étapes : le calcul de tous les transferts à effectuer entre les secteurs, puis la mise en œuvre des transferts.

4.1 Sectorisation équilibrée

La sectorisation équilibrée consiste en la création de secteurs de « taille » égale. La taille est dans ce cas une quantité à partager : surface, population, clients, … De plus, la localisation des secteurs n'est pas définie à l'avance. La sectorisation équilibrée est toutefois un problème largement étudié en théorie des graphes sous le nom de partitionnement de graphe (*Graph Partitioning*).

Il existe de nombreuses méthodes de partitionnement de graphe pour réaliser une sectorisation équilibrée [*SKK03*]. Nous ne les présenterons pas ici, car nous allons seulement présenter l'une des méthodes les plus efficaces, ainsi que ses évolutions. Il s'agit de la bipartition récursive. Cette méthode découpe itérativement le territoire en morceaux de plus en plus petits. Par exemple, pour une partition en 5 secteurs d'un territoire de taille 100, l'algorithme procède d'abord à un partage entre un grand secteur A de taille 40 et un autre secteur B de taille 60. Puis le secteur A est découpé à nouveau en deux secteurs A1 et A2 de tailles 20. On procède de même pour le secteur B. Le schéma suivant montre les ce partitionnement récursive.

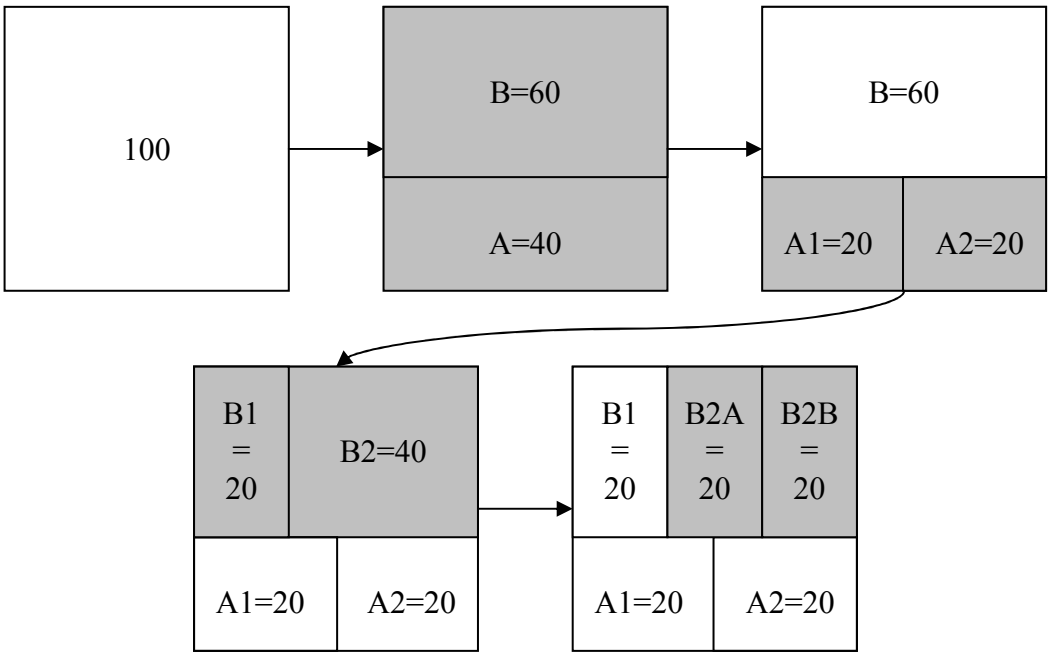


Fig. 37 –Exemple de bissection (bi-partitionnement) récursive d'un territoire en 5 secteurs égaux.

Avant d'aller plus loin, nous allons définir les notions suivantes :

- Un *nœud* du graphe représente un objet qui sera regroupé avec d'autres dans une partition (secteur)
- Une *arête* du graphe lie deux nœuds voisins. Si deux noeuds ne sont pas voisins, il n'y a pas d'arête entre eux. La sectorisation coupe les arêtes du graphe.
- Le *poids d'un nœud* représente la quantité contenue par le nœud. La quantité totale d'une partition est la somme des poids des nœuds appartenant à cette partition. Une bonne sectorisation (partitionnement) équilibrée doit faire que la quantité soit la même dans chaque secteur (ou au moins s'en approche le plus possible).
- Le *poids d'une arête* représente le degré de voisinage entre deux nœuds. Un poids de 0 équivaut à l'absence d'une arête, c'est-à-dire à l'absence de lien de voisinage. Au contraire, un poids très élevé correspond à une forte proximité entre les nœuds qui sont alors fortement liés. Une bonne sectorisation fait que les nœuds fortement liés restent ensemble. C'est toutefois un critère beaucoup moins important que l'équilibre des quantités entre les secteurs.

Les méthodes de bissection (bi-partitionnement) récursive diffèrent dans la méthode de partitionnement d'un secteur en deux. Les méthodes de bissection les plus utilisées sont les suivantes :

- La *bissection spectrale* (*Spectral Bisection*) [*PSL90*] calcule les vecteurs propres (*eigenvectors*) afin d'ordonner les nœuds (les objets géographiques dans notre cas). Il suffit ensuite de prendre les nœuds les uns après les autres dans l'ordre jusqu'à ce que la somme des pondérations des nœuds atteigne la quantité désirée. Ces premiers

nœuds forment alors le premier secteur tandis que les nœuds restant forment l'autre secteur.

- La *croissance de graphe* (*Graph Growing Algorithm*) [*KK99*] choisit un nœud de la bordure au hasard et ajoute les nœuds voisins dans l'ordre de leur degré de voisinage avec le nœud initial jusqu'à ce que la somme des pondérations des nœuds atteigne la valeur souhaitée. Cela forme le premier secteur tandis que les nœuds restant forme l'autre secteur.
- La *croissance de graphe améliorée* (*Greedy Graph Growing Algorithm*) [*KK99*] choisit un nœud de la bordure au hasard et ajoute le nœud voisin ayant le meilleur gain selon la formule de Fiduccia-Mattheyses. Le voisinage est mis à jour et les gains sont aussi mis à jour. On réitère le processus jusqu'à ce que la somme des pondérations des nœuds ajoutés atteigne la valeur souhaitée. Cela forme le premier secteur tandis que les nœuds restant forme l'autre secteur. Le gain d'un nœud est la différence entre le nombre de voisins déjà ajoutés et le nombre de voisins non ajoutés.

On remarquera que la *croissance de graphe* et la *croissance de graphe adaptée* conduisent systématiquement à des secteurs sans trous, c'est-à-dire qu'il n'existe pas de nœuds isolés. De plus, la *croissance de graphe adaptée* conduit naturellement à ce que les deux secteurs adoptent une forme compacte.

La technique de la bissection récursive est cependant inapplicable pour des « grands graphes ». La solution la plus utilisée consiste en l'ajout d'une étape de « compactage » du graphe avant le partitionnement et d'une étape de « décompactage » après le partitionnement. Le compactage consiste à agréger ensemble plusieurs nœuds et le décompactage est l'opération inverse. Le partitionnement est ainsi effectué sur un « petit graphe ». Le schéma suivant illustre ce mécanisme.

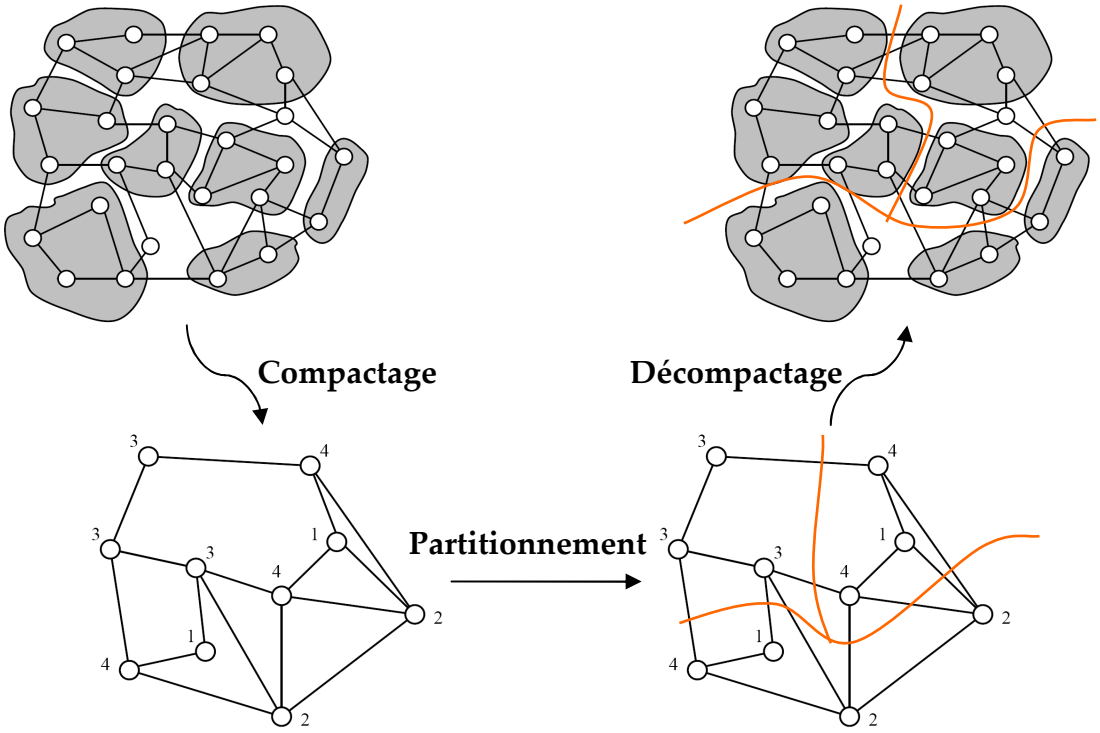


Fig. 38 –Exemple de partitionnement d'un graphe en trois secteurs égaux à partir de son graphe compacté.

L'étape de compactage peut se faire selon plusieurs méthodes, dont les plus utilisées sont les suivantes :

- Le *groupement aléatoire* (*Random Matching*) [*HL95*] choisit aléatoirement un nœud et le regroupe avec un de ses voisins qui est aussi choisi au hasard. Les nœuds déjà groupés ne sont plus utilisables pour un regroupement ultérieur. Il est donc possible que certains nœuds ne puissent être regroupés car tous leur voisins sont déjà « pris » dans un autre groupe. Pour réduire, le graphe jusqu'à la taille désiré, il est nécessaire de réitérer cette méthode plusieurs fois car la taille du graphe est divisée à peu près par deux à chaque itération.
- Le *groupement des arêtes les plus lourdes* (*Heavy Edge Matching*) [*KK99*] suit le même principe que l'algorithme précédent mais lorsqu'un nœud est choisi, il ne choisit pas le nœud voisin au hasard : il prend celui dont le poids de l'arête commune entre les deux nœuds est le plus grand. Le but est ici de garder grouper les nœuds fortement liés.

Une évolution de ce schéma est la bissection récursive multi-niveaux (*Multilevel Recursive Bisection*) qui gère la « propagation » de la bissection jusqu'au niveau non compacté en l'améliorant à chaque étape du décompactage. Le schéma suivant illustre ce mécanisme.

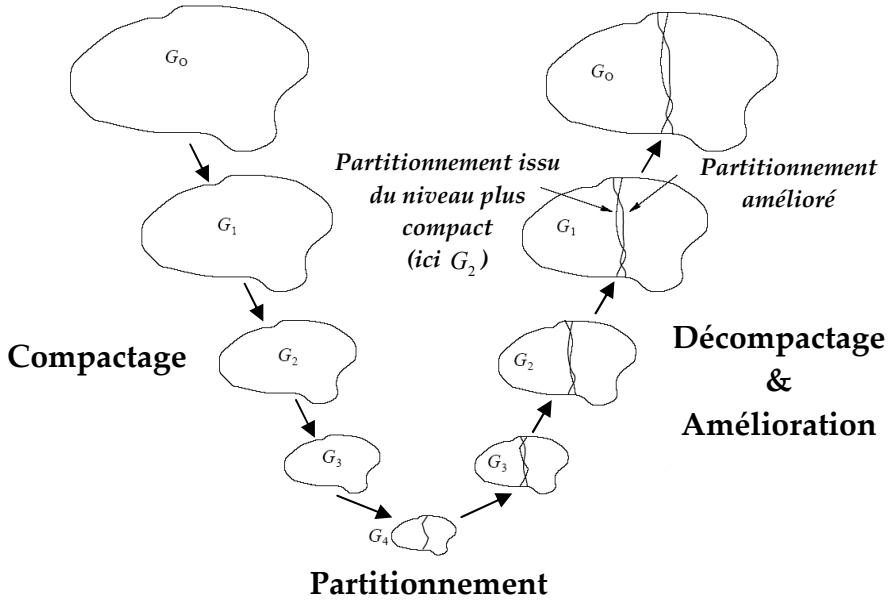


Fig. 39 –Illustration d'une bissection multi-niveaux (5 niveaux ici) avec propagation et amélioration lors du décompactage.

L'une des méthodes d'amélioration les plus utilisées est l'amélioration de la forme (compacité) selon le gain de Fiduccia-Mattheyses. Il s'agit de transférer des objets entre les deux secteurs afin d'améliorer la *compacité*. Cette méthode calcule pour chaque objet situé à la séparation entre les deux secteurs son *gain de transfert*. Ainsi, on transfère l'objet ayant le meilleur gain. On met à jour les gains et on réitère le processus jusqu'à ce que les gains soient nuls.

Une variante extrêmement rapide est le partitionnement multi-niveaux direct en k partitions au niveau le plus compacté (*Multilevel k-way Partitioning*) [*KK98*].

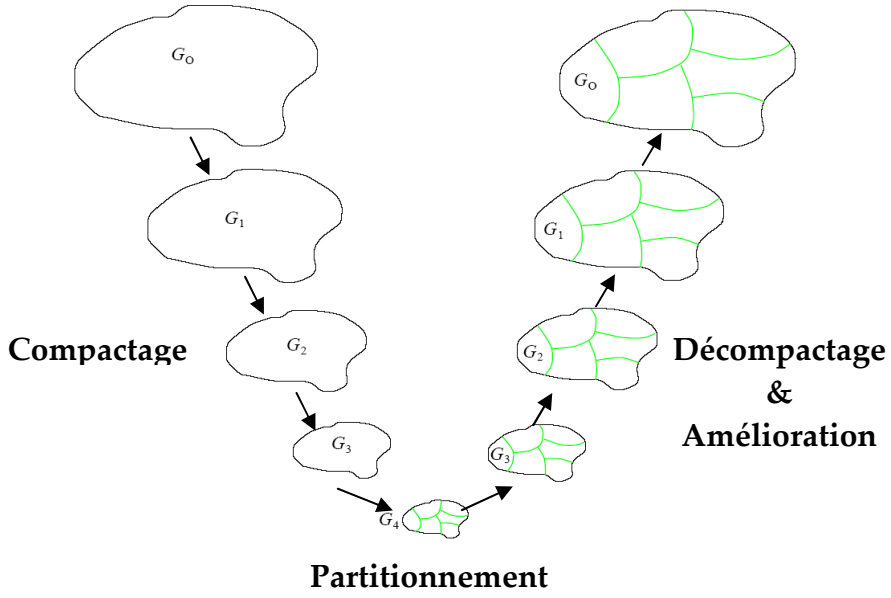


Fig. 40 –Illustration du partitionnement multi-niveaux direct en k partitions au niveau le plus compacté (5 niveaux ici) avec propagation et amélioration lors du décompactage.

C'est ce dernier algorithme que nous réutiliserons pour réaliser la sectorisation équilibrée de données géographiques.

4.2 Rééquilibrage des secteurs

Lorsque nous disposons d'une sectorisation existante, il est parfois plus intéressant d'améliorer cette sectorisation plutôt que d'en calculer une nouvelle. Le rééquilibrage des secteurs vise donc à « améliorer » les quantités de chaque secteur en transférant des objets géographiques entre les secteurs. Ceci dans le but que la quantité de chaque secteur se rapproche de la quantité idéale.

Ce problème peut être modélisé de différentes façons [*SKK97*]. On peut classer les méthodes entre les méthodes de diffusion non dirigées (locales) et celles dirigées (globales). Les méthodes non dirigées réalisent directement les transferts entre secteurs voisins et tente ainsi itérativement d'atteindre l'équilibre. Au contraire, les méthodes dirigées calculent d'abord les transferts à réaliser avant de les mettre en œuvre. Cette dernière méthode est la plus intéressante car elle permet de minimiser les quantités transférées et évite donc de trop modifier la forme des secteurs. C'est ce dernier type de méthode que nous allons étudier.

4.2.1 Calcul des quantités à transférer

Ce problème peut être modélisé de différentes façons : les calculs les plus utilisées sont d'une part, le calcul de la solution minimisant la somme des quantités transférées [*OR94*] et d'autre part, le calcul de la solution minimisant la somme des carrés des quantités transférées [*HB95*]. Cette dernière méthode a notre préférence car elle est plus simple à mettre en œuvre. En

effet, il s'agit de résoudre un système d'équations linéaires dont la solution unique est facilement trouvée par un solveur d'équations. De plus, les transferts à effectuer sont de plus faible quantité (mais plus nombreux).

Nous définissons dans un premier temps les contraintes liées aux transferts. D'abord, nous établissons pour chaque secteur, les secteurs immédiatement voisins, car pour des raisons de contiguïté, un transfert ne peut être effectué qu'entre secteurs voisins. Nous notons T_{ij} la quantité à transférer entre le secteur S_i et le secteur S_j. Les contraintes sur les transferts sont les suivantes :

- La somme des transferts arrivant vers un secteur S_i doit permettre d'atteindre la quantité désirée pour ce secteur :

$$\sum_{S_j \in Voisinage(S_i)} T_{ji} = D_i$$

  Avec D_i la différence entre la quantité actuelle et la quantité souhaitée du Secteur S_i

- Les transferts entre deux secteurs sont opposés :

$$T_{ij} = -T_{ji}$$

Nous allons maintenant voir la méthode la plus intéressante qui permet de minimiser la somme des carrés des quantités transférées, c'est-à-dire qui vérifie que $\sum (T_{ji})^2$ est minimum. La modélisation est la suivante. Chaque transfert T_{ij} est considéré comme la différence entre deux variables x_i et x_j dépendant respectivement du secteur S_i et du secteur S_j, et on pose $T_{ij} = x_i - x_j$. Ainsi, le système de contraintes devient le système linéaire à *n* inconnues x_i et à *n* équations :

$$\sum_{S_j \in Voisinage(S_i)} x_i - x_j = D_i \Leftrightarrow n \times x_i - \sum_{S_j \in Voisinage(S_i)} x_j = D_i$$

Ces équations sont liées car la somme des *n* équations est nulle car $\sum_{i=1}^{n} D_j = 0$. Nous supprimons donc la première équation et nous fixons $x_1 = 0$ afin d'obtenir un système linéaires de *n-1* équations indépendantes avec *n-1* inconnues. Ce système linéaire admet une solution unique que nous trouvons informatiquement à l'aide d'un solveur d'équations. Nous déduisons alors tous les transferts à partir des valeurs des x_i.

L'exemple suivant montre le calcul des transferts entre les secteurs A, B, C et D. Les relations de voisinages sont symbolisées par des liens. De plus, l'excédent (ou le déficit) de chaque secteur est indiqué sous son nom.

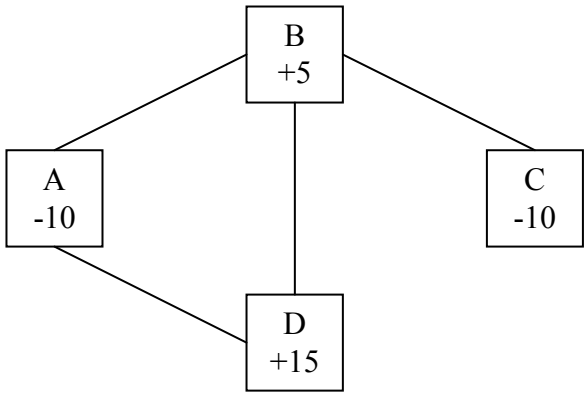


Fig. 41 –Secteurs symbolisés avec leur déficit (ou excédent)

À partir de ce graphe, nous pouvons modéliser le système à *n=4* équations et autant d'inconnues :

$$\begin{array}{lllll} \text{Pour A :} & 2\times x_A & -1\times x_B & +0\times x_C & -1\times x_D & =-10 \\ \text{Pour B :} & -1\times x_A & +3\times x_B & -1\times x_C & -1\times x_D & =+5 \\ \text{Pour C :} & 0\times x_A & -1\times x_B & +1\times x_C & +0\times x_D & =-10 \\ \text{Pour D :} & 1\times x_A & -1\times x_B & +0\times x_C & +2\times x_D & =+15 \end{array}$$

La première équation vient du Secteur A qui a un déficit de 10 et a pour voisins deux autres secteurs ($2\times x_A$) qui sont le secteur B ($-1\times x_B$) et le secteur D ($-1\times x_D$).

En éliminant la première ligne et en posant x_A=0, le système à résoudre devient :

$$\begin{array}{llll} +3\times x_B & -1\times x_C & -1\times x_D & =+5 \\ -1\times x_B & +1\times x_C & +0\times x_D & =-10 \\ -1\times x_B & +0\times x_C & +2\times x_D & =+15 \end{array}$$

La résolution informatique de ce système avec un solveur donne : $x_B = 1{,}66$, $x_C = -8{,}33$ et $x_D = 8{,}33$. Les transferts se déduisent en faisant la différence entre les valeurs. En les reportant dans le schéma et en ne gardant que les flux positifs, cela donne :

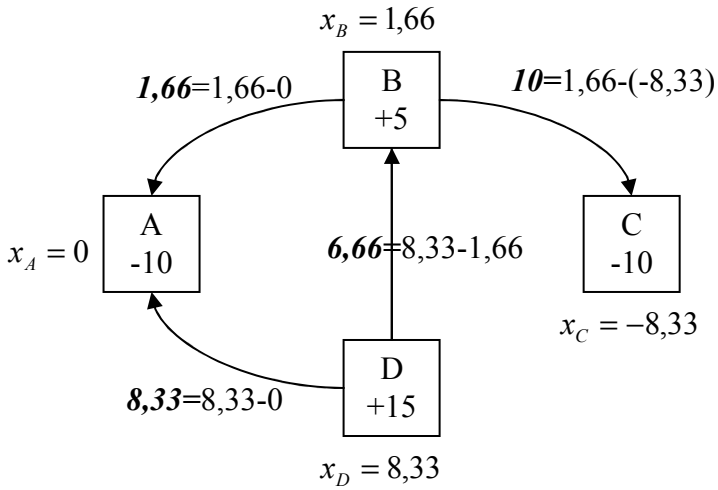


Fig. 42 –Secteurs symbolisés avec leur déficit (ou excédent) et les transferts à effectuer afin d'atteindre l'équilibre

Tous les objectifs sont atteints, par exemple, pour le secteur B, il reçoit $T_{DB} = 6{,}66$ et donne 11,66 ($T_{BA} = 1{,}66$ et $T_{BD} = 10$) et a donc perdu 5 au total (6,66 – 11,66) ce qui était l'objectif. On observe que certains secteurs tel le secteur B, sont des secteurs d'échange car il reçoive et donne beaucoup.

4.2.2 Mise en œuvre des transferts

Une fois les quantités à transférer établies, il faut alors mettre ces transferts en œuvre. Nous allons voir les méthodes existantes.

Dans la littérature, les méthodes mettant en œuvre le rééquilibrage sont nombreuses [*AK95*]. Ces méthodes sont pour la plupart itératives, transférant les objets géographiques un par un entre les secteurs. Ces méthodes peuvent donc fonctionner indifféremment dans le cadre d'une diffusion dirigée ou non dirigées. Dans le cadre, d'une diffusion dirigée, il s'agit de réaliser les transferts calculés tandis que dans le cadre d'une diffusion non dirigée, il s'agit d'atteindre l'équilibre pour les secteurs. Bien évidemment, les résultats sont meilleurs et trouvés plus rapidement dans le cadre de la diffusion dirigée. Les méthodes diffèrent par la façon dont elles choisissent l'objet qui doit migrer entre deux secteurs. Les plus utilisées pour mettre en œuvre la diffusion sont les suivantes :

- L'*algorithme Fiduccia-Mattheyses* [*FM82*] trie les objets situés à la frontière en fonction de leur « gain ». L'objet ayant le meilleur gain migre. L'intérêt de cette méthode est qu'elle permet au secteur de garder une forme compacte. Cependant, son principal défaut est que si l'objet ayant le meilleur gain ne peut pas migrer (si sa quantité est trop grande) alors, on est obligé d'arrêter l'algorithme. En effet, si l'on décide de migrer un autre objet, on risque alors de créer un trou dans le secteur d'arrivé, ce trou étant formé justement par l'objet que l'on avait pas pu faire migrer.
- L'*algorithme de Krishnamurthy* [*Kris84*] reprend le précédent algorithme en essayant de déterminer plus efficacement les objets à migrer. Pour cela, il calcule un gain sur plusieurs coups à l'avance, c'est-à-dire que le gain est calculé pour une série d'objets à faire migrer. La taille de la série permet de faire varier « l'intelligence » de l'algorithme.

Pour une longueur de 1, il est identique à l'algorithme de Fiduccia-Mattheyses. Cette stratégie permet de limiter l'arrêt prématuré.

On peut noter qu'il est possible d'utiliser la stratégie multi-niveaux en complément des algorithmes précédents afin de rééquilibrer rapidement des sectorisations comprenant un très grand nombre d'objets. C'est ce que réalise notamment JOSTLE [*WC00*].

Nous verrons ultérieurement la mise en œuvre de ces algorithmes en répondant aux questions suivantes : Quel ordre pour les transferts ? Faut-il réaliser les transferts en totalité ou pas à pas ? Comment rééquilibrer les secteurs ayant un centre ?

Conclusions

Nous avons vu l'état de l'art pour chacun des sujets suivants : la *Classification de Données*, la *Visualisation de Classifications*, le *Lissage Spatial* et la *Sectorisation*.

En ce qui concerne la *Classification de Données*, nous avons constaté qu'il existe un grand nombre d'algorithmes de classification efficaces, dont certains incrémentaux. Cependant, ils ne sont pas forcément simples à paramétrer. D'autres part, les données mixtes posent des problèmes quant à la définition de la distance à utiliser. Pour ces raisons, nous verrons dans le chapitre suivant l'algorithme que nous avons développé pour la classification de grands volumes de données mixtes.

La partie consacrée à la *Visualisation de Classifications,* nous a montré les différentes visualisations existantes sous forme de résumés et de tableau. Cependant, les graphiques de résumés sont souvent trop chargés en « chiffres ». D'autre part, l'affichage en tableau ne permet pas l'affichage de la hiérarchie des classes (ce qui est important pour l'analyse et l'exploration dans le cadre d'une classification hiérarchique). Pour ces raisons, nous verrons dans le chapitre suivant les solutions que nous avons proposées : la hiérarchie évoluée et le tableau évolué. La hiérarchie évoluée se base sur des profils de classes qui permettent d'afficher sous forme purement graphique (sans aucun texte) les principales informations relatives aux variables d'une classe. De plus, ces visualisations intègrent l'optimisation de l'ordre des classes et des variables, ce qui les rend encore plus faciles à exploiter.

Nous avons ensuite observé que le *Lissage Spatial* est un outil puissant pour mettre en lumière les tendances se produisant à grande échelle spatiale. Toutefois, le résultat fourni n'est pas synthétique et ne permet donc pas de comparer simplement les différentes cartes de lissage obtenues pour plusieurs échelles. Ainsi, dans le chapitre suivant nous détaillerons les méthodes de détermination des pôles et de leur hiérarchisation qui permettent justement de créer une synthèse des résultats de lissage spatial sous forme de hiérarchie.

Dans la dernière partie, nous avons abordé des problématiques liées à la *Sectorisation*. Nous avons vu les méthodes de partition des graphes utilisables pour réaliser une *Sectorisation Équilibrée*. Cependant, les résultats produits sont de qualité variable d'un essai à l'autre : présence ou non de secteurs non contigus, équilibre plus ou moins bien atteint. D'autre part, le rééquilibrage d'une sectorisation permet d'optimiser une sectorisation existante, mais la mise en œuvre des transferts est problématique : Quel ordre pour les transferts ? Faut-il réaliser les transferts en totalité les uns après les autres ou progressivement en parallèle ? Ainsi, dans le chapitre suivant, nous proposons une amélioration de la sectorisation équilibrée en utilisant un algorithme itératif générant plusieurs solutions de sectorisation et mesurant la qualité de celles-ci (à partir d'un indicateur de qualité que nous avons mis au point). Nous verrons aussi notre proposition d'un algorithme de *Sectorisation à partir de Centres* qui répond à une problématique courante en sectorisation mais qui n'est pas traitée par la *Sectorisation Équilibrée*. De plus, nous avons aussi proposé un algorithme itératif et progressif mettant en œuvre les transferts de façon satisfaisante dans la plupart des cas.

Ce chapitre nous a donc permis de définir les techniques existantes relativement à ces quatre sujets : la *Classification de Données*, la *Visualisation de Classifications*, le *Lissage Spatial* et la *Sectorisation.* Nous allons donc voir par la suite les améliorations et contributions que nous avons proposées.

Chapitre 2

2. CONTRIBUTIONS

Introduction

Dans le chapitre précédent, nous avons vu l'état de l'art qui concerne les quatre sujets suivants : la *Classification de Données*, la *Visualisation de Classifications*, le *Lissage Spatial* et la *Sectorisation.* Dans ce chapitre, nous allons voir les améliorations et contributions que nous avons réalisées. Cela concerne les points suivants : la *Classification de Données pour de grands volumes de données mixtes*, la *Visualisation de Classifications*, la *Détermination de Pôles et leur Hiérarchisation*, la *Sectorisation*.

La première partie concerne la C*lassification de Données pour de grands volumes de données mixtes*, c'est-à-dire comportant des variables (attributs) de nature quantitative (numérique) et aussi de nature qualitative (modale). Nous expliquerons tout d'abord la mesure de dissimilarité que nous avons retenue et qui fonctionne avec les données mixtes. Puis nous aborderons l'algorithme de classification proprement dit que nous avons développé : la Classification Ascendante Approximative (CAA) en montrant qu'il répond bien aux critères d'un algorithme de classification incrémentale. Nous verrons aussi son extension « naturelle » que nous avons appelée Classification Ascendante Hiérarchique Approximative (CAHA) et qui permet de réaliser une classification hiérarchique. Nous aborderons aussi la version parallélisée de ces algorithmes.

La deuxième partie est consacrée à la *Visualisation de Classifications* et traite trois aspects de ce problème : la visualisation sous forme de *Hiérarchie Évoluée des profils de classes*, la visualisation sous forme de *Tableau Évolué des profils de classes*, l'*optimisation de l'ordre des variables et des classes*. Nous verrons en premier la *Hiérarchie Évoluée des profils de classes*. Notre approche innovante combine un graphique hiérarchique habituel avec des profils de classes très détaillés. Ces profils ne contiennent aucun chiffre et l'analyse peut alors être effectuée « graphiquement » sans avoir recours aux informations chiffrées. Cela rend la lecture très simple et très rapide avec un peu d'entraînement. Nous verrons ensuite le *Tableau Évolué des profils de classes*. Il combine informations symboliques et informations chiffrées afin d'offrir deux niveaux de lectures simultanés : le premier donnant les informations les plus importantes pour une vue globale du tableau et l'autre donnant plus de détails pour une lecture approfondie. Nous verrons enfin

l'*optimisation de l'ordre des variables et des classes*. Nous montrerons l'approche originale que nous avons utilisée pour l'optimisation de l'ordre des variables à partir de la matrice de corrélation des variables.

La troisième partie présente le travail réalisé pour la *Détermination et la Hiérarchisation de Pôles*. Ce travail s'appuie sur les techniques de lissage spatial détaillées dans le chapitre précédent concernant l'état de l'art. Nous verrons d'abord comment il est possible de résumer une carte de lissage spatial en extrayant ses pôles. La technique que nous proposons permet de décrire très rapidement et brièvement une carte de lissage spatial. Nous montrerons ensuite l'intérêt de superposer sur une même carte les pôles correspondant à des échelles différentes en les incluant dans une hiérarchie. Nous obtenons alors une carte très synthétique qui a aussi une valeur explicative importante : les pôles de l'échelle spatiale la plus courte permettent d'expliquer les pôles des échelles spatiales plus grandes.

La quatrième partie traite des problèmes relatifs à la *Sectorisation*, il s'agit de la *Sectorisation Équilibrée*, de la *Sectorisation à partir de Centres* et du *Rééquilibrage de Sectorisations.* Nous allons tout d'abord voir la transformation des données géographiques en un graphe et l'évaluation de la qualité d'une sectorisation. À partir de ces deux prérequis, nous définirons tout d'abord l'algorithme itératif permettant de réaliser une *Sectorisation Équilibrée*. Il s'appuie sur un algorithme éprouvé de partitionnement de graphe vu dans le chapitre consacré à l'état de l'art. Ensuite nous traiterons de la *Sectorisation à partir de Centres.* Dans une première approche du problème, nous présentons un algorithme basique de *Sectorisation à partir de Centres* que nous faisons ensuite évolué vers un algorithme itératif donnant des résultats plus satisfaisants. Nous finirons avec le *Rééquilibrage de Sectorisations* en nous intéressant à la partie consacrée à la mise en œuvre des transferts. Il s'agit de l'étape qui intervient après que les quantités à transférer aient été déterminées (cela a été vu dans le chapitre sur l'état de l'art). Nous proposons un algorithme itératif qui effectue les transferts « en parallèle » afin d'aller progressivement vers une solution de rééquilibrage.

1 Classification de Données pour de grands volumes de données mixtes

Nous avons défini en premier une mesure de dissimilarité fonctionnant avec les données mixtes. L'utilisation de cette mesure nécessite quelques prétraitements sur les données : une disjonction suivie d'une normalisation. L'intérêt de ces opérations est aussi de rendre homogène les variables afin de faciliter leur représentation ultérieurement dans la *Visualisation de Classifications*.

Dans la partie suivante, nous exposerons la méthode innovante de classification que nous avons mise au point : la Classification Ascendante Approximative (CAA) et son extension la Classification Ascendante Hiérarchique Approximative (CAHA). Il s'agit d'une méthode de classification incrémentale simple à mettre en œuvre et à paramétrer. Elle est très rapide car sa complexité est linéaire. Nous expliquerons d'abord la démarche de création de notre algorithme incrémental et sa parenté avec l'algorithme de la Classification Ascendante Hiérarchique (CAH). Puis nous présenterons ensuite la version parallélisée de cet algorithme et enfin nous verrons « l'effet de dérive » qui gêne la classification et nous verrons comment y remédier.

1.1 Dissimilarité utilisée

La dissimilarité que nous utilisons pour quantifier la dissemblance entre les individus s'applique sur des données mixtes quelconques. Notre approche consiste à transformer les données qualitatives en données quantitatives et à appliquer ensuite une distance fonctionnant uniquement avec les données quantitatives.

Pour cela, nous appliquons successivement deux prétraitements sur les données : la disjonction et la normalisation. Nous obtenons à la suite de ces prétraitements des paramètres que nous intégrons dans la formule de la distance.

1.1.1 Prétraitements des données

Les deux prétraitements s'appliquent de manière successive. Il s'agit d'abord de réaliser une disjonction des variables qualitatives pour les transformer en variables binaires (que l'on peut donc considérer comme quantitative). Puis nous normalisons les données en « centrant-réduisant » chaque variable.

L'opération de disjonction des variables qualitatives permet de transformer une variable qualitative en plusieurs variables quantitatives. Cette opération crée autant de variables quantitatives qu'il y a de modalités. L'intérêt de cette méthode est de n'avoir à gérer par la suite que des données quantitatives. Cependant, afin que les variables ayant beaucoup de modalités ne soient pas avantagées par rapport aux autres variables, nous pondérons chacune des $V_{k1},\ldots,V_{km}$ variables issues d'une même variable V_k qualitative par $1/m$ pour que le poids global de ces variables fasse 1. Cette pondération est alors automatiquement reprise dans le paramétrage de la distance.

La normalisation des données consiste simplement à centrer-réduire les données (dont les variables créées à partir des variables qualitatives). Pour cela, nous devons calculer la moyenne M_k et l'écart-type σ_k de chaque variable V_k. Cette opération est réalisée en une seule lecture des données. La moyenne et l'écart-type sont ensuite indiqués en tant que paramètres de la distance afin de pouvoir calculer la distance entre les individus à partir de leurs variables normalisées.

1.1.2 Définition de la distance utilisée et de ses paramètres

Comme nous avons transformé par disjonction les variables qualitatives en variables quantitatives, nous pouvons utilisé une distance fonctionnant exclusivement avec des variables quantitatives. De plus, nous devons intégrer pour chaque variable un paramètre de pondération issu de la disjonction que nous appellerons le poids de « disjonction ». Toutefois, si la variable n'est pas issue d'une disjonction, ce poids vaut 1.

Nous utilisons donc la distance euclidienne pondérée :

$$\forall o_i, o_j \; d(o_i, o_j) = \sqrt{\sum_{k=1}^{m} \left(PD_k \times \left(CR_k(o_i) - CR_k(o_j) \right) \right)^2}$$

Avec m variables quantitatives V_k

Et $CR_k(o_i)$, la valeur centrée-reduite de la variable V_k pour l'objet o_i

Et PD_k, le poids de « disjonction » de la variable V_k

Pour rappel, la valeur centrée-réduite est $\forall o_i \;\; CR_k(o_i) = \dfrac{V_k(o_i) - M_k}{\sigma_k}$

Avec $V_k(o_i)$, la valeur de la variable V_k pour l'objet o_i

Et M_k la moyenne de la variable V_k

Et σ_k l'écart-type de la variable V_k

Par ailleurs, nous souhaitons que l'utilisateur puisse, s'il le souhaite, spécifier pour chaque variable V_k un poids P_k afin de moduler l'influence de cette variable dans la distance :

- Par défaut, P_k =1, c'est le poids standard
- P_k =0, la variable n'intervient plus dans le calcul de la distance : elle est passive.
- P_k >1, la variable intervient de façon plus intensive dans le calcul de la distance. Par exemple, pour P_k =2, la variable intervient comme si elle était présente deux fois.
- Une autre valeur intéressante est $P_k = \sigma_k$ pour chaque variable. Cela à pour effet d'annuler l'action de normalisation et les variables interviennent dans la distance comme si elles n'étaient pas centrées-réduites. Cette pondération n'est intéressante que dans le cas où toutes les variables sont quantitatives à la base et qu'elles sont de même nature.

Nous reviendrons ultérieurement sur les choix de pondération effectués par l'utilisateur.

Finalement, nous utilisons la distance suivante :

$$\forall o_i, o_j \; d(o_i, o_j) = \sqrt{\sum_{k=1}^{m} \left(P_k \times PD_k \times \left(\frac{V_k(o_i) - V_k(o_j)}{\sigma_k} \right) \right)^2}$$

Avec m variables quantitatives V_k

Et P_k, le poids défini par l'utilisateur de la variable V_k

Et PD_k, le poids de « disjonction » de la variable V_k

Et σ_k l'écart-type de la variable V_k

Et $V_k(o_i)$, la valeur de la variable V_k pour l'objet o_i

On remarquera que la moyenne M_k n'intervient pas dans la distance car les termes $M_k(o_i)$ et $M_k(o_j)$ s'annulent.

1.2 Méthode de la Classification Ascendante Approximative (CAA)

Nous allons maintenant voir notre algorithme de la Classification Ascendante Approximative (CAA) qui va nous servir pour classifier des grands volumes de données mixtes. Dans un premier temps, nous expliquerons la démarche d'élaboration de notre algorithme incrémental et sa parenté avec l'algorithme de la Classification Ascendante Hiérarchique (CAH). Nous présenterons ensuite une version parallélisée de cet algorithme. Nous verrons finalement un effet gênant inhérent aux algorithmes incrémentaux : « l'effet de dérive » et nous verrons comment y remédier.

1.2.1 Principe

Pour mettre au point notre algorithme, nous sommes partis des contraintes que doit respecter un algorithme incrémental tout en essayant de conserver un algorithme simple à mettre en œuvre, avec peu de paramètres. Nous avons ainsi défini que l'algorithme ne devait pas dépasser *k* classes (ou résumés) tout en permettant la création de nouvelles classes pour les individus atypiques. Ainsi comme l'a défini M. Charikar et d'autres [*CCFM97*], le problème que doit résoudre notre algorithme incrémental à *k* classes maximum est le suivant : pour une classification en *k* classes existantes, l'arrivée d'un nouvel individu à classer se traduit soit par l'inclusion de l'individu dans une classe existante, soit par la création d'une nouvelle classe contenant cet individu et la fusion de deux classes existantes afin de maintenir le nombre de classes égal à *k*. L'obligation d'avoir à fusionner deux classes entre elles place notre algorithme potentiel dans la catégorie des algorithmes ascendants (agglomératifs), la construction d'une hiérarchie étant une conséquence secondaire. Dans la famille des algorithmes ascendants se trouve la Classification Ascendante Hiérarchique (CAH) [*JD88*] et en étudiant le fonctionnement de la CAH, nous sommes arrivés à la conclusion que l'algorithme de la CAH répond à notre problème moyennant quelques adaptations. Tout d'abord, il convient de repenser l'algorithme de la CAH en le modifiant pour qu'il devienne un algorithme incrémental. « L'ajout d'un individu » peut se définir ainsi : l'individu est incorporé dans une nouvelle classe, puis comme dans la CAH classique, les deux classes les plus semblables sont agrégées. Cette méthode à l'avantage de ne pas centrer le problème sur le nouvel individu arrivant (au contraire d'autres algorithmes tels que BIRCH [*ZRL96*]), par contre la recherche des deux classes les plus semblables peut être coûteux. Le schéma suivant illustre le problème de l'arrivée d'un nouvel individu dans une classification existante comportant 3 classes.

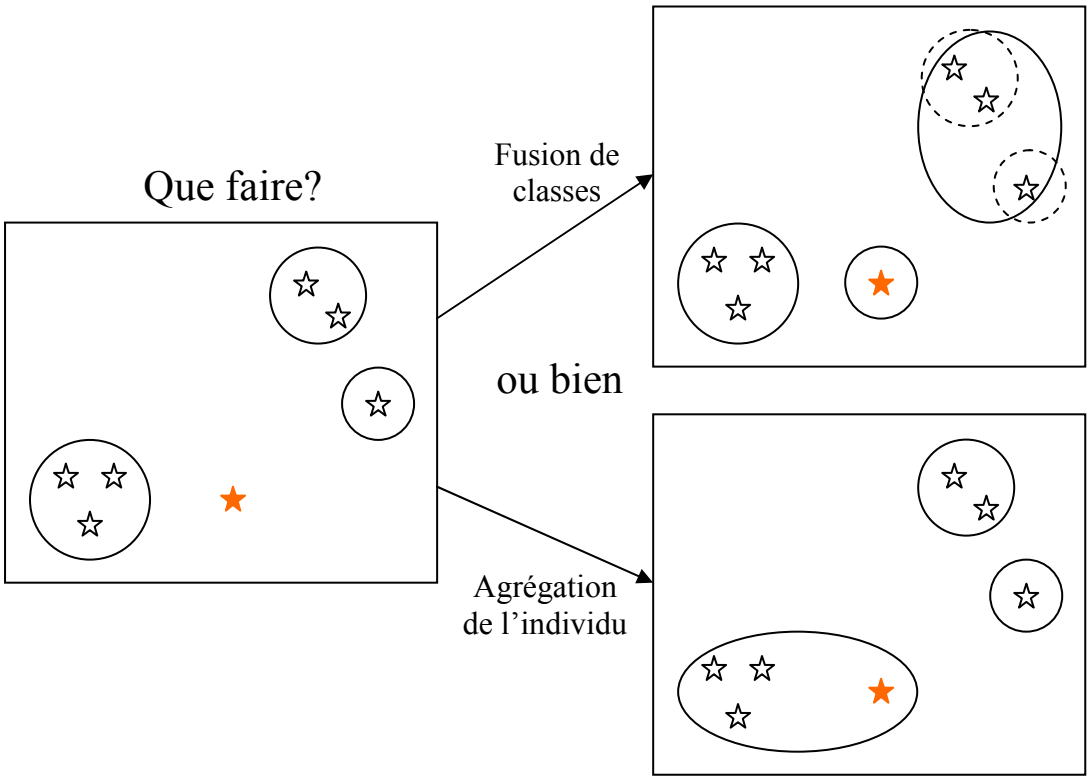


Fig. 43 – Les deux façons possibles d'intégrer un nouvel individu

Le schéma suivant illustre la méthode utilisée par la CAA pour gérer l'arrivée d'un nouvel individu dans une classification existante comportant 3 classes.

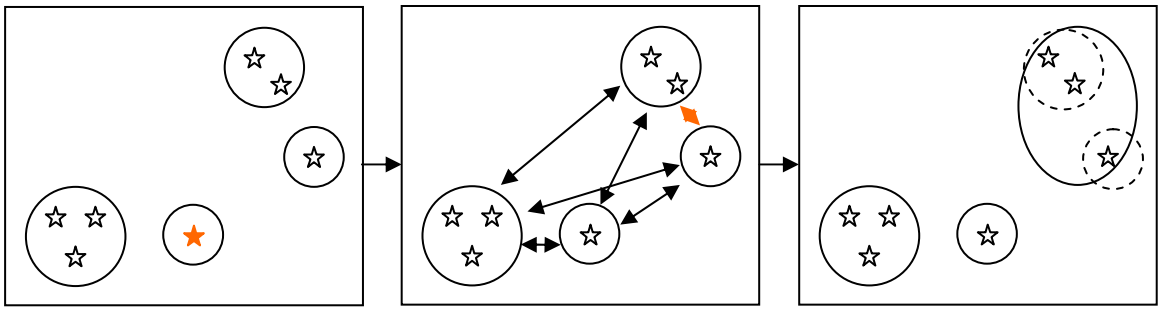


Fig. 44 – Intégration d'un nouvel individu dans la CAA

Nous définissons donc notre algorithme, la Classification Ascendante Approximative (CAA), de la manière suivante, avec pour paramètre k le nombre de résumés maximum et d la mesure de distance ou de dissimilarité choisie :

- Etape 1 : La liste des classes est vide: $LC := \varnothing$
- Etape 2 : Tant qu'il y un individu *o* à classer faire :
  - Créer une nouvelle classe *C* pour cet individu : $C := \{o\}$
  - Ajouter cette classe à la liste des classes : $LC := LC \cup \{C\}$
  - Si il y a plus de *k* classes Alors
    - Trouver les deux classes C_x et C_y les plus semblables :
      $d(C_x, C_y) = \min(d(X,Y))$ avec $X \in R$ et $Y \in R$
    - Les fusionner en une nouvelle classe : $C_n := C_x \cup C_y$
    - Ajouter la nouvelle classe et enlever les deux anciennes classes :
      $LC := LC \cup \{C_n\} - \{C_x, C_y\}$
  - Sinon ne rien faire

Fig. 45 – Algorithme de la CAA

Une variante de cet algorithme, la Classification Ascendante Hiérarchique Approximative (CAHA), rajoute une ultime étape. On continue à fusionner les classes les plus semblables jusqu'à obtenir la classe contenant tous les individus. Cependant, on mémorise la hiérarchie *H* et les nœuds *N* de cette hiérarchie. Cette dernière étape permet de construire l'arbre de classification des classes-résumés LC. Il s'agit en fait d'une Classification Ascendante Hiérarchique standard réalisée sur les résumés avec l'utilisation de la distance *d* :

- Etape 3 :
  - Initialisation des nœuds : $N := LC$
  - Initialisation de la hiérarchie : $H := \varnothing$
  - Initialisation de la liste des classes restantes : $LCr := LC$
- Etape 4 :
  - Tant que LCr contient plus d'une classe ($Card(LCr) > 1$) :
    - Trouver les deux classes C_x et C_y les plus semblables
      $d(C_x, C_y) = \min(d(X,Y))$ avec $X \in LCr$ et $Y \in LCr$
    - Les fusionner en une nouvelle classe : $C_n := C_x \cup C_y$
    - Mettre à jour la liste des nœuds *N* : $N := N \cup \{C_n\}$
    - Mettre à jour la hiérarchie *H* : $H := H \cup \{(C_n \Rightarrow C_x), (C_n \Rightarrow C_y)\}$
    - Ajouter la nouvelle classe et enlever les deux anciennes classes de *LCr*:
      $LCr := LCr \cup \{C_n\} - \{C_x, C_y\}$
  - Fin Tant Que

Fig. 46 – Extension de la CAA pour obtenir l'algorithme de la CAHA

Le choix du nombre de classes-résumés *k* permet de paramétrer la précision de l'algorithme. En prenant *k* aussi grand ou plus grand que le nombre d'individus à traiter, notre algorithme est équivalent à la CAH. En effet, durant l'étape 2, il n'y a aucune fusion et chaque classe résume un seul individu. L'étape 3 et l'étape 4 réalisent alors une CAH sur la totalité des données. Le choix de *k* offre donc une alternative entre un algorithme rapide et approximatif, d'une part, et un algorithme lent (voir très lent) et précis, d'autre part. Cette précision dépend ainsi de l'adéquation entre le nombre d'individus *n* et le nombre de résumés *k*. Le choix de *k*

peut être établi en fonction de considérations portant sur l'utilisation des résumés. Généralement, en partant de l'analyse habituelle faite d'une classification, on s'aperçoit qu'une personne ne souhaite pas avoir plus de 20 classes à analyser. Dans un souci de précision de l'algorithme, nous établissons donc le niveau de résumés à un niveau 5 fois supérieur en prenant *k=100* résumés. C'est un nombre de résumés suffisamment grand pour obtenir 20 classes acceptables à partir des résumés et suffisamment petit pour obtenir un résultat rapidement.

La complexité de l'algorithme peut être appréhendée via la matrice des distances permettant de calculer la distance minimale entre les résumés. Cette matrice symétrique a une taille $k \times k$. Lors de l'ajout d'un nouveau résumé, il est nécessaire de calculer *k* distances. Par ailleurs, la recherche de la distance minimale nécessite le parcours de toute la matrice. De fait, les étapes d'ajout et de recherche étant réalisées *n* fois durant l'algorithme, la complexité temporelle de l'algorithme est $o\left(n\left(k+k^2\right)\right)=o\left(nk+nk^2\right)=o\left(nk^2\right)$. La complexité temporelle de l'algorithme est donc bien linéaire par rapport au nombre d'individus *n*. Lorsque *n=k*, on remarque que comme attendu, la complexité temporelle est $o\left(n^3\right)$ comme dans l'algorithme de la CAH standard.

Afin d'optimiser la recherche de la distance minimale, nous avons utilisé un index (arbre binaire équilibré) sur la matrice des distances avec la distance pour clef d'indexation. Ainsi, la recherche de la distance minimale est très rapide et s'effectue en $o(\log(k))$ au lieu de $o\left(k^2\right)$. La contrepartie est le coût d'insertion des distances dans l'index : chaque insertion a aussi un coût de $o(\log(k))$ et il y en a *k* à réaliser à chaque ajout d'un individu. La complexité finale est ainsi $o(n(k\log(k)+\log(k)))=o(nk\log(k)+n\log(k))=o(nk\log(k))$. La complexité est ainsi moindre suivant le nombre de résumés *k*. Lorsque *n=k*, on remarque que la complexité temporelle est $o\left(n^2\log(n)\right)$ comme dans l'algorithme de la CAH optimisée.

1.2.2 Algorithme de la CAA Parallèle

Cet algorithme est facilement parallélisable. L'intérêt d'une version parallèle est bien entendu de diminuer le temps de calcul tout en évitant de perdre en qualité. Nous proposons un algorithme naïf en deux étapes. Soit *m* machines parallèles disponibles, les *n* individus sont partitionnés de manière quelconque en *m* partitions de taille identique. La première étape distribue de manière équitable les données entre les machines et chaque machine effectue une CAA des données reçues : sur chaque machine n/m individus sont réduits à *k* résumés. Dans la deuxième étape, les résumés sont rassemblés sur une seule machine qui les résume à nouveau : les $m \times k$ résumés sont réduits à *k* résumés. Nous verrons dans les tests les avantages et les inconvénients de la version parallèle par rapport à la version classique : la version parallèle est effectivement plus rapide mais elle souffre d'un « effet de superposition » qui dégrade les résultats de façon conséquente.

Nous allons maintenant voir la complexité de cet algorithme parallèle. Tout d'abord, la complexité temporelle finale est identique au fait d'avoir utilisé la CAA sur n/m individus pour *k* résumés dans un premier temps et $m \times k$ « individus » pour *k* résumés ensuite. La complexité temporelle de l'algorithme parallèle est donc en $o\left(\frac{n}{m}k\log(k)+\left(m\times k\times k\log(k)\right)\right)$

$= o\left(\left(\frac{n}{m} + m \times k\right) k \log(k)\right)$. Nous posons $n_s = \frac{n}{m} + m \times k$, le nombre d'individus équivalents pour l'algorithme standard. En effet, l'algorithme standard traite n_s individus en autant de temps que l'algorithme parallèle le fait pour *n* individus. Nous posons aussi $g = \frac{n}{n_s}$, le gain réalisé par l'utilisation de l'algorithme parallèle par rapport à l'algorithme standard, en négligeant les coûts de communication des données et $e = \frac{n}{m \times n_s} = \frac{g}{m}$, l'efficacité qui est le gain obtenu par machine. L'étude de la fonction de gain en fonction de *m*, montre que le gain est maximum pour $m = \sqrt{\frac{n}{k}}$ et que le gain est alors de $\frac{1}{2}\sqrt{\frac{n}{k}} = \frac{m}{2}$ et l'efficacité seulement de $\frac{1}{2}$. Par ailleurs, lorsque *m* est très petit ($m << \sqrt{\frac{n}{k}}$), nous avons $g = \frac{n}{\frac{n}{m} + m \times k} \approx \frac{n}{\frac{n}{m}} = m$ ce qui est très proche du cas idéal et $e \approx 1$, qui est l'efficacité maximale.

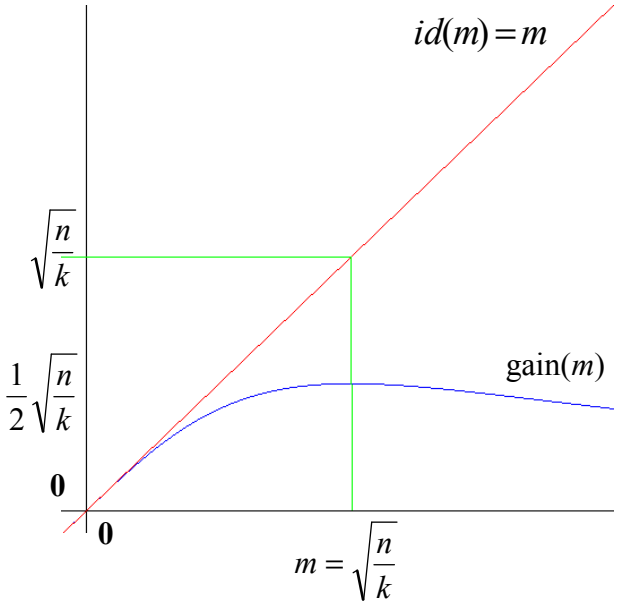


Fig. 47 –Courbe du gain en fonction du nombre de machines m

Ainsi, les cas où l'algorithme est le plus efficace sont ceux où $m << \sqrt{\frac{n}{k}} \Leftrightarrow m^2 k << n$. Ainsi pour *m=10* machines et la précision *k=100*, l'algorithme sera d'autant plus efficace que le nombre d'individus *n* est grand par rapport à $10^2 \times 100 = 10000$. Par exemple, pour une base de données contenant 1 000 000 d'individus nous permet d'avoir une complexité équivalente à $n_s = \frac{1000000}{10} + 100 \times 10$, $= 100000 + 1000 = 101000 \approx 100000$ soit un gain de temps

$g = \frac{1000000}{101000} \approx m = 10$ et l'efficacité $e \approx 1$ est presque maximale. Avec 100 machines, nous aurions le gain maximum car $\sqrt{\frac{n}{k}} = \sqrt{\frac{100000}{100}} = \sqrt{10000} = 100$, soit $g = \frac{m}{2} = 50$. Par contre l'efficacité serait seulement de $\frac{1}{2}$.

1.2.3 Définition de l'effet de dérive

Lorsque *k* le nombre de résumés utilisés est faible par rapport au nombre de classes « réelles », la CAA subit un fort effet de « dérive ». Cet effet est dû à l'ordre « mauvais » dans lequel les individus sont pris en compte. Cela se traduit par le fait que certains individus sont plus proches d'une autre classe que de la classe à laquelle ils appartiennent. Cet effet est commun à tous les algorithmes purement incrémentaux car ils ne remettent jamais en cause l'appartenance d'un individu à une classe. Pour éliminer cet effet sans utiliser de poste-traitement, il faudrait introduire les individus les plus proches entre eux en premier pour finir par les individus les plus éloignés. Malheureusement, cela nécessiterait de calculer en pré-traitement la matrice des distances entre tous les individus et ensuite de trier les individus en commençant par les plus proches entre eux pour finir par les plus éloignés. Ce qui serait finalement équivalent à réaliser une CAH standard.

Un exemple de dérive est le suivant : on utilise *k*=2 résumés et les individus arrivent dans l'ordre donné par le schéma. Les individus n'ont qu'une seule variable et sont donc représentés par une étoile sur un axe. La distance utilisée est la distance au barycentre des classes. Les classes sont les rectangles arrondis et la position du barycentre de la classe est donnée par le cercle. On commence avec un seul individu. Chaque étape combine l'ajout d'un individu et la fusion des 2 classes les plus proches (à part pour les 2 premières étapes car le nombre de classes n'a pas atteint le nombre de résumés souhaités, il n'y a donc pas de fusion).

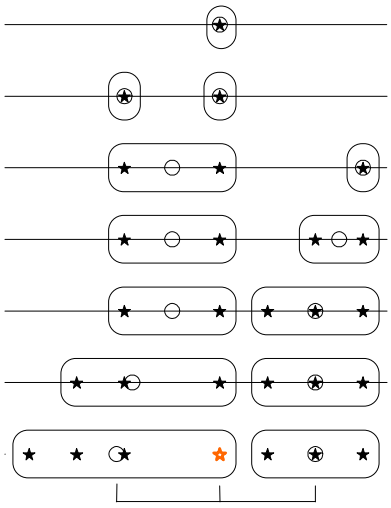


Fig. 48 –Illustration de l'effet de dérive

On voit sur le graphique que l'individu en rouge n'appartient pas à la classe dont il est le plus proche du barycentre, ce dernier ayant trop dérivé vers la gauche, tandis que le barycentre de l'autre classe s'est rapproché. Pour corriger ces effets de dérive, le mécanisme le plus simple est de rattacher chaque individu à la classe la plus proche. C'est justement ce que réalise les k-moyennes que nous utiliserons donc en post-traitement de la CAA. Cela permettra de réduire rapidement le nombre d'individus mal classés tout en améliorant globalement la qualité des résumés. Nous établirons les modalités de l'utilisation des k-moyennes dans la partie expérimentation. Cependant, dans le cadre des gros volumes de données, il n'est généralement pas possible d'utiliser les k-moyennes car chaque itération de cet algorithme nécessite une lecture complète des données.

L'autre solution comme nous le verrons est d'augmenter la valeur de k (le nombre de résumés utilisés) à un niveau satisfaisant et nous déterminerons dans la partie expérimentation les valeurs nécessaires que doit prendre k.

1.3 Conclusion

Nous avons mis au point un algorithme de classification incrémentale efficace : la Classification Ascendante Approximative (CAA). Il est très simple à mettre en œuvre et à paramétrer (un seul paramètre). Il est de plus très rapide avec une complexité linéaire. Son extension, la Classification Ascendante Hiérarchique Approximative (CAHA) permet de construire une classification hiérarchique pour les grands volumes de données à partir de leurs résumés. Ces deux méthodes, en dehors de la mesure de dissimilarité, ne nécessitent que le choix du paramètre k fixant le nombre de résumés utilisés en mémoire. C'est un paramètre permettant de faire l'équilibre entre la rapidité (valeurs faibles de k) et la précision (valeurs élevées de k). En effet, pour des valeurs de k identiques ou supérieures au nombre d'objets traités, la CAHA produit des résultats identiques à la CAH, algorithme reconnu pour sa fiabilité. Nous avons aussi défini et utilisé une mesure de dissimilarité permettant d'utiliser l'algorithme sur des données mixtes (à la fois quantitatives et qualitatives). Cette mesure se traduit par la transformation des variables qualitatives en variables quantitatives par disjonction, puis par la normalisation de toutes les variables. Le principal inconvénient actuel de notre algorithme de la CAA est le paramétrage de la distance via des pondérations qui ne sont pas forcément adaptées. C'est pourquoi nous envisageons d'étudier le remplacement de la distance euclidienne pondérée actuellement utilisée par la distance de Mahalanobis [*Maha36*].

Nous verrons dans le chapitre consacré aux expérimentations et aux applications que notre algorithme permet d'obtenir de bonnes classifications en utilisant la perte d'inertie comme mesure de qualité.

2 Visualisation de Classifications

Nous abordons dans une première partie le problème de la visualisation d'une hiérarchie évoluée comportant des profils de classes. Nous montrerons l'approche que nous avons retenue, combinant un graphique hiérarchique habituel complété de profils de classes très

détaillés. L'analyse peut alors être effectuée « graphiquement » sans avoir recours aux informations chiffrées contrairement aux approches classiques (boites à moustaches en parallèle [*Tuke77*], visualisation en coordonnées parallèle [*ID90*]).

La deuxième partie est consacrée à la visualisation d'une classification sous forme d'un tableau. Nous exposerons notre approche qui combine informations symboliques et informations chiffrées.

La troisième partie traite de la recherche de l'ordre optimal pour les classes et les variables. Nous expliquerons l'adaptation de la technique retenue pour trouver l'ordre optimal des classes. Puis nous détaillerons la méthode originale que nous avons développée pour trouver l'ordre optimal des variables à partir de la matrice de corrélation des variables.

2.1 Hiérarchie évoluée des profils de classes

Notre visualisation d'une hiérarchie de classes est un graphique représentant la hiérarchie dans laquelle sont intégrés les profils de classes. Ces derniers sont un croisement de la visualisation en coordonnées parallèles et des boites à moustaches en parallèles. Ainsi le principal point fort de cet affichage est l'intégration au sein de l'arbre de classification de la description des classes sous forme de profils très détaillés mais compréhensibles facilement et rapidement. Un exemple d'arbre est le suivant :

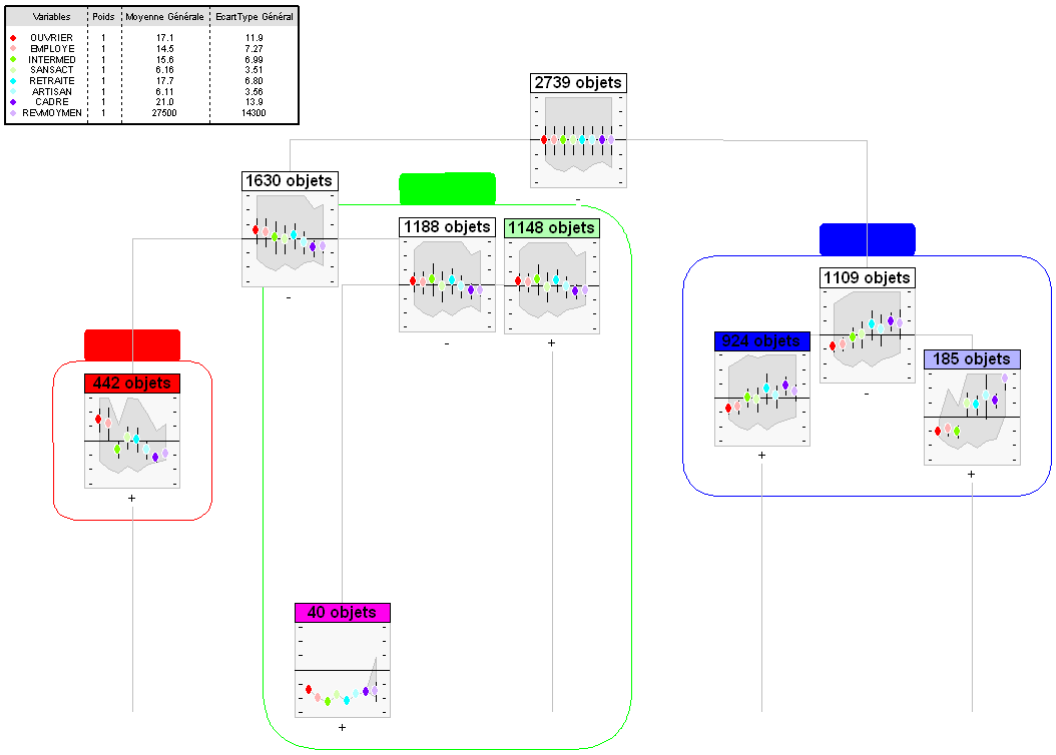


Fig. 49 –Visualisation hiérarchique évoluée

Le choix des classes à interpréter est grandement facilité par la présence des profils.

Un exemple de profil en détail est le suivant :

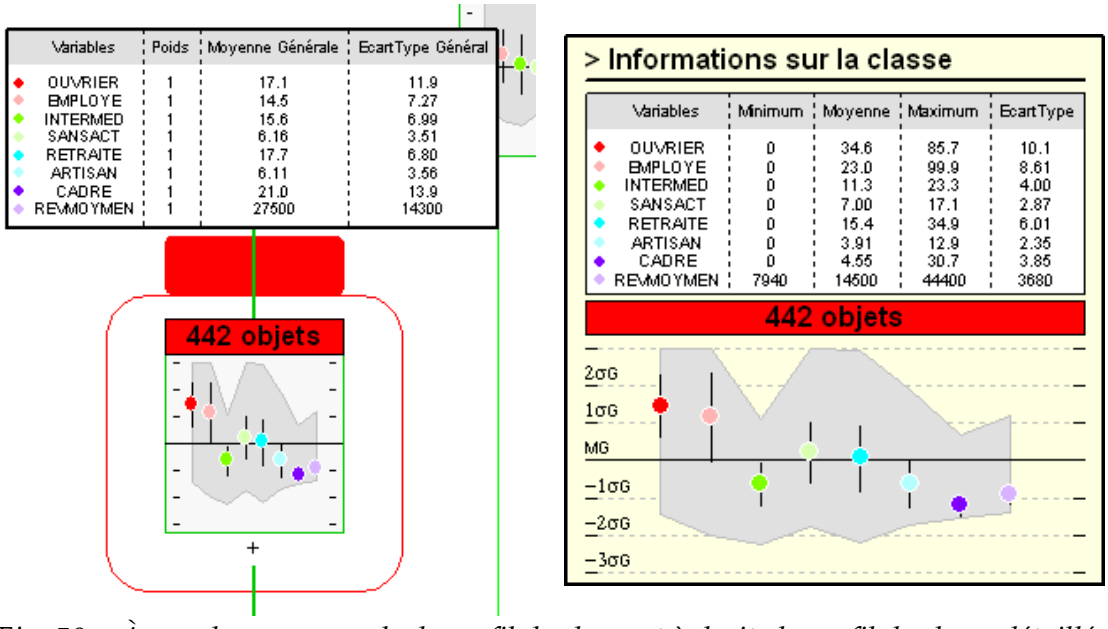


Fig. 50 – À gauche, un exemple de profil de classe et à droite le profil de classe détaillé

La ligne horizontale (MG) sur le graphique correspond à la moyenne générale de chaque variable et l'échelle de représentation (σG) donnée sur l'axe vertical est l'écart type général de chaque variable. Cela permet de créer une échelle commune à toutes les variables et de pouvoir ainsi comparer les variables entre elles. La légende permet de lire les valeurs. En effet, pour retrouver une valeur il suffit de lire la valeur en écart-type dans le profil, de la multiplier par l'écart-type général correspondant et finalement d'ajouter la moyenne générale. Par exemple, nous lisons dans la classe donnée en exemple précédemment, que le taux d'ouvriers moyen dans la classe est à environ +1,5 écart-type au dessus de la ligne moyenne. Comme il est indiqué dans la légende que l'écart-type général de cette variable est d'environ 12 (11,9 exactement) et que sa moyenne générale est d'environ 17 (17,1 exactement), on en déduit que le taux d'ouvriers moyen dans la classe est 1,5 écart-type x 12 + 17 = 35 %, car la variable a pour unité les pourcents. Le profil détaillé est là pour éviter de faire les calculs mentaux, il indique que le taux d'ouvrier moyen dans la classe est précisément de 34,6 %.

Pour chaque variable, le profil donne les principales informations sur la classe :

- la moyenne de la variable dans la classe (représentée par le point de couleur) ;
- l'écart type de la variable dans la classe (représenté par les barres verticales de part et d'autre du point de couleur) ;
- le minimum de la variable dans la classe (représenté par la délimitation inférieure de la zone grisée) ;
- le maximum de la variable dans la classe (représenté par la délimitation supérieure de la zone grisée).

Ces informations donnent en fait trois niveaux de lecture simultanément :

- le premier niveau est la valeur moyenne : on considère que les individus de la classe ont une valeur voisine de la moyenne ;

- le second niveau est l'intervalle formé par l'écart-type autour de la moyenne : on considère que les individus sont majoritairement compris dans cet intervalle. Si l'on fait l'hypothèse que la distribution de la variable est de type gaussienne, on a 70 % des valeurs compris dans cet intervalle. Quoiqu'il en soit, plus cet intervalle est petit, plus on est assuré que l'hypothèse du premier niveau est « vraie », et au contraire, plus l'intervalle est grand, plus cette hypothèse est « fausse » ;
- Le troisième niveau est formé par l'intervalle des valeurs, il s'étend de la valeur minimum et la valeur maximum. Il s'interprète comme le niveau précédent mais il assure qu'il n'y a aucune valeur en dehors de l'intervalle compris entre le minimum et le maximum.

Ces trois niveaux de lecture servent à moduler la fiabilité que l'on peut accorder à une interprétation.

D'autres informations telles la médiane, les quartiles ou encore la distribution des valeurs pour chaque variable sont assez difficile à ajouter car le profil deviendrait alors trop chargé en information. En outre, ces informations peuvent être considérées comme redondantes dans la mesure où l'interprétation de la médiane est voisine de celle de la moyenne et l'intervalle formé par l'écart-type autour de la moyenne peut s'interpréter de manière similaire à l'intervalle compris entre le premier et le dernier quartile. La distribution des valeurs via un histogramme est quand à elle trop volumineuse graphiquement.

Nous verrons dans le chapitre sur les expérimentations et les applications que cette visualisation permet de réaliser une analyse de données facilement et rapidement.

2.2 Tableau évolué des profils de classes

Nous avons repris l'idée du tableau des profils de classes pour des données n'utilisant que des variables quantitatives. En effet, s'il y avait des variables qualitatives, elles ont été transformées en variables quantitatives par disjonction.

Ce graphique est un tableau donnant pour chaque variable dans chaque classe :

- La valeur moyenne sous forme symbolique ;
- La taille de l'intervalle de valeurs sous forme symbolique ;
- La valeur moyenne ;
- La valeur maximum ;
- La valeur minimum.

Le codage de la valeur moyenne sous forme symbolique se fait de la manière suivante :

- « ++ » si la moyenne de la classe est très au-dessus la moyenne générale (supérieure à la moyenne générale de plus de 1.5 écarts types généraux) ;
- « + » si elle est légèrement au-dessus (comprise entre 0.5 et 1.5 écarts types généraux au dessus de la moyenne générale) ;
- « = » si elle en est proche (comprise entre -0.5 et 0.5 écarts types généraux autour de la moyenne générale) ;
- « - » si elle est légèrement au-dessous (comprise entre -0.5 et -1.5 écarts types généraux au dessous de la moyenne générale) ;

- « -- » si elle est très au-dessous la moyenne générale (inférieure à la moyenne générale au delà de -1.5 écarts types généraux).

La taille de l'intervalle des valeurs est un indicateur de dispersion, son codage se fait de la façon suivante :

- En vert, la représentativité de la valeur moyenne de la classe est considérée comme fiable. La taille de l'intervalle de valeurs est inférieure ou égale à 0.5 écarts types généraux ;
- En orange, la représentativité de la valeur moyenne de la classe est probable. La taille de l'intervalle de valeurs est comprise entre 0.5 et 1 écarts types généraux ;
- En rouge, la représentativité de la valeur moyenne de la classe est incertaine. La taille de l'intervalle de valeurs est supérieure à 1 écart type général.

Il convient toutefois de garder à l'esprit que ces codages ne sont que des aides à l'interprétation. Ils peuvent se révéler inexact dans la mesure où la personne qui procède à l'interprétation juge que l'écart-type général d'une variable est trop faible ou trop grand. Si l'écart-type général est jugé comme trop faible, il convient alors de ne pas interpréter la variable et de considérer qu'elle ne varie pas, sa valeur restant sa valeur moyenne. Au contraire, si l'écart-type général semble trop grand, les indices de dispersion sont faussés et une valeur moyenne indiquée comme « fiable » peut ne pas l'être. Pour ces raisons, pour chaque variable sont indiqués les moyennes générales et surtout les écart-types généraux.

L'exemple suivant montre le tableau des classes correspondant à l'arbre de classification décrit précédemment. Chaque case comporte la valeur maximum en haut à gauche en gris, la valeur minimum en bas à gauche en gris, la valeur moyenne au milieu à droite et la valeur moyenne sous forme symbolique au milieu.

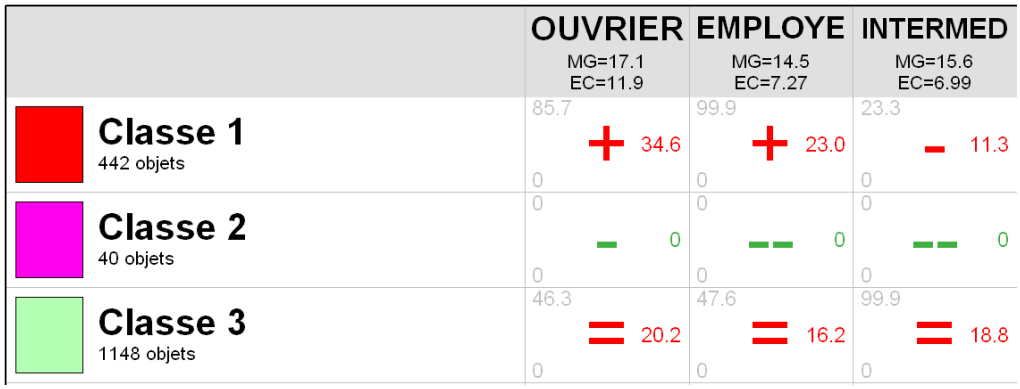


Fig. 51 –Détail d'un tableau des profils de classes évolué

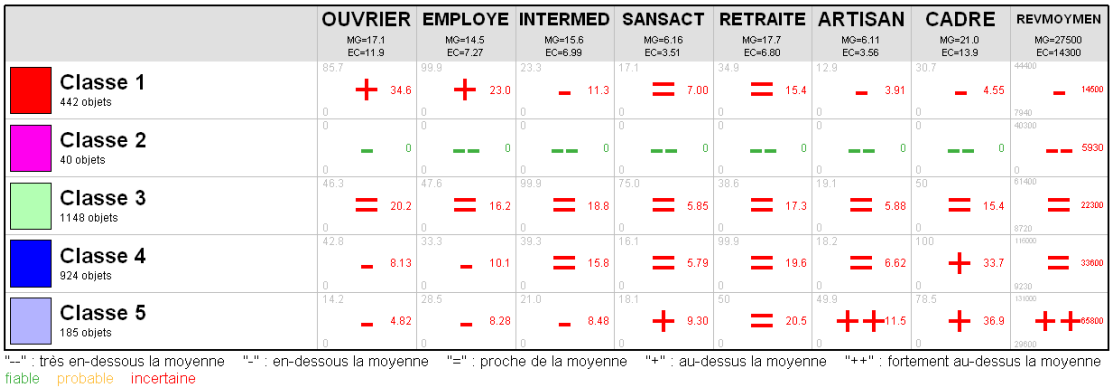


Fig. 52 –Tableau des profils de classes évolué

2.3 Optimisation de l'ordre des variables et des classes

Au vu des résultats précédents, nous avons décidé d'optimiser l'ordre des résumés obtenus en utilisant l'algorithme de calcul de l'ordre optimal de Z. Bar-Joseph et d'autres [*BDGHJS02*] à partir de la classification hiérarchique. Nous injectons donc dans cet algorithme la hiérarchie des résumés obtenue à l'aide de la CAHA et nous utilisons notre distance entre résumés (aussi utilisées par la CAHA).

Pour trier les variables, nous souhaitons appliquer le même algorithme d'optimisation. Cependant, comme nous traitons un grand volume de données, nous ne pouvons travailler directement sur le tableau de données. Pour cette raison, nous utilisons une matrice de similarité (ou de dissimilarité) entre les variables. Elle est construite en même temps que les résumés. Cette matrice est petite car le nombre de variables est petit (de l'ordre de vingtaine au maximum). Elle est aussi symétrique et donne la similarité (ou la dissimilarité) entre chacune des variables.

Nous avons choisi deux types de similarité ou de dissimilarité :

- La corrélation qui est une similarité. Dans ce cas, la similarité maximum vaut *1*.
- La distance euclidienne normalisée qui est une dissimilarité. Dans ce cas, la dissimilarité minimum vaut *0*.

Soit deux variables V_X et V_Y, nous avons donc:

$$Corrélation(V_X, V_Y) = \frac{1}{n}\sum_{i=1}^{n} CR_X(o_i) \times CR_Y(o_i)$$

$$DistanceNormalisée(V_X, V_Y) = \sqrt{\sum_{i=1}^{n} (CR_X(o_i) - CR_Y(o_i))^2}$$

Avec n le nombre d'objets o_i

Et $CR_k(o_i)$, la valeur centrée-reduite de la variable V_k pour l'objet o_i

Pour rappel, on a : $\forall o_i \;\; CR_k(o_i) = \dfrac{V_k(o_i) - M_k}{\sigma_k}$

Avec $V_k(o_i)$, la valeur de la variable V_k pour l'objet o_i

Et M_k la moyenne de la variable V_k

Et σ_k l'écart-type de la variable V_k

Ainsi, pour les deux type de similarité (ou de dissimilarité), la contribution d'un objet o_i à la similarité (ou la dissimilarité) entre deux variables V_X et V_Y est :

- Pour la corrélation : $ContributionCorrélation(V_X, V_Y) = CR_X(o_i) \times CR_Y(o_i)$
- Pour la distance euclidienne normalisée :

$$ContributionDistanceNormalisée(V_X, V_Y) = (CR_X(o_i) - CR_Y(o_i))^2$$

Ainsi, lorsqu'un objet o_i est lu, chaque case de la matrice est incrémentée par la contribution de l'objet à la similarité (ou la dissimilarité) entre les deux variables concernées. Lorsque tous les objets ont été lus, il ne reste plus qu'à finaliser la matrice en effectuant les opérations suivantes : Pour la corrélation, il faut diviser la valeur contenue dans chaque case par le nombre d'objets *n*; pour la distance euclidienne normalisée, il faut remplacer la valeur contenue dans chaque case par sa racine carrée. On remarquera toutefois que cette dernière opération est facultative car elle ne modifie pas l'ordre des valeurs.

L'exemple suivant montre la construction itérative de la matrice.

Soit les 2 objets *A* et *B* caractérisés par trois variables V_X, V_Y et V_Z, on a :

	CR_X	CR_Y	CR_Z
A	-1	1	0
B	1	-1	1

On calcule la matrice des distances normalisées :

	V_X	V_Y	V_Z
V_X	0	0	0
V_Y	0	0	0
V_Z	0	0	0

Initialisation

	V_X	V_Y	V_Z
V_X	0	4	1
V_Y	4	0	1
V_Z	1	1	0

car :

$(CR_X(A) - CR_Y(A))^2 = 4$

$(CR_X(A) - CR_Z(A))^2 = 1$

$(CR_Y(A) - CR_Z(A))^2 = 1$

Ajout de A

	V_X	V_Y	V_Z
V_X	0	8	1
V_Y	8	0	5
V_Z	1	5	0

car :

$(CR_X(B) - CR_Y(B))^2 = 4$

$(CR_X(B) - CR_Z(B))^2 = 0$

$(CR_Y(B) - CR_Z(B))^2 = 4$

Ajout de B

	V_X	V_Y	V_Z
V_X	0	$\sqrt{8}$	1
V_Y	$\sqrt{8}$	0	$\sqrt{5}$
V_Z	1	$\sqrt{5}$	0

Finalisation

Fig. 53 – Exemple de construction de la matrice des distances euclidiennes normalisées entre les variables

Nous construisons ensuite directement une hiérarchie sur cette matrice en utilisant la distance euclidienne entre les individus et la distance de Ward entre les classes.

L'exemple suivant illustre le fonctionnement de cette méthode. À partir des données sur les Régions nous avons construit la matrice des corrélations entre les variables en utilisant la méthode précédente.

Variables	Rev_Moy	Agri	Artisan	Cadre	Interméd	Employé	Ouvrier	Retraite	Sans_Act
Rev_Moymen	1,00	-0,61	-0,20	0,93	0,81	0,19	-0,05	-0,69	-0,46
Agriculteur	-0,61	1,00	0,27	-0,57	-0,60	-0,41	-0,13	0,78	-0,11
Artisan	-0,20	0,27	1,00	0,01	-0,35	0,48	-0,83	0,45	0,50
Cadre	0,93	-0,57	0,01	1,00	0,72	0,39	-0,33	-0,57	-0,25
Intermédiaire	0,81	-0,60	-0,35	0,72	1,00	-0,17	0,22	-0,68	-0,60
Employé	0,19	-0,41	0,48	0,39	-0,17	1,00	-0,67	-0,14	0,67
Ouvrier	-0,05	-0,13	-0,83	-0,33	0,22	-0,67	1,00	-0,41	-0,45
Retraite	-0,69	0,78	0,45	-0,57	-0,68	-0,14	-0,41	1,00	0,16
Sans_Activité	-0,46	-0,11	0,50	-0,25	-0,60	0,67	-0,45	0,16	1,00

Fig. 54 –Matrice des corrélations des variables des données sur les Régions

L'exemple suivant montre l'ordre optimal des variables obtenus à partir de la hiérarchie.

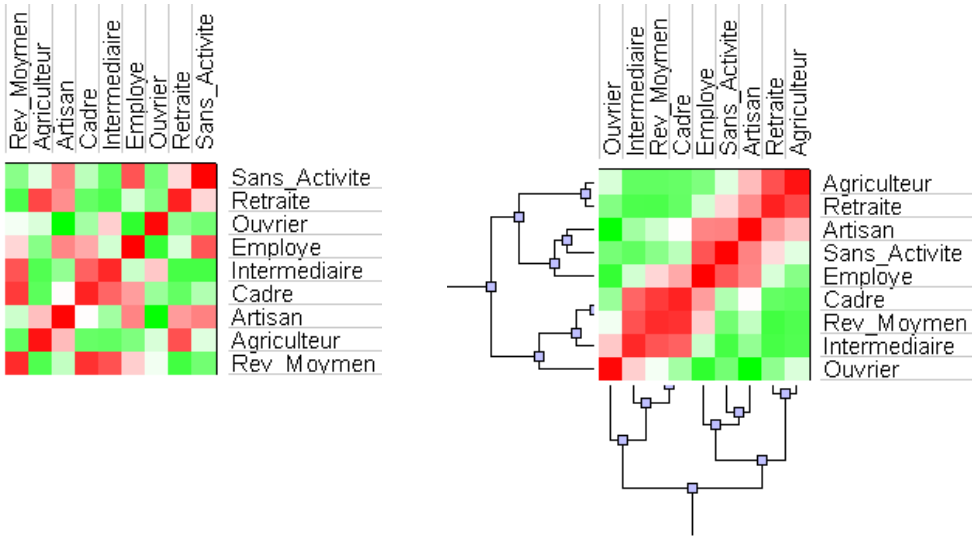


Fig. 55 –Ordre des variables obtenus à partir de la matrice des corrélations

On remarquera que l'ordre trouvé est presque identique à celui trouvé précédemment dans le chapitre sur l'état de l'art (paragraphe I.2.3.1.5) en construisant la hiérarchie des variables directement sur les données. La seule petite différence est le changement de position d'*Ouvrier* qui vient se positionner à côté d'*Intermédiaire* avec lequel il est très légèrement corrélé.

Ainsi, le fait d'utiliser un ordre optimal ou quasi-optimal tant pour les classes que pour les variables permet aux outils de visualisation d'être beaucoup plus clairs et compréhensibles. Nous verrons ultérieurement qu'il est alors généralement possible de grouper visuellement certaines variables ou certaines classes qui sont semblables. De plus, les différences entre les variables ou les classes ressortent d'autant mieux. D'une manière générale, le fait de réordonner les variables et les classes facilite la compréhension des visualisations.

2.4 Conclusion

Nous avons élaboré une visualisation hiérarchique évoluée qui est une représentation hiérarchique classique complétée des profils de classes. Contrairement aux représentations classiques, ces profils sont très lisibles car ils sont purement graphiques et ne comportent aucune information chiffrée. La légende commune permet de retrouver approximativement les valeurs chiffrées. Cependant, dans le cadre d'une analyse détaillée, les valeurs chiffrées sont disponibles via le *profil détaillé* accessible dynamiquement par un clic de souris sur le profil désiré.

Nous avons aussi élaboré un tableau évolué afin de compenser une lacune de la visualisation hiérarchique évoluée : la difficulté de comparer rapidement les informations disponibles pour une même variable. A la différence des tableaux classiques, nous offrons deux niveaux de lecture, l'un pour une lecture globale et rapide de l'ensemble du tableau et l'autre pour une lecture minutieuse. La lecture rapide correspond aux informations présentes sous forme symbolique (valeur moyenne et écart) et la lecture minutieuse correspond aux informations chiffrées (valeur moyenne, valeur maximale et valeur minimale). Ainsi, le tableau évolué est complémentaire de la visualisation hiérarchique évoluée.

Nous avons aussi traité le problème de l'ordre optimal des variables et des résumés. Nous avons réutilisé une méthode existante pour la recherche de l'ordre optimal dans une hiérarchie et nous l'avons adaptée afin de réorganiser les variables. Nous verrons que cette méthode peut être utilisée à des fins d'analyse dans le chapitre traitant des applications.

Toutefois, l'ordre optimal des résumés n'entraîne pas forcément un ordre optimal des classes construites par la hiérarchie. Nous étudions donc la possibilité de calculer l'ordre des classes de la hiérarchie durant son exploration.

3 Détermination et Hiérarchisation de pôles

Nous avons vu dans l'état de l'art (partie I.3) que le lissage spatial permet de mettre en évidence les tendances se produisant aux grandes échelles. Cependant, le résultat n'est pas très synthétique et les zones intéressantes qui sont celles de forte valeur ne sont pas mises en évidence. Nous proposons donc notre méthode de détermination et de hiérarchisation des pôles qui permet de résumer des cartes de lissage spatial par quelques pôles et ensuite de les hiérarchiser afin de mettre en évidence les relations entre les différentes échelles de lissage.

La première partie montre comment il est possible de résumer une carte de lissage spatial en extrayant les pôles. Cette technique permet de décrire très rapidement et brièvement une carte de lissage spatial. Cette méthode se déroule en trois étapes successives : le calcul du graphe de voisinage, la détermination des pôles potentiels et l'affinage des pôles afin de les faire évoluer jusqu'à leur emplacement final et de supprimer les pôles redondants. Une dernière section est consacrée à la description automatique des pôles.

Dans la seconde partie, nous montrerons l'intérêt de superposer sur une même carte les pôles correspondant à des échelles différentes en les incluant dans une hiérarchie. Nous

obtenons une carte très synthétique et de plus, les pôles de l'échelle la plus courte permettent d'expliquer les pôles des échelles plus grandes.

3.1 Détermination de pôles

Les pôles sont des lieux où une valeur lissée est maximale par rapport à son entourage immédiat. Ils permettent par ailleurs de résumer une carte de lissage spatial. L'étape préalable à la détermination des pôles est le calcul des valeurs lissées que nous avons vu dans le chapitre sur l'état de l'art (partie I.3). La détermination des pôles se déroule en trois étapes. La première étape est la construction des relations de voisinage immédiat entre les différents lieux, il s'agit du calcul du graphe de voisinage. L'étape suivante est très simple, il s'agit de l'extraction des pôles potentiels. La dernière étape permet d'affiner ces pôles. Il s'agit principalement de les faire évoluer jusqu'à leur emplacement final et de supprimer les pôles redondants. La dernière partie est consacrée à la description automatique des pôles à partir des noms des objets y contribuant le plus.

3.1.1 Calcul du graphe de voisinage

Le graphe de voisinage se déduit en principe des relations de contact entre les objets géographiques. Toutefois, lorsque la relation de contact fonctionne mal ou fait défaut (par exemple, pour des objets ponctuels), il convient d'utiliser des méthodes alternatives. Ainsi, à partir des centroïdes (centres de gravité) de chaque zone, nous calculons la triangulation de Delaunay qui donne les relations de voisinage et le diagramme de Voronoï associé qui donne un pavage de la carte avec des polygones [*Aure91*]. L'algorithme utilisé est celui développé par S. Fortune [*Fort86*] dont la complexité est faible car en $o(n\log(n))$ selon n le nombre d'objet géographiques. Cependant, des lieux très éloignés mais situés au bord peuvent être considérés à tort comme des voisins immédiats. Pour supprimer une grande partie de ces relations de voisinage aberrantes nous utilisons la règle suivante : si le trait symbolisant la relation de voisinage entre deux lieux ne passe pas par la frontière commune de leur polygone de Voronoï alors cette relation de voisinage est supprimée. Les figures ci-dessous (fig 56, 57) montrent la triangulation de Delaunay pour les îlots du département de la Loire Atlantique. Les relations de voisinage en pointillés bleus ont été éliminées grâce à la simplification décrite précédemment.

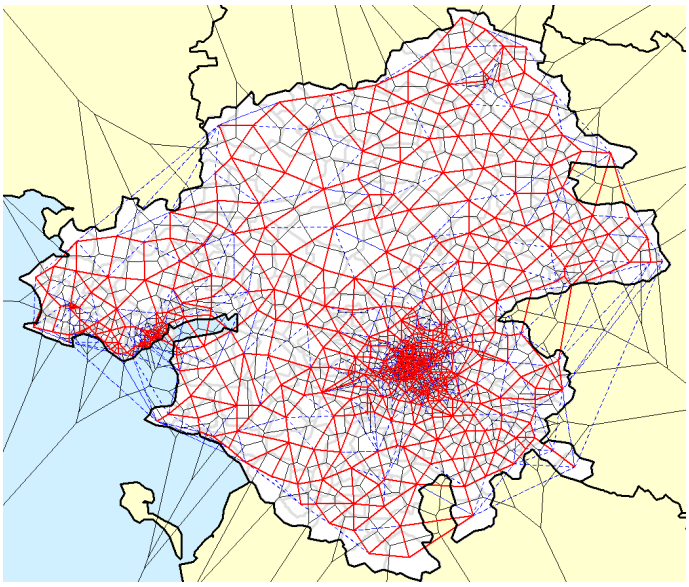


Fig. 56 –Voisinage réduit des îlots avec le pavage de polygones associé pour les îlots du département de la Loire Atlantique

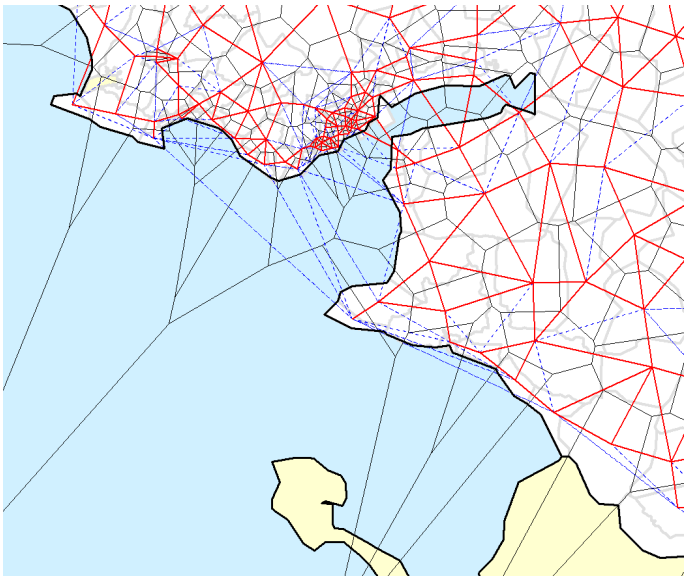


Fig. 57 – Détail de la carte précédente montrant la côte Atlantique

Grâce aux relations de voisinage, nous pouvons déterminer les emplacements où les valeurs lissées sont localement maximales.

3.1.2 Recherche des pôles potentiels

Un lieu sera considéré comme un maximum local (pôle potentiel) si tous les lieux immédiatement voisins ont une valeur inférieure à la sienne. Ainsi, nous regardons si pour chaque lieu, la valeur lissée est maximale par rapport aux lieux immédiatement voisins. Nous partons de la carte lissée de la densité de supermarché et d'hypermarché. Le lissage avait été

effectué pour un rayon de 10 km dans la partie consacrée à l'état de l'art (partie I.3.3). Le résultat obtenu à partir de cette carte lissée est le suivant :

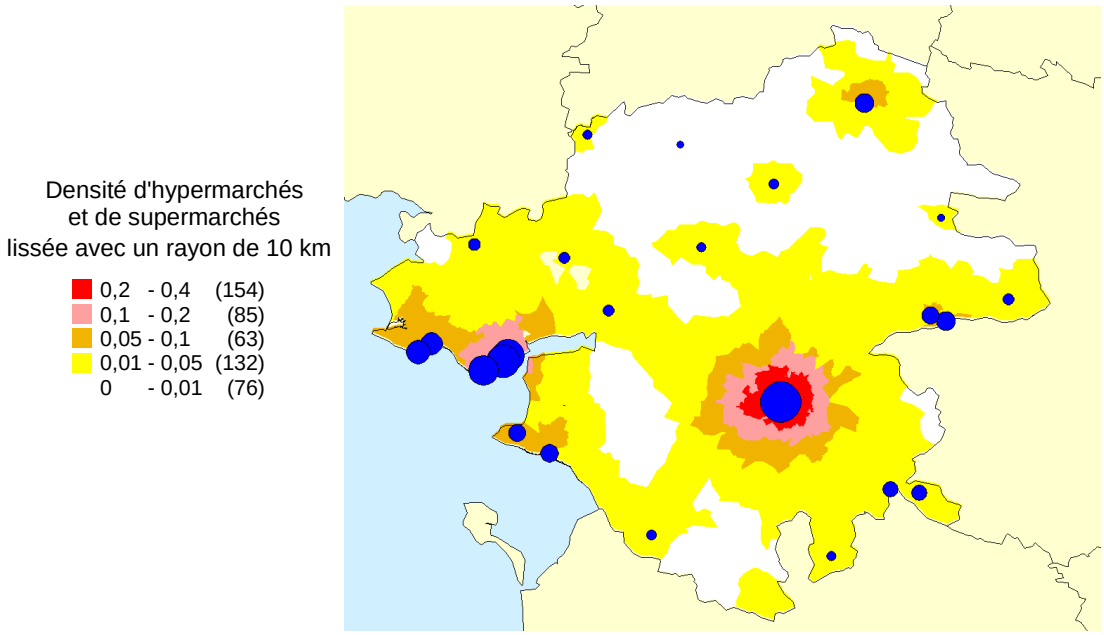


Fig. 58 –Carte des pôles potentiels (lieux où les valeurs lissées sont localement maximales)

Les cercles représentant les pôles ont une surface proportionnelle à la valeur lissée qu'ils représentent. On remarque que certains pôles sont très proches et qu'il serait préférable de les fusionner. L'étape suivante va justement permettre de supprimer les pôles redondants et d'améliorer les autres.

3.1.3 Affinage des pôles

À cause d'un maillage trop irrégulier ou trop imprécis. Les pôles déterminés précédemment souffrent d'un double handicap. D'une part, certains pôles sont mal positionnés et devrait être décalés vers une meilleure position ou la valeur lissée sera véritablement maximale localement. D'autres part, certains pôles sont redondants avec d'autres pôles mais il est difficile de déterminer ceux à garder et ceux à supprimer. Nous utilisons donc une heuristique proche de la méthode des nuées dynamiques (k-moyennes ou *k-means*) qui permet de faire glisser progressivement les pôles vers des lieux où la valeur est plus forte. Il s'agit de la méthode Mean-Shift [*CM02*]. Ainsi, les pôles seront bien placés et les pôles redondants glisseront progressivement vers une position commune et il suffira de garder un seul des pôles pour une même position. Notre heuristique détermine pour chaque pôle sa nouvelle position en calculant le barycentre des valeurs lissées. Le *Lieu Pondéré Lissé* s'exprime de la façon suivante :

$$LieuPdrLiss_{f,R}(X) = \frac{\sum_{j=1}^{n}\left(CtrbPdr_{f,R}(X,O_j)\times Lieu(O_j)\right)}{\sum_{j=1}^{n} CtrbPdr_{f,R}(X,O_j)}$$

Avec $CtrbPdr_{f,R}(X,O_j)$ qui est la contribution pondérée d'un objet O_j à la valeur lissée du lieu X (Cette formule est détaillée dans la partie sur le lissage spatial dans le chapitre sur l'état de l'art).

L'évolution d'un pôle *P* peut ainsi se voir comme une suite de lieux étapes P_1, P_2, … jusqu'à l'infini avec $P_{n+1} = LieuPdrLiss_{f,R}(P_n)$. Nous arrêtons les itérations lorsque nous atteignons un nombre suffisant d'itérations (par exemple, 100 itérations), ou lorsque la distance entre une étape et la suivante est faible par rapport au rayon *R*, c'est-à-dire lorsque $dist(P_n, P_{n+1}) < R/100$. Il faut toutefois remarquer la méthode Mean-Shift ne garantit pas que la valeur lissée du nouveau lieu du pôle qui soit supérieure à celle de l'ancien lieu du pôle. Cependant, dans les faits, il est rare d'observer une décroissance qui est alors très faible.

Une fois que les pôles potentiels ont achevé leur migration, nous supprimons les pôles redondants. Pour cela, nous comparons les pôles deux à deux et si la distance les séparant est très faible, nous en supprimons un :

Si $dist(P_X, P_Y) < \frac{R}{1000}$ alors P_Y est supprimé.

La figure suivante montre les trajectoires suivies par les pôles potentiels ainsi que les pôles potentiels redondants supprimés. Il s'agit toujours de l'exemple portant sur la densité de supermarchés et d'hypermarchés dans le département de Loire Atlantique.

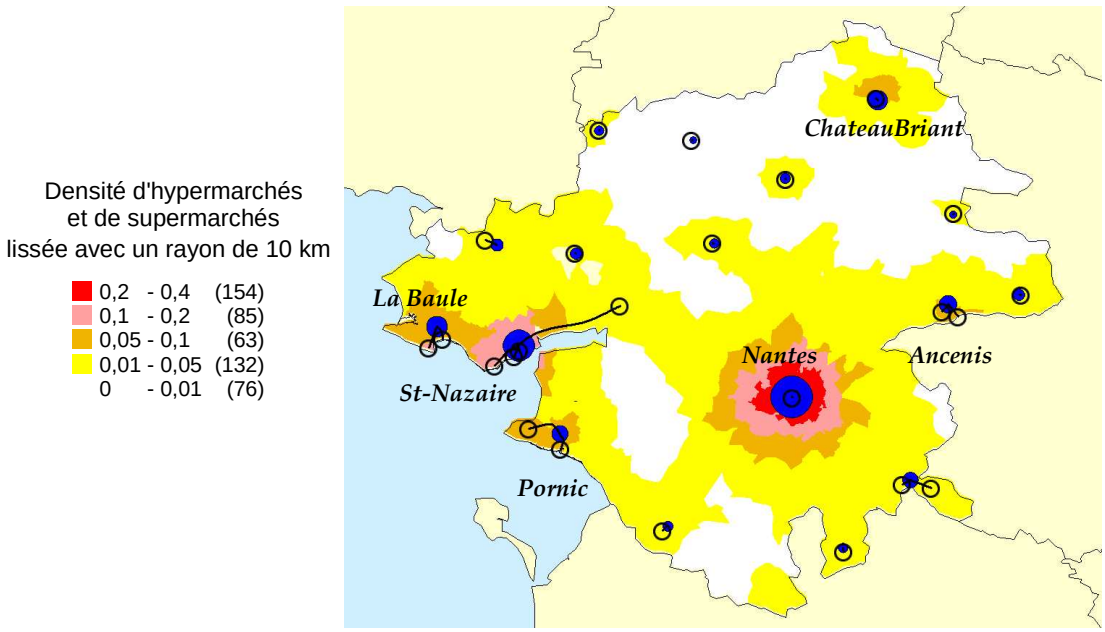


Fig. 59 –Carte des pôles finaux et trajectoires des pôles potentiels.

On remarque que le résultat est proche de celui que l'on pouvait espérer, sauf peut-être pour un pôle potentiel situé initialement assez loin de St-Nazaire. Il a migré loin de sa position d'origine et a donc été supprimé alors qu'il fallait peut-être le conservé.

Nous allons maintenant définir une manière de décrire automatiquement ces pôles.

3.1.4 Description des pôles

Nous pouvons identifier les objets ayant de grandes contributions grâce aux formules de la *Contribution* ou de la *Contribution Pondérée*. Cependant nous avons plutôt besoin de caractériser un pôle par les objets y contribuant le plus. Ainsi, pour chaque pôle, nous listons par ordre décroissant les contributions des différents objets. Pour faciliter la compréhension, nous utilisons la *Contribution Pondérée Relative* qui permet de voir les plus grandes contributions des objets O_j relativement à la totalité des contributions pour un lieu X (un pôle).

$$CtrbPdrRltv_{f,R}(X,O_j) = \frac{CtrbPdr_{f,R}(X,O_j)}{\sum_{j=1}^{n} CtrbPdr_{f,R}(X,O_j)}$$

Afin de ne retenir que les objets dont la contribution est significative, nous restreignons la liste aux premiers objets dont la contribution relative totale permet d'atteindre un seuil de *80%* par défaut. Une conséquence intéressante est la possibilité de nommer les pôles à partir des noms des objets ayant les contributions relatives les plus importantes.

La carte suivante indique en rayures et pour chaque pôle, les secteurs contribuant au total à plus de 80%. On remarque que pour les secteurs ruraux, la contribution provient souvent d'une seule zone, tandis que pour les zones urbaines et notamment Nantes, les zones contribuant sont beaucoup plus nombreuses. Le cercle autour de chaque pôle correspond à un rayon de 10 km. Les zones contribuant le plus sont localisées à moins de 5 km du pôle auquel elles contribuent. Ainsi, la dispersion reste tout de même faible contrairement à l'impression donnée par la cartographie de la densité seule : les deux zones en orange indiquant une forte densité de magasins sont très étendues alors que les objets contribuant aux pôles dans ces zones restent concentrés.

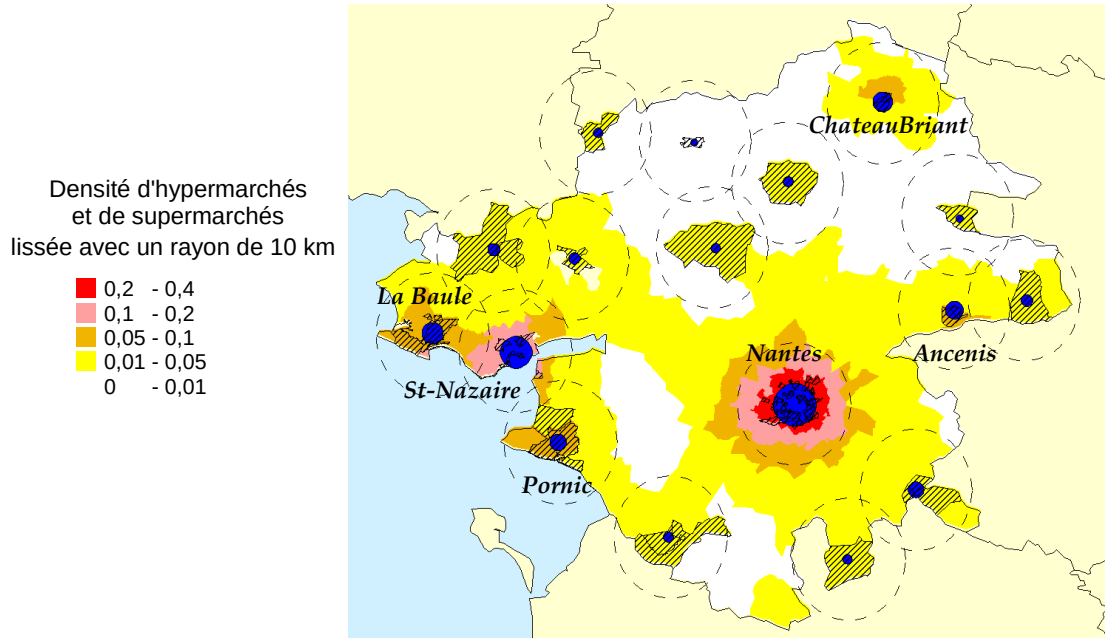


Fig. 60 –Carte des pôles et les zones y contribuant le plus à hauteur de 80 % en cumulé par pôle.

Le tableau suivant est un extrait de la table des contributions. Cet extrait concerne le pôle numéro 11 qui est en fait celui situé sur Nantes. On observe qu'aucune zone n'a de rôle prépondérant, la plus forte contribution relative étant d'environ 4 %. Les centres commerciaux tels ceux de la Beaujoire ou Atlantis ne se dégagent pas du reste. Cela est dû au fait que c'est le nombre de supermarchés et d'hypermarchés qui a été pris en compte. Un résultat plus vraisemblable pourrait être obtenu en prenant plutôt en compte le Chiffre d'Affaire ou la Surface de Vente.

Nom de la zone	Nombre d'hypers et de supers	Pole	Contrib	Contrib Cumulée	Contrib Relative (%)	Contrib Relative Cumulée (%)
ZA ATOUT SUD	2	11	1,60	1,60	4,05	4,05
LAURIERS	2	11	1,44	3,03	3,64	7,70
BEAUJOIRE HALVEQUE	3	11	1,35	4,39	3,43	11,13
WALDECK - SULLY	1	11	0,97	5,35	2,45	13,59
QUAI DE LA FOSSE	1	11	0,94	6,29	2,38	15,97
REPUBLIQUE LES PONTS	1	11	0,94	7,23	2,38	18,34
SAINT FELIX	1	11	0,94	8,16	2,37	20,72
ATLANTIS	2	11	0,93	9,09	2,36	23,08
GRILLAUD - PROCE	1	11	0,93	10,02	2,35	25,43
ROND POINT DE RENNES	1	11	0,91	10,93	2,32	27,76
…	…	…	…	…	…	…
METAIRIE	1	11	0,61	31,34	1,56	79,56
ZA CHEVIRE	1	11	0,56	31,90	1,41	80,97

Fig. 61 –Extrait de la table des contributions

On remarque que la contribution et la contribution cumulée s'expriment aussi en nombre d'hypermarchés et de supermarchés. Par rapport à la valeur initiale du nombre d'hypermarchés et de supermarchés, la contribution est systématiquement minorée par la fonction d'interaction spatiale en fonction de la distance au pôle.

La figure suivante permet de voir la localisation des zones contribuant le plus au pôle 11.

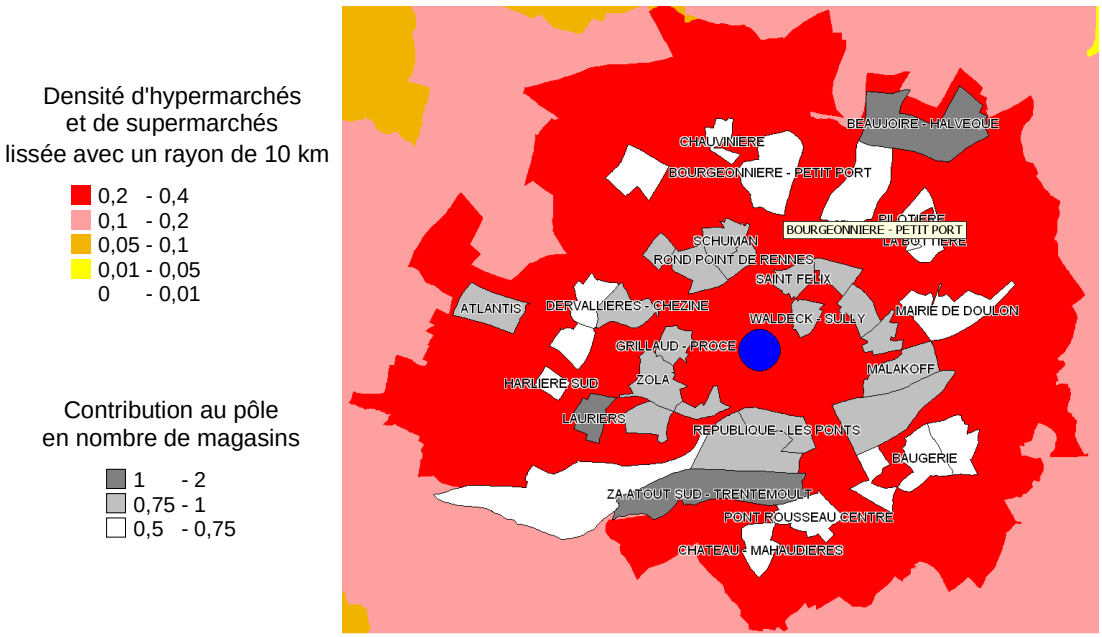


Fig. 62 –Carte des îlots contribuant le plus au pôle de Nantes.

3.1.5 Conclusion

La détermination de pôles à partir des valeurs lissées d'une variable permet de décrire rapidement et simplement un territoire pour cette variable. En examinant les objets contribuant le plus à ces pôles, il est alors possible de préciser la localisation des fortes valeurs à l'intérieur de l'aire d'influence de chaque pôle. De plus, la description obtenue est valable pour l'échelle donnée par le rayon. Ainsi, en faisant varier le rayon, nous obtenons différentes descriptions correspondant aux différentes échelles (de l'échelle locale à l'échelle globale).

3.2 Hiérarchisation des pôles

L'intérêt de la hiérarchisation des pôles est de superposer sur une même carte les pôles correspondant à des échelles différentes en les incluant dans une hiérarchie. Nous obtenons une carte très synthétique et de plus, les pôles de l'échelle la plus courte permettent d'expliquer les pôles des échelles plus grandes. Nous allons d'abord voir comment définir le lien hiérarchique afin qu'il dépende à la fois de la distance et de la valeur de pôles. Puis nous verrons son application pour visualiser la localisation des supermarchés et hypermarchés de la Loire-Atlantique à deux échelles différentes simultanément.

3.2.1 Définition du lien hiérarchique

Le lien hiérarchique s'exprime par la création de liens entre les pôles de deux cartes de lissage différentes : les pôles de la carte ayant le rayon de lissage le plus grand sont en haut de

la hiérarchie tandis que les pôles de la carte ayant le rayon de lissage le plus petit sont en bas de la hiérarchie.

Par ailleurs, pour construire la hiérarchie, il faut caractériser la force du lien entre deux pôles quelconques. Dans un premier temps, nous avions envisagé d'utiliser la distance : plus il sont proches, plus le lien est fort. Cependant, il était difficile de venir compléter cette première approche en y intégrant les valeurs lissées des pôles. Pour cette raison, nous avons utilisé une autre approche. On peut considérer un pôle comme un ensemble d'objets ayant chacun une valeur de contribution. Nous avons défini la *Contribution Commune* qui est la contribution la plus faible commune au deux pôles. La *Contribution Commune* d'un objet O_j à deux pôle *X* et *Y* est donc :

$$CtrbCmmn_{f,R}(X,Y,O_j) = \min(Ctrb_{f,R}(X,O_j), Ctrb_{f,R}(Y,O_j))$$

La version pondérée est la suivante :

$$CtrbPdrCmmn_{f,R}(X,Y,O_j) = \min(CtrbPdr_{f,R}(X,O_j), CtrbPdr_{f,R}(Y,O_j))$$

À partir des *Contribution Commune*, nous caractérisons donc la force de la liaison hiérarchique comme étant la somme des contributions communes aux deux pôles pour l'ensemble des objets. La formule de la *Force de Liaison* peut ainsi se voir comme une adaptation de la fonction d'intersection $\cap$ en théorie ensembliste et surtout de la fonction *ET* en logique floue [*Bezd93*]. La formule de la *Force de Liaison* entre deux pôles *X* et *Y* est :

$$\begin{aligned} ForceLiaison_{f,R}(X,Y) &= \sum_{j=1}^{n} CtrbCmmn_{f,R}(X,Y,O_j) \\ &= \sum_{j=1}^{n} \min(Ctrb_{f,R}(X,O_j), Ctrb_{f,R}(Y,O_j)) \end{aligned}$$

La version pondérée est la suivante :

$$\begin{aligned} ForceLiaisonPdr_{f,R}(X,Y) &= \sum_{j=1}^{n} CtrbPdrCmmn_{f,R}(X,Y,O_j) \\ &= \sum_{j=1}^{n} \min(CtrbPdr_{f,R}(X,O_j), CtrbPdr_{f,R}(Y,O_j)) \end{aligned}$$

D'une part, la *Force de Liaison* traduit bien la distance entre les pôles car plus ils sont éloignés plus elle est faible. D'autre part, elle dépend aussi de la valeur des pôles (plus leur valeur lissée est importante, plus la *Force de Liaison* est importante). De plus, elle s'exprime dans la même unité que la *Contribution*.

L'exemple suivant montre le détail du calcul de la *Force de Liaison* pour le pôle de « niveau 30 km » de St-Nazaire et le pôle de « niveau 10 km » de La Baule (situé à gauche de St-Nazaire). Pour chaque zone, le nombre de magasins présents est indiqué en haut en noir. La *Contribution* de la zone au pôle de La Baule est indiquée en bleu à droite et la *Contribution* au pôle de St-Nazaire est indiquée en rouge à gauche. La *Contribution Commune* est quant à elle indiquée en bas en gris en italique.

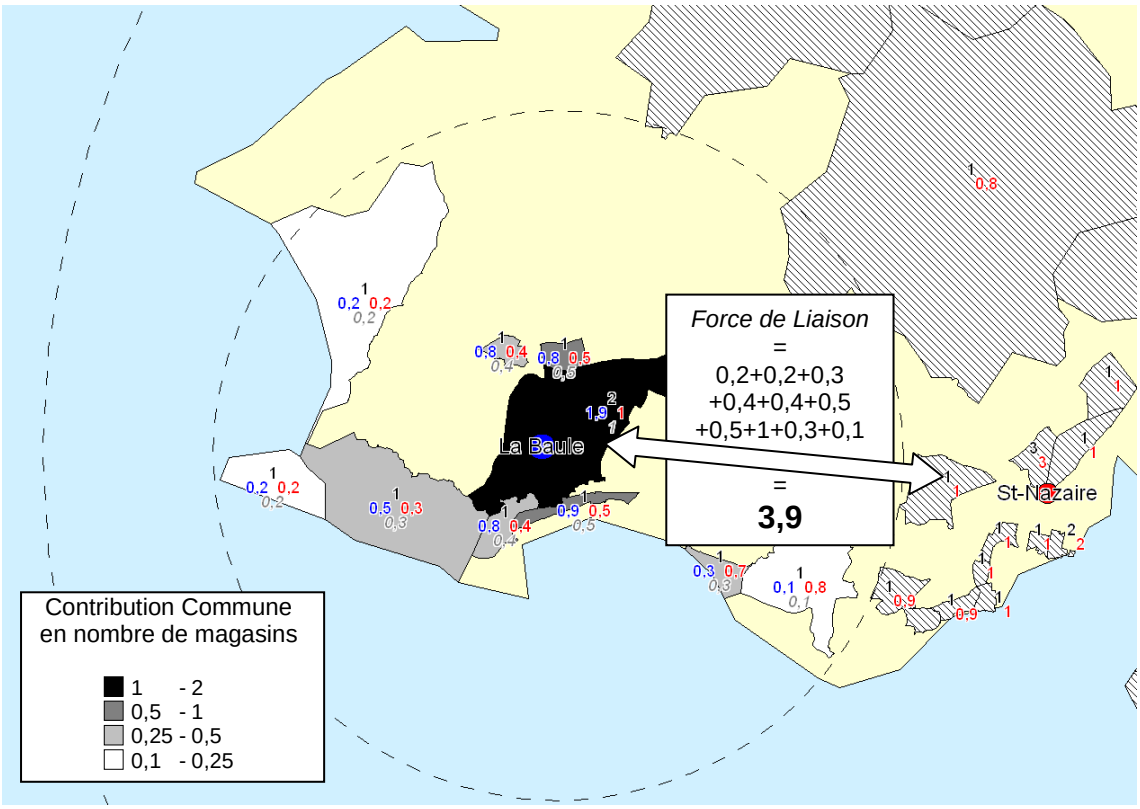


Fig. 63 –Carte des Contributions Communes entre le pôle de St-Nazaire (en rouge) et le pôle de La Baule (en bleu).

La *Force de Liaison* se déduit des *Contributions Communes* en les sommant. Dans le cas de St-Nazaire et de La Baule, La *Force de Liaison* a une valeur de *4.9* « magasins ». Cette liaison n'est qu'une liaison parmi d'autres et nous allons maintenant voir la hiérarchie complète entre les pôles de « niveau 30 km » et ceux de « niveau 10 km ».

3.2.2 Exemple de hiérarchie

L'exemple suivant montre les pôles de « niveau 30 km » (en rouge) et ceux de « niveau 10 km » (en bleu). Le fond de carte est le lissage obtenu avec un rayon de 10 km.

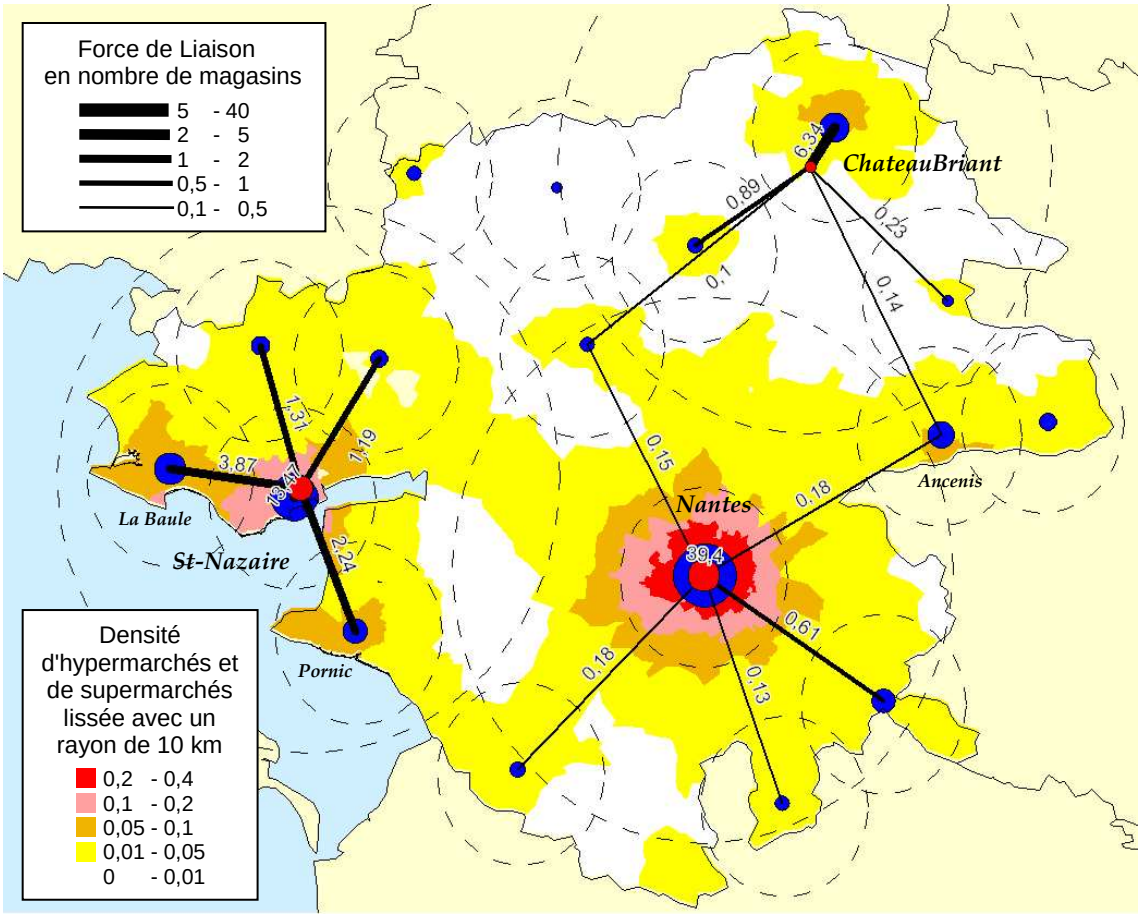


Fig. 64 –Hiérarchie entre les pôles de « niveau 30 km » et ceux de « niveau 10 km » avec en fond de carte, les densités lissées du « niveau 10 km ».

Grâce à la hiérarchie, nous pouvons analyser les pôles du « niveau 30 km » en nous servant des pôles du « niveau 10 km ». Ainsi, on peut dire que le pôle de *Nantes* est organisé de façon centrale, c'est-à-dire que la grande majorité des magasins sont répartis dans toute la proche périphérie à moins de 10 km du centre-ville. Au contraire, le pôle de *St-Nazaire* est organisé le long de la côte Atlantique et il englobe principalement deux pôles de « niveau 10 km » qui sont celui de *Pornic* au sud et *La Baule* à l'ouest. Comparativement à *Nantes* et même *St-Nazaire*, Le pôle de *Châteaubriant* est beaucoup moins important en nombre de magasins et comme pour *Nantes*, ses magasins sont concentrés autour de la ville même de *Châteaubriant.*

On remarque aussi que le pôle de « niveau 10 km » d'*Ancenis* est comparable à *Chateaubriant* mais que contrairement à ce dernier, il est situé assez près de *Nantes*, à 30 km environ. Ainsi, le pôle de *Nantes* de « niveau 30 km » arrive a « éclipser » *Ancenis* mais pas *Chateaubriant* qui est situé beaucoup plus loin, à plus de 50 km de *Nantes*.

3.2.3 Conclusion

La hiérarchie des pôles permet d'afficher sur une même carte les pôles correspondant à des échelles différentes. De plus, les pôles de l'échelle la plus courte permettent d'expliquer les pôles des échelles plus grandes, ce qui est très important en analyse spatiale.

3.3 Conclusion

La détermination des pôles permet de résumer une carte de lissage spatial. Les pôles sont les valeurs maximales locales et elles permettent de décrire très rapidement et brièvement une carte de lissage spatial. Cette méthode nécessite une phase d'affinage au cours de laquelle les pôles évoluent jusqu'à leur emplacement final permettant ainsi de supprimer les pôles redondants. Le résultat obtenu est ainsi particulièrement satisfaisant. De plus, nous avons créé une description automatique des pôles à partir des noms des objets y contribuant le plus. De ce fait, il est possible de donner automatiquement un nom aux pôles.

La hiérarchisation des pôles permet de superposer sur une même carte les pôles correspondant à des échelles différentes en les incluant dans une hiérarchie. Nous obtenons ainsi une carte très synthétique et riche en information car les pôles de l'échelle la plus courte permettent d'expliquer les pôles des échelles plus grandes.

Nous souhaitons améliorer la qualité des résultats en utilisant les distances routières (en temps de parcours) plutôt que les distances à vol d'oiseau. Nous étudions aussi les méthodes de lissage adaptatif afin de détecter des pôles car il apparaît qu'il se forme systématiquement des zones vides de pôles autour des pôles les plus forts. Ce phénomène est assez analogue à la lumière d'un phare qui empêche de voir les bougies allumées à proximité.

4 Sectorisation

Nous traitons deux types de sectorisations distinctes :

- La S*ectorisation Équilibrée* consiste en la création de secteur de « taille » égale. La taille est dans ce cas une quantité à partager : surface, population, clients, … La localisation des secteurs n'est pas définie à l'avance.
- La *Sectorisation à partir de Centres* permet de créer des secteurs localisés autour de « centres ». Un centre peut être un magasin, un dépôt, une ville, … De plus, il est possible de spécifier la « taille » à atteindre pour chaque secteur.

Nous allons voir dans une première partie les méthodes communes aux deux types de sectorisation. Il s'agit d'une part, de la création du graphe de voisinage permettant de connaître, pour chaque objet, ses voisins immédiats et d'autre part, de l'évaluation de la qualité d'une sectorisation afin de pouvoir comparer les sectorisations entre elles et de garder la meilleure.

Dans la seconde partie, nous aborderons la *Sectorisation Équilibrée* en nous intéressant au partitionnement de graphes (*Graph Partitioning*) qui l'équivalent de ce problème en théorie des graphes. A partir d'une méthode existante de partitionnement de graphe, et nous créerons son adaptation pour la sectorisation des données géographiques. Nous établirons un algorithme itératif afin de tester plusieurs solutions et de retenir la meilleure.

Dans la partie suivante, nous traiterons la *Sectorisation à partir de Centres*. Nous verrons dans un premier temps la méthode basique que nous avons retenue. Nous verrons ensuite une version améliorée de l'algorithme : il s'agit d'un algorithme itératif se basant à chaque étape sur

la sectorisation précédente pour déterminer des nouveaux paramètres afin de produire une meilleure sectorisation à l'étape suivante.

Nous verrons aussi, dans une dernière partie, le *Rééquilibrage* qui permet d'optimiser une sectorisation existante. En effet, il est parfois plus intéressant d'améliorer une sectorisation existante que d'en calculer une nouvelle. Le rééquilibrage des secteurs vise donc à améliorer les quantités de chaque secteur en transférant des objets géographiques entre les secteurs. La technique de rééquilibrage utilisée se décompose en deux étapes : le calcul de tous les transferts à effectuer entre les secteurs et leur mise en oeuvre. Nous nous intéresserons à cette dernière phase en exposant l'algorithme itératif que nous avons developpé afin de réaliser les transferts progressivement et en parallèle, afin de converger vers la solution de rééquilibrage.

4.1 Méthodes communes aux deux types de sectorisation

Les deux types de sectorisation ont en commun les méthodes suivantes :

- La transformation des données géographiques en graphe afin de permettre leur sectorisation
- L'évaluation de la meilleure solution parmi toutes les sectorisations calculées. En effet, chacune des deux méthodes de sectorisation génère plusieurs solutions et il faut garder la meilleure comme résultat final.

4.1.1 Transformation des données géographiques en un graphe

La notion au cœur du problème de sectorisation est la relation de voisinage immédiat. C'est à dire qu'il faut au préalable, pour chaque objet géographique, connaître ses voisins immédiats. Cela se traduit par la construction d'un graphe de voisinage. Nous avons déjà vu, dans le cadre du la *Détermination de Pôles* que nous pouvions obtenir un tel graphe à l'aide de la triangulation de Delaunay. Une fois les données transformées en graphe, le problème de sectorisation des données géographiques devient un problème de partition de graphe. De plus, cette méthode permet de construire un graphe de voisinage pour des données géographiques ponctuelles.

Nous obtenons ainsi pour le graphe les caractéristiques suivantes :

- Un *nœud* du graphe représente un objet géographique qui sera regroupé avec d'autres dans une partition (secteur).
- Une *arête* du graphe lie deux nœuds voisins. Si deux noeuds ne sont pas voisins, il n'y a pas d'arête entre eux. Une sectorisation coupe certaines arêtes du graphe.
- Le *poids d'un nœud* représente la quantité contenue par le nœud. La quantité totale d'une partition est la somme des poids des nœuds appartenant à cette partition. Une bonne sectorisation (partitionnement) équilibrée doit faire que la quantité soit la même dans chaque secteur (ou au moins s'en approche).

- Le *poids d'une arête* représente le degré de voisinage entre deux nœuds. Un poids de 0 équivaut à l'absence d'une arête, c'est-à-dire à l'absence de lien de voisinage. Au contraire, un poids très élevé correspond à une forte proximité entre les nœuds qui sont alors fortement liés. Une bonne sectorisation fait que les nœuds fortement liés restent ensembles. C'est toutefois un critère beaucoup moins important que l'équilibre des quantités entre les secteurs.

Nous définissons le *poids d'une arête* à partir de la distance entre les objets géographiques de cette arête, il vaut l'inverse de la distance.

L'exemple suivant montre la transformation en un graphe pondéré. Les objets géographiques sont des communes et la quantité indiquée est la population.

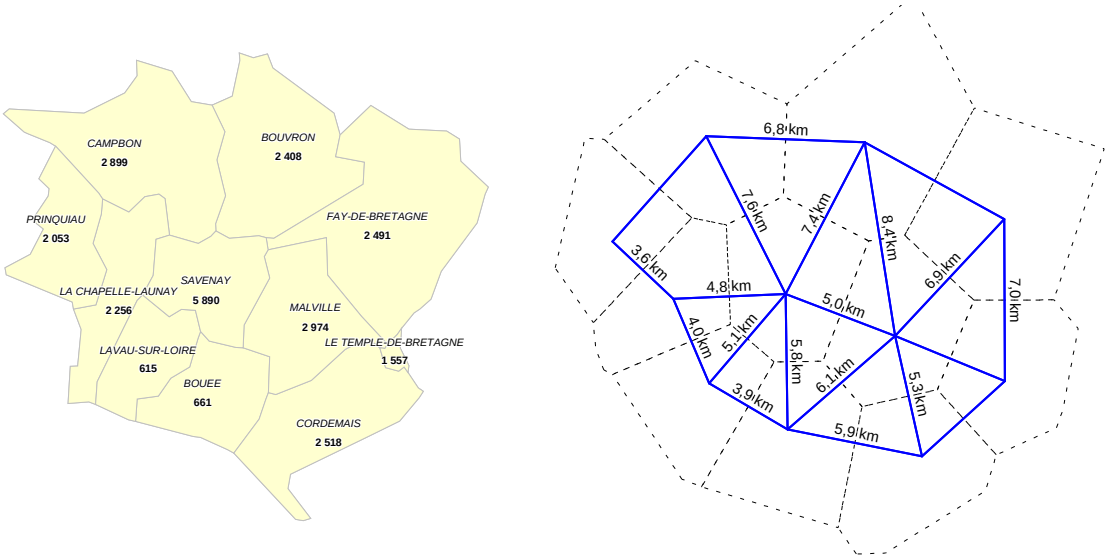


Fig. 65 –Données géographiques de départ (à droite) et le graphe de voisinage obtenu par la triangulation de Delaunay

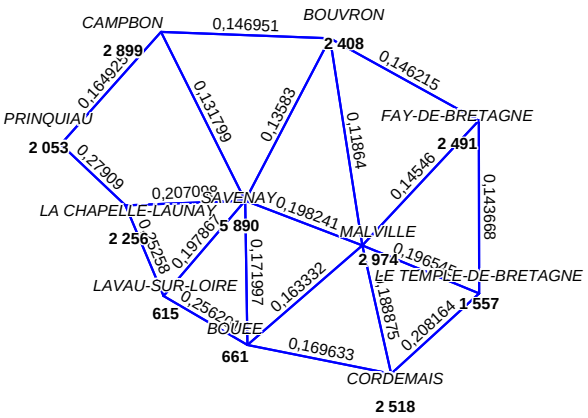


Fig. 66 –Graphe de voisinage final avec les noeuds pondérés et les liens pondérés

Le graphe est alors prêt à être sectorisé.

4.1.2 Mesure de la qualité d'un partitionnement

La mesure de la qualité d'un partitionnement est importante dans le cas où plusieurs partitionnement doivent être évalués afin de choisir le meilleur, ce qui est le cas avec les deux méthodes de sectorisation que nous avons développées.

La qualité d'un partitionnement dépend globalement de deux critères :

- Le **critère d'équilibre** mesure l'adéquation entre la quantité obtenue et la quantité désirée pour chaque secteur.
- Le **critère de compacité** mesure la compacité des secteurs obtenus.

La mesure du critère d'équilibre s'exprime en fonction de chacun des ratios entre la quantité obtenue et la quantité désirée pour chaque secteur. On peut le considérer comme la moyenne de ces ratios ou dans un cas plus restrictif comme le pire ratio. Au contraire, on mesure souvent le critère de compacité de manière global, on utilise généralement le nombre d'arcs coupés (*edge cut*), c'est-à-dire le nombre d'arcs reliant des objets situés dans des secteurs différents, les arcs entre les objets inclus dans un même secteur étant ignorés. En effet, moins les formes des secteurs sont compacts plus le nombre d'arcs coupés augmente.

Dans le cadre des algorithmes que nous allons utiliser, le critère de compacité est pour ainsi dire « directement minimiser » lors de la construction des secteurs. Nous nous intéressons donc plutôt au critère d'équilibre, en mesurant l'**écart maximal** et l'**écart moyen**. Toutefois, il arrive malheureusement que les secteurs produits ne soient pas d'un seul bloc, nous avons donc rajouté en priorité un test de **contiguité** des secteurs.

Ainsi, la mesure de la qualité se fonde sur plusieurs critères de mesure de qualité d'une solution *Sol.* Nous les avons classé par ordre d'importance décroissante :

- La **contiguïté** $Cont$: chaque secteur doit être formé d'un seul tenant, il n'y a donc pas de trous non plus. Si c'est le cas $Cont = 0$ sinon $Cont = 1$
- L'**écart maximal** $Ec_{\max}$: le plus grand écart $Ec_{\max} = \max_{S_i \in Secteurs(Sol)} (Ec(S_i))$ avec $Ec(S_i)$ l'écart du secteur S_i de la sectorisation *S*
- L'**écart moyen** $Ec_{moy} = \frac{1}{nbSecteurs(Sol)} \sum_{S_i \in Secteurs(Sol)} Ec(S_i)$

L'écart $Ec(S_i)$ d'un secteur S_i peut se définir comme le ratio entre $Quant\_réel(S_i)$ la quantité réellement contenue dans le secteur S_i et $Quant\_théo(S_i)$ la quantité que doit contenir le secteur S_i :

Si $Quant\_réel(S_i) > Quant\_théo(S_i)$ $\quad Ec(S_i) = \frac{Quant\_réel(S_i)}{Quant\_théo(S_i)} - 1$

Sinon $\quad Ec(S_i) = \frac{Quant\_théo(S_i)}{Quant\_réel(S_i)} - 1$

On a donc toujours $Ec(S_i) \geq 0$. Ainsi, Tous les critères ainsi définis sont d'autant meilleurs que leur valeur est faible et proche de zéro.

L'algorithme suivant permet de comparer la qualité de deux sectorisations en tenant compte de la hiérarchie des critères de qualité. Ainsi, il procède de la manière suivante : La meilleure solution est celle qui est contiguë. Sinon en cas d'égalité la meilleure solution est celle qui a l'écart maximal le plus faible. Sinon, en cas de nouvelle égalité, la meilleure solution est celle ayant l'écart moyen le plus faible. Sinon, en cas de nouvelle égalité, les deux solutions sont de qualité équivalente. Cet algorithme s'écrit de la façon suivante :

ComparerSolutions(Q_1 : Qualité, Q_2 : Qualité, *Résultat* : {inferieur, égal, superieur})
{

- Si $Q_1.Cont < Q_2.Cont$ Alors
  - $Résultat \leftarrow inférieur$
- FinSi
- Si $Q_1.Cont > Q_2.Cont$ Alors
  - $Résultat \leftarrow supérieur$
- FinSi
- Si $Q_1.Cont = Q_2.Cont$ Alors
  - Si $Q_1.Ec_{\max} < Q_2.Ec_{\max}$ *Alors*
    - $Résultat \leftarrow inférieur$
  - FinSi
  - Si $Q_1.Ec_{\max} > Q_2.Ec_{\max}$ *Alors*
    - $Résultat \leftarrow supérieur$
  - FinSi
  - Si $Q_1.Ec_{\max} = Q_2.Ec_{\max}$ *Alors*
    - Si $Q_1.Ec_{moy} < Q_2.Ec_{moy}$ *Alors*
      - $Résultat \leftarrow inférieur$
    - FinSi
    - Si $Q_1.Ec_{moy} > Q_2.Ec_{moy}$ *Alors*
      - $Résultat \leftarrow supérieur$
    - FinSi
    - Si $Q_1.Ec_{moy} = Q_2.Ec_{moy}$ *Alors*
      - $Résultat \leftarrow égal$
    - FinSi
  - FinSi
- FinSi

}

Fig. 67 –Méthode permettant d'évaluer la meilleure sectorisation entre deux sectorisations.

4.2 Sectorisation équilibrée

La sectorisation équilibrée consiste en la création de secteurs de « taille » égale. La taille est dans ce cas une quantité à partager : surface, population, clients, … De plus, la localisation des secteurs n'est pas définie à l'avance. La sectorisation équilibrée est toutefois un problème largement étudié en théorie des graphes sous le nom de partitionnement de graphes (*Graph Partitioning*). Nous verrons l'adaptation d'une méthode existante pour la sectorisation des données géographiques. Ce faisant, nous établirons un algorithme itératif afin de tester plusieurs solutions et de retenir la meilleure.

Notre algorithme de base se résume en trois étapes :

- Transformation des données géographiques en un graphe.
- Partitionnement équilibré avec le partitionnement multi-niveaux direct en *k* partitions au niveau le plus compacté (*Multilevel k-way Partitioning*).
- Affectation des objets géographiques au secteur correspondant.

Nous utilisons le programme *ParMETIS* [*KSK03*] et sa fonction *ParMETIS_V3_PartKway* pour réaliser la sectorisation. La particularité de ce programme est que l'amélioration de la forme est prise en charge par une autre méthode *ParMETIS_V3_RefineKway*. Ainsi, nous pourrons utiliser l'amélioration de la forme sur une sectorisation existante quelconque comme nous le verrons ultérieurement dans le chapitre relatif aux expérimentations et applications.

Cependant, cette première version ne nous a pas donné entièrement satisfaction. L'inconvénient majeur de cette méthode est l'existence d'un grand nombre de bonnes sectorisations très différentes. De ce fait, en initialisant aléatoirement l'algorithme, les résultats obtenus d'une fois sur l'autre sont très différents, ce qui est assez déroutant pour un utilisateur s'attendant à une solution unique. Cela peut être contrôlé en initialisant l'algorithme toujours de la même façon. De plus, pour mieux tirer parti de la possibilité d'obtenir des solutions variées, nous avons modifié notre algorithme initial afin de tester plusieurs solutions (100 itérations par défaut) et de garder la meilleure. De plus, nous pouvons spécifier une qualité à atteindre pour éviter de tester d'autres solutions. Par défaut, cette qualité définie que la sectorisation doit être contiguë, possèder au plus un écart maximal de 10 % et au plus un écart moyen de 5% pour la quantité à partager.

Notre nouvel algorithme est le suivant :

- Etape 1 :
  - Transformation des données géographiques en un graphe.
- Etape 2 : Trouver la meilleure sectorisation
  - Initialisation des paramètres :
    - *SolMeilleure* : la meilleure solution avec *SolMeilleure.Qualite* sa qualité.
    - *QualitéMin* : La qualité minimale à atteindre
    - *MaxIteration* : le nombre maximal de solutions évaluées
    - *Compteur* : le compteur
    - *SolMeilleure := Vide*
    - *QualitéMin := {Contigu,* $Ec_{max} = 10\%$, Ec_{moy}=5%*}*
    - *MaxIteration :=100*
    - *Compteur :=1*
  - Tant que *Compteur< MaxIteration* et *SolMeilleure.Qualite< QualitéMin* Faire
    - *SolAct* : la solution actuelle
    - Faire un partitionnement équilibré avec le partitionnement multi-niveaux direct en *k* partitions au niveau le plus compacté (*Multilevel k-way Partitioning*) et mettre le résultat dans *SolAct*.
    - Calculer la qualité de la solution actuelle, c'est-à-dire *SolAct.Qualite*
    - Si *SolAct.Qualite> SolMeilleure.Qualite* Alors
      - *SolMeilleure := SolAct*
    - Fin Si
  - Fin TantQue
- Etape 3 :
  - Affectation des objets géographiques aux secteurs correspondant de la meilleure sectorisation.

Fig. 68 –Algorithme itératif permettant d'évaluer plusieurs sectorisations issues de partitionnements multi-niveaux et de garder la solution ayant la meilleure qualité.

Nous exposerons les résultats obtenus avec cet algorithme dans le chapitre sur les expérimentations et applications.

4.3 Sectorisation à partir de centres

La sectorisation à partir de centres permet de créer des secteurs localisés autour de « centres ». Un centre peut être un magasin, un dépôt, une ville, … De plus, il est possible de spécifier la « taille » à atteindre pour chaque secteur. Nous verrons dans un premier temps la méthode basique que nous avons retenue pour effectuer ce type de sectorisation. Nous verrons ensuite une version améliorée de l'algorithme, il s'agit d'un algorithme itératif se basant sur la sectorisation de l'étape précédente pour déterminer des nouveaux paramètres ayant pour but de produire une meilleure sectorisation à l'étape suivante. Les sectorisations sont alors de meilleure qualité que celles obtenues avec la méthode basique.

4.3.1 Généralités

Contrairement à la *Sectorisation Équilibrée* qui est sans contrainte sur le positionnement des secteurs, ce problème part de la contrainte de positionnement du centre des secteurs. De fait,

c'est un sujet assez peu traité. L'approche la plus commune est la suivante. Chaque secteur de départ est vide. On initialise chaque secteur en lui assignant l'objet géographique le plus proche de son centre. Puis itérativement, on assigne à chaque secteur ses voisins immédiats. Ainsi, la contrainte de centralité est très forte et la forme des secteurs est en principe compacte. Par contre, le respect de la quantité à atteindre passe au second plan. C'est pourquoi, à partir de cet algorithme de base, il convient de greffer diverses techniques et astuces afin d'essayer de respecter le critère de quantité.

Nous allons maintenant voir la méthode originale que nous avons développé afin de réaliser une sectorisation à partir de centres.

4.3.2 Algorithme

Le but est de construire des secteurs à partir de centres. Notre algorithme est une évolution de l'algorithme de base décrit précédemment. La principale modification vient du découpage de l'algorithme en plusieurs étapes (par défaut une centaine) qui assurent une croissance homogène et progressive des secteurs. En effet, le principal inconvénient de l'algorithme de base est qu'il ne « vérifie » pas si la croissance des secteurs est homogène. C'est-à-dire que, la quantité étant inégalement répartie sur le territoire, la quantité de certains secteurs peut croître très vite tandis que d'autres peinent à voir leur quantité augmenter. Ce qui, au final aboutira à de très grandes disparités, certains secteurs ayant une quantité trop importante et d'autre trop faible. Ainsi, pour assurer une augmentation homogène de la quantité dans chaque secteur, nous découpons l'algorithme en plusieurs étapes.

Nous avons *m* objets géographiques O_i, *n* centres C_j et *n* secteurs vides S_j. Nous avons aussi :

- $QuantitéThéo(S_j)$ est la quantité à atteindre (quantité théorique) pour un secteur S_j. C'est une donnée du problème.
- $Quantité(O_i)$ est la quantité d'un objet géographique O_i. C'est aussi une donnée du problème.
- $QuantitéRéelle(S_j)$ est la quantité actuelle (quantité réelle) pour un secteur S_j. On a : $QuantitéRéelle(S_j) = \sum_{O_i \in S_j} Quantité(O_i)$

Ainsi, à chaque étape e_k nous définissons la quantité théorique à atteindre pour le secteur S_j :

$$QuantitéThéoEtape(S_j, e_k) = QuantitéThéo(S_j) \times \frac{k}{nb\_étapes}$$

Avec $nb\_étapes$, le nombre d'étapes.

Par exemple, si nous avons $nb\_étapes = 100$, alors à l'étape 50, chaque secteur devra atteindre la moitié (50%) de la quantité final à atteindre. Ainsi, chaque secteur peut croître tant qu'il n'est pas trop proche de la quantité à atteindre pour l'étape en cours (cela peut être un peu au dessus ou un peu au dessous). L'algorithme de création des secteurs est donc le suivant :

- Initialisation
  - Initialisation des *n* secteurs S_j qui sont vides : $S_j := \varnothing$
  - Initialisation de la frontière des secteurs F_j : $F_j := \{O\}$ avec *O* l'objet géographique le plus proche du centre C_j
  - Initialisation des objets non attribués NA, c'est-à-dire tous : $NA := \{O_1,...,O_m\}$
  - Initialisation de l'étape *e* : $e := 1$
- Itération des étapes
  - TantQue $NA \neq \varnothing$ Faire
    - Pour $j = 1$ à n Faire
      - Initialisation d'une variable booléenne *arret* : $arret := faux$
      - TantQue $(F_j \neq \varnothing)$ et $(arret = faux)$ Faire
        - Trouver l'objet *O* de la frontière F_j le plus proche du centre : c'est-à-dire tel que $O \in F_j$ et $dist(O, C_j) = \min_{O_k \in F_j}(dist(O_k, C_j))$
        - Calcul de l'écart actuel : $EcartActuel := QuantitéThéoEtape(S_j, e) - QuantitéRéelle(S_j)$
        - Calcul de l'écart après l'ajout : $EcartAprès := QuantitéRéelle(S_j \cup \{O\}) - QuantitéThéoEtape(S_j, e)$
        - Si $(EcartActuel > EcartAprès)$ Alors
          - L'attribuer au secteur S_j : $S_j := S_j \cup \{O\}$
          - L'enlever des objets non attribués NA : $NA := NA - \{O\}$
          - L'enlever des frontières de tous les secteurs : $\forall k$, $F_k := F_k - \{O\}$
          - Calculer la nouvelle frontière F_j en y ajoutant les objets non attribués ayant O pour voisin : c'est-à-dire $\forall O_l \in NA$, Si O_l voisin de O alors $F_j := F_j \cup \{O_l\}$
        - Sinon
          - Il ne faut plus ajouter d'objet pour le moment : $arret := vrai$
        - Fin Si
      - Fin TantQue
    - Fin Pour
    - $e := e + 1$
  - Fin TantQue

Fig. 69 –Algorithme de la sectorisation à partir de centres procédant à un découpage en étape

On remarquera plusieurs choses. Premièrement, notre algorithme est rapide en raison de sa complexité linéaire par rapport au nombre d'objets car à chaque itération, un objet est incorporé à un secteur. Deuxièmement, le processus s'arrête quand il n'y a plus d'objets non attribués, car le nombre d'étapes effectuées est presque systématiquement supérieur au nombre d'étapes théoriques. En effet, certains secteurs se retrouveront bloqués (il n'y a plus d'objets disponibles à leur frontière) et ils seront déficitaires, tandis que d'autres secteurs disposeront encore d'objets disponibles à leur frontière après la « dernière étape » et ils devront les absorber. Ces secteurs seront excédentaires.

4.3.3 Utilisation de la pondération des centres

Plutôt que de spécifier pour chaque secteur la quantité à atteindre, nous utilisons des pondérations affectées à chaque centre. Ainsi la quantité à atteindre pour chaque secteur est proportionnelle au poids de son centre. Cela permet d'être sûr de partager effectivement toute la quantité disponible.

Nous avons donc :

- Pour chaque centre C_j, $Poids(C_j)$ est le poids du centre.
- Le poids total $PoidsTotal = \sum_{j=1}^{n} Poids(C_j)$
- La quantité totale à partager $QuantitéTotale = \sum_{i=1}^{m} Quantité(O_i)$

À partir de là, nous pouvons définir la part de chaque secteur :

- $Part(S_j) = \frac{Poids(C_j)}{PoidsTotal}$

Et donc la quantité à atteindre pour chaque secteur est :

- $QuantitéThéo(S_j) = Part(S_j) \times QuantitéTotale$

Nous allons maintenant voir une amélioration de notre algorithme initial.

4.3.4 Algorithme itératif aléatoire

Le but de cet algorithme est de modifier de manière en partie aléatoire la quantité demandée pour chaque secteur afin d'obtenir une sectorisation de meilleure qualité. En effet, il est très fréquent que la quantité obtenue pour chaque secteur diffère significativement de la quantité demandée. Pour remédier à ce problème, nous envisageons de « corriger » les quantités demandées en analysant le résultat d'une première sectorisation:

- Si le secteur est excédentaire (par rapport à la quantité demandée), alors nous allons diminuer la quantité demandée.
- Si le secteur est déficitaire, alors nous allons augmenter la quantité demandée.

En faisant l'hypothèse qu'il existe une relation de proportionnalité entre l'objectif de taille et la quantité obtenue pour chaque secteur, nous pouvons définir le ratio entre la quantité demandée et la quantité obtenue. Puis à partir de ce ratio nous pouvons définir une nouvelle quantité demandée idéale permettant d'atteindre l'objectif de taille. On a donc :

$$Ratio_k(S_j) = \frac{QuantitéDemandée_k(S_j)}{QuantitéObtenue_k(S_j)}$$

$$QuantitéDemandéeIdéale_{k+1}(S_j) = QuantitéThéo(S_j) \times Ratio_k(S_j)$$

$$= \frac{QuantitéThéo(S_j) \times QuantitéDemandée_k(S_j)}{QuantitéObtenue_k(S_j)}$$

On remarque que ce processus est itératif. L'initialisation se fait en demandant directement l'objectif de taille :

$$QuantitéDemandée_1(S_j) = QuantitéThéo(S_j)$$

Ainsi, par exemple, dans le cadre d'une sectorisation de la population de la France autour de 8 centres urbains, nous pouvons calculer la quantité à demander idéale pour le secteur de *Lille* :

$$QuantitéDemandéeIdéale_2(Lille) = QuantitéThéo(Lille) \times Ratio_1(Lille)$$
$$= 7252364 \times 1{,}15 = 8340218 \text{ habitants}$$

Ainsi, pour que le secteur de *Lille* atteigne 7 252 364 habitants, nous devrions demander 8 340 218 habitants.

Cependant, la quantité disponible étant fixe, les secteurs se comportent comme des vases communiquant et si un secteur prend plus (ou moins), ce sont les secteurs voisins qui en subiront les conséquences en ayant réciproquement moins (ou plus). Toutefois, nous avons expérimenté que dans la majorité des cas, afin d'atteindre la quantité souhaité, il faut prendre une valeur comprise entre la *Quantité Théorique* et la *Quantité Demandée Idéale*, cette dernière pouvant être vue comme une borne à ne pas dépasser.

Par ailleurs, afin que chaque secteur tende vers la quantité théorique tout en essayant d'éviter les effets perturbateurs. Nous allons essayer de privilégier les valeurs proches de la *Quantité Théorique* « tout en allant dans la direction » de la *Quantité Demandée Idéale*. Pour cela, nous utilisons une variable aléatoire *X* dont les valeurs sont comprises entre 0 et 1 :

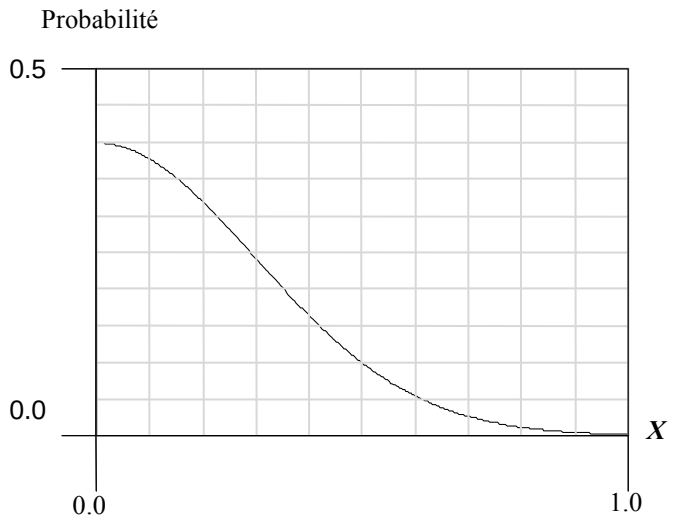


Fig. 70 –Fonction de probabilité des valeurs de la variable X

Nous pouvons alors déterminer pour chaque secteur la quantité demandée aléatoirement.

Ainsi, pour *X=0*, évènement le plus probable, nous aurons :

$$QuantitéDemandéeAléatoire_{k+1}(S_j) = QuantitéThéo(S_j)$$

Et, pour *X=1*, évènement le moins probable, nous aurons :

$$QuantitéDemandéeAléatoire_{k+1}(S_j) = QuantitéDemandéeIdéale_{k+1}(S_j)$$

La formule générale en fonction de la valeur de *X* est la suivante :

$$QuantitéDemandéeAléatoire_{k+1}(S_j)$$
$$= QuantitéThéo(S_j) + X \times (QuantitéDemandéeIdéale_{k+1}(S_j) - QuantitéThéo(S_j))$$
$$= QuantitéThéo(S_j) + X \times (QuantitéThéo(S_j) \times Ratio_k(S_j) - QuantitéThéo(S_j))$$
$$= QuantitéThéo(S_j) \times (1 + X \times (Ratio_k(S_j) - 1))$$

Toutefois, nous ne sommes pas assurés que les quantités demandées correspondent à la quantité totale à partager qui est disponible. C'est pourquoi nous devons les normaliser :

$$QuantitéDemandée_{k+1}(S_j)$$
$$= QuantitéDemandéeAléatoire_{k+1}(S_j) \times \frac{QuantitéTotale}{QuantitéDemandéeAléatoireTotale_{k+1}}$$

Avec $QuantitéDemandéeAléatoireTotale_k = \sum_{j=1}^{n} QuantitéDemandéeAléatoire_k(S_j)$

À partir des nouvelles Quantités Demandées, nous calculons la sectorisation correspondante. Ce processus peut être itéré autant que souhaité. L'algorithme itératif de la sectorisation à partir de centres est le suivant :

- Initialisation
  - Initialisation de la meilleure Sectorisation : $SectoMeilleure := \emptyset$
  - Initialisation des Quantités Demandés pour chaque secteur S_j :
    $$QuantitéDemandée(S_j) = QuantitéThéo(S_j)$$
- Itération des itérations
  - Pour *i=1 à max_itérations* Faire
    - Calculer la Sectorisation Actuelle $SectoAct$ à partir des Quantités Demandées
    - Si $SectoAct.Qualité > SectoMeilleure.Qualité$ Alors
      - Remplacer $SectoMeilleure$ par $SectoAct$: $SectoMeilleure := SectoAct$
    - Fin Si
    - Calculer les nouvelles Quantité Demandées pour chaque secteur à partir de la méthode proposée
  - Fin Pour

Fig. 71 –Algorithme itératif aléatoire de la sectorisation à partir de centres

4.3.5 Conclusion

Dans le cadre de la sectorisation à partir de centres, nous avons mis au point un nouvel algorithme efficace. Cet algorithme permet d'obtenir des résultats acceptables à partir de données réelles telles que la population. Nous exposerons les résultats obtenus avec cet algorithme dans le chapitre sur les expérimentations et applications.

4.4 Rééquilibrage des secteurs

Lorsque nous disposons d'une sectorisation existante, il est parfois plus intéressant d'améliorer cette sectorisation plutôt que d'en calculer une nouvelle. Le rééquilibrage des secteurs vise donc à « améliorer » les quantités de chaque secteur en transférant des objets géographiques entre les secteurs. Ceci dans le but que la quantité de chaque secteur se rapproche de la quantité idéale.

Nous avons vu dans le chapitre concernant l'état de l'art une méthode efficace pour calculer les transferts à réaliser entre les secteurs. Nous allons maintenons détailler l'algorithme que nous avons conçu pour effectuer les transferts de façon progressive en convergeant vers la solution de rééquilibrage.

4.4.1 Algorithme utilisé

Pour effectuer le rééquilibrage, nous utilisons le même type d'algorithme que pour la sectorisation à partir de centres : il s'agit d'un algorithme procédant par étape, à chaque étape on transfert une fraction de la quantité totale à transférer. L'intérêt de cet algorithme est d'éviter d'avoir à ordonner les transferts à effectuer car ils sont tous simultanément et progressivement effectués, étape par étape.

Ainsi, à chaque étape e_k, nous définissons la quantité transférée à atteindre entre les secteurs S_i et S_j. Cette quantité est proportionnelle à l'étape en cours et à la quantité T_{ij} à transférer entre du secteur S_i vers le secteur S_j :

$$T_{ij}Fraction(e_k) = T_{ij} \times \frac{k}{nb\_étapes}$$

Il y a *n* secteurs et le rééquilibrage s'effectue de la manière suivante en transférant les objets ayant le meilleure gain (nous allons définir par la suite le gain) :

- Initialisation
  - Initialisation des quantités transférées : $\forall i , \forall j , T_{ij}Effectuée := 0$
- Pour *e :=1* à *nb_étapes* Faire
  - Pour i *:=1* à *n Faire*
    - Pour *j = 1* à *n* Faire
      - Calculer la frontière F_{ij} qui contient les objets transférables entre S_i et S_j, c'est-à-dire les objets de S_i en contact avec des objets de S_j
      - Initialisation d'une variable booléenne *arret* : $arret := faux$
      - Tant Que $(F_{ij} \neq \varnothing)$et $(arret = faux)$ Faire
        - Sélectionner l'objet *O* de la frontière F_{ij} ayant le meilleur gain, c'est-à-dire tel que $O \in F_{ij}$ et $Gain_{ij}(O) = \min_{O_k \in F_j}(Gain_{ij}(O_k))$
        - Si $(T_{ij}Effectuée + O.Quantité < T_{ij}Fraction(e))$ Alors
          - L'ajouter au secteur S_j : $S_j := S_j \cup \{O\}$
          - L'enlever du secteur S_i : $S_i := S_i - \{O\}$
          - L'enlever de la frontière commune: $F_{ji} := F_{ji} - \{O\}$
          - Mettre à jour la quantité transférée entre les deux secteurs: $T_{ij}Effectuée := T_{ij}Effectuée + O.Quantité$
        - Sinon
          - Il ne faut plus ajouter d'objet pour le moment : $arret := vrai$
        - Fin Si
      - Fin Si
    - Fin Pour
  - Fin Pour
- Fin Pour

Fig. 72 –Algorithme de la mise en œuvre du rééquilibrage procédant par étapes

L'algorithme a une complexité linéaire par rapport au nombre de transferts non nul à effectuer ($T_{ij} > 0$). Le nombre de transferts non nul à effectuer est lui-même proportionnel au nombre de secteurs et varie entre 2 et 3 fois le nombre de secteurs.

Nous allons maintenant voir les deux méthodes utilisées pour calculer le gain. La première est celle de Fiduccia-Mattheyses car elle est simple à mettre en œuvre et est rapide. La deuxième est une modification de cette première méthode et utilise les distances à vol d'oiseau par rapport aux centres. Bien évidemment, cette dernière méthode n'est utilisable que dans le cas d'une sectorisation avec des centres.

Afin de bien saisir le calcul des gains, nous utilisons un exemple simple composé de 2 secteurs, le secteur Est et le secteur Nord. Ces secteurs sont composés de départements et l'on doit transférer 1 500 000 habitant du secteur Est vers le secteur Nord. La carte suivante montre que seuls deux objets géographiques de la frontière (ceux rayés) sont initialement transférables.

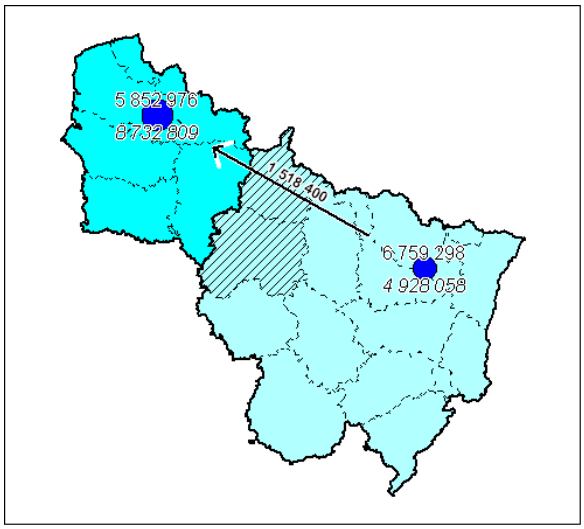


Fig. 73 –Objets transférables du secteur Est au secteur Nord

La méthode de Fiduccia-Mattheyses a pour but de transférer les objets géographiques beaucoup plus liés au secteur d'arrivé qu'au secteur de départ : le gain est la différence entre le nombre de voisins de l'objet géographique dans le secteur d'arrivé et le nombre de voisins dans le secteur de départ.

L'exemple suivant montre le calcul du gain pour les départements situés à la frontière.

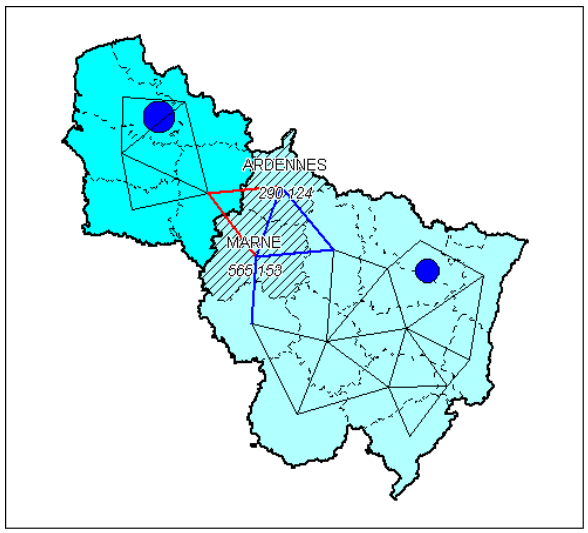


Fig. 74 –Liens entre les objets transférables et le secteur Est et le secteur Nord

Les liens avec le secteur d'arrivée sont indiqués en rouge et les liens avec le secteur de départ en bleu, les gains sont les suivants :

GainVois(Ardennes) = 1 - 2 = -1

GainVois(Marne) = 1 - 3 = -2

Ainsi, pour le transfert, les Ardennes sont choisies.

La méthode du gain selon la distance aux centres n'est utilisable que pour des secteurs ayant un centre. Le but est alors de mettre en avant le critère de centralité des secteurs dans le choix du meilleur objet transférable. Le gain d'un objet géographique est alors égal à la différence entre sa distance au centre du secteur de départ et sa distance au centre du secteur d'arrivée.

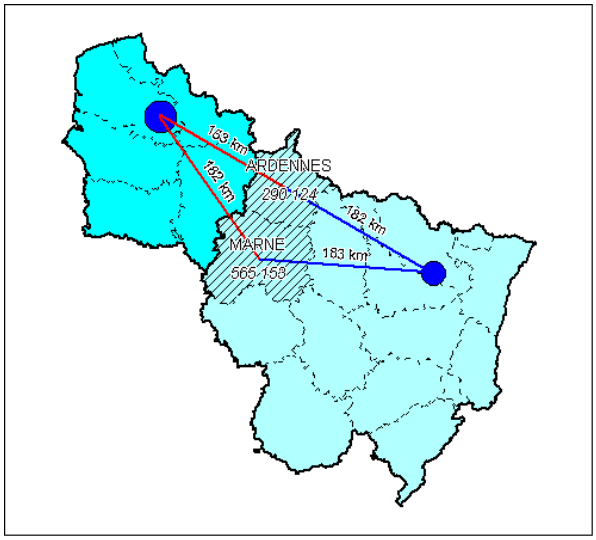


Fig. 75 –Distances entre les objets transférables et le centre du secteur Est et le centre du secteur Nord

Les gains sont les suivants :

GainDist(Ardennes) = 182 km – 163 km = 19 km

GainDist(Marne) = 183 km – 182 km = 1 km

De nouveau le département des Ardennes possède le meilleur Gain.

Dans le chapitre concernant les expérimentations et les applications, nous verrons les résultats produits par notre algorithme de rééquilibrage.

4.4.2 Conclusion

La méthode de rééquilibrage permet d'améliorer les sectorisations décrites précédemment. Nous avons innové en développant une méthode de réauilibrage progressif et en prenant en compte les centres des secteurs. Ce dernier point permet d'adapter le rééquilibrage aux résultats de la sectorisation à partir de centres. Une amélioration possible est l'utilisation de l'algorithme de Krishnamurthy. Une autre solution est d'effectuer les transferts dans un ordre précis en commençant par les transferts des secteurs donneurs pour finir par les transferts des secteurs receveurs en passant par les transferts des secteurs qui donnent et reçoivent.

L'exemple suivant montre l'ordre « optimal » des transferts pour le cas de figure envisagé :

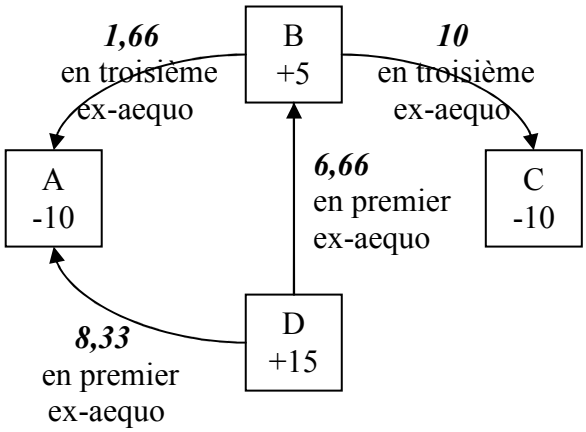


Fig. 76 –Ordre « optimal » des transferts

4.5 Conclusion

La *Sectorisation Équilibrée* tire profit de l'existant en s'appuyant sur une méthode de partitionnement récursive et multi-échelles ayant fait ses preuves dans le domaine du partitionnement de graphe.

Quant à la *Sectorisation à partir de Centres*, nous avons défini un algorithme itératif assez simple mais offrant tout de même d'assez bons résultats.

Les résultats produits par ces méthodes ne sont pas toujours bons. Cela nous a naturellement amené aux méthodes de *Rééquilibrage* qui permettent d'atteindre l'équilibre en évitant au maximum de toucher à la forme des secteurs. Nous avons innové en développant une méthode progessive spécialement adaptée aux sectorisations à partir de centres.

Finalement, la *Sectorisation Équilibrée*, la *Sectorisation à partir de Centres* et le *Rééquilibrage de Sectorisations* sont des méthodes complémentaires permettant de répondre à la plupart des problématiques de sectorisation.

Nous envisageons d'améliorer les résultats des sectorisations en nous focalisant sur le rééquilibrage. Dans ce cas de figure, nous considérerons que les solutions données par la *Sectorisation Équilibrée* et la *Sectorisation à partir de Centres* donnent des sectorisations « brutes » et que le résultat final est donné après le *Rééquilibrage*. C'est pourquoi nous pensons utiliser l'algorithme de Krishnamurthy afin de créer un outil de *Rééquilibrage des Secteurs* plus efficace.

Conclusions

Dans ce chapitre, nous avons vu les améliorations et contributions que nous avons réalisées. Cela concerne les points suivants : la *Classification de Données pour de grands volumes de données mixtes*, la *Visualisation de Classifications*, la *Détermination de Pôles et leur Hiérarchisation*, la *Sectorisation*.

Tout d'abord, nous avons mis au point un algorithme de classification incrémentale efficace : la Classification Ascendante Approximative (CAA) et son extension, la Classification Ascendante Hiérarchique Approximative (CAHA). Ces méthodes ont une complexité linéaire par rapport au nombre d'objets à classer. En dehors de la mesure de dissimilarité, ces deux méthodes ne nécessitent que le choix du paramètre k fixant le nombre de résumés utilisés en mémoire, ce qui les rend très simple à paramétrer. Nous avons aussi défini et utilisé une mesure de dissimilarité permettant d'utiliser l'algorithme sur des données mixtes (à la fois quantitatives et qualitatives). Le principal inconvénient actuel de notre algorithme de la CAA est le paramétrage de la distance via des pondérations qui ne sont pas forcément adaptées, c'est pourquoi nous envisageons d'étudier le remplacement de la distance euclidienne pondérée actuellement utilisée par la distance de Mahalanobis. Nous verrons dans le chapitre suivant, les expérimentations permettant de tester la CAA avec des données fictives et des données réelles.

En ce qui concerne la *Visualisation de Classifications*, nous avons élaboré une visualisation hiérarchique évoluée utilisant des profils très lisibles car ils sont purement graphiques et ne comportent aucune information chiffrée. Nous avons aussi élaboré un tableau évolué afin de compenser une lacune de la visualisation hiérarchique évoluée : la difficulté de comparer rapidement les informations disponibles pour une même variable. Ainsi, elle est complémentaire de la visualisation hiérarchique évoluée. Nous avons aussi traité le problème de l'ordre optimal à choisir pour les variables et les classes. Nous verrons dans le chapitre suivant des exemples d'applications sur des données réelles.

Pour la *Détermination des Pôles et leur Hiérarchisation*, nous avons d'abord présenté la *Détermination des Pôles* qui permet de résumer une carte de lissage spatial par quelques pôles. Nous avons développé une méthode d'affinage au cours de laquelle les pôles évoluent jusqu'à leur emplacement final et les pôles redondants sont supprimés. Grâce à la méthode de description automatique des pôles, il est possible de donner automatiquement un nom aux pôles. Ensuite, la *Hiérarchisation des Pôles* permet de superposer sur une même carte les pôles correspondant à des échelles différentes en les incluant dans une hiérarchie. La carte obtenue est alors très synthétique et la hiérarchie fournit de nouvelles informations explicatives : les pôles de l'échelle la plus courte permettent d'expliquer les pôles des échelles plus grandes. Nous verrons dans le chapitre suivant un exemple d'applications sur la répartition géographique de la population française.

Pour la partie traitant de la *Sectorisation*, nous avons adaptée une méthode de partitionnement de graphes éprouvée et l'avons ensuite encapsulée dans un algorithme itératif afin d'obtenir plusieurs sectorisations et de garder la meilleure. Nous avons ensuite développé la *Sectorisation à partir de Centres* qui permet de construire des secteurs autour de centres. Il s'agit là aussi d'un algorithme itératif assez simple : l'algorithme de création des secteurs construit les secteurs progressivement en se basant sur des paramètres initiaux, il est rapide car sa complexité est linéaire par rapport au nombre d'objets ; l'algorithme itératif recommence

autant que nécessaire ce processus en modifiant les paramètres initiaux et garde la meilleure sectorisation obtenue. Nous verrons dans le chapitre suivant que cela permet d'obtenir d'assez bons résultats. Pour finir, la méthode de *Rééquilibrage de Sectorisations* que nous avons présentée permet d'atteindre l'équilibre en évitant au maximum de toucher à la forme des secteurs. Cette méthode est rapide car sa complexité est linéaire par rapport au nombre de secteurs à rééquilibrer. Nous avons innové en concevant un algorithme permettant de réaliser progressivement et en parallèle les transferts. Nous verrons dans le chapitre suivant des applications pour la sectorisation de la population française et constaterons la complémentarité de ces trois approches.

Nous allons maintenant passer au chapitre suivant qui traite des expérimentations et applications de ces méthodes.

Chapitre 3

3. Expérimentations et Applications

Introduction

Dans le chapitre précédent, nous avons vu nos contributions concernant les quatre domaines suivants : la *Classification de Données pour de grands volumes de données mixtes*, la *Visualisation de Classifications*, la *Détermination de Pôles et leur Hiérarchisation*, la *Sectorisation.* Dans ce chapitre, nous allons voir les expérimentations et les applications concernant les trois points suivants : la *Classification de Données & la Visualisation de Classifications*, le *Lissage Spatial & la Détermination des Pôles*, et la *Sectorisation*.

Dans la partie traitant de la *Classification de Données & de la Visualisation de Classifications*, nous expérimentons d'abord notre algorithme de classification de grands volumes de données mixtes. Nous montrons que notre algorithme produit de bons résultats. Ensuite, nous présentons la *Classification de Variables* qui est une extension très intéressante de la *Classification de Données & de la Visualisation de Classification*, elle permet de réaliser une analyse des variables qui est complémentaire de l'analyse des données. Nous montrerons ensuite l'intérêt de la *Classification de Données* et la *Classification de Variables* en procédant à l'analyse des données socioprofessionnelles de Paris et de sa Petite Couronne. Nous terminerons cette partie par la *Classification Hiérarchique Spatiale* qui est une adaptation de la *Classification de Données*. Elle permet de découper de façon hiérarchique un territoire en secteurs situés autour de zones de fortes valeurs, il s'agit d'une méthode de sectorisation complémentaire de celles discutées dans la partie concernant la *Sectorisation*.

La partie suivante est consacrée aux applications du *Lissage Spatial & de la Détermination des Pôles.* Tout d'abord, nous utilisons la *Détermination et la Hiérarchisation des Pôles* pour analyser la répartition de la population française. Le résultat comporte les pôles pour l'échelle locale (50 km) et l'échelle nationale (130 km) avec la hiérarchie qui permet d'expliquer les pôles de l'échelles nationale à partir de ceux de l'échelle locale. Ensuite, nous montrons que le *Lissage Spatial* des données en prétraitement de la *Classification de Données* permet d'analyser les

données pour des échelles spatiales plus grandes et d'obtenir une carte des classes plus facile à analyser.

Enfin, la dernière partie concerne les applications des méthodes de *Sectorisation* : la *Sectorisation Équilibrée*, la *Sectorisation à partir de Centres* et le *Rééquilibrage de Sectorisations*. Nous abordons en premier le partage du territoire français en 22 secteurs de population égale et montrons que les résultats obtenus sont de qualité. Puis à l'aide de la *Sectorisation à partir de Centres*, nous constatons que notre méthode donne de bons résultats en terme de contiguïté et d'objectif de taille, mais que la compacité des secteurs n'est pas toujours optimale. L'amélioration de la compacité se faisant au détriment des objectifs de taille, nous montrons que le *Rééquilibrage de Sectorisations* permet d'atteindre les objectifs de taille sans trop altérer la compacité. C'est pourquoi son utilisation est intéressante sur des sectorisations bonnes en terme de compacité et de contiguïté, mais « moyennes » en terme d'objectif de taille.

1 Classification de Données & Visualisation de Classifications

Nous expérimenterons d'abord notre algorithme de Classification de grands volumes de données mixtes et nous le comparerons avec un autre algorithme de référence : les k-moyennes (ou K-Means) qui est lui aussi simple à mettre en œuvre et à paramétrer. Nous réaliserons des comparaisons en utilisant des données réelles (données socioprofessionnelles relatives à Paris et sa Petite Couronne) et des données simulées. Puis nous établirons ensuite la pertinence de l'utilisation de la CAHA par rapport à la CAA dans le cadre de la réduction du nombre de résumés. Finalement, nous testerons la CAA en version parallèle et nous montrerons qu'elle souffre d'un effet de « superposition », ce qui la rend peu efficace.

La partie suivante montre une extension de la *Classification de Données* et de sa visualisation : il s'agit de la *Classification de Variables*. Elle est basée sur la méthode développée pour la recherche de l'ordre optimal des variables décrite dans le chapitre précédent. Nous montrerons son intérêt dans une démarche d'analyse des variables qui est complémentaire de l'analyse des objets réalisée par la classification de données.

Nous montrerons ensuite l'intérêt des techniques de visualisation de classifications en procédant à une étude de cas à l'aide de la *Classification de Données* et la *Classification de Variables*: l'analyse des données socioprofessionnelles de Paris et de sa Petite Couronne.

La dernière partie présente la *Classification Hiérarchique Spatiale* qui est une adaptation de la *Classification de Données*. Elle permet de découper de façon hiérarchique un territoire en secteurs situés autour de zones de fortes valeurs.

1.1 Expérimentations et Comparaison avec les K- moyennes

Nous définirons en premier les indicateurs que nous allons utiliser pour mesurer la qualité des classifications. Puis nous effectuerons une première comparaison entre la CAA et les k-moyennes en utilisant des données réelles (données socioprofessionnelles relatives à Paris et sa Petite Couronne). Nous utiliserons ensuite des données simulées. Puis nous établirons ensuite la pertinence de l'utilisation de la CAHA par rapport à la CAA dans le cadre de la réduction du nombre de résumés. Finalement, nous testerons la CAA en version parallèle et nous montrerons qu'elle souffre d'un effet de « superposition » qui est très gênant et plus difficile à atténuer que l'effet de « dérive » déjà évoqué.

1.1.1 Indicateurs mesurant la qualité de la classification

Pour mesurer les performances de la CAA, nous la comparons avec l'algorithme standard des Nuées Dynamiques (k-moyennes ou *k-means*) car il est lui aussi simple à mettre en œuvre et à paramétrer, et fournit une référence pour la comparaison. Les variables utilisées sont des variables quantitatives (numériques) et nous pouvons donc utiliser les mesures de qualités suivantes :

- l'inertie perdue par les classes-résumés (Ip) : plus elle est importante, plus la classification est mauvaise.
- l'inertie maximale d'une classe-résumé (Imx) : elle indique l'inertie de la classe ayant la plus grande inertie. A priori, plus elle est importante, plus la classe en question résume « mal » les individus qu'elle contient. Cependant cela peut aussi être dû au fait qu'elle contient un très grand nombre d'individus.
- la variance maximale d'une classe-résumé (Vmx) : il s'agit de la variable ayant la plus grande variance dans une classe donnée.
- la distance interne maximale d'une classe-résumé (Dmx) : il s'agit de la classe ayant la plus grande distance entre son centre et un individu.
- le nombre d'individus mal classés (MC) qui traduit l'effet de dérive
- le temps de calcul (T) en secondes.

Tous ces indicateurs ont en commun la propriété suivante : plus ils sont petits, meilleure est la classification et réciproquement, plus ils sont grands, moins bonne est la classification.

Dans les résultats suivants, nous n'avons gardé que l'inertie perdue (Ip) et le nombre d'individus mal classés (MC) car ils sont les indicateurs les plus intéressants.

La CAA et les K-moyennes utilisent tous les deux la distance euclidienne pour calculer la distance entre les individus. Cependant, la CAA utilise la distance de Ward comme distance entre les classes et les K-moyennes utilisent la distance euclidienne entre le centre de la classe et un individu.

1.1.2 Test sur des données réelles

Les mesures ci-dessous ont été faites sur des données socioprofessionelles des zones de Paris et de sa Petite Couronne. Il y a 2739 zones (les individus) et 27 variables (qui forment une dizaine de groupes de variables corrélées). Les variables ont été centrées et réduites. L'Inertie Totale est donc de 73953 = 27 x 2739.

Les algorithmes testés sont :

- La CAA seule
- Les K-moyennes avec 5 itérations seulement
- La CAA avec en post-traitement les K-moyennes avec 5 itérations
- Les K-moyennes avec 20 itérations
- La CAA avec en post-traitement les K-moyennes avec 20 itérations

On remarquera que la CAA ne lit les données qu'une seule fois tandis que les K-moyennes nécessite une lecture de la totalité des données pour **chaque itération**, ce qui n'est généralement pas faisable sur les très grands volumes de données, mais que nous avons fait dans le cadre de nos tests.

Pour le test actuel, l'indicateur le plus pertinent est la perte d'inertie, les autres indicateurs ne faisant que confirmer celui-ci. Nous avons donc les résultats suivants pour différents niveaux de précision (c'est-à-dire le nombre de résumés demandés). Dans chaque cellule, on trouve en haut l'inertie perdue (Ip), sa valeur relative en pourcent entre parenthèses, et en bas est indiqué le nombre d'individus mal classés (MC).

Précision (ou nb de Résumés)	CAA	k-moy 5 itérs	CAA + k-moy 5 itérs	k-moy 20 itérs	CAA + k-moy 20 itérs
20	Ip = 33255 (45%) MC = 368	31885 (43%) 468	28787 (39%) 137	***27366 (37%)*** 33	27113 (37%) 36
50	23111 (31%) 417	25282 (34%) 468	20201 (27%) 123	21087 (29%) 33	***19552 (26%)*** 59
100	16350 (22%) 379	20920 (28%) 370	***14787 (20%)*** 88	17187 (23%) 63	14464 (20%) 6
200	***10935 (15%)*** 322	16580 (22%) 301	10278 (14%) 32	14908 (20%) 40	10218 (14%) 0
400	***7225 (10%)*** 171	12144 (16%) 191	6980 (9%) 10	11301 (15%) 9	6965 (9%) 0

Fig. 77 –Mesure de la qualité des résumés avec la perte d'inertie comme indicateur pour plusieurs valeurs de précision sur les données socioprofessionnelles de Paris et de sa Petite Couronne

Ce premier test montre que pour un nombre de résumés faibles (jusqu'à 20 résumés au moins), les K-moyennes sont meilleures que la CAA et que l'association de la CAA et des K-moyennes est la meilleure solution. En augmentant la précision, les k-moyennes n'apporte plus rien :

- Pour 50 résumés, le meilleur résultat est obtenu avec la CAA avec les K-moyennes avec 20 itérations
- Pour 100 résumés, la CAA avec les K-moyennes avec 5 itérations suffit.
- Pour 200 et 400 résumés, la CAA seule suffit.

De plus, pour un nombre d'individus mal classés équivalent, le résultat de la CAA est souvent meilleur, surtout pour des valeurs de *k* supérieures ou égales à 100.

1.1.3 Test sur des données simulées

Les tests suivants sont effectués sur des jeux de données générés. Chacun de ces jeux comporte 50 variables indépendantes et 10000 individus organisés de façon différente :

- Dans le jeu *A*, ils sont organisés en 50 classes,
- Dans le jeu *B*, ils sont organisés en 400 classes,
- Et dans le jeu *C*, ils sont organisé en 10000 classes (il y a en moyenne une classe par individu, c'est-à-dire qu'il n'y a pas de classification existante à redécouvrir).

Les individus d'une classe sont distribués selon une loi de probabilité Normale centrée sur le centre de la classe. Les classes n'ont pas exactement le même nombre d'individus. Les variables ont été centrées et réduites. L'Inertie Totale est donc de 500000 = 50 x 10000.

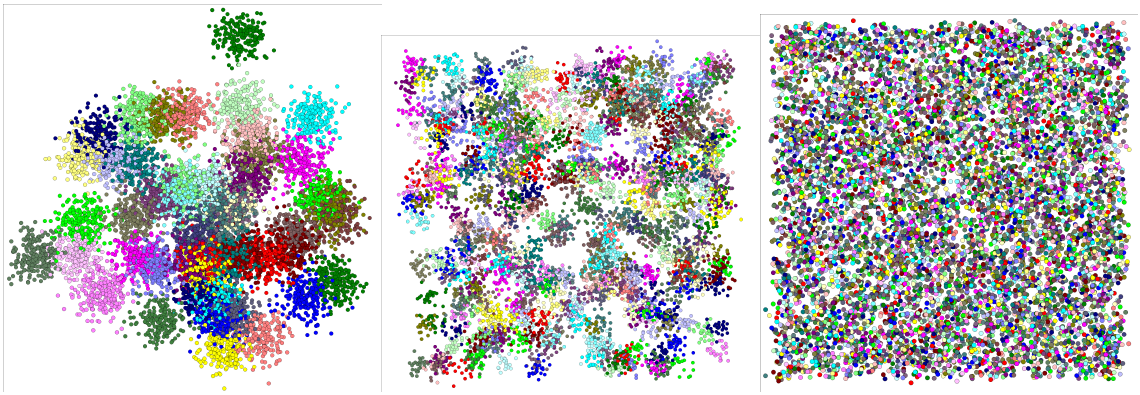


Fig. 78 –De gauche à droite, les visualisations selon les 2 premières variables (en tant que coordonnées) parmi les 50 variables des jeux de données A (50 classes), B (400 classes) et C (10000 classes)

Les résultats sont les suivants :

Jeu de données	Précision (ou nb de Résumés)	CAA	k-moy 5 itérs	CAA + k-moy 5 itérs	k-moy 20 itérs	CAA + k-moy 20 itérs
Jeu A 50 classes	100	Ip = **11623 (2%)** MC = 255	73397 (15%) 148	11595 (2%) 29	73373 (15%) 0	11592 (2%) 0
Jeu B 400 classes	100	**281071 (56%)** 22	374464 (75%) 229	280757 (56%) 0	360495 (72%) 172	280757 (56%) 0
Jeu C 10000 classes	100	**407919 (82%)** 825	427235 (85%) 1800	404225 (81%) 101	416338 (83%) 62	403617 (81%) 3

Fig. 79 –Mesure de la qualité des résumés avec la perte d'inertie comme indicateur pour les jeux de données A, B et C

Dans tous les cas, la CAA reste la meilleure. De plus, dans ce cas, l'utilisation des K-moyennes en post-traitement est inutile. La principale explication (voir l'annexe pour plus de détails) est que les K-moyennes « s'initialisent » avec de « mauvais individus » tandis que la CAA, via le mécanisme des fusions, s'affranchit rapidement de son initialisation et arrive ainsi à de meilleurs résultats.

1.1.4 Tests de la réduction du nombre de classes-résumés

Nous avons aussi cherché à savoir quelle était la meilleure façon d'obtenir des résumés :

- La première méthode consiste à utiliser comme précédemment la CAA ou les K-moyennes avec une précision égale au nombre de résumés souhaités.
- La seconde méthode (meilleure a priori) consiste à utiliser la CAA ou les k-moyennes avec une précision élevée puis à réduire le nombre de résumé jusqu'au nombre désiré en utilisant une CAH.

On remarquera que dans le cas où il s'agit d'utiliser la CAH sur le résultat de la CAA, nous réalisons en fait une CAHA décrite précédemment dans le chapitre des contributions.

Les tests ont été effectués avec une précision *k=100* pour un nombre de résumés souhaités *R=40*. Les données utilisées pour le test sont les données socioprofessionnelles de Paris et de sa Petite Couronne et le jeu de données A (50 classes).

Les résultats sont les suivants :

	CAA	K-moyennes (20 itérs)	CAA + K-moyennes (20 itérs)	CAA	K-moyennes (20 itérs)	CAA + K-moyennes (20 itérs)
Précision	100	100	100	40	40	40
Nb de résumés final	40	40	40	40	40	40
Paris	Ip = ***22491*** MC = 476	***21400*** 322	21542 297	25415 491	22352 52	***21208*** ***45***
Jeu A	***69856*** 0	94196 0	69856 0	81854 0	234846 9	_ _

Fig. 80 –Mesure de l'influence de la réduction pour les données IRIS de Paris et sa petite couronne et le jeu de données générées A.

On observe que comme attendu, il vaut mieux généralement passer par une étape où la précision est importante (CAA ou K-moyennes) puis de réduire le nombre de résumés obtenus jusqu'au nombre désiré en utilisant la CAH.

1.1.5 Test de la CAA parallélisée

La version parallélisée est testée avec 10 processus parallèles. Les tests suivants sont effectués sur le jeu de données généré *A* (50 classes) qui comporte 10000 individus et les données socioprofessionnelles de Paris et sa Petite Couronne qui comporte 2739 individus. Ainsi, pour le jeu de données A, les 10 processus parallèles vont résumer chacun 1000 individus (un dixième des individus) dans 100 classes-résumés. Cela fera un total de 1000 classes-résumés qui seront à leur tour résumées par un 11ème processus.

	CAA	K-moyennes (5 itérs)	CAA + K-moyennes (5 itérs)	CAA **parallèle**	CAA + K-moyennes (5 itérs) **parallèle**	CAA **parallèle**
précision	100	100	100	100	100	200
résumés	100	100	100	100	100	100
Jeu A 2 variables	Ip = 213 MC = 425	240 373	202 127	311 **2903**	200 110	260 **1962**
Jeu A 50 variables	11623 255	73397 148	11595 29	11738 **762**	11596 37	11711 **1003**
Paris	16350 379	20920 370	14787 88	16474 **622**	14062 18	16191 **472**

Fig. 81 –Mesure de la qualité de l'algorithme de la CAA parallèle avec comme indicateur l'inertie perdue (Ip) et le nombre de mal classés (MC)

Le principal défaut de l'algorithme parallèle est le nombre d'individus mal classés qui est très grand par rapport à l'algorithme standard de la CAA. Pour résoudre ce problème, nous utilisons les k-moyennes en post traitement de chaque processus de CAA parallèle, cela permet effectivement d'obtenir des résultats aussi bons ou meilleurs qu'avec l'algorithme standard. Si les volumes de données sont trop importants, l'emploi des K-moyennes n'est pas souhaitable, l'autre solution est alors d'augmenter la précision. Cependant, cela ne n'améliore pas ou très peu les résultats. La principale raison vient du fait que les résumés trouvés par les processus en parallèle se recoupent beaucoup. On peut appeler cela l'effet de « superposition ». Cet effet de « superposition » est alors le principal responsable des individus mal classés. Le schéma suivant illustre l'effet de « superposition ». Les individus sont répartis aléatoirement en deux groupes. Chaque groupe d'individu est classé en 3 classes. Lorsque l'on réunit les classifications, on s'aperçoit que certaines classes se superposent et que donc certains individus sont mal classés.

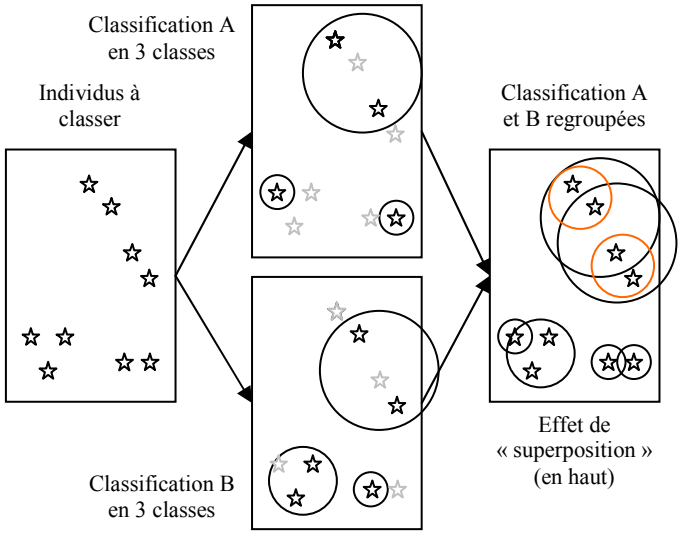


Fig. 82 –Illustration de l'effet de superposition

Une solution envisageable est l'utilisation des k-moyennes immédiatement après le regroupement des classifications. Cependant, cela n'est pas applicable pour les gros volumes de données. Ainsi, à précision équivalente, la version parallèle de la CAA donne de moins bons résultats que la version standard à cause de l'effet de « superposition ». Cet effet peut toutefois être en parti compensé par une augmentation de la précision (le nombre de classes-résumés demandés).

1.1.6 Conclusion des tests

Les expérimentations ont montré que notre algorithme produit des résultats meilleurs que l'algorithme des k-moyennes. Nous avons aussi établi que dans le cadre d'une utilisation normale, une précision *k=100* est largement suffisante pour réaliser une classification fiable comprenant au final moins d'une vingtaine de classes. De plus, l'effet de « dérive » constaté n'est important que pour des faibles valeurs de *k* (40 et moins). Par contre, les tests ont montré que notre version parallèle de la CAA souffre fortement de l'effet de « superposition », ce qui la rend peu intéressante.

1.2 Autre application : la Classification de Variables

Dans le chapitre sur les contributions, pour trouver l'ordre optimal des variables, nous avons eu recours à une classification hiérarchique effectuée sur la matrice de corrélation des variables. Il nous est clairement apparu que l'analyse de la hiérarchie construite peut servir à analyser les variables. Nous allons donc maintenant explorer cette hiérarchie en utilisant la *Hiérarchie Évoluée*.

Le graphique suivant est ainsi la *Hiérarchie Évoluée* obtenue à partir de la classification réalisée sur la matrice de corrélation des données socioprofessionnelles sur les Régions.

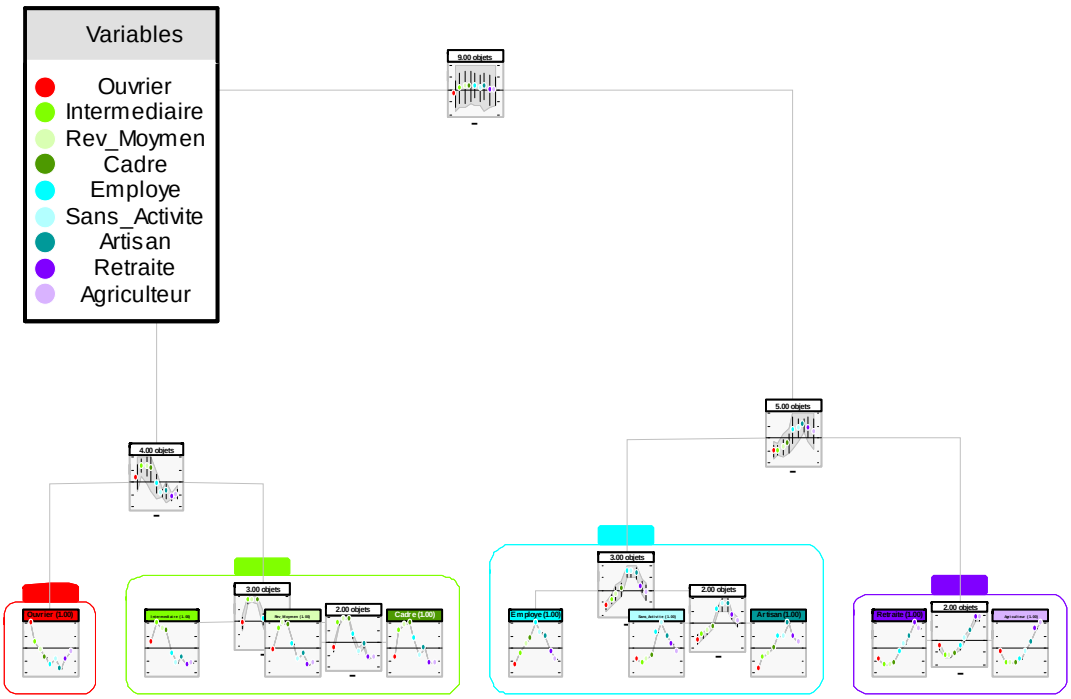


Fig. 83 – Visualisation hiérarchique évoluée de la classification de la matrice des corrélations

Grâce, à cette visualisation, nous avons pu faire quatre groupes de variables :

- *Ouvrier*
- *Intermédiaire*, *Revenu_Moyen* et *Cadre*
- *Employé*, *Sans_Activité* et *Artisan*
- *Retraite* et *Agriculteur*

Si on s'intéresse à un groupe en particulier, par exemple, celui constitué par *Intermédiaire*, *Revenu_Moyen* et *Cadre*. Son profil est le suivant :

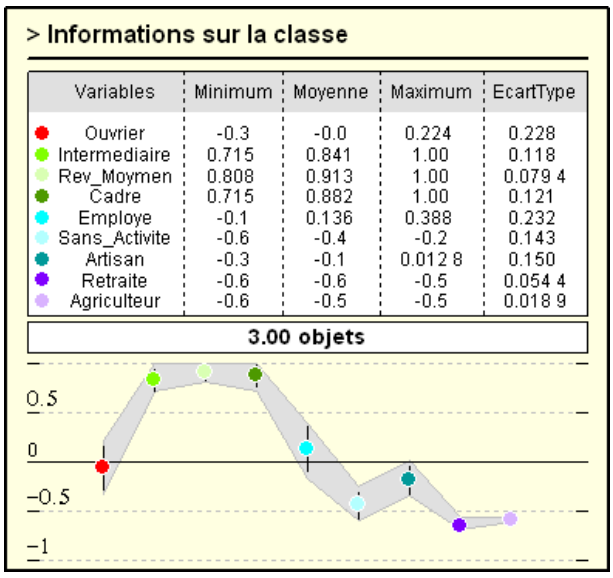


Fig. 84 – Détail du profil du groupe « Intermédiaire, Revenu_Moyen *et* Cadre »

On remarque que ce groupe est opposé au groupe *Retraite* et *Agriculteur* (corrélations négatives). Il est aussi plutôt opposé à *Sans_Activité*. Il est par contre plutôt indépendant des autres variables comme *Ouvrier*, *Employé* et *Artisan*. On remarque que le groupe est correcte valide car les variables appartenant à ce groupe (*Intermédiaire*, *Revenu_Moyen* et *Cadre)* sont toutes corrélées positivement.

Ainsi la visualisation hiérarchique évoluée permet d'analyser les variables et les relations entre celles-ci. De plus, il permet d'analyser les groupements de variables.

Nous allons maintenant voir un exemple d'analyse de données effectué avec la visualisation hiérarchique évoluée et le tableau évolué.

1.3 Analyse des données socioprofessionnelles de Paris et de sa Petite Couronne

Nous allons analyser les 2800 zones de Paris (département 75) et de sa petite couronne (départements des Hauts-de-Seine 92, Seine-St-Denis 93 et Val-de-Marne 94). Elles sont décrites par des variables socioprofessionnelles (le pourcentage de population dans chaque catégorie socioprofessionnelle) et le revenu moyen.

Nous allons d'abord effectuer l'analyse des variables afin de préparer l'analyse de la classification de données. Le résultat de la classification de la matrice de corrélation est donné par la figure suivante.

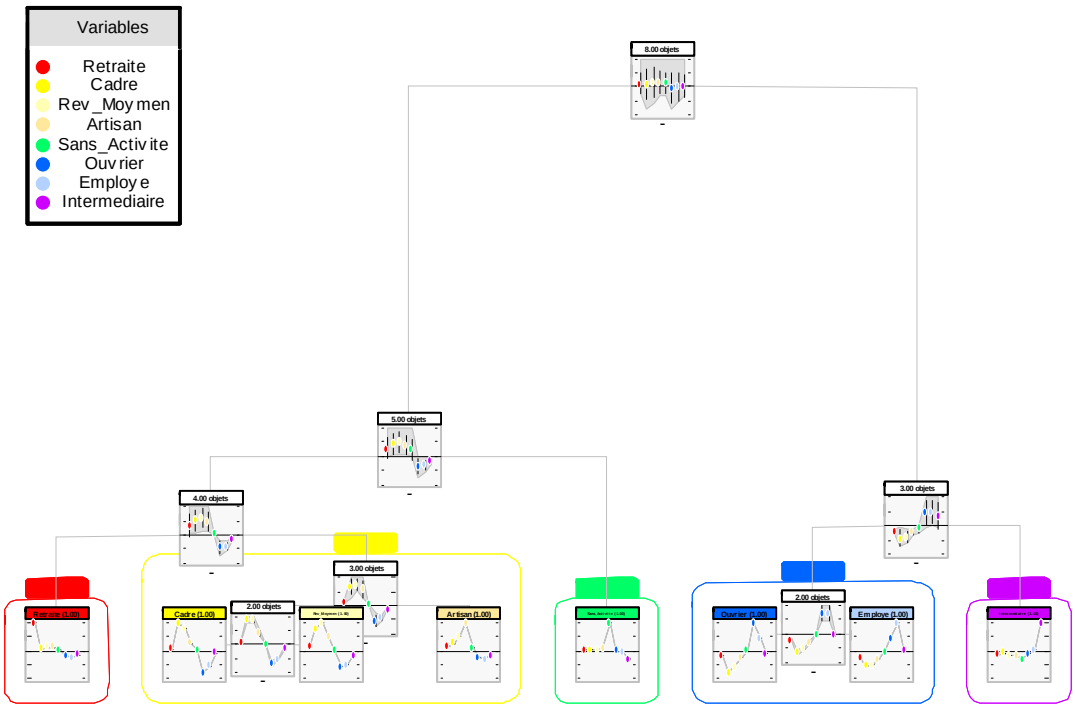


Fig. 85 – Analyse des variables

L'analyse des variables nous donne les résultats suivants :

- *Retraite*, *SansActivité* et *Intermédiaire* sont indépendantes et ne semblent pas corrélées avec d'autres variables
- *Cadre*, *RevenuMoyen* et *Artisan* sont plutôt liées et opposées au groupe « *Employé* et *Ouvrier* »
- Inversement *Employé* et *Ouvrier* sont liées et opposées au groupe « *Cadre*, *RevenuMoyen* et *Artisan* »

Ainsi, on s'attend à ce que la classification de données s'effectue autour d'une opposition entre « *Cadre*, *RevenuMoyen* et *Artisan* » d'un côté et « *Employé* et *Ouvrier* » de l'autre côté.

Nous allons maintenant effectuer la classification des zones selon ces mêmes variables. Les résultats sont les suivants :

Variables	Poids	Moyenne Générale	EcartType Général
Intermediaire	1.00	15.9	4.91
Employe	1.00	14.8	6.51
Ouvrier	1.00	17.5	11.5
Sans_Activite	1.00	6.26	2.95
Retraite	1.00	17.9	5.21
Artisan	1.00	6.13	2.96
Rev_Moymen	1.00	27 300	13 100
Cadre	1.00	21.3	13.1

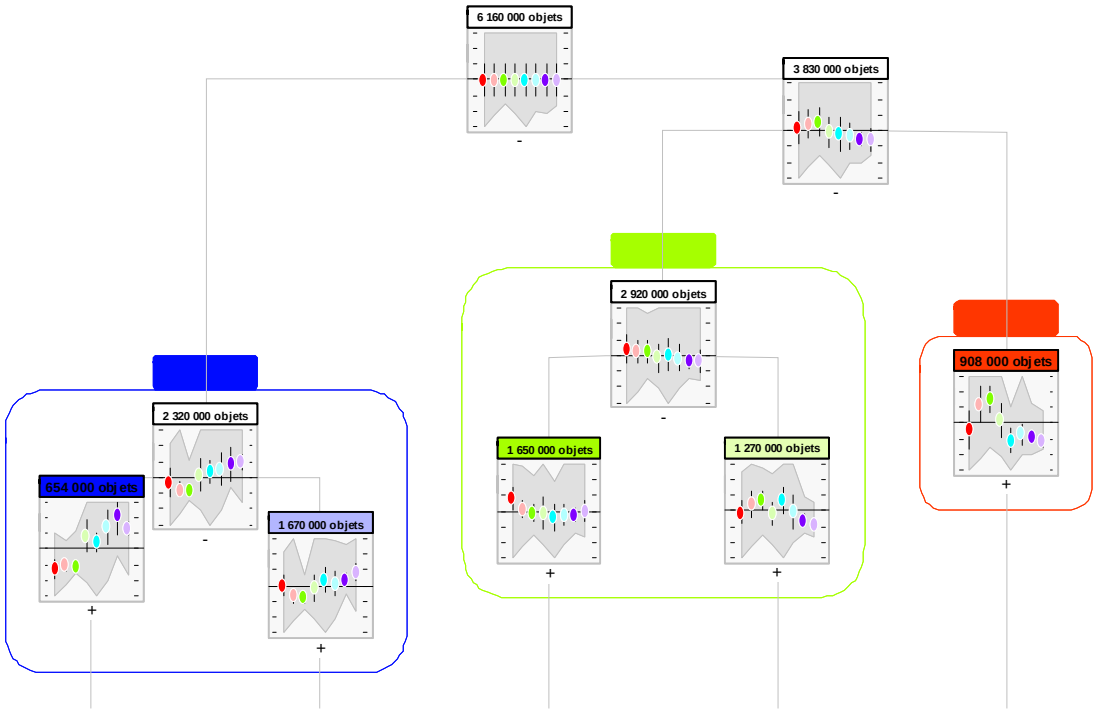


Fig. 86 – Hiérarchie évoluée montrant 5 classes

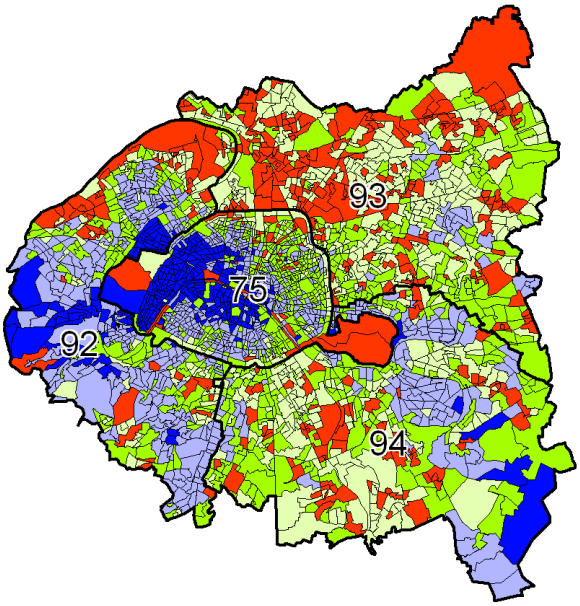


Fig. 87 – Carte correspondant aux 5 classes précédentes

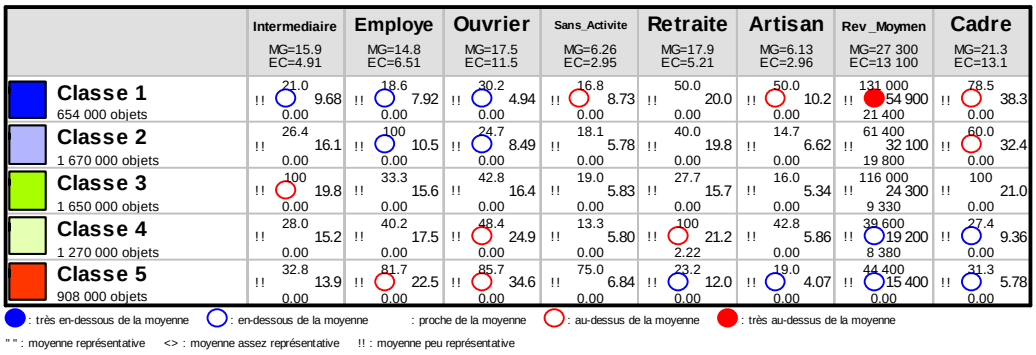


Fig. 88 – Tableau évolué correspondant aux 5 classes précédentes

Nous effectuons d'abord une analyse rapide en trois groupes :

- Groupe Bleu : Il s'agit des quartiers plutôt favorisés, le taux de cadres et le revenu y sont généralement plus élevés que la moyenne, le taux d'ouvriers et d'employés est par contre plus faible que la moyenne. On constate que géographiquement, il s'agit de Paris-Ouest et des Hauts-de-Seine.
- Groupe Vert : Il s'agit des quartiers dans la moyenne.
- Groupe Rouge : Il s'agit des quartiers plutôt moins aisés, les taux d'ouvriers et d'employés sont généralement plus élevés que la moyenne et les taux de retraités, d'artisans et de cadres sont inférieurs à la moyenne ainsi que le revenu.

Si l'on fait une analyse plus en détail, avec 5 classes, nous avons :

- Pour le groupe Bleu :
  - La classe Bleu Foncé montre les quartiers généralement très favorisés, le revenu moyen y est très élevé. Les autres particularités significatives sont un taux d'artisans plus élevé que la moyenne et taux de sans activités plus élevé que la moyenne. On peut avancer les explications suivantes :
    - l'artisanat et les professions libérales s'adressent préférentiellement à des gens à fort pouvoir d'achat et il est donc logique qu'ils exercent exercent dans ces quartiers.
    - Il y a plus de personnes sans activités dans ces quartiers car elles n'ont pas besoin de travailler car elles ont suffisamment d'argent pour vivre.
  - La classe Bleu Clair correspond presque au groupe Bleu.
- Pour le groupe Vert :
  - La classe Verte indique les quartiers situés dans la moyenne pour toutes les variables excepté le taux d'intermédiaires qui y est plus élevé que la moyenne
  - La classe Vert Pâle se rapproche par certain aspect des quartiers moins aisés (Groupe Rouge) : il y a plutôt peu de cadres, le revenu est plutôt faible et il y a beaucoup d'ouvriers. Cependant, on peut noter une différence majeure : il y a beaucoup de retraités.
- Pour le groupe Rouge : on peut analyser plus en détail la différence avec la classe Vert Pâle qui lui est toutefois assez semblable par certains aspects. On remarque que le taux de retraités est assez faible et que le taux d'ouvriers est assez élevé par rapport à la classe Vert Pâle. On peut avancer l'explication suivante : la classe Vert Pâle correspond

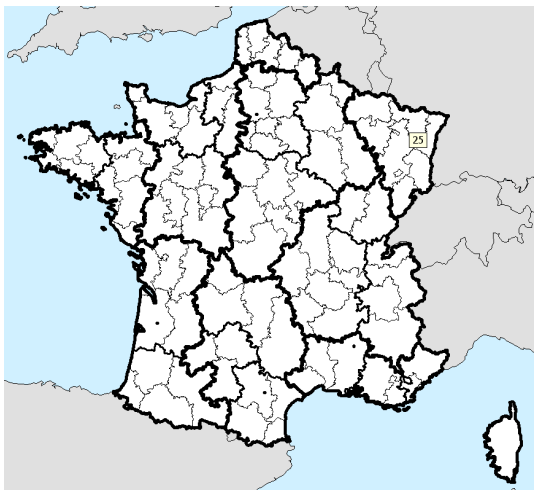
à des quartiers populaires plus anciens (plus de retraités) tandis que la classe Rouge correspond à des quartiers populaires récents (peu de retraité, beaucoup d'actifs).

Cet exemple illustre ainsi les possibilités d'analyse offertes par les outils de visualisation que nous avons développés.

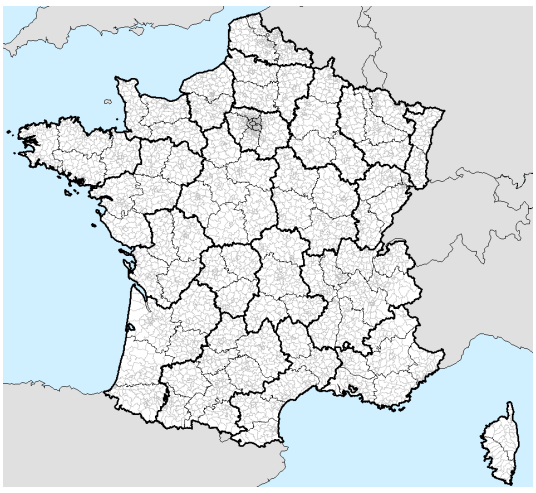
1.4 Autre application : la Classification Hiérarchique Spatiale

Nous avons adapté notre algorithme de la CAHA afin de réaliser la *Classification Hiérarchique Spatiale*. La modification réalisée est très simple, il s'agit d'utiliser comme variables les deux variables de coordonnées : la latitude et la longitude. Dans le cas d'objets surfaciques, nous prenons pour point de référence le centre de l'objet. Cependant, cet outil n'est intéressant que si une variable de pondération est utilisée. Ainsi, les groupes seront construits de façon préferentielle autour de zones de fort poids.

L'exemple suivant montre les résultats obtenus en classifiant les 3700 cantons de France en utilisant la population comme variable de pondération. À titre de comparaison, le découpage hiérarchique administratif est aussi présenté. Par ailleurs, comme le nombre de résumés réellement utilisés est de l'ordre de la centaine (le nombre de départements), plus élevé que les 20 résumés généralement utilisés dans le cas classique, nous avons du relever le niveau de précision de la CAHA à *k=400* au lieu de *k=100* en standard. La figure suivante montre les différences entre les niveaux créés par la CAHA et les niveaux administratifs existants. Cependant, on ne note pas de différence très notable au niveau des formes et des tailles.



Niveau	Nombre
Non présent	3700
Niveau 1	95
Niveau 2	21
Niveau 3	9



Niveau	Nombre
Cantons	3700
Départements	95
Régions	21
Non disponible	9

Fig. 89 –Comparaison entre les niveaux hiérarchiques créés par la CAHA et les niveaux administratifs existants

La figure suivante montre les densités de population pour les différents niveaux.

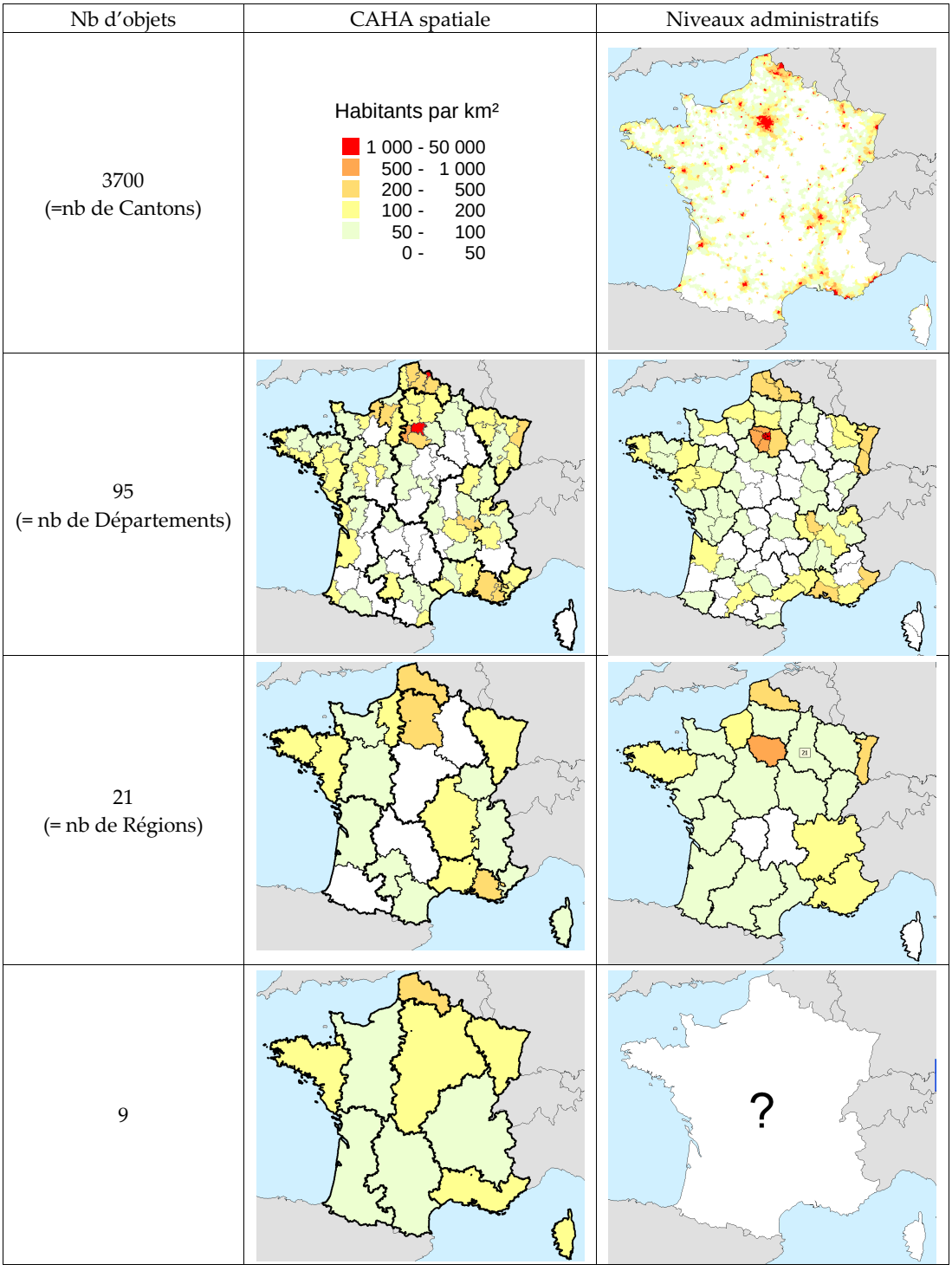


Fig. 90 –Comparaison de la densité de population entre les niveaux hiérarchiques créés par la CAHA et les niveaux administratifs existants

On remarque que les différences sont plus marquées dans le cas de la Classification Spatiale, ce qui est normal car le but de la classification est de regrouper ensemble les zones qui se

ressemblent. Ainsi, au niveau comportant 21 objets, la CAHA spatiale met bien en évidence la fameuse « diagonale du vide » allant du Nord-Est au Sud-Ouest.

Cependant, pour des niveaux supérieurs comme celui comportant 9 classes, les résultats ne sont pas très satisfaisants. C'est pourquoi il est préférable d'utiliser une autre méthode, le Lissage Spatial qui donne de bons résultats aux grandes échelles.

1.5 Conclusion

Les expérimentations ont montré que notre algorithme de classification (la CAA dans sa version standard) produit des résultats meilleurs que l'algorithme des k-moyennes si l'on utilise une précision suffisante. Nous avons aussi établi que notre version parallèle de cet algorithme n'est pas très intéressante à cause de l'effet de « superposition » décrit précédemment.

Nous avons vu une extension très utile de la *Classification de Données* et de sa visualisation : la *Classification de Variables*. Cette méthode permet une analyse globale des relations entre les variables, ce qui est fort utile pour préparer l'analyse d'une *Classification des Données*.

Finalement, nous avons illustré la puissance de ces visualisations dans le cadre de l'analyse des données socioprofessionnelles de Paris et sa petite Couronne. Nous avons d'abord utilisé l'outil d'analyse des variables afin d'anticiper les principaux axes d'analyse de la classification de données qui a suivie. Cette dernière a été effectuée à partir de seulement trois visualisations complémentaires : la *Hiérarchie Évoluée*, le *Tableau Évolué* et la carte correspondante. Nous avons ainsi pu décrire les principales caractéristiques de la population de Paris et de sa petite Couronne et mettre en lumière certaines caractéristiques des différentes populations (par exemple, un fort taux de sans activités parmi la population très aisée et leur localisation dans l'Ouest de Paris).

Nous avons aussi adapté la CAHA pour réaliser la *Classification Hiérarchique Spatiale*. Nous obtenons ainsi un outil de sectorisation complémentaire car, outre la hiérarchie des secteurs, cette méthode permet de construire les secteurs de façon préferentielle autour de zones de fortes valeurs.

2 Lissage Spatial, Détermination et Hiérarchisation de Pôles

Tout d'abord, nous verrons un exemple d'analyse spatiale montrant les pôles de population française pour l'échelle locale (50 km) et l'échelle nationale (130 km) et leur hiérarchisation. Le résultat est très synthétique et donne un bon résumé de la répartition de la population sur le territoire.

Nous verrons ensuite l'intérêt du *Lissage Spatial* dans le cadre de la *Classification de Données*. Nous montrerons que le *Lissage Spatial* des données en prétraitement de la *Classification de Données* permet d'analyser les données pour des échelles spatiales plus grandes. Une autre conséquence est que la carte des classes (qui sont issues de la classification) est plus facile à analyser.

2.1 Analyse spatiale de la population pour la France entière

À partir de la population de chacun des 3700 cantons de la France, nous calculons les Sommes Lissées de la population avec la fonction d'interaction spatiale gaussienne pour des rayons de 130 km et 50 km. Nous en extrayons les pôles pour chacune des cartes de lissage spatial. Nous obtenons 9 pôles pour le niveau de 130 km et 51 pôles pour le niveau de 50 km.

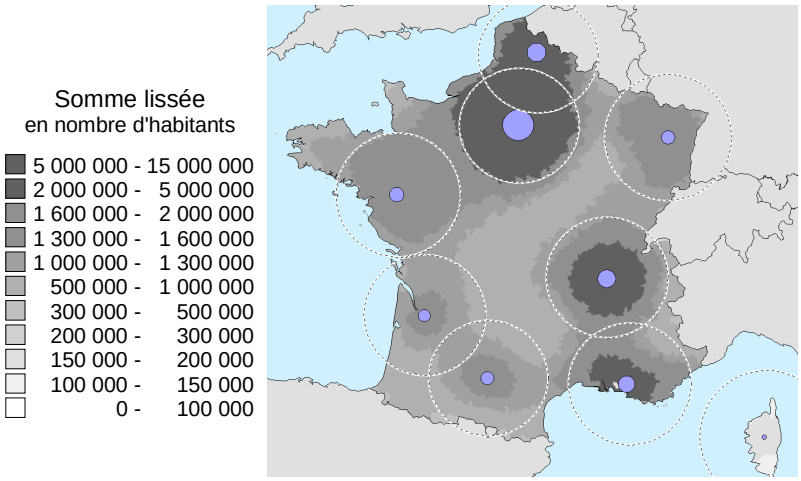


Fig. 91 –Pôles pour la population lissée en utilisant la sommation avec un rayon de 130 km

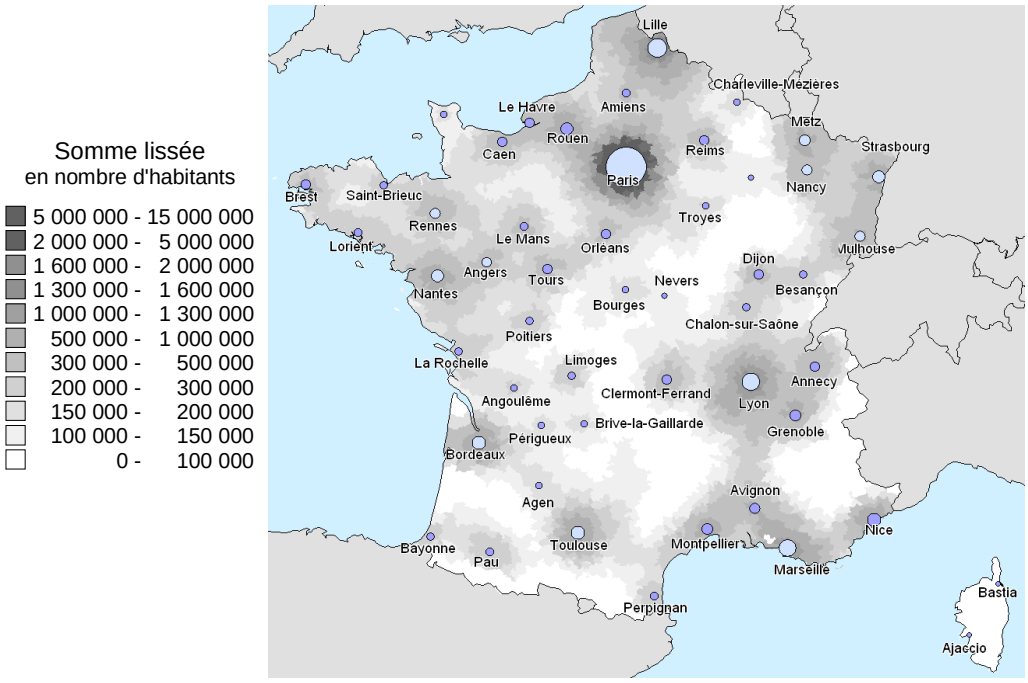


Fig. 92 –Pôles pour la population lissée en utilisant la sommation avec un rayon de 50 km

En analysant, les pôles du niveau 50 km, nous nous sommes aperçus qu'ils étaient pratiquement tous localisés sur des agglomérations qui étaient d'ailleurs leur principal contributeur. Pour cette raison, nous avons pu nommer la quasi-totalité des pôles de ce niveau.

Nous construisons ensuite la hiérarchie entre les deux niveaux.

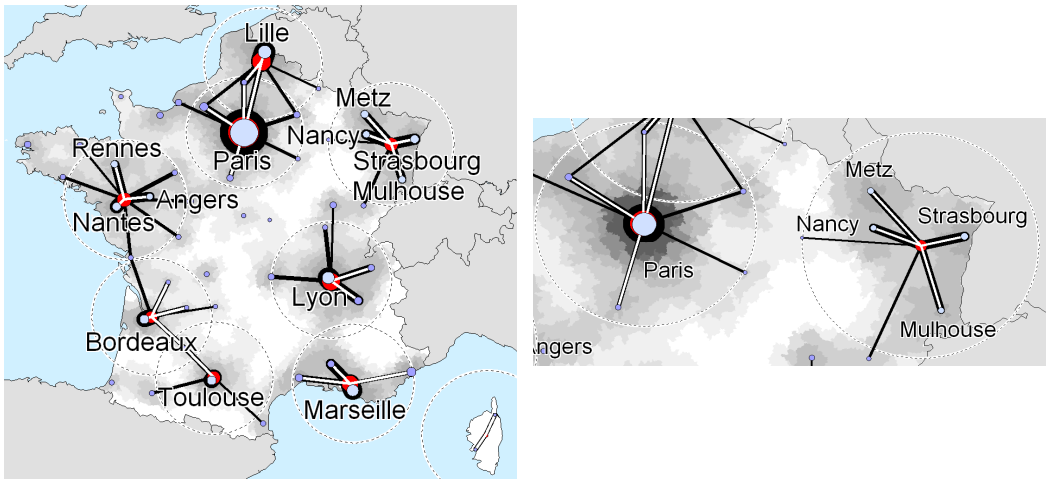


Fig. 93 – À gauche, Hiérarchie des pôles entre les niveaux 50 km et 130 km
À droite, détail sur le pôle au nord-est et celui de Paris.

Nous remarquons que pour le niveau de 130 km, nous avons deux types de pôles :

- Les pôles simples sont, comme ceux du niveau 50 km, localisés sur une agglomération qui y contribue donc fortement. On trouve **Paris, Lyon, Marseille, Lille, Toulouse et Bordeaux**.

- Les pôles composites sont formés de plusieurs agglomérations et ces pôles sont souvent localisés au milieu d'une zone vide. Il s'agit du pôle **Nantes-Rennes-Angers** et du pôle **Metz-Nancy-Strasbourg-Mulhouse** et la **Corse.**

Cet exemple montre ainsi qu'il est possible de décrire la localisation de la population française à l'échelle 130 km en utilisant 9 pôles dont la description peut-être faite le cas échéant à partir des pôles du niveau inférieur, c'est-à-dire 50 km. Nous utiliserons ultérieurement cette description dans le cadre d'un exemple mettant en œuvre la *Sectorisation à partir de Centres*.

2.2 Utilisation du Lissage Spatial en prétraitement de la Classification de Données

Nous allons refaire l'analyse des 2800 zones de Paris (département 75) et de sa petite couronne (départements des Hauts-de-Seine 92, Seine-St-Denis 93 et Val-de-Marne 94). Ces zones sont décrites par des variables socioprofessionnelles (le pourcentage de population dans chaque catégorie socioprofessionnelle) et le revenu moyen. Cette analyse a déjà été effectuée dans la partie consacrée à la visualisation des classifications de données.

Nous effectuons donc un lissage spatial avant de réaliser la classification des données. Le lissage utilisé est la Moyenne Pondérée en utilisant la population comme pondération et un rayon de lissage de 1,5 km. L'exemple suivant montre les différences obtenues lorsque l'on applique le lissage. Il s'agit de la variable donnant le revenu moyen.

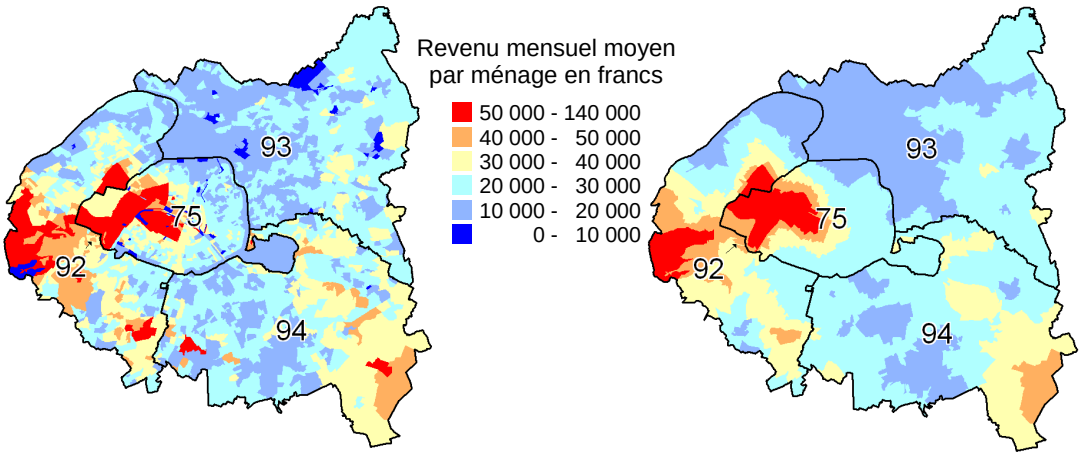


Fig. 94 – À gauche. Carte du revenu moyen par zone
À droite. Carte du revenu moyen par zone avec un lissage utilisant la Moyenne Pondérée et un rayon de 1,5 km

On s'aperçoit que la carte du revenu moyen est alors beaucoup plus simple à interpréter : les forts revenus sont situés dans Paris-Ouest et à l'ouest de Paris. Nous allons maintenant passer à la classification.

La classification des variables donne des résultats identiques, il n'est donc pas intéressant d'en parler et nous passons directement à la classification des données dont les résultats sont les suivants.

Variables	Poids	Moyenne Générale	EcartType Général
Lissage_Intermediaire	1.00	15.7	2.92
Lissage_Employe	1.00	14.7	3.84
Lissage_Ouvrier	1.00	17.2	9.56
Lissage_Sans_Activite	1.00	6.38	2.05
Lissage_Retraite	1.00	17.9	2.33
Lissage_Artisan	1.00	6.19	1.65
Lissage_Rev_Moymen	1.00	27 700	11 200
Lissage_Cadre	1.00	21.6	10.9

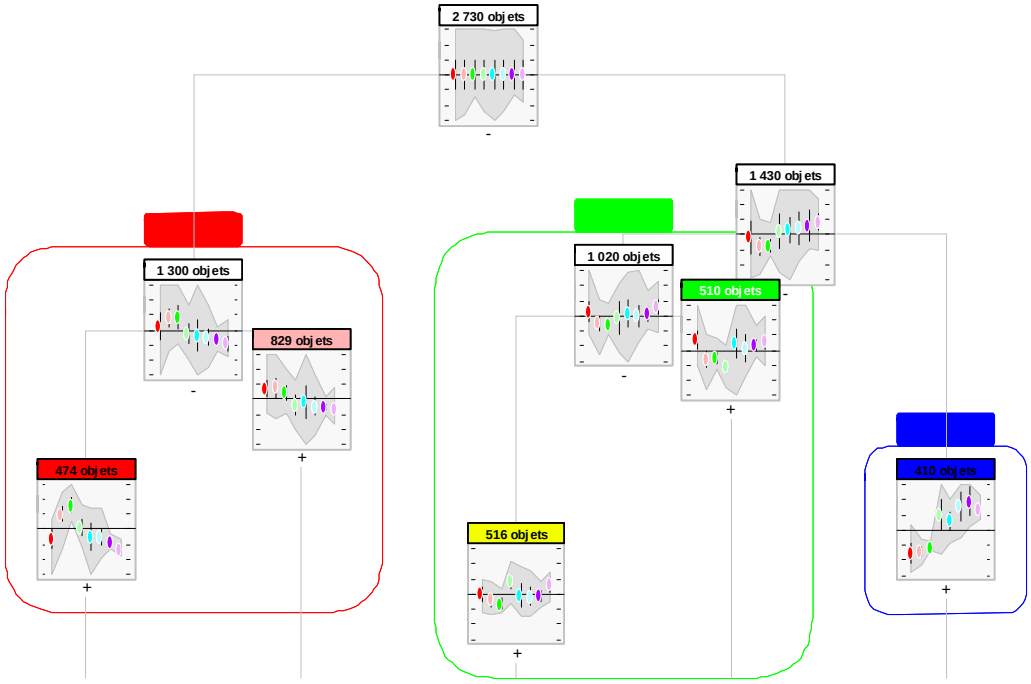


Fig. 95 –Classification hiérarchique des données lissées avec un rayon de 1,5 km.

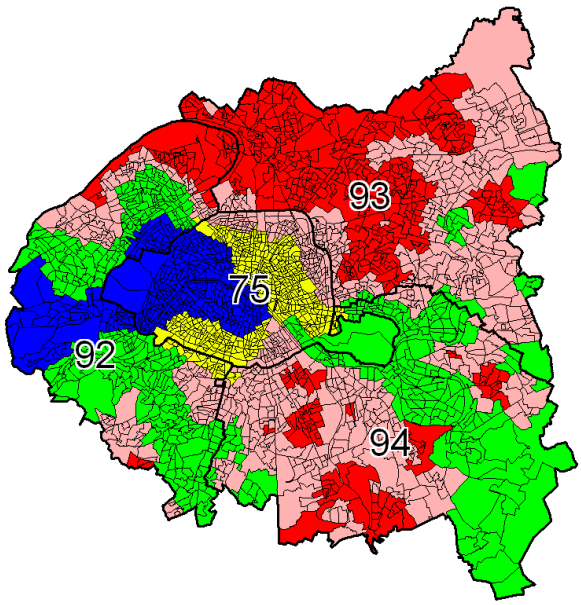


Fig. 96 –Carte correspondant à la classification.

Comme attendu, les zones sont beaucoup plus homogènes en raison du lissage effectué. Les classes trouvées ne sont pas exactement les mêmes que celles trouvées sans lissage. On conserve cependant les mêmes tendances générales :

- Groupe Bleu : Il s'agit des quartiers plutôt favorisés. Le taux de cadres et le revenu y sont généralement plus élevés que la moyenne, le taux d'ouvriers, d'employés et d'intermédiaires est par contre plus faible que la moyenne. On constate que géographiquement, il s'agit de Paris-Ouest et de la partie des Hauts-de-Seine immédiatement à l'ouest de Paris.
- Groupe Vert : Il s'agit des quartiers dans la moyenne. On peut toutefois remarquer que la partie concernant Paris (en jaune) est marquée par un taux de sans-activité plus élevé que la moyenne, tandis que l'autre partie (en vert) est marquée par un taux de sans-activité plus faible et plus d'intermédiaires que la moyenne.
- Groupe Rouge : Il s'agit des quartiers plutôt moins aisés, les taux d'ouvriers et d'employés sont généralement plus élevés que la moyenne et les taux de retraités, d'artisans et de cadres sont inférieurs à la moyenne ainsi que le revenu. On peut noter que les quartiers en rouge foncé ont moins d'intermédiaires que la moyenne et par conséquent plus d'ouvriers et d'employés. Le revenu y est ainsi généralement plus faible.

Ainsi, l'utilisation du lissage avant la classification de données permet l'étude des tendances spatiales à de plus grandes amplitudes (ici de l'ordre de 1,5 km) que la taille des objets étudiés (le rayon moyen des zones de Paris est de l'ordre de 200 m). Ce faisant elle produit des classifications ayant une localisation spatiale sous forme de zones beaucoup plus homogènes et évite ainsi l'aspect mosaïque de la cartographie de la classification standard.

2.3 Conclusion

Nous avons illustré le fonctionnement de la méthode de *Détermination et de Hiérarchisation des Pôles* pour la recherche et la description des pôles de la population française pour l'échelle locale (50 km) et l'échelle nationale (130 km). La carte obtenue est alors très synthétique et donne une bonne image de la répartition géographique de la population française.

Nous avons aussi montré la pertinence de l'utilisation du *Lissage Spatial* avant de réaliser une *Classification de Données* : cela permet d'étudier les tendances spatiales à de plus grandes amplitudes que la taille des objets étudiés. Une conséquence directe est que les classes trouvées forment des zones beaucoup plus homogènes, évitant ainsi l'aspect mosaïque de la cartographie de la classification standard. Les cartes sont alors plus simples à interpréter et plus esthétiques.

3 Sectorisation

Nous allons voir d'abord l'application de notre méthode de *Sectorisation Équilibrée* pour le partage du territoire français en 22 secteurs de population égale. Nous montrerons que les résultats obtenus sont de qualité en répondant aux contraintes de contiguïté, d'équilibre et de compacité. Toutefois, nous mettrons en évidence que les résultats obtenus ont un « biais » caché : les zones de forte population sont très souvent partagées entre plusieurs secteurs.

Dans la partie suivante, nous utiliserons la *Sectorisation à partir de Centres*. Il s'agira de créer des secteurs autour de centres qui sont des grandes agglomérations (ou groupes d'agglomérations). Nous verrons dans un premier temps comment la méthode simple permet d'obtenir un résultat déjà intéressant. Ensuite, nous montrerons que la méthode itérative que nous avons mise au point permet d'obtenir de meilleurs résultats. Toutefois, même si les résultats sont très satisfaisants quant au respect des objectifs de taille et de contiguïté, la compacité des secteurs n'est pas toujours au rendez-vous et il faut alors utiliser une méthode d'amélioration de la compacité au détriment du respect des objectifs de taille.

Nous montrons donc dans la partie suivante que le *Rééquilibrage de Sectorisations* permet justement d'atteindre les objectifs de taille en partant d'une sectorisation bonne en terme de compacité et de contiguïté, mais « moyenne » en terme d'objectifs de taille. Ainsi, en partant des résultats de la *Sectorisation à partir de Centres*, nous montrons que le rééquilibrage permet d'obtenir de bons résultats respectant les trois critères : compacité, taille et contiguïté.

3.1 Sectorisation équilibrée de la population française en 22 secteurs

Dans l'exemple ci-dessous, la population française a été partagée en 22 nouvelles régions de même population. Les objets géographiques sont les départements. La population moyenne théorique à atteindre était 2 600 000 habitants par région. Le résultat obtenu est le meilleur parmi 100 solutions générées à partir d'initialisations aléatoires. Les caractéristiques de sa qualité sont :

- Les secteurs sont contigus.
- L'écart maximal $Ec_{\max}$ est de 33 % à cause du secteur de moins de 2 millions d'habitants dans le nord de la France.
- L'écart moyen Ec_{moy} est de 12 %.

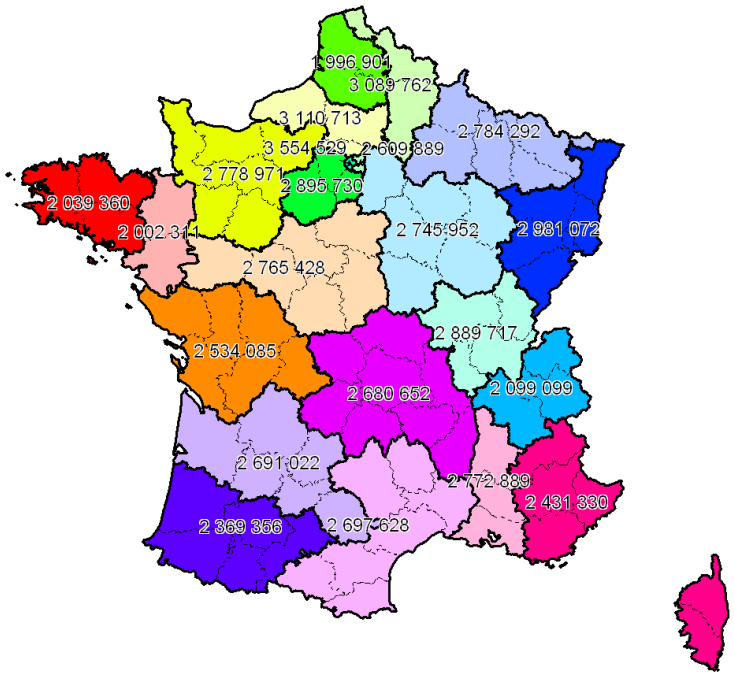


Fig. 97 –Sectorisation des départements en 22 secteurs de population égale

Dans l'exemple suivant, la population française a été aussi partagée en 22 nouvelles régions de même population. Les objets géographiques sont plus petits, il s'agit des 3600 cantons. La population moyenne théorique à atteindre était aussi 2 600 000 habitants par région. La qualité de cette sectorisation est meilleure :

- Les secteurs sont contigus.
- L'écart maximal $Ec_{\max}$ est de 16 %.
- L'écart moyen Ec_{moy} est de 8 %.

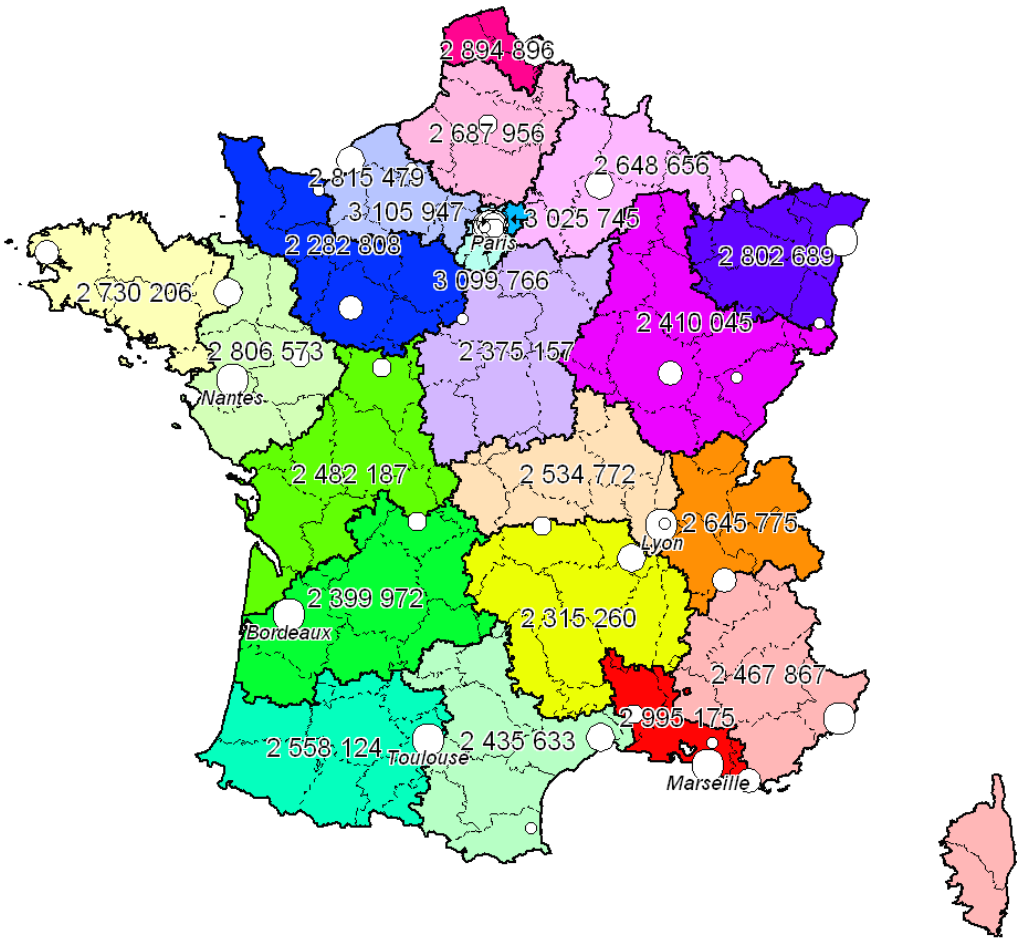


Fig. 98 –Sectorisation des Cantons en 22 secteurs de population égale avec les principales agglomérations indiquées

Le résultat obtenu est meilleur, ce qui est normal car le nombre d'objets est beaucoup plus grand, ce qui autorise un partitionnement plus fin. Par contre, le principal problème que rencontre cette méthode est qu'il est très fréquent qu'une frontière passe par une grande ville ou juste à côté. C'est le cas pour *Paris* et *Lyon* qui sont partagé entre plusieurs secteurs. Il y a aussi *Bordeaux*, *Toulouse* et *Rennes* qui sont en bordure de secteur.

3.2 Sectorisation de la population française à partir de 9 centres

Nous réutilisons le résultat obtenu par la *Détermination des Pôles*. Il s'agissait de la population par canton. La méthode de lissage était la sommation avec un rayon de 130 km en utilisant la fonction d'interaction spatiale.

L'exemple suivant montre le calcul de la quantité de chaque secteur en fonction des pondérations.

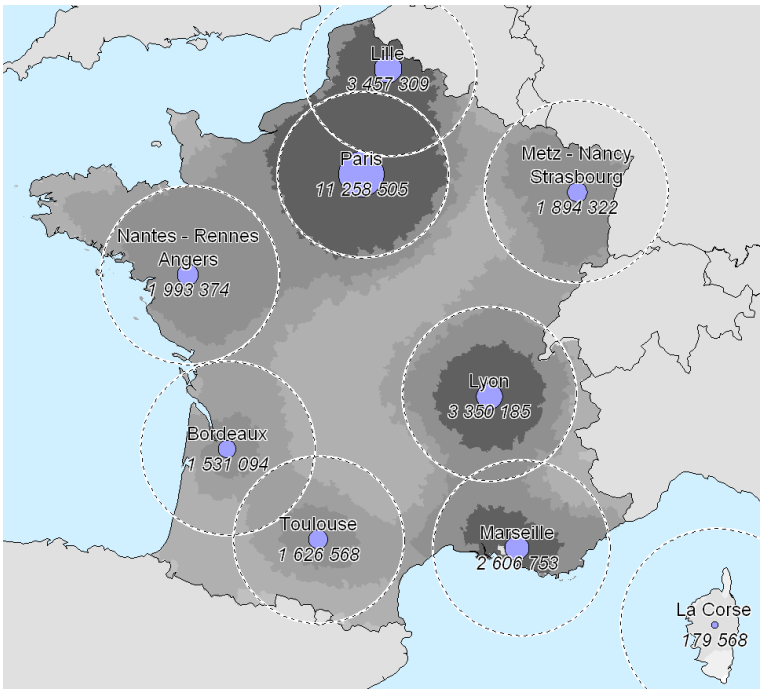


Fig. 99 –Pôles de population lissée avec un rayon de 130 km (en nb d'habitants)

Nous souhaitons donc prendre pour centre ces pôles en utilisant comme poids leur valeur lissée. Nous calculons donc la part de chaque secteur puis nous en déduisons sa taille car la population totale à partager est de 58 millions d'habitants.

Nom	Poids	Part	Taille du secteur
La Corse	179 568	0,7%	433 139
Bordeaux	1 531 094	6,3%	3 693 185
Toulouse	1 626 568	6,7%	3 923 479
Metz-Nancy-Strasbourg	1 894 322	7,8%	4 569 334
Nantes-Rennes-Angers	1 993 374	8,2%	4 808 260
Marseille	2 606 753	10,7%	6 287 804
Lyon	3 350 185	13,8%	8 081 053
Paris	11 258 506	46,4%	27 156 884
Total	**24 260 800**	**100%**	**58 520 000**

Fig. 100 –Calcul de la taille des secteurs à partir des poids des centres

Nous allons maintenant voir la mise en pratique de la sectorisation avec ces centres.

3.2.1 Exemple avec l'algorithme simple

Les cartes suivantes montrent les différents stades de la sectorisation en fonction des étapes exprimées en pourcentage. La population actuelle des secteurs est indiquée au-dessus du centre, la population théorique à atteindre est indiquée en dessous en italique, et le ratio entre la population réelle et la population à atteindre est indiqué en pourcentage au milieu.

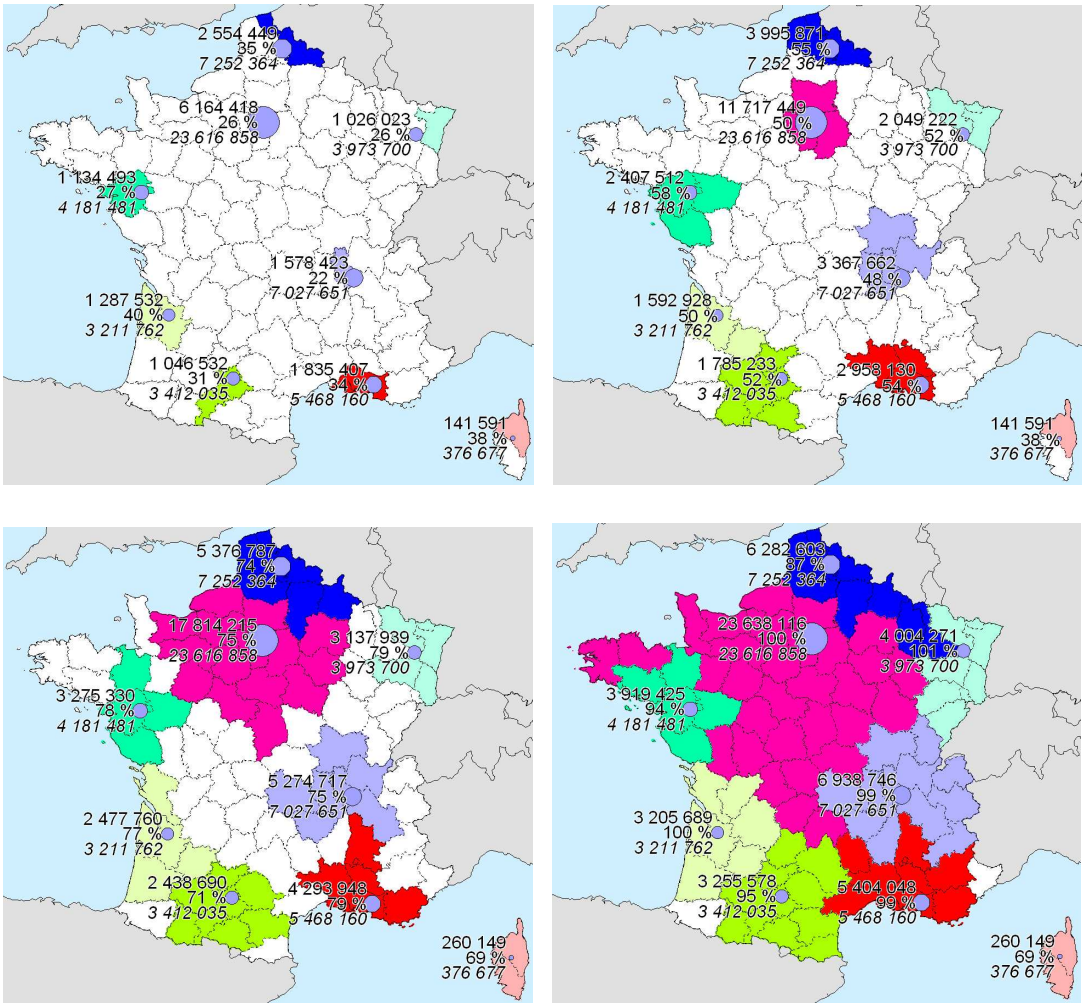


Fig. 101 –Progression de la taille des secteurs aux « étapes » 25 %, 50 %, 75 % et 100%

Nous voyons que la croissance des secteurs ne pose pas de problème tant que les secteurs ont de la « place » pour croître, c'est-à-dire jusqu'à « l'étape 75 % » dans l'exemple. À partir de là, les secteurs sont au contact et la « place » libre se fait rare. Ainsi, à « l'étape 100 % », le secteur de *Lille* est déficitaire tandis qu'au Sud, les secteurs sont à l'équilibre et qu'il reste deux départements non attribués. Finalement, à partir de « l'étape 120 % », il n'y a plus de départements non attribués et le résultat final est le suivant.

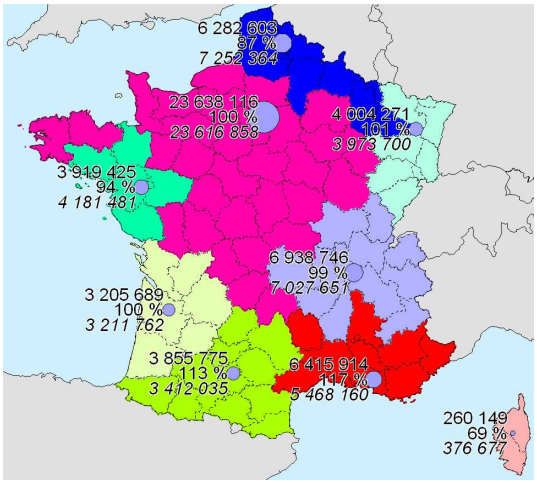


Fig. 102 –Secteurs finaux (120%)

Nous observons que, hormis la Corse, les objectifs de taille des secteurs sont atteints à plus ou moins 15 %. De plus, ce que nous avons anticipé à « l'étape 100 % » est confirmé : le secteur de *Lille* est déficitaire tandis que les secteurs du Sud comme celui de *Toulouse* et de *Marseille* sont excédentaires. La qualité est la suivante : l'écart maximal $Ec_{\max}$ est de 44 % (à cause de la Corse) et l'écart moyen Ec_{moy} est de 11 %.

La forme des secteurs n'étant pas très bonne, il est possible d'effectuer une amélioration de la forme selon la méthode décrite dans la *Sectorisation Equilibrée* en utilisant la fonction *ParMETIS_V3_RefineKway*. Le résultat est alors le suivant :

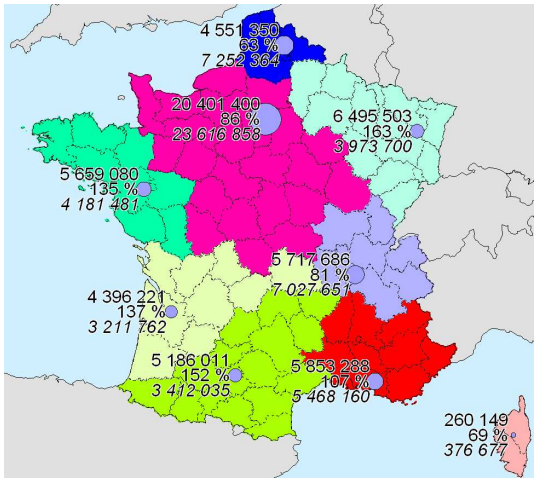


Fig. 103 –Secteurs finaux dont la forme a été améliorée.

Le résultat obtenu est alors bien meilleur pour la forme mais les tailles des secteurs sont alors très éloignées des objectifs. La qualité est maintenant : l'écart maximal $Ec_{\max}$ est de 63 % (à cause du Nord-Est) et l'écart moyen Ec_{moy} est de 37 %.

Nous allons maintenant voir une amélioration de notre algorithme initial.

3.2.2 Exemple avec l'algorithme itératif

Nous avons repris les données de population utilisées précédemment et nous avons calculé itérativement 100 sectorisations. Nous avons aussi calculé sa forme améliorée en utilisant la fonction *ParMETIS_V3_RefineKway* expliquée dans la *Sectorisation Equilibrée*. Pour rappel, nous indiquons en premier la sectorisation obtenue précédemment et son amélioration :

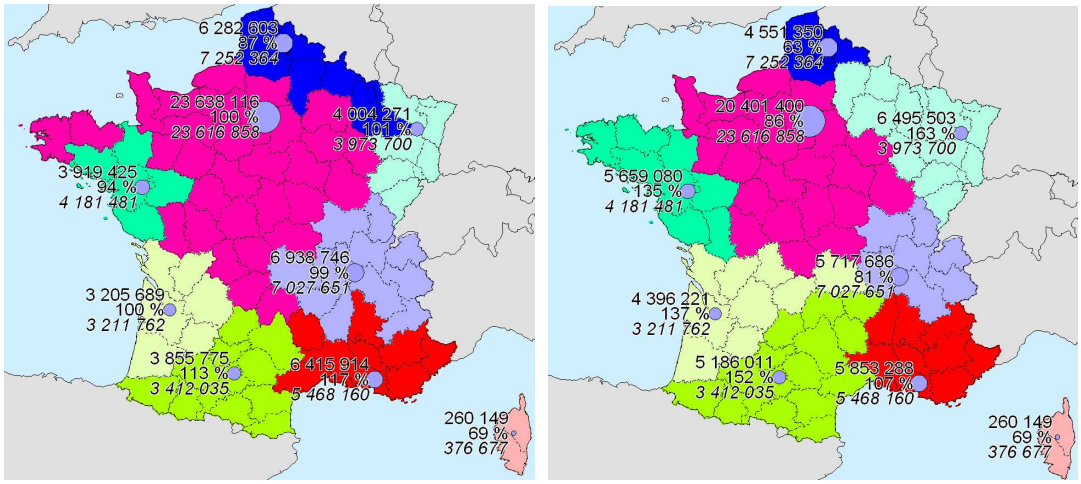


Fig. 104 – À gauche, la sectorisation calculée avec l'algorithme standard. À droite, sa forme améliorée

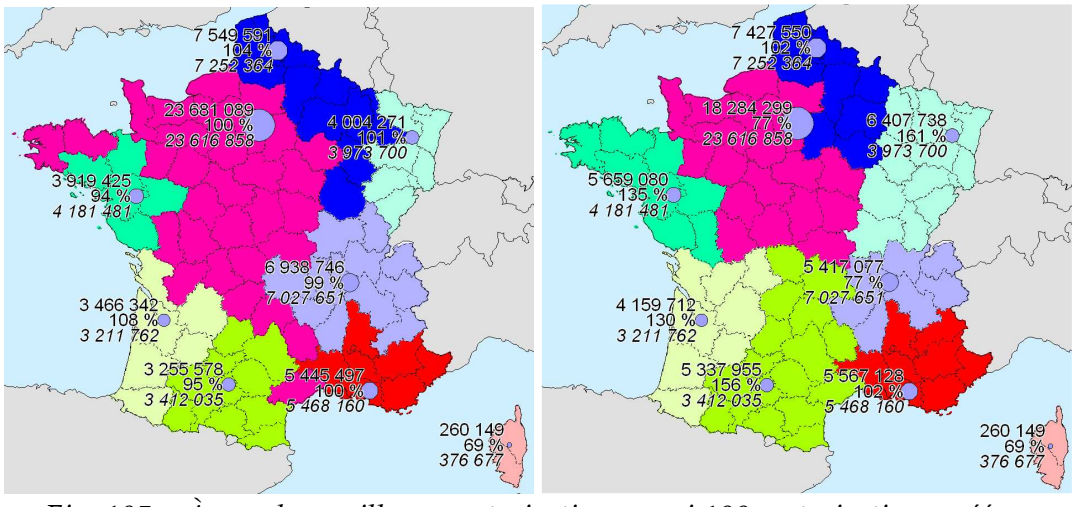


Fig. 105 – À gauche, meilleure sectorisation parmi 100 sectorisations créées À droite, sa forme améliorée.

La qualité de cette sectorisation est meilleure que celle trouvée par l'algorithme simple, nous avons toujours l'écart maximal $Ec_{\max}$ qui vaut 44 % (à cause de la Corse), mais l'écart moyen Ec_{moy} est seulement de 8 % (au lieu de 11%). La sectorisation dont la forme est améliorée est aussi meilleure que la précédente amélioration puisque nous avons maintenant $Ec_{\max}$ = 61 % (au lieu de 63 %) et Ec_{moy} = 32 % (au lieu de 37 %)

3.3 Rééquilibrage de la sectorisation précédente

L'exemple suivant montre la sectorisation à partir de centres dont la forme a été améliorée au détriment de l'équilibre. Pour chaque secteur, les quantités actuelles (en nombre d'habitants) sont indiquées au dessus du centre du secteur et la quantité souhaitée est indiquée en dessous en italique. Les transferts entre les secteurs ont été calculés et sont indiqués sur la carte. On observe ainsi que le secteur Nord (Lille) devra recevoir beaucoup du Centre (Paris) et de l'Est.

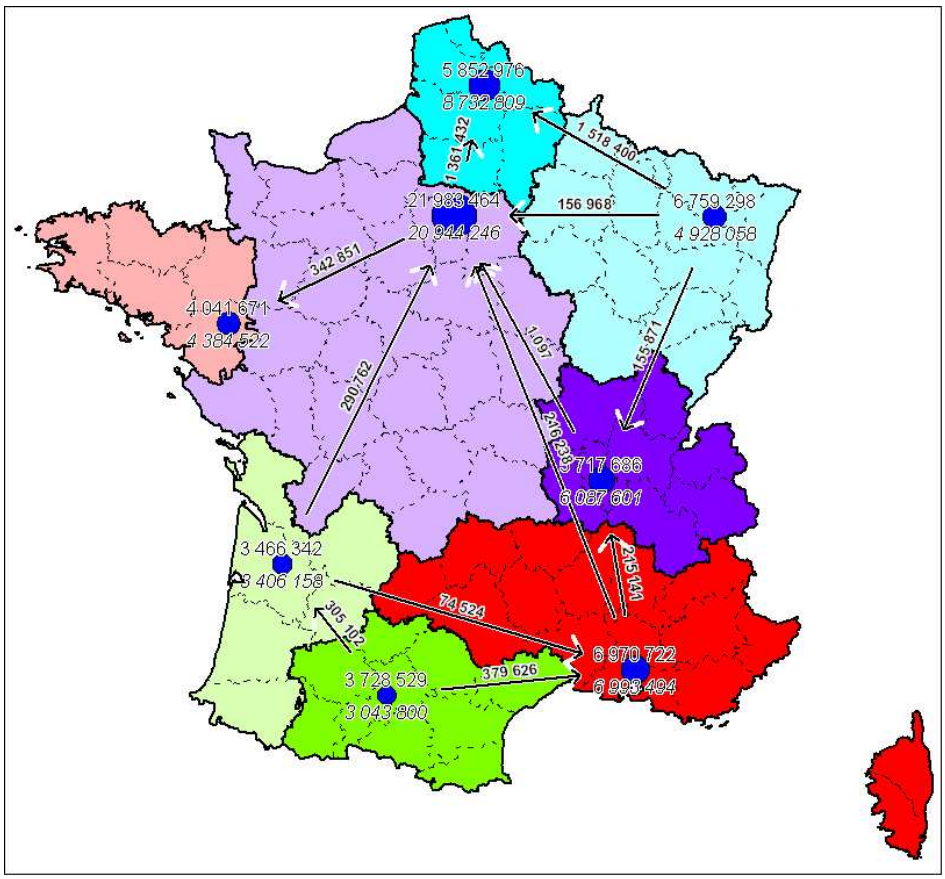


Fig. 106 –Quantités à transférer entre les secteurs

La carte suivante montre les transferts effectués entre les différents secteurs dans le cadre de l'algorithme de rééquilibrage. Le secteur ayant le plus bénéficié de ces transferts est le secteur Nord. Les départements transférés sont indiqués en hachuré.

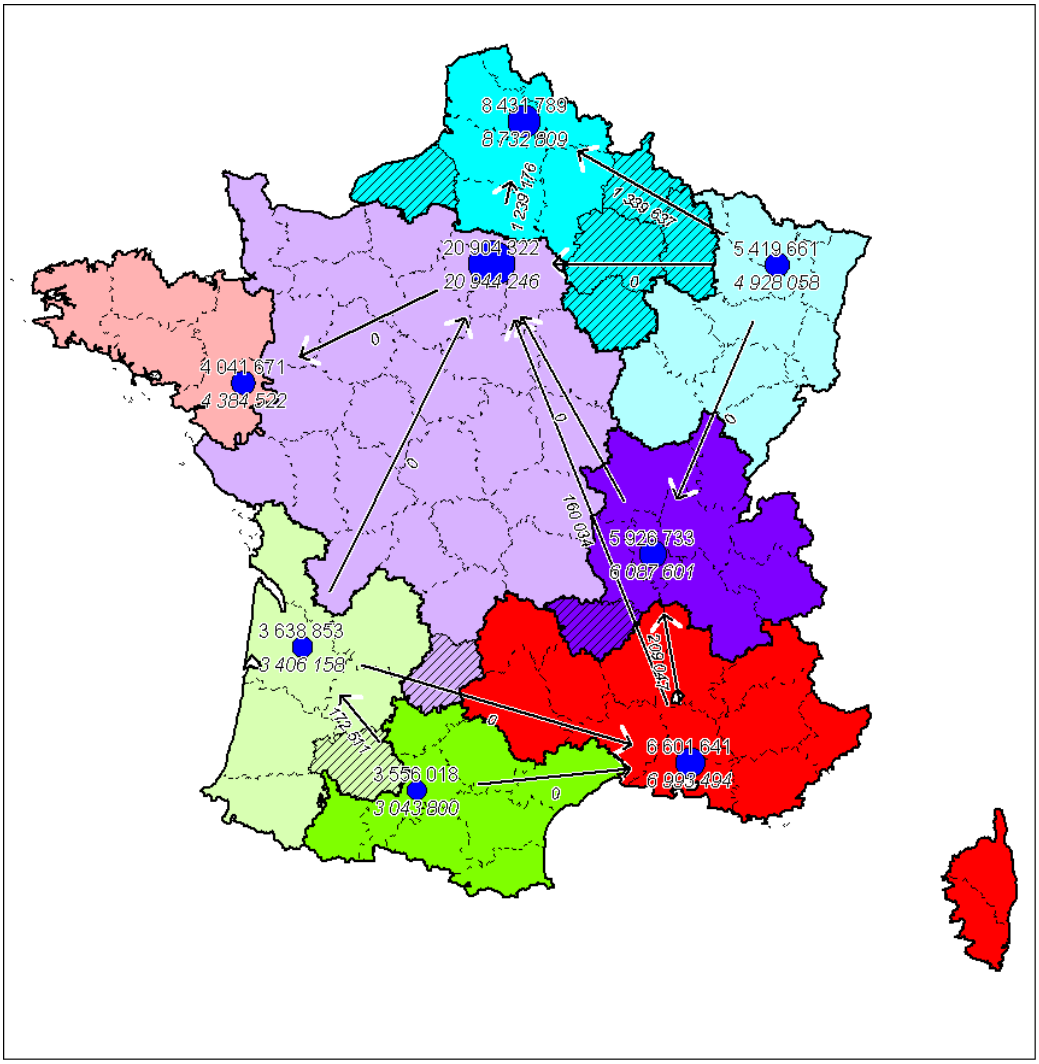


Fig. 107 –Secteurs obtenus après la réalisation des transferts

Dans cet exemple, le rééquilibrage a permis d'atteindre les objectifs à quelques pourcent près tout en conservant une forme assez compacte pour chacun des secteurs.

3.4 Conclusion

La sectorisation équilibrée de la population française donne de bons résultats en terme de compacité, de contiguité et d'équilibre entre les secteurs. Toutefois, l'absence de contrainte sur la localisation des secteurs conduit fréquemment à ce que les frontières des secteurs passent par les zones de fortes valeurs, ce qui peut-être gênant.

Au contraire, la sectorisation à partir de centres permet de construire des secteurs autour de centres. Nous avons montré que notre algorithme itératif offre d'assez bons résultats. La compacité de la forme étant moins bien respectée, il est souvent utile de procéder à une amélioration de la forme des secteurs. Cependant, le critère d'équilibre est dans ce cas mis à mal.

Cela nous a naturellement amené à utiliser notre méthode de rééquilibrage qui permet d'atteindre l'équilibre en évitant au maximum de toucher à la forme des secteurs. Nous avons ainsi montré que cette méthode de rééquilibrage permet d'améliorer les sectorisations existantes.

Finalement, la *Sectorisation Équilibrée*, la *Sectorisation à partir de Centres* et le *Rééquilibrage de Sectorisations* sont des méthodes complémentaires permettant de répondre à la plupart des problématiques de sectorisation.

Conclusions

Dans ce chapitre, nous avons vu les expérimentations et applications concernant les trois points suivants : la *Classification de Données & la Visualisation de Classifications*, le *Lissage Spatial & la Détermination des Pôles*, et la *Sectorisation*.

Dans la première partie concernant la *Classification de Données & la Visualisation de Classifications*, nos expérimentations ont montré que notre algorithme de classification (la CAA dans sa version standard) produit des résultats meilleurs que ceux obtenus avec l'algorithme des k-moyennes en nous basant sur la perte d'inertie et le nombre d'objets mal classés comme indicateurs de qualité. Ensuite, nous avons exposé la *Classification de Variables* qui est une extension de la *Classification des Données* et qui permet d'analyser aisément les variables. Nous avons illustré la puissance de ces méthodes de visualisation dans le cadre de l'analyse des données socioprofessionnelles de Paris et de sa petite Couronne. Nous avons terminé cette partie en présentant la *Classification Hiérarchique Spatiale* qui est une adaptation de notre algorithme de *Classification de Données*. Cette méthode permet de construire les secteurs autour de zones de fortes valeurs et il s'agit donc d'une méthode de sectorisation complémentaire.

La partie suivante est dédiée aux applications du *Lissage Spatial & de la Détermination des Pôles*. En premier, nous avons illustré le fonctionnement de la méthode de *Détermination et de Hiérarchisation des Pôles*. La carte obtenue est alors très synthétique et donne une bonne image de la répartition géographique de la population française. Ensuite, nous avons aussi montré la pertinence de l'utilisation du lissage spatial avant de réaliser une classification de données : cela permet de réaliser une analyse des données pour des échelles spatiales plus grandes que la taille des objets étudiés. Les classes trouvées forment alors spatialement des zones beaucoup plus homogènes et la carte est alors plus simple à interpréter et plus esthétique.

La dernière partie traite des exemples d'application de nos méthodes de *Sectorisation* : la *Sectorisation Équilibrée*, la *Sectorisation à partir de Centres* et le *Rééquilibrage de Sectorisations*. Nous avons tout d'abord présenté le partage du territoire français en 22 secteurs égaux à l'aide de la *Sectorisation Équilibrée* et montré que les résultats sont très bons si l'on ne tient pas compte du fait que les frontières des secteurs passent assez souvent par les zones de fortes valeurs. Ensuite, nous avons vu que la *Sectorisation à partir de Centres* permet de construire des secteurs autour de centres de bonne qualité. Cependant, la compacité de la forme étant moins bien respectée, il est souvent utile de procéder à une amélioration de la forme des secteurs qui dans ce cas dégrade les objectifs de taille des secteurs. Nous sommes alors dans le cas d'une sectorisation existante compacte mais dont les objectifs de taille des secteurs sont « moyennement » atteints. Nous avons alors montré que notre algorithme de *Rééquilibrage de Sectorisations* permet alors d'atteindre les objectifs de taille en évitant au maximum de toucher à la forme des secteurs et contribue ainsi à l'amélioration de sectorisations existantes.

Tout au long de ce chapitre, nous avons aussi montré que les méthodes pouvaient avoir des applications très différentes (Par exemple, la *Classification de Variables* et la *Classification Hiérarchique Spatiale* sont issues de la *Classification de Données)* et aussi être complémentaires (Par exemple, l'utilisation du *Lissage Spatial* en prétraitement de la *Classification de Données*).

CONCLUSION GÉNÉRALE ET PERSPECTIVES

Notre démarche de recherche dans le cadre de l'ECD et des SIG a permis l'étude de méthodes d'ECD classique (la *Classification de Données* et la *Visualisation de Classifications*) et aussi des méthodes d'ECD propres aux SIG (le *Lissage Spatial* et la *Sectorisation*). Comme nous avons effectué notre recherche au sein de la société GÉOBS spécialisée dans l'analyse de données géographiques, nous avons étudié ces méthodes pour résoudre des problématiques liées au Géomarketing, à la Logistique, à l'Aménagement, à l'Environnement et à la Santé. De ce fait, toutes les méthodes ont pu être validées avec des données réelles.

Le **premier chapitre** sur l'**état de l'art** nous a permis de voir les techniques existantes relativement à ces quatre sujets : la *Classification de Données*, la *Visualisation de Classifications*, le *Lissage Spatial* et la *Sectorisation.* En ce qui concerne la *Classification de Données*, nous avons constaté qu'il existe un grand nombre d'algorithmes de classifications efficaces, dont certains incrémentaux mais qui ne sont pas forcément simples à paramétrer et à mettre en oeuvre. D'autres part, les données mixtes posent des problèmes quant à la définition de la distance à utiliser. Ensuite, dans la partie de ce chapitre consacrée à la *Visualisation de Classifications*, nous avons étudié les différentes possibilités de visualisation existantes sous forme de résumés et de tableaux. Nous avons remarqué que l'optimisation de l'ordre des classes et des variables rend les visualisations plus facilement lisibles. Puis nous avons ensuite observé que le *Lissage Spatial* permet de mettre en lumière les tendances se produisant aux grandes échelles spatiales, mais que le résultat fourni n'est pas synthétique et ne permet donc pas de comparer simplement les différentes cartes de lissage obtenues pour plusieurs échelles. Dans la dernière partie de ce chapitre, nous avons abordé des problématiques liées à la *Sectorisation*. Nous avons vu que les méthodes de partitionnement de graphes sont utilisables pour réaliser une *Sectorisation Équilibrée*, mais elles doivent être adaptées aux données géographiques. D'autre part, dans le cadre du rééquilibrage, la mise en œuvre des transferts est problématique (Quel ordre pour les transferts ? Faut-il réaliser les transferts en totalité ou pas-à-pas ?).

Dans le **second chapitre**, nous avons vu les **améliorations et contributions** que nous avons réalisées concernant les points suivants : la *Classification de Données pour de grands volumes de données mixtes*, la *Visualisation de Classifications*, la *Détermination de Pôles et leur Hiérarchisation*, la *Sectorisation*. Dans la première partie de ce chapitre sur la *Classification de Données*, nous avons mis au point un algorithme de classification incrémentale efficace : la Classification Ascendante Approximative (CAA) et son extension, la Classification Ascendante Hiérarchique Approximative (CAHA). Elles sont très rapides en raison de leur complexité linéaire par rapport au nombre d'objets à classer et très robustes en raison de leur simplicité (contrairement à d'autres méthodes mettant en jeu de multiples mécanismes d'éclatement, de fusion, de descente,…). Elles sont de plus très faciles à paramétrer à l'aide d'un seul paramètre k fixant le

nombre de résumés utilisés en mémoire. Nous avons aussi défini une mesure de dissimilarité permettant d'utiliser l'algorithme sur des données mixtes (à la fois quantitatives et qualitatives). Ensuite, en ce qui concerne la *Visualisation de Classifications*, nous avons élaboré la *Hiérarchie Évoluée* qui utilise des profils très lisibles car ils sont purement graphiques et ne comportent aucune information chiffrée. Nous avons aussi élaboré le *Tableau Évolué* qui est complémentaire de la *Hiérarchie Évoluée* : il permet de comparer rapidement les informations disponibles pour une même variable, ce que ne permet pas la *Hiérarchie Évoluée*. Nous avons aussi traité le problème de l'ordre optimal des variables d'une manière originale permettant de réaliser une analyse des variables. Dans la troisième partie de ce chapitre traitant de la *Détermination des Pôles et leur Hiérarchisation*, nous avons d'abord présenté la *Détermination des Pôles* qui permet de résumer une carte de lissage spatial par quelques pôles et aussi de décrire automatiquement les pôles. Ensuite, la *Hiérarchisation des Pôles* permet de superposer sur une même carte les pôles correspondant à des échelles différentes en les incluant dans une hiérarchie. La carte obtenue est alors très synthétique et la hiérarchie fournit de nouvelles informations explicatives : les pôles de l'échelle spatiale la plus courte permettent d'expliquer les pôles des échelles plus grandes. Nous achevons ce chapitre avec la dernière partie traitant de la *Sectorisation.* Nous avons adapté une méthode de partitionnement de graphes éprouvée et l'avons ensuite encapsulée dans un algorithme itératif réalisant la *Sectorisation Équilibrée* en calculant plusieurs sectorisations et gardant la meilleure. Nous avons ensuite développé la *Sectorisation à partir de Centres* qui permet de construire des secteurs autour de centres. L'algorithme de création des secteurs construit les secteurs progressivement en se basant sur des paramètres initiaux et l'algorithme itératif recommence autant que nécessaire ce processus en modifiant les paramètres initiaux et sélectionne la meilleure sectorisation obtenue. Pour finir, la méthode de *Rééquilibrage de Sectorisations* que nous avons définie permet d'atteindre progressivment l'équilibre en évitant au maximum de toucher à la forme des secteurs.

Le **troisième chapitre** présente les **expérimentations et applications** concernant les trois points suivants : la *Classification de Données & la Visualisation de Classifications*, le *Lissage Spatial & la Détermination des Pôles*, et la *Sectorisation*. Dans la première partie de ce chapitre concernant la *Classification de Données & la Visualisation de Classifications,* nos expérimentations ont montré que notre algorithme de classification (la CAA dans sa version standard) produit des résultats souvent meilleurs que ceux obtenus avec l'algorithme des k-moyennes. Puis nous avons exposé la *Classification de Variables* qui est une extension de la *Classification des Données* et qui permet d'analyser aisément les variables. Nous avons illustré la puissance de ces méthodes de visualisation dans le cadre de l'analyse des données socioprofessionnelles de Paris et de sa Petite Couronne. Nous avons ensuite présenté la *Classification Hiérarchique Spatiale* qui est une adaptation de notre algorithme de *Classification de Données* permettant de construire les secteurs autour de zones de fortes valeurs (il s'agit donc d'une méthode de sectorisation complémentaire). La partie suivante de ce dernier chapitre est dédiée aux applications du *Lissage Spatial & de la Détermination des Pôles*. En premier, nous avons utilisé la méthode de *Détermination et de Hiérarchisation des Pôles* pour obtenir une carte très synthétique donnant une bonne image de la répartition géographique de la population française. Ensuite, nous avons aussi montré la pertinence de l'utilisation du *Lissage Spatial* avant de réaliser une *Classification de Données* : cela permet de réaliser une analyse des données pour des échelles spatiales plus grandes que la taille des objets étudiés et la cartographie des classes est alors plus homogène et donc à la fois plus simple à interpréter et plus esthétique. Finalement, la dernière partie de ce dernier chapitre traite des exemples d'application de nos méthodes de *Sectorisation* : la *Sectorisation Équilibrée*, la *Sectorisation à partir de Centres* et le *Rééquilibrage de Sectorisations*. Nous avons tout d'abord présenté le partage du territoire français en 22 secteurs égaux à l'aide de la *Sectorisation Équilibrée* et obtenu de très bons résultats (si l'on n'accorde pas d'importance au fait que les frontières des secteurs passent assez souvent par les zones de forte population). Ensuite,

nous avons vu que la *Sectorisation à partir de Centres* permet de construire des secteurs autour de centres de bonne qualité, mais il est souvent utile de procéder à une amélioration de la forme des secteurs qui, en contrepartie, dégrade les objectifs de taille des secteurs. En partant de ce dernier résultat, nous avons alors montré que notre algorithme de *Rééquilibrage de Sectorisations* permet d'atteindre les objectifs de taille en évitant au maximum de toucher à la forme des secteurs et contribue ainsi à l'amélioration de sectorisations existantes.

Comme ces travaux de recherche ont été effectués dans le cadre de la phase Recherche du projet R&D de la société GÉOBS et à l'intérieur de cette entreprise, et, compte tenu de l'expertise et de la spécialisation de GÉOBS en SIG et en analyse de données géographiques, ces méthodes ECD ont été expérimentées, appliquées et vérifiées sur des jeux de données fournis par GÉOBS et liés à des problématiques de Développement Économique, de Géomarketing, d'Analyse de Risque, d'Environnement, de Santé, etc.

Pour conclure, l'étude des méthodes d'analyse retenues par GÉOBS dans le cadre de son projet R&D sur le thème de l'Extraction de Connaissances à partir de Données (ECD) dans le cadre des Systèmes d'Information Géographiques (SIG) nous a permis d'acquérir une vision globale, en termes d'ECD, de plusieurs problématiques et des méthodes d'analyse appropriées. Cela n'aurait pas été possible en se focalisant sur une seule problématique particulière ou une seule méthode particulière. Ainsi, nous avons vu que certaines méthodes étaient apparentées dans leur fonctionnement. Par exemple, la *Classification de Variables* et la *Classification Hiérarchique Spatiale* sont issues de la *Classification de Données*. Certaines méthodes sont complémentaires dans leur fonctionnement. Par exemple, la *Détermination de Pôles* se base sur les résultats du *Lissage Spatial* et la *Hiérarchisation des Pôles* se base sur les résultats de la *Détermination des Pôles*, et aussi, la *Sectorisation Équilibrée* et la *Sectorisation à partir de Centres* utilise le même indicateur pour mesurer la qualité des sectorisation obtenues et garder la meilleure. Nous avons aussi montré que certaines méthodes étaient complémentaires dans leur utilisation. Par exemple, l'utilisation du *Lissage Spatial* en prétraitement d'une *Classification de Données*, ou encore l'utilisation de la *Sectorisation à partir de Centres* avec les pôles déterminés par l'*Extraction de Pôles*, ou encore l'utilisation conjointe de la *Classification de Variables* et de la *Classification de Données* pour réaliser une bonne interprétation. Nous avons aussi vérifié que ces méthodes étaient bien complémentaires dans leur utilisation métier (Développement Économique, Géomarketing, Analyse de Risque, Environnement, Santé, etc.). Ainsi, certaines méthodes sont connectées les unes avec les autres soit du point de vue technique, soit du point de vue de leur utilisation.

Les perspectives et axes de recherche sont assez nombreux. En ce qui concerne la *Classification des Données*, dans le cadre d'une amélioration de notre algorithme de la CAA, nous envisageons de remplacer la distance euclidienne pondérée actuellement utilisée par la distance de Mahalanobis. Pour la *Visualisation de Classifications*, nous avons constaté que l'ordre optimal des résumés n'entraîne pas forcément un ordre optimal des classes construites par la hiérarchie et donc nous pensons explorer la possibilité de calculer dynamiquement l'ordre des classes de la hiérarchie afin d'avoir toujours une bonne lecture. Pour la *Détermination des Pôles*, nous souhaitons améliorer la qualité des résultats en utilisant les distances routières (ou le temps de parcours) à la place des distances à vol d'oiseau et nous étudions aussi les méthodes de lissage adaptatif afin d'améliorer la détection des pôles. Pour les méthodes de *Sectorisation*, nous souhaitons prendre en compte la compacité moyenne des secteurs dans le critère de qualité utilisé pour déterminer la meilleure sectorisation. Nous envisageons aussi d'améliorer les résultats en considérant que les solutions données par la *Sectorisation Équilibrée* et la *Sectorisation à partir de Centres* sont des sectorisations « brutes » et que le résultat final est donné après le *Rééquilibrage*. De plus, afin de créer un outil de *Rééquilibrage des Secteurs* plus efficace, nous

pensons utiliser l'algorithme de Krishnamurthy, ce qui par la même occasion améliorera les résultats des deux autres outils de sectorisation.

ANNEXE

Introduction

Cette annexe est consacrée à des méthodes non centrales dans notre travail de recherche, elles concernent l'*Arbre de Décision*, l'*Autocorrélation Spatiale* et la *Modélisation des Flux*.

Dans la première partie, nous nous intéressons à la méthode de l'*Arbre de Décision* (ou de segmentation) en considérant deux problématiques : la première est la recherche du meilleur partitionnement binaire concernant les modalités d'une variable qualitative et la seconde est de favoriser les règles fiables (c'est-à-dire toujours vraies). Tout d'abord, nous montrons que l'algorithme de la CAHA (déjà exposé dans la *Classification de Données*) permet de trouver très rapidement des solutions aussi bonne que l'algorithme SLIQ, qui est dans ce cas beaucoup plus lent en raison de sa complexité algorithme supérieure. Puis, nous exposerons notre algorithme favorisant les règles fiables et illustrerons son intérêt par un exemple.

La partie suivante traite de l'*Autocorrélation Spatiale* qui mesure le degré de ressemblance entre les objets proches. Nous montrerons tout d'abord que les différents coefficients existants ont un certain « biais » car ils ne traitent pas tous les objets de façon identique car ils pénalisent les objets situés en bordure. A partir de ce constat, nous proposons nos propres coefficients d'autocorrélation et nous démontrons qu'ils corrigent effectivement ce biais.

Dans la dernière partie, nous abordons la *Modélisation des Flux*. Nous partons du modèle de Tobler dont les résultats produits souffrent d'un défaut : les flux résultats peuvent être négatifs, ce qui est impossible dans la réalité. Nous proposons alors une amélioration et montrons que comme attendu, les corrections effectuées sont minimes et ne perturbent pas la modélisation des flux.

1 Contributions à la méthode de l'Arbre de Décision

Dans cette partie, nous nous intéressons aux arbres de décision (ou de segmentation). Cet outil cherche des règles afin d'expliquer les valeurs d'une variable cible. La première partie expose tout d'abord les méthodes de construction d'arbres de décision les plus communes. Puis nous verrons plus en détail les caractéristiques des arbres de décision binaires. Ce qui nous

permettra de soulever le problème de la recherche du meilleur partitionnement binaire concernant les modalités d'une variable qualitative. Ce problème est l'objet de la partie suivante.

Nous exposons une méthode utilisant un algorithme de classification hiérarchique (la CAH) permettant de trouver une solution approchée à ce problème. La première étape consiste en une modélisation des données et la définition de la distance utilisée. Nous montrons que, dans un premier temps, notre méthode n'est pas plus efficace qu'une autre méthode de référence (la méthode SLIQ [*MAR96*]). Dans un second temps, nous définissons plusieurs autres distances et nous montrons que dans les cas extrêmes, c'est-à-dire lorsque le nombre de modalités est très grand (au-delà de la centaine), l'algorithme de la CAHA (déjà exposé précédemment) permet de trouver très rapidement des solutions aussi bonnes que l'algorithme SLIQ, qui est dans ce cas beaucoup plus lent.

À partir des nombreuses expérimentations effectuées, nous avons mis en évidence que dans certains cas, les règles trouvées n'étaient pas les plus pertinentes du point de vue de l'utilisateur car il avait besoin de règles fiables et les règles comportant des exceptions ne lui étaient pas utiles (par exemple, dans le cas de la comestibilité des champignons). Pour cette raison, nous définissons donc un coefficient permettant à l'algorithme de favoriser les règles fiables. Nous illustrons ensuite son intérêt avec l'exemple de la comestibilité des champignons.

1.1 Rappels

Nous allons d'abord voir les généralités concernant la construction des arbres de décision et les méthodes les plus utilisées. Puis nous nous pencherons sur le cas particulier des arbres de décision binaires que nous allons utiliser par la suite.

1.1.1 Arbre de Décision

L'arbre de décision est une méthode permettant d'expliquer les modalités d'une variable en fonction des modalités d'autres variables dites « explicatives ». Son principal intérêt est de proposer les règles explicatives sous la forme d'un arbre. La construction de l'arbre se fait de manière descendante et récursive. La structure globale de l'algorithme est commune à la plupart des méthodes d'arbre de décision. Un « nœud » est une structure comportant :

- *I* des « individus »,
- *C* un « choix »,
- *Q* un « nombre » représentant la qualité de ce choix,
- *P* une « liste de nœuds » représentant la partition des individus *I* selon le choix *C*.

Un nœud ne comporte initialement aucun individu, aucune explication et aucune partition associée. L'algorithme s'initialise en appelant la fonction *CreerSousArbre(R)* avec *R* le nœud racine contenant tous les individus.

```
CreerSousArbre(N : nœud)
{
    • Etape 1 : Trouver la meilleure explication et les partitions associées :
        o TrouverMeilleureExplicationEtPartitionsAssociées(N)
    • Etape 2 : le cas échéant, itérer le processus pour toutes les partitions de N.P associées au
      choix N.C:
        o Si (taille de N.P > 1) Alors
            ▪ Pour i := 1 à taille de N.P Faire
                • CreerSousArbre(N.P(i))
            ▪ Fin Pour
        o Fin Si
}
```

Fig. 108 – Algorithme principal de l'Arbre de Décision

L'algorithme principal utilise l'algorithme de recherche du meilleur choix et des partitions associées. Nous décrivons ici une version très générale de cet algorithme. En effet, dans les versions optimisées, le calcul de certains types de choix peut être optimisé en se servant du calcul des choix effectués précédemment. Cependant, l'avantage de l'algorithme naïf est de mettre en évidence les points suivants :

- le calcul de la liste des choix pour le nœud en cours « ChoixDisponibles(N : nœud) »
- la création des partitions en fonction d'un choix « CreerPartitions(N : nœud, C : choix) »
- la mesure de la qualité du partitionnement « MesurerQualite(P : liste de nœuds) »

```
TrouverMeilleureExplicationEtPartitionsAssociées (N : nœud)
{
    • Etape 1 : Initialiser la liste des choix à tester
        o LC : « liste de choix »
        o LC=ChoixDisponibles(N)
    • Etape 2 : Trouver le meilleur choix
        o Pour i = 1 à taille de LC Faire
            ▪ Qact : «nombre »
            ▪ Pact : « liste de nœuds »
            ▪ Pact := CreerPartitions(N,LC(i))
            ▪ Qact := MesurerQualite(N.P)
            ▪ Si (Qact>N.Q) Alors
                • N.Q := Qact
                • N.P := Pact
            ▪ Fin Si
        o Fin Pour
}
```

Fig. 109 – Version naïve de l'algorithme de recherche du meilleur choix et des partitions associées

Nous allons maintenant voir les différentes caractéristiques des méthodes de création d'arbre de décision. Les plus anciennes ne prennent en charge que les variables qualitatives, le choix se portant sur une seule variable et ses modalités. Un premier tri entre les méthodes peut se faire sur le nombre de partitions créées par le choix. Une première méthode, CART [*BFOS84*] conduit à un partitionnement minimal, en deux partitions se partageant les modalités. Au contraire, les méthodes CLS [*HMQ66*], ID3 et C4.5 [*Quin93*] conduisent à un partitionnement

maximal, chaque partition correspondant à une modalité particulière de la variable choisie. Une autre famille comportant les méthodes AID [*MS63*], CHAID [*Kass80*] et leurs dérivées produit un nombre de partitions variant entre le partitionnement minimal et le partitionnement maximal. La prise en charge des variables quantitatives (continues ou ordinales) nécessite dans ce cas leur transformation en variables qualitatives par discrétisation.

La prise en charge directe des variables quantitatives conduit à l'utilisation d'un autre type de choix : il s'agit alors de trouver la valeur de coupure de la variable permettant de séparer les individus dont la valeur est inférieure à la valeur de coupure et ceux dont la valeurs est supérieure. Cette méthode est utilisée par C4.5 et CART.

Outre les méthodes de choix, le critère souvent mis en avant par les méthodes est la fonction permettant de calculer la qualité (ou pureté) du partitionnement. Les plus connus sont : le gain d'information (utilisé par ID3), le gain relatif d'information (utilisé par C4.5), l'indice de Gini (utilisé par CART) et le test du Chi2 (utilisé par CHAID).

Les algorithmes décrits précédemment sont univariés car le choix ne porte que sur une seule variable. Il existe des algorithmes multivariés ne fonctionnant généralement qu'avec des variables quantitatives : ils construisent une variable quantitative synthétique qui est une combinaison linéaire de variables quantitatives. Le choix est alors le point de coupure idéal de cette variable synthétique. Un exemple de méthode est LMDT [*BU92*].

Pour un état de l'art plus exhaustif, il convient de se reporter à l'article de S. Murthy [*Murt98*].

Pour conclure ce rapide tour d'horizon, la qualité globale d'un arbre de décision se mesure généralement avec deux indicateurs évoluant de manière contraires : le taux d'erreur et le nombre de règles utilisées. En effet, un faible taux d'erreurs est plus facilement obtenu par un arbre comportant un très grand nombre de règles et réciproquement un faible nombre de règles tend à augmenter considérablement le taux d'erreurs. Pour ces raisons, différentes stratégies d'élagage permettent de trouver un équilibre entre le taux d'erreurs et le nombre de règles en coupant les règles jugées inutiles. Ainsi l'élagage a pour but principal de diminuer de manière importante le nombre de règles (et leur longueur) tout en minimisant l'augmentation du taux d'erreur.

Nous allons maintenant nous intéresser au cas particulier des arbres de décision binaires.

1.1.2 Arbre de Décision Binaire

De manière générale les méthodes de construction d'arbre non binaires conduisent à la construction d'arbres larges, tandis que les méthodes de construction d'arbre binaires conduisent à la construction d'arbres longs. Cependant, à partir des observations établies par différents chercheurs [*LS97*, *Quin93*], les méthodes de construction d'arbres non binaires peuvent conduire à la construction de branches « inutiles » ou « redondantes » qui auraient dues êtres groupées. Ce phénomène de « fragmentation » est bien plus faible avec les méthodes de construction d'arbre binaires. Ainsi, le nombre de nœuds des arbres binaires est généralement inférieur aux autres arbres.

Dans le cadre des arbres binaires, la méthode de partitionnement des variables quantitatives est la recherche du point de coupure idéale, déjà utilisée par les arbres de décision classique. En ce qui concerne le choix du partitionnement binaire pour une variable qualitative, il existe plusieurs méthodes. La recherche exhaustive (comme dans CART) est très coûteuse car il y a environ 2^m choix à tester avec m le nombre de modalités de la variable. Ainsi, cette méthode n'est applicable que pour une variable ayant un petit nombre de modalités. D'autres méthodes,

telles que PC [*CHH99*], calculent un ordre d'agrégation des modalités quasi-optimal, cette méthode ne nécessite donc que seulement $2 \times m$ choix à tester. Elle se base sur une analyse en composantes principales de la matrice des probabilités des modalités et l'ordre est donné par la valeur des coefficients des modalités pour l'axe principal. Cette méthode n'est toutefois efficace que si les modalités sont corrélées. Une autre méthode est SLIQ [*MAR96*] qui nécessite de tester environ $\frac{m^2}{2}$ choix. Cette méthode s'initialise par une partition vide ne contenant aucune modalité et une partition pleine, contenant toutes les modalités. L'algorithme SLIQ fait alors passer progressivement toutes les modalités de la partition pleine à la partition vide, en choisissant à chaque étape de transférer la modalité qui donne la meilleure qualité pour les partitions créées. Finalement, le partitionnement final correspond à l'étape intermédiaire au cours de laquelle la qualité a été la meilleure.

Ces algorithmes de recherche du meilleur partitionnement sont trop coûteux ou bien ils nécessitent des conditions particulières (corrélation des variables par exemple). Nous allons donc voir par la suite notre méthode de recherche du meilleur partitionnement qui est plus rapide que les autres méthodes tout en produisant des résultats de qualité similaire.

1.2 Recherche du meilleur partitionnement binaire pour une variable qualitative

En nous basant sur les avis de plusieurs chercheurs, nous nous plaçons dans le cadre des arbres de décision binaires. Notre méthode de construction d'arbre de décision réutilise les fonctionnalités suivantes :

- la mesure de la qualité du partitionnement est le gain d'information.
- les choix sont univariés, c'est-à-dire qu'ils ne portent que sur une seule variable à la fois.
- pour une variable quantitative, on cherche le point de coupure idéal.

Notre innovation réside dans le choix du partitionnement binaire pour une variable qualitative.

Tout d'abord, nous allons développer notre méthode de choix de partitionnement qui est théoriquement plus exhaustive que celle de SLIQ en utilisant la CAH (Classification Ascendante Hiérarchique) et la CAHA (CAH Approximative développée dans le chapitre « Classification de Données »). Ces méthodes de classification utilisent dans ce cas une « pseudo distance » appelée la distance du Gain d'Impuretés en Partition Binaire (GIPB).

Nous comparerons ensuite cette distance avec d'autres distances classiques dans le cadre de la CAH (et de la CAHA) et nous montrerons quelles sont les meilleures distances pour la recherche du partitionnement binaire pour une variable qualitative.

Nous allons tout d'abord définir le codage des données afin que la CAH (ou la CAHA) puissent réaliser la recherche des meilleures partitions.

1.2.1 Codage des modalités

Le choix du partitionnement porte sur une variable cible *C* ayant les modalités C_1, …, C_n et une variable explicative *E* ayant les modalités E_1, …, E_m. L'exemple suivant porte sur des fruits et des légumes et montre les données initiales concernant la variable explicative *Couleur* (dont les modalités sont *Jaune*, *Rouge, Vert* et *Orange*) et la variable cible *Forme* (dont les modalités sont *Rond* et *Allongé*).

Fruit	Couleur	Forme
Pomme Golden	Jaune	Rond
Haricot vert	Vert	Allongé
Cerise	Rouge	Rond
…	…	…
Banane	Jaune	Allongé

Fig. 110 – Tableau de données sur les fruits et légumes

À partir des données, la matrice de contingence est construite. Il s'agit d'un tableau de comptage. Chaque case E_iC_j du tableau de contingence à l'intersection de la ligne E_i et de la colonne C_j donne le nombre d'exemples vérifiant à la fois la modalité E_i et la modalité C_j. Nous avons en outre besoin pour chaque modalité E_i et C_j du « Total », qui correspond aux nombre d'exemples où apparaît cette modalité : $Total(E_i) = \sum_{j=1}^{n} E_iC_j$ et $Total(C_j) = \sum_{i=1}^{m} E_iC_j$. Nous utilisons aussi le nombre total d'exemples $TotalGlobal = \sum_{i=1}^{m} Total(E_i) = \sum_{j=1}^{n} Total(C_j) = \sum_{i=1}^{m}\sum_{j=1}^{n} E_iC_j$.

Le tableau utilisé est ainsi les suivant :

	C_1	…	C_n	Total (Poids)
E_1	E_1C_1	…	E_1C_n	$Total(E_1)$
…	…	…	…	…
E_m	E_mC_1	…	E_mC_n	$Total(E_m)$

Fig. 111 – Tableau de contingence enrichi

L'exemple suivant est le tableau de contingence issu des données sur les fruits et légumes :

	Rond	Allongée	Total (Poids)
Rouge	10	1	11
Orange	5	1	6
Vert	1	5	6
Jaune	3	4	7
Total	19	11	30

Fig. 112 – Tableau de contingence sur la Couleur et la Forme à partir des données sur les fruits et légumes

Ensuite nous remplaçons chaque case par sa probabilité relative, la probabilité C_j sachant que E_i, $pE_iC_j = \frac{E_iC_j}{Total(E_i)}$

	C_1	…	C_n	Total (Poids)
E_1	pE_1C_1	…	pE_1C_n	$Total(E_1)$
…	…	…	…	…
E_m	pE_mC_1	…	pE_mC_n	$Total(E_m)$
Total	$Total(C_1)$	…	$Total(C_n)$	$TotalGlobal$

Fig. 113 – Tableau des probabilités

Le tableau de contingence de l'exemple devient le tableau des probabilités suivant:

	Rond	Allongée	Total
Rouge	0,9=10/11	0,1=1/11	11
Orange	0,8=5/6	0,2=1/6	6
Vert	0,2=1/6	0,8=5/6	6
Jaune	0,4=3/7	0,6=4/7	7
Total	19	11	30

Fig. 114 – Tableau des probabilités sur la Couleur et la Forme à partir des données sur les fruits et légumes

On remarque que, dès lors, le codage des données est achevé et qu'il est maintenant possible de réaliser une classification des variables explicatives. Intuitivement, un regroupement acceptable peut se faire en regroupant les modalités ayant des probabilités proches. Par exemple, regrouper *Rouge*=(0,9 ; 0,1) et *Orange*=(0,8 ; 0,2) est un bon choix. Dans ce cas, on aura plutôt une règle du type : Si le fruit est *Rouge* ou *Orange* alors il y a de fortes chances pour qu'il soit *Rond.*

La fusion de deux variables explicatives en une seule est la moyenne pondérée de ces deux variables. Soit $E_z = E_x \cup E_y$, nous avons :

pour $1 \le i \le n$: $E_zC_i = E_xC_i + E_yC_i$

d'où $pE_zC_i = \frac{E_zC_i}{Total(E_z)}$

$$= \frac{E_x C_i + E_y C_i}{Total(E_z)}$$

$$= \frac{pE_x C_i \times Total(E_x) + pE_y C_i \times Total(E_y)}{Total(E_x) + Total(E_y)}$$

L'exemple suivant montre le résultat de la fusion de *Rouge* et *Orange.* Nous avons:

$$E_{RougeEtOrange} C_{Rond} = E_{Rouge} C_{Rond} + E_{Orange} C_{Rond} = 10 + 5 = 15$$

$$E_{RougeEtOrange} C_{Allongé} = E_{Rouge} C_{Allongé} + E_{Orange} C_{Allongé} = 1 + 1 = 2$$

	Rond	Allongée	Total
Rouge	0,9=**10**/11	0,1=**1**/11	**11**
Orange	0,8=**5**/6	0,2=**1**/6	**6**
Vert	0,2=1/6	0,8=5/6	6
Jaune	0,4=3/7	0,6=4/7	7
Total	19	11	30

→

	Rond	Allongée	Total
Rouge ou Orange	0,9=**15**/17	0,1=**2**/17	**17**
Vert	0,2=1/6	0,8=5/6	6
Jaune	0,4=3/7	0,6=4/7	7
Total	19	11	30

Fig. 115 –Exemple de fusion de 2 modalités en une seule

La règle formée est la suivante : Si le fruit est *Rouge* ou *Orange* alors il est *Rond* dans 90 % des cas et *Allongé* dans 10 % des cas.

Une fois la fusion réalisée, on peut réitérer le processus en cherchant les deux modalités les plus semblables pour les fusionner et ainsi de suite.

1.2.2 Fonctions utilisées

Nous définissons ici les fonctions usuelles utilisées dans les arbres de décision. Il s'agit de :

- La fonction d'information donne la quantité d'information moyenne codée par une variable. On note la fonction d'information $Info(X)$ d'une variable *X* ayant X_1, …, X_t modalités sous différentes formes : la forme semi-développée est $Info(p_{X_1},....,p_{X_t})$ avec p_{X_i} la probabilité de la modalité X_i et la forme compacte est $Info(p_{X_i}, 1 \le i \le t)$.
- La fonction d'impureté qui donne la quantité d'information totale codée par une variable dans les données: $Impureté(X) = occurences(X) \times Info(X)$

Les fonctions d'information les plus utilisées sont l'entropie, $Entropie(p_{X_i}, 1 \le i \le n) = \sum_{i=1}^{n} - p_{X_i} \log(p_{X_i})$ et l'indice de Gini, $Gini(p_{X_i}, 1 \le i \le t) = 1 - \sum_{i=1}^{n} {p_{X_i}}^2$. Nous utilisons l'entropie comme fonction d'information.

Par ailleurs, dans le cadre de l'utilisation de ces fonctions dans le tableau des probabilités défini précédemment, nous avons :

- $Info(C\_sachant\_que\_E_i) = Info(C/E_i) = Info(pE_iC_j, 1 \le j \le n)$
- $Impureté(C/E_i) = Total(E_i) \times Info(pE_iC_j, 1 \le j \le n)$

1.2.3 Comparaison entre SLIQ et la CAH (et la CAHA)

Nous avons précédemment défini le codage des données dans le tableau des probabilités adéquats et aussi l'expression des fonctions usuelles des arbres de décision dans ce tableau des probabilités. Nous allons d'abord définir les distances simples que nous allons utiliser pour le regroupement des modalités. Puis nous définirons la « pseudo-distance » du Gain d'Impuretés en Partition Binaire afin de comparer l'algorithme SLIQ avec la CAH (et la CAHA). Nous la mettrons en œuvre dans le cadre de la CAH et de la CAHA afin de comparer les performances avec SLIQ.

1.2.3.1 Définition de la distance du Gain d'Impuretés en Partition Binaire

La distance du Gain d'Impuretés en Partition Binaire (GIPB) entre deux classes X et Y correspond au Gain d'Information obtenu par le partitionnement en une partition $Z = (X \cup Y)$ et sa partition complémentaire $\overline{Z} = (\overline{X \cup Y})$. Le calcul de la pseudo-distance GIPB est très rapide car les caractéristiques du complémentaire $\overline{Z}$ se déduisent simplement et immédiatement de Z.

Le Gain d'Information pour le partitionnement entre les modalités $X_1,...,X_n$ de la variable *X* pour la variable cible *C* s'exprime de la façon suivante :

$$\begin{aligned} Gain_C(X_1,...,X_n) &= Info(C) - \sum_{i=1}^{n} pX_i \times Info(C/X_i) \\ &= Info(C) - \sum_{i=1}^{n} \frac{Total(X_i) \times Info(C/X_i)}{TotalGlobal} \\ &= \frac{TotalGlobal \times Info(C) - \sum_{i=1}^{n} Total(X_i) \times Info(C/X_i)}{TotalGlobal} \\ &= \frac{1}{TotalGlobal} \times \left(Impureté(C) - \sum_{i=1}^{n} Impureté(C/X_i) \right) \end{aligned}$$

On remarquera que $Gain_C(X_1,...,X_n)$ varie entre 0 et $GainMax_C \left(= \frac{Impureté(C)}{TotalGlobal} \right)$

Nous définissons la pseudo distance GIPB entre E_x et E_y de la façon suivante :

$$DistGIPB(E_x, E_y) = Impureté(C/E_x \cup E_y) + Impureté(C/\overline{E_x \cup E_y})$$

Nous avons ainsi:

$$Gain_C\left(E_x \cup E_y, \overline{E_x \cup E_y}\right) = \frac{1}{TotalGlobal} \times \left(Impureté(C) - DistGIPB\left(E_x, E_y\right)\right)$$

Cela signifie que lorsque la pseudo-distance GIPB entre E_x et E_y est minimale, le gain de la partition entre $E_z = E_x \cup E_y$ et $\overline{E_z} = \overline{E_x \cup E_y}$ est maximal, ce qui est le but recherché. Les calculs de la distance GIPB se ramène ainsi à la construction d'une matrice des probabilités compactée en 2 modalités $E_z = E_x \cup E_y$ et $\overline{E_z} = \overline{E_x \cup E_y}$ comme le montre la figure suivante.

	C_1	…	C_n	*Total*
E_1	pE_1C_1	…	pE_1C_n	$Tot(E_1)$
…	…	…	…	…
E_x	pE_xC_1	…	pE_xC_n	$Tot(E_x)$
…	…	…	…	…
E_y	pE_yC_1	…	pE_yC_n	$Tot(E_y)$
…	…	…	…	…
E_m	pE_mC_1	…	pE_mC_n	$Tot(E_m)$
E_0	0	…	0	0

$\Rightarrow$

	C_1	…	C_n	*Total*
E_z	pE_zC_1	…	pE_zC_n	$Tot(E_z)$
$\overline{E_z}$	$\overline{pE_zC_1}$	…	$\overline{pE_zC_n}$	$Tot(\overline{E_z})$

Fig. 116 –Compactage des modalités en deux modalités $E_z = E_x \cup E_y$ et $\overline{E_z}$

On remarquera la présence d'une modalité E_0 qui représente la modalité vide. Son intérêt est de permettre le calcul de la pseudo distance GIPB entre une modalité E_i et rien (c'est-à-dire la modalité E_0) afin de ne pas oublier de calculer le gain entre E_i et $\overline{E_i}$. En effet, cette opération correspond au calcul du Gain du partitionnement binaire entre $E_i \cup E_0$ et $\overline{E_i \cup E_0}$, c'est-à-dire entre E_i et $\overline{E_i}$.

Nous détaillons ensuite le calcul de la distance entre E_x et E_y en fonction des informations de $E_z = E_x \cup E_y$ et $\overline{E_z} = \overline{E_x \cup E_y}$. Les informations sur E_z s'obtiennent par sommation car E_x et E_y sont des modalités bien distinctes et $\overline{E_z}$ s'obtient à partir de E_z :

$$Total(E_z) = Total(E_x) + Total(E_y)$$
$$Total(\overline{E_z}) = TotalGlobal - Total(E_z)$$

Et pour $1 \le i \le n$: $E_zC_i = E_xC_i + E_yC_i$

$$\overline{E_zC_i} = Total(C_i) - E_zC_i$$

Les probabilités se déduisent immédiatement :

pour $1 \le i \le n$: $pE_zC_i = \frac{E_zC_i}{Total(E_z)}$

$$\overline{pE_zC_i} = \frac{\overline{E_zC_i}}{Total\left(\overline{E_z}\right)}$$

L'exemple suivant montre la matrice des probabilités compactée correspondant au calcul de la pseudo-distance GIPB entre *Rouge* et *Orange* pour les données sur les fruits et légumes.

	Rond	Allongée	Total
Rouge	0,9=**10**/11	0,1=**1**/11	**11**
Orange	0,8=**5**/6	0,2=**1**/6	**6**
Vert	0,2=**1**/6	0,8=**5**/6	**6**
Jaune	0,4=**3**/7	0,6=**4**/7	**7**
Total	19	11	30

→

	Rond	Allongée	Total
Rouge ou Orange	0,9=**15**/17	0,1=**2**/17	**17**
Non (Rouge ou Orange)	0,3=**4**/13	0,7=**9**/13	**13**
Total	19	11	30

Fig. 117 –Exemple de compactage en 2 modalités $E_{RougeEtOrange}$ et $\overline{E_{RougeEtOrange}}$

Nous avons pour $E_{RougeEtOrange}$:

$$E_{RougeEtOrange}C_{Rond} = E_{Rouge}C_{Rond} + E_{Orange}C_{Rond} = 10 + 5 = 15$$
$$E_{RougeEtOrange}C_{Allongé} = E_{Rouge}C_{Allongé} + E_{Orange}C_{Allongé} = 1 + 1 = 2$$
$$Total\left(E_{RougeEtOrange}\right) = E_{RougeEtOrange}C_{Rond} + E_{RougeEtOrange}C_{Allongé} = 15 + 2 = 17$$

Et nous avons pour $\overline{E_{RougeEtOrange}}$:

$$\overline{E_{RougeEtOrange}C_{Rond}} = Total\left(C_{Rond}\right) - E_{RougeEtOrange}C_{Rond} = 19 - 15 = 4$$
$$\overline{E_{RougeEtOrange}C_{Allongé}} = Total\left(C_{Allongé}\right) - E_{RougeEtOrange}C_{Allongé} = 11 - 2 = 9$$
$$Total\left(\overline{E_{RougeEtOrange}}\right) = TotalGlobal - Total\left(E_{RougeEtOrange}\right) = 30 - 17 = 13$$

D'où :

$$DistGIPB\left(E_{Rouge}, E_{Orange}\right) = 17 \times Info\left(\frac{15}{17}, \frac{2}{17}\right) + 13 \times Info\left(\frac{4}{13}, \frac{9}{13}\right)$$

1.2.3.2 Expérimentations

Chaque algorithme (SLIQ, CAH ou CAHA) réalise la fusion progressive des modalités explicatives. Dans le cadre de la CAH (et de la CAHA), à chaque itération, les deux modalités E_x et E_y les plus proches du point de vue de la pseudo distance GIPB sont fusionnées et remplacer par $E_z = E_x \cup E_y$ dont on calcule les caractéristiques. Lorsque l'algorithme est terminé, la modalité E_z et la modalité complémentaire $\overline{E_z}$ correspondant à la distance du Gain d'Information minimum trouvée donne la meilleure partition. Nous sommes obligés de rechercher la meilleure distance GIPB parmi toutes les fusions réalisées car cette pseudo

distance est composée de $Impureté\left(C/E_x \cup E_y\right)$ qui croit en fonction des fusions et de $Impureté\left(C/\overline{E_x \cup E_y}\right)$ qui décroît en fonction des fusions. Ainsi, lors du processus, la pseudo-distance GIPB commence généralement par décroître pour atteindre son minimum puis recommence à croître. Il peut en outre exister plusieurs minima locaux.

Les premiers résultats sont décevants car l'utilisation de la CAH ou la CAHA n'améliore pas les résultats par rapport à SLIQ contrairement à ce qui avait été espéré. L'explication de l'échec de la CAHA est que la pseudo-distance GIPB est très sélective. En effet, elle a tendance à réutiliser la dernière modalité qu'elle vient de créer par fusion. Or le but de la CAH était justement de ne pas être obligé de réutiliser la modalité qui vient juste d'être créée car c'est ce que fait déjà l'algorithme SLIQ. Ainsi, au final, le chemin des fusions réalisées par la CAH est très souvent identique à celui réalisé par SLIQ.

Le bilan de cette étude est que la pseudo-distance GIPB n'est pas intéressante dans le cadre d'une utilisation avec la CAH ou la CAHA. Par la suite, nous allons tester d'autres distances dans le cadre de la CAH et de la CAHA dans le but de trouver les résultats de qualité équivalente mais de manière plus rapide qu'avec SLIQ.

1.2.4 Recherche des meilleures distances pour le partitionnement binaire en utilisant la CAH (et la CAHA)

Nous allons d'abord définir les distances qui seront utilisées par la CAH (et la CAHA) pour la recherche du meilleur partitionnement binaire. Puis nous réaliserons les expérimentations.

1.2.4.1 Définition des distances utilisées

Nous définissons les distances dont nous allons comparer les performances par la suite. Il s'agit de :

- La distance de Kullback-Leiber utilisant la divergence:

$$DistDiv\left(E_x, E_y\right) = KL\left(E_x, E_y\right) + KL\left(E_y, E_x\right)$$

Avec la divergence de Kullback-Leiber :

$$KL\left(E_x, E_y\right) = \sum_{i=1}^{n} pE_x C_i \log\left(\frac{pE_x C_i}{pE_y C_i}\right)$$

On remarquera que l'on peut écrire cette distance sous les formes suivantes qui sont toutefois plus lourdes à manipuler:

$$DistDiv\left(E_x, E_y\right) = \sum_{i=1}^{n} \left(pE_x C_i - pE_y C_i\right) \log\left(\frac{pE_x C_i}{pE_y C_i}\right)$$

ou bien $$DistDiv\left(E_x, E_y\right) = \sum_{i=1}^{n} \left(pE_y C_i - pE_x C_i\right) \log\left(\frac{pE_y C_i}{pE_x C_i}\right)$$

- La distance de Divergence Pondérée qui prend en compte le nombre d'occurrences et pas seulement les probabilités :

$$DistDivPond(E_x, E_y) = Total(E_x \cup E_y) \times DistDiv(E_x, E_y)$$

- La distance du Gain d'Impureté donne l'augmentation de l'impureté causée par la fusion de deux modalités en une seule. Elle permet ainsi de minimiser l'augmentation de l'impureté. Il y a un parallèle évident avec la distance de Ward qui permet de minimiser la perte d'inertie :

$$\begin{aligned} &DistGainImp(E_x, E_y) \\ &= Impureté(C/E_x \cup E_y) - (Impureté(C/E_x) + Impureté(C/E_y)) \end{aligned}$$

- La distance de Ward est très utilisée pour résumer l'information, elle minimise la perte d'inertie :

$$DistWard(E_x, E_y) = \frac{Total(E_x) \times Total(E_y)}{Total(E_x) + Total(E_y)} \times \sum_{i=1}^{n} (pE_x C_i - pE_y C_i)^2$$

1.2.4.2 Expérimentation et comparaison des distances

Nous utilisons une CAH sur la matrice des probabilités en utilisant les distances définies précédemment. Dans le cadre de l'utilisation de la pseudo-distance GIPB, nous prenons pour partition directe le nœud dont les fils ont la pseudo-distance GIPB la plus faible et nous en déduisons la partition complémentaire. À part pour la pseudo-distance GIPB, le meilleur partitionnement binaire trouvé par la CAH correspond généralement à la dernière fusion réalisée. À part pour la pseudo-distance GIPB, nous étudierons pour chaque distance les deux cas de figure suivant :

- La partition retenue correspond à la dernière fusion réalisée : il s'agit des deux nœud fils du nœud racine de la hiérarchie.
- La partition retenue est celle correspondant au nœud E_z de la hiérarchie *H* donnant le meilleur gain avec son complémentaire, c'est-à-dire tel que $Gain(E_z, \overline{E_z}) = \max_{E_x \in H} (Gain(E_x, \overline{E_x}))$

On peut ainsi résumer la recherche de la partition binaire par les deux étapes suivantes :

- Calcul de la hiérarchie en utilisant une des distances proposées.
- Sélection de la partition binaire selon une des deux méthodes décrites précédemment.

D'après les résultats obtenus, il est préférable de choisir la partition en explorant la hiérarchie plutôt qu'en prenant celle correspondant à la racine. Les analyses suivantes seront donc faites à partir des résultats obtenus en cherchant dans la hiérarchie la partition ayant le meilleur gain.

Le tableau suivant compare les distances pour un même jeu de données et un taux d'erreur identique. La qualité des résultats obtenus pour chaque distance est déterminée par le nombre de règles nécessaires et la profondeur atteinte (longueur des règles) pour obtenir le taux d'erreur choisi. Le nombre de règles est indiqué en haut dans chaque case tandis que la profondeur est indiquée en bas de chaque case. Les données *Brest-Cancer* (sur le cancer des poumons), *Mushroom* (sur la comestibilité des champignons) et *Autos* (sur la risque d'accident des types de voitures) sont des jeux de données classiques des Arbres de Décision et sont disponibles à l'adresse suivante : http://www.datalab.uci.edu/data/mldb-sgi/data/.

Données	Nb de mod. expl.	Nb de mod. cibl.	Taux d'erreurs	*Div*	*DivPond*	*GIPB*	*GainImp*	*Ward*
Simulées A	100	10	36 %	16 4	32 6	16 4	15 4	16 4
Simulées B	1000	10	0 %	11 4	10 4	11 4	10 4	11 4
Brest-Cancer	variable	2	19 %	12 4	12 4	12 4	12 4	12 4
Mushroom	variable	2	Aucune erreur	9 6	6 5	8 6	8 6	8 6
Autos	variable	6	13 %	15 à 24 4 à 5	15 4	15 à 26 4 à 5	15 à 22 4 à 5	14 à 24 4 à 5

Fig. 118 –Comparaison de la qualité des résultats en fonction de la distance utilisée

On observe peu de différence entre les performances de ces distances. Cela est en partie due à la méthode de recherche de la meilleur partition dans la hiérarchie construite : en effet, les différences sont plus grandes lorsqu'on prend pour partition la racine de la hiérarchie. De cette première étude, il ressort que le choix de la distance n'est pas crucial pour la qualité du partitionnement en utilisant la CAH.

Le tableau suivant donne les résultats des tests obtenus en remplaçant la CAH par la CAHA afin d'obtenir un algorithme très rapide.

Données	Nb de mod. expl.	Nb de mod. cibl.	Taux d'erreurs	Précision *k*	*Div*	*DivPond*	*GIPB*	*GainImp*	*Ward*
Simulées A	100	10	36 %	10	26 5	34 6	***31 à 48*** ***5 à 6***	15 4	16 4
Simulées A	100	10	36 %	100	16 4	32 6	16 4	15 4	16 4
Simulées B	1000	10	0 %	10	11 4	10 4	***+ de 9*** ***+ de 5***	10 4	11 4
Simulées B	1000	10	0 %	100	11 4	10 4	***+ de 8*** ***+ de 5***	10 4	11 4
Simulées B	1000	10	0 %	1000	11 4	10 4	11 4	10 4	11 4

Fig. 119 –Comparaison de la qualité des résultats en fonction de la distance utilisée et de la précision k de la CAHA

On observe que la qualité des résultats dépend des distances utilisées et de la précision utilisée. La distance la plus sensible à la précision est la pseudo-distance GIPB qui nécessite que le paramètre de précision soit très proche de la taille des données traitées, c'est-à-dire proche du nombre de modalités de la variable explicative. Les autres distances supportent très bien des faibles valeurs de précision de l'ordre de *k=10*.

Ainsi, dans le cas, où le nombre de modalités des variables explicatives est faible (inférieur à 10, par exemple), il n'y a pas de réel intérêt à utiliser la CAHA même si elle permet pour trouver de bons résultats. Au contraire, si le nombre de modalités des variables explicatives est élevé (supérieur à 10), il convient d'utiliser la CAHA avec une précision faible pour que les calculs soient rapides. Dans ce cas, il est préférable d'utiliser la distance de Ward ou la distance du Gain d'Impureté afin que les résultats restent de bonne qualité.

Pour cette raison, nous utilisons la CAHA avec la distance du Gain d'Impuretés et une précision *k*=10 qui donne de bons résultats dans tous les cas. Ainsi, notre algorithme de création d'Arbre de Décision est très rapide et gère très bien les variables qualitatives ayant un très grand nombre de modalités (plus de 1000 modalités par exemple). En effet, la complexité en temps est minimale : elle est linéaire en fonction du nombre de modalités *m* de la variable explicative. Pour rappel dans le cadre de la CAH, la complexité en temps est de l'ordre de m^3, dans le cadre de SLIQ, elle est de l'ordre de m^2, et pour la recherche exhaustive, elle est de l'ordre de 2^m. De plus, contrairement à l'algorithme rapide PC, elle ne nécessite pas que les modalités soient corrélées.

1.3 Facteur de correction avantageant la création de partitions pures

Nous avons remarqué que dans certains cas, l'arbre de décision donnait des règles assez longues concernant un grand nombre d'individus alors qu'il existait des règles courtes pour expliquer la majeure partie de ces individus. Nous allons donc proposer un facteur de correction permettant de favoriser les règles courtes.

1.3.1 Définition du facteur de correction

Ce coefficient correcteur s'intègre dans la formule calculant la quantité d'impuretés afin d'avantager les fusions créant des partitions pures. Ainsi, ce coefficient correcteur n'intervient pas dans la construction de la hiérarchie mais seulement au moment de la recherche de la meilleure partition binaire dans la hiérarchie. La nouvelle formule du Gain Corrigé d'un nœud E_z créé par la fusion de E_x et E_y est la suivante :

$$GainCorrigé_C\left(E_z, \overline{E_z}\right)$$
$$= \frac{1}{TotalGlobal} \times \left(Impureté(C) - CoeffPB\left(E_z, \overline{E_z}\right) \times \left(Impureté\left(C/E_z\right) + Impureté\left(C/\overline{E_z}\right)\right)\right)$$

Avec $CoeffPB(E_z, \overline{E_z})$ $= \min(Coeff(E_z), Coeff(\overline{E_z}))$ si E_z ou $\overline{E_z}$ est pure
$= 1$ sinon

Nous définissons $Coeff$ de manière à ce qu'il varie entre 0 et 1 et dépende du rapport entre le nombre d'individus contenus dans la partition pure et du nombre d'individus total. Ainsi plus la partition pure contient un grand pourcentage des individus, plus le coefficient sera faible et plus le Gain Corrigé sera grand. Ainsi nous aurons :

- si E_z ou $\overline{E_z}$ est pure, $GainMax_C > GainCorrigé_C(E_z, \overline{E_z}) > Gain_C(E_z, \overline{E_z}) > 0$ et plus la partition pure concernera d'individus, plus le gain corrigé sera proche de $GainMax_C$.
- sinon $GainCorrigé_C(E_z, \overline{E_z}) = Gain_C(E_z, \overline{E_z})$

On pose donc :

$$Coeff(E_z) = \left(1 - \frac{Total(E_z)}{TotalEx}\right)^{\alpha} = (1-x)^{\alpha}$$

Avec α un coefficient d'intensité positif valant 1 par défaut.

$x = \frac{Total(E_z)}{TotalEx}$ qui varie entre 0 lorsque la partition pure est vide et 1 lorsque la partition pure contient tout les individus.

Le graphique ci-dessous montre l'évolution du coefficient en fonction de x pour différentes valeurs de α.

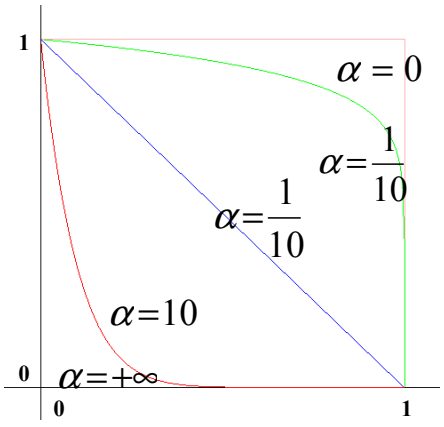


Fig. 120 –Différentes fonction de correction en fonction du nombre d'individus concernés donné par x variant de 0 à 1

Ainsi, pour un coefficient α nul, le coefficient vaut 1 et le Gain Corrigé est identique au Gain standard. Pour un coefficient α infini, le coefficient est nul et le Gain Corrigé est égale au Gain maximal si la partition est pure, ce qui l'avantage systématiquement par rapport aux autres partitions non pures. Par défaut α vaut 1 et le coefficient décroît linéairement, par exemple, si la partition pure contient la moitié des individus, la quantité d'impuretés prises en

compte sera divisée par deux et le Gain Corrigé augmenté d'autant par rapport au Gain standard.

Le paramétrage de la valeur de α permet ainsi d'avantager plus ou moins les partitions pures par rapport aux autres partitions compte tenu du nombre d'individus qu'elles contiennent.

1.3.2 Expérimentations

L'exemple suivant montre l'intérêt que peut avoir l'utilisation de ce coefficient. Les données concernent les champignons (fichier *MUSHROOMS*).

Données	Nb de mod. expl.	Nb de mod. cibl.	Taux d'erreurs	Avec facteur de correction	*Div*	*DivPond*	*GIPB*	*GainImp*	*Ward*
Mushroom	?	2	Aucune erreur	Non	9 6	***6*** ***5***	8 6	8 6	8 6
Mushroom	?	2	Aucune erreur	Oui	***6*** ***5***	***6*** ***5***	***6*** ***5***	***6*** ***5***	***6*** ***5***

Fig. 121 –Comparaison de l'influence du facteur de correction sur la qualité des résultats en fonction de la distance utilisée

On constate que quelque soit la distance, l'utilisation du facteur de correction conduit au même résultat permettant de trier les champignons en seulement 6 règles dont les 2 premières permettent de trier avec certitude 91 % des champignons.

La figure suivante est l'un des résultats initiaux en 8 règles.

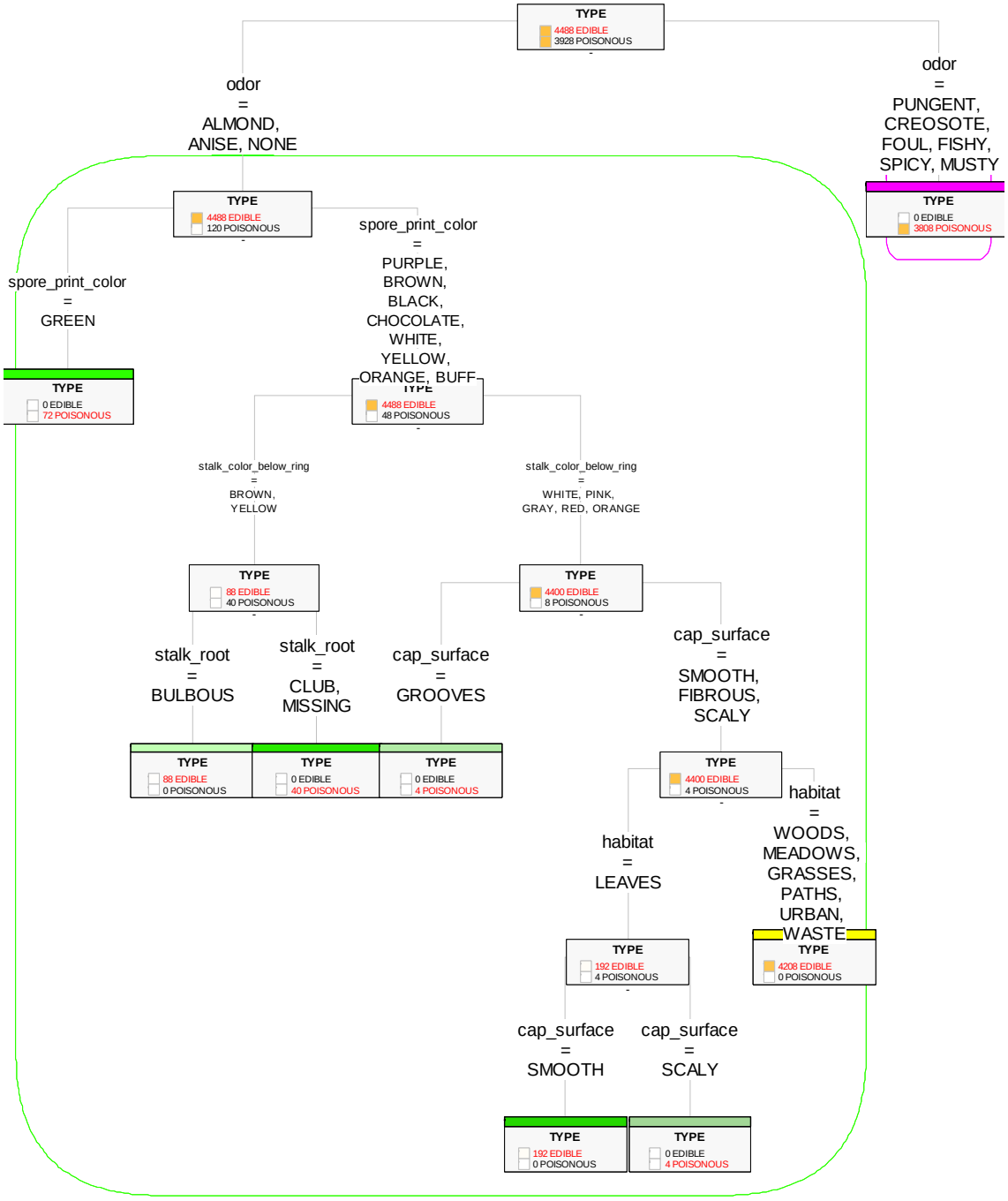


Fig. 122 –Arbre de décision obtenu sur les données Mushrooms avec la distance de Ward sans le facteur de correction

On remarque que la plupart des champignons non comestibles sont triés avec certitude dès le début (règle en mauve), tandis que la règle opposée (en vert) ne trie pas avec certitude les champignons comestibles. De plus la règle permettant d'obtenir la plupart des champignons comestibles est très longue (règle en jaune) :

- Règle mauve :

  Si odor = PUNGENT, CREOSOTE, FOUL, FISHY, SPICY, MUSTY alors

dans 100% des cas (3808/3808) TYPE = POISONOUS

- Règle verte :

  Si odor = ALMOND, ANISE, NONE alors

  dans 97% des cas (4488/4608) TYPE = EDIBLE

  dans 2% des cas (120/4608) TYPE = POISONOUS

- Règle jaune :

  _ Si odor = ALMOND, ANISE, NONE et spore_print_color = PURPLE, BROWN, BLACK, CHOCOLATE, WHITE, YELLOW, ORANGE, BUFF et stalk_color_below_ring = WHITE, PINK, GRAY, RED, ORANGE et cap_surface = SMOOTH, FIBROUS, SCALY et habitat = WOODS, MEADOWS, GRASSES, PATHS, URBAN, WASTE alors

  dans 100% des cas (4208/4208) TYPE = EDIBLE

L'arbre obtenu avec le facteur de correction est beaucoup plus intéressant, car il permet de trier dès le départ et avec certitude 91 % des champignons.

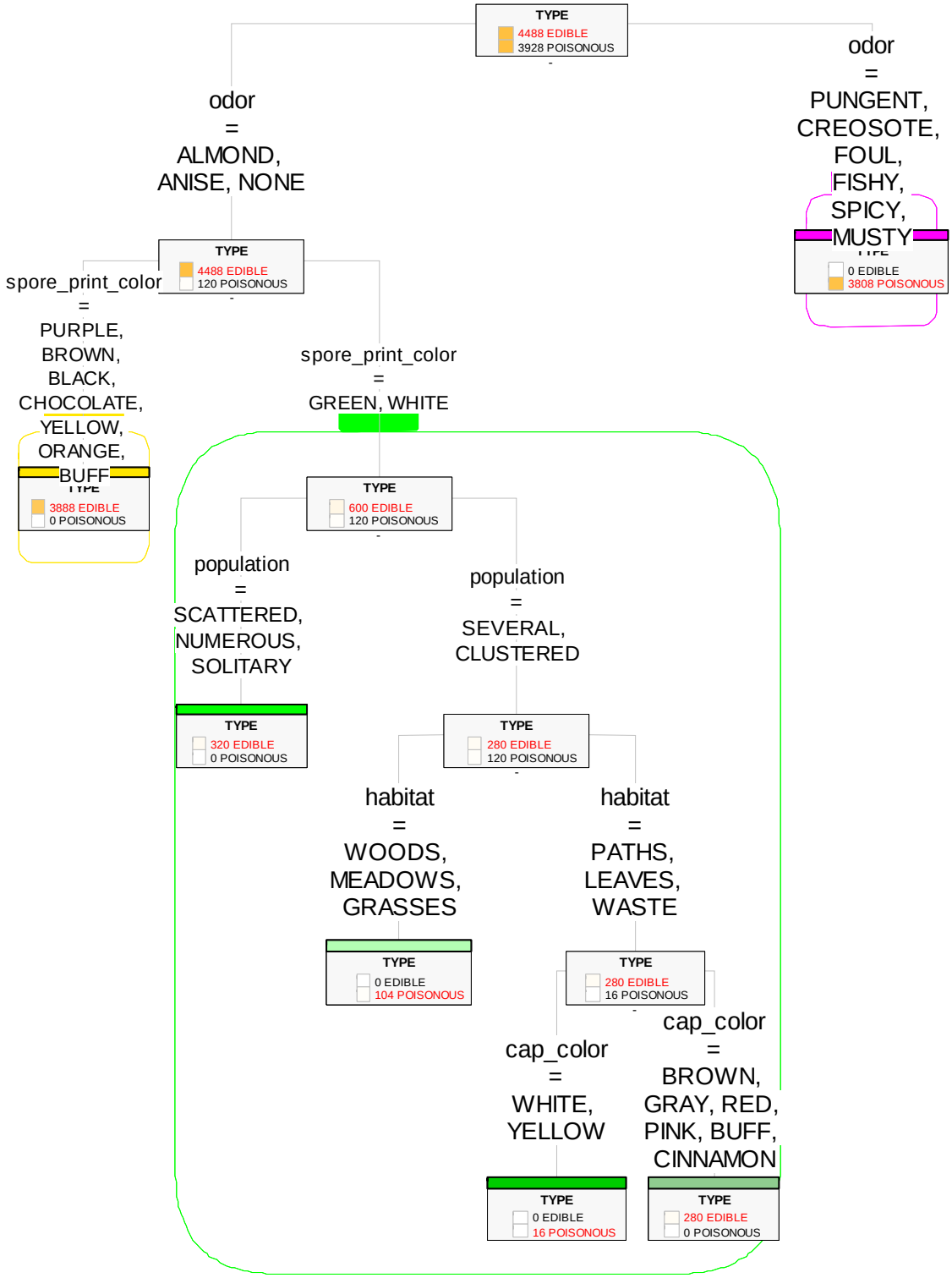


Fig. 123 –Arbre de décision obtenu sur les données Mushrooms avec la distance de Ward avec le facteur de correction

On remarque que cette fois-ci, il existe une règle (en jaune) permettant de trier avec certitude la plupart des champignons comestibles. Les 4 autres règles (en vert) permettent de trier les champignons restants (10% du total). La règle jaune est la suivante :

- Si odor = ALMOND, ANISE, NONE et spore_print_color = PURPLE, BROWN, BLACK, CHOCOLATE, YELLOW, ORANGE, BUFF alors dans 100% des cas (3888/3888) TYPE = EDIBLE

On remarquera que paradoxalement, cette règle est bien plus courte que la règle jaune précédente mais qu'elle concerne moins de champignons comestibles : seulement 3888 au lieu de 4208.

Cet exemple illustre l'utilité de la méthode avantageant les partitions pures ayant un grand nombre d'individus. Nous avons testé ce coefficient sur de nombreux autres exemples mais les améliorations procurées par cette méthode sont alors mitigées. D'une manière générale, les différences avec la méthode standard sont alors très faibles car les partitions pures sont souvent peu intéressantes en raison du faible nombre d'objet concernés. De plus, ce léger bénéfice est contrebalancé par une moins bonne qualité des partitions non pures. Ce qui fait que globalement le taux d'erreurs est plus élevé qu'avec la méthode standard pour un même nombre de règles. C'est d'ailleurs le cas avec l'exemple sur les champignons car pour les 3 règles les plus courtes, la méthode standard donne 48 erreurs au total tandis que la nouvelle méthode donne 120 erreurs. Par contre dans ce cas, les données incertaines étaient de 4536 (48 non comestibles parmi 4536 champignons) contre 600 (120 non comestibles parmi 600 champignons) avec la nouvelle méthode d'où son intérêt lorsque l'on cherche en priorité des règles courtes et fiables concernant un grand nombre d'individus.

1.4 Conclusion

Dans le cadre des arbres de décision, nous avons traité le problème de la recherche d'un partitionnement binaire quasi-optimal des modalités d'une variable qualitative. Nous avons mis au point une méthode rapide et efficace basée sur l'algorithme de la CAHA (en utilisant la distance de Ward ou la distance du Gain d'Impuretés). Son principal avantage est que sa complexité est linéaire et qu'elle ne nécessite aucun pré-requis sur les modalités (comme par exemple, une bonne corrélation entre les modalités). Ainsi, elle convient particulièrement aux cas dans lesquels les modalités sont très nombreuses (supérieures à la centaine) car les autres méthodes sont alors trop lentes pour être utilisées.

Nous avons aussi créé un facteur de correction permettant d'avantager les règles fiables. Il n'est réellement efficace que dans certains cas où ces règles existent, comme par exemple pour la détermination de la comestibilité des champignons (données *Mushroom*). Dans les autres cas, il n'y a pas de réel bénéfice car la nécessité de fiabilité est moins forte et/ou de telles règles fiables sont rares ou trop longues. Dans ces cas, les résultats sont même un peu moins bons concernant le taux d'erreurs. Nous cherchons donc d'autres jeux de données où de telles règles fiables et courtes existent afin d'illustrer davantage l'intérêt de ce facteur de correction.

2 Contributions à l'amélioration des coefficients d'Autocorrélation Spatiale

L'autocorrélation spatiale mesure le degré de ressemblance entre les objets proches. Il existe différents coefficients d'autocorrélation spatiale. Les plus communs sont les coefficients globaux : le résultat est un chiffre concernant tous les objets géographiques. Le résultat indique donc une tendance générale. Cependant, les plus intéressants sont les coefficients locaux car ils sont calculés pour chaque objet et peuvent donc être cartographiés.

Nous allons voir dans une première partie les coefficients d'autocorrélation spatiale existants et nous illustrerons leur intérêt pour analyser les variations du taux d'agriculteurs au niveau départemental en France. Nous étudierons ensuite les points faibles de ces indicateurs dont le principal est le traitement inégal entre les objets ayant beaucoup de voisins et ceux en ayant peu.

Nous montrerons comment les indicateurs que nous proposons évitent ces écueils. Nous illustrerons leur fonctionnement sur le même exemple en insistant sur la facilité de la cartographie des nouveaux indicateurs locaux.

2.1 Etat de l'art

Nous allons voir en premier les coefficients d'autocorrélation spatiale globaux, puis les coefficients locaux définis à partir des coefficients globaux. Nous illustrerons finalement leur fonctionnement et leur intérêt pour analyser les variations du taux d'agriculteurs au niveau départemental.

2.1.1 Autocorrélation spatiale globale

Les coefficients globaux les plus utilisés sont le coefficient de Moran [*Mora50*] et le coefficient de Geary [*Gear50*]. Ils sont calculés pour une variable quantitative donnée. Ces deux coefficients sont discutés et comparés dans le livre de D. Pumain et T. Saint-Julien [*PS97*]. Nous allons toutefois résumer brièvement leur fonctionnement. Les deux coefficients ont des échelles de valeurs distinctes :

- À la manière d'une corrélation, le coefficient de *Moran Global* varie à peu près entre -1 et 1 et la valeur pivot est 0.
- Comme un ratio d'écart-type, le coefficient de *Geary Global* varie entre 0 et l'infini et la valeur pivot est 1.

Cependant, indépendamment de leur signification respective, ces deux coefficients offre une même interprétation du biais spatial :

- Si le coefficient de *Geary Global* est inférieur à 1 ou le coefficient de *Moran Global* est positif (proche de 1) : les objets proches sont plus semblables que les objets pris au hasard et il y a donc un biais spatial.
- Si le coefficient de *Geary Global* est proche de 1 ou le coefficient de *Moran Global* est nul : Les objets proches ne sont ni plus dissemblables, ni plus semblables que les objets pris au hasard. Le voisinage ne joue donc aucun rôle quant à la prédictibilité des valeurs et il n'y a donc pas de biais spatial.
- Si le coefficient de *Geary Global* est supérieur à 1 ou le coefficient de *Moran Global* est négatif (proche de -1) : les objets proches sont plus dissemblables que les objets pris au hasard et il y a donc un biais spatial.

Dans le détail, l'interprétation du coefficient de *Moran Global* est la suivante :

- Si le coefficient de *Moran Global* est positif (proche de 1) : les objets proches sont « plutôt situés du même côté de la moyenne ». C'est-à-dire qu'ils sont plutôt inférieures à la moyenne de la variable quantitative ou bien plutôt supérieures à la moyenne.
- Si le coefficient de *Moran Global* est nul : les objets proches sont plutôt proches de la moyenne et/ou « situés indifféremment de part et d'autre de la moyenne »
- Si le coefficient de *Moran Global* est négatif (proche de -1) : les objets proches sont « situés à l'opposé de chaque côté de la moyenne ». C'est-à-dire qu'un lieu plutôt inférieur à la moyenne aura la plupart de ses voisins plutôt supérieures à la moyenne et inversement.

L'interprétation du coefficient de *Geary Global* est la suivante :

- Si le coefficient de *Geary Global* est proche de l'unité (proche de 1) : la dispersion moyenne entre chaque objet et son voisinage est la même que la dispersion globale.
- Si le coefficient de *Geary Global* est supérieur à l'unité (supérieur à 1) : la dispersion moyenne entre chaque objet et son voisinage est plus élevée que la dispersion globale.
- Si le coefficient de *Geary Global* est inférieur à l'unité (inférieur à 1) : la dispersion moyenne entre chaque objet et son voisinage est plus faible que la dispersion globale.

Nous allons maintenant voir dans le détail les formules des coefficients globaux.

Soit N le nombre d'unités spatiales élémentaires de l'ensemble géographique observé

X la variable étudiée

X_i la valeur de l'objet O_i pour la variable X

$\overline{X}$ la moyenne du caractère étudié pour les N unités spatiales

l_{ij} le lien entre i et j (1 s'il y a contiguïté, 0 sinon)

Nous avons :

- la variance globale de la variable X : $V(X) = \frac{1}{N}\sum_{i=0}^{N}(X_i - \overline{X})^2$
- la somme des liens : $\mathrm{L} = \sum_{i=1}^{N}\sum_{j=1, i\neq j}^{N} l_{ij}$

Les formules des coefficients sont les suivantes :

$$MoranGlobal(X) = \frac{1}{V(X)\times L}\times\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} l_{ij}\times(X_i - \overline{X})\times(X_j - \overline{X})$$

Et $$GearyGlobal(X) = \frac{1}{V(X) \times 2 \times L} \times \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} l_{ij} \times (X_i - X_j)^2$$

Nous reviendrons ultérieurement plus en détail sur les formules. Mais elles sont nécessaires pour présenter l'autocorrélation spatiale locale qui suit.

2.1.2 Autocorrélation spatiale locale

Le principal indicateur d'autocorrélation spatiale locale est celui issu de la formule du coefficient de Moran. Il s'agit du LISA (*Local Indicators of Spatial Association*) [*Anse95*]. Techniquement, il isole la partie concernant un objet et ses voisins de manière à ce que la somme des indicateurs locaux soit égale au coefficient de Moran global.

Nous avons donc le coefficient de Moran local pour l'objet O_i et la variable X définit par :

$$MoranLocal(X, O_i) = \frac{N}{V(X) \times L} \times \sum_{j=1, j \neq i}^{N} \left(l_{ij} \times (X_i - \overline{X}) \times (X_j - \overline{X}) \right)$$

d'où $$MoranGlobal(X) = \frac{1}{N} \sum_{i=1}^{N} MoranLocal(X, O_i)$$

De fait, contrairement au coefficient de Moran global, ce coefficient local peut être inférieur à -1 ou supérieur à 1. Par contre, la valeur pivot reste toujours 0 et l'interprétation est sensiblement identique. Ainsi, l'interprétation du coefficient de Moran local d'un objet O_i peut être la suivante :

- Si son coefficient de *Moran Local* est négatif alors sa valeur X_i est globalement opposée à celles de son entourage :
  - Si X_i est supérieur à la moyenne $\left(X_i > \overline{X}\right)$ alors les valeurs dans son entourage sont globalement inférieures à la moyenne.
  - Sinon, X_i est inférieur à la moyenne et les valeurs dans son entourage est globalement supérieures à la moyenne.
- Si son coefficient de *Moran Local* est positif alors sa valeur X_i va globalement « dans le même sens » que celles de son entourage :
  - Si X_i est supérieur à la moyenne $\left(X_i > \overline{X}\right)$ alors les valeurs dans son entourage sont aussi globalement supérieures à la moyenne.
  - Sinon, X_i est inférieur à la moyenne et les valeurs dans son entourage sont aussi globalement inférieures à la moyenne.

Ce coefficient a été utilisé pour la détection des zones où il se produit le plus d'accidents de la route [*Flah01*]. Une autre application de ce coefficient est l'analyse de la distribution spatiale des cancers sur l'île de Long Island [*JG03*]. On peut toutefois remarquer que dans les deux cas, le but principal était de faire apparaître des zones de fortes valeurs, ainsi il aurait mieux valu utiliser le lissage spatial qui été plus adapté pour cette tâche précise.

2.1.3 Exemple

L'exemple suivant permet d'illustrer l'intérêt et le fonctionnement des coefficients d'autocorrélation globaux et locaux. Les données sont les taux d'agriculteurs en fonction de la population active pour les départements français.

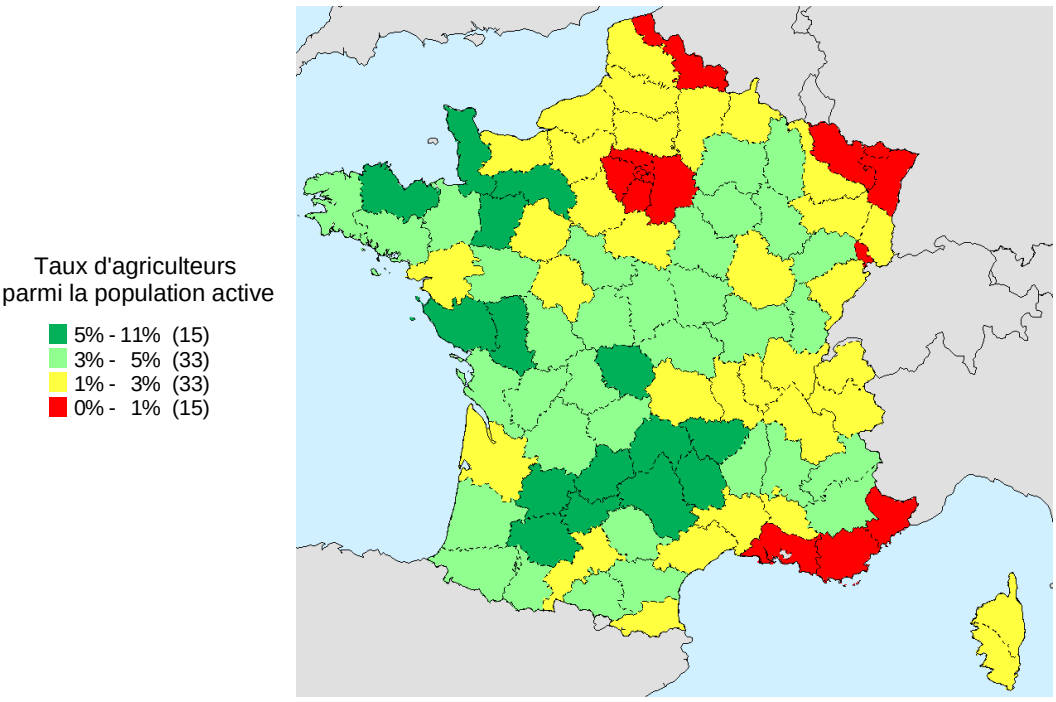


Fig. 124 –Carte des taux d'agriculteurs par départements

Sachant que la moyenne nationale du taux d'agriculteurs est 3 %, nous avons les départements très ruraux (5% à 10%) en vert foncé, assez ruraux (3% à 5%) en vert clair, assez urbanisés (de 1% à 3%) en jaune et très urbanisés (moins de 1%) en rouge.

Les coefficients globaux sont :

- Coefficient de *Moran Global* = 0,49
- Coefficient de *Geary Global* = 0,57

Le coefficient de *Moran Global* étant compris entre 0 et 1, on en déduit que les objets proches ont tendance à se ressembler. De même le coefficient de *Geary Global* est compris entre 0 et 1 et confirme que les objets proches ont tendance à se ressembler.

La carte suivante montre les coefficients de *Moran Locaux*.

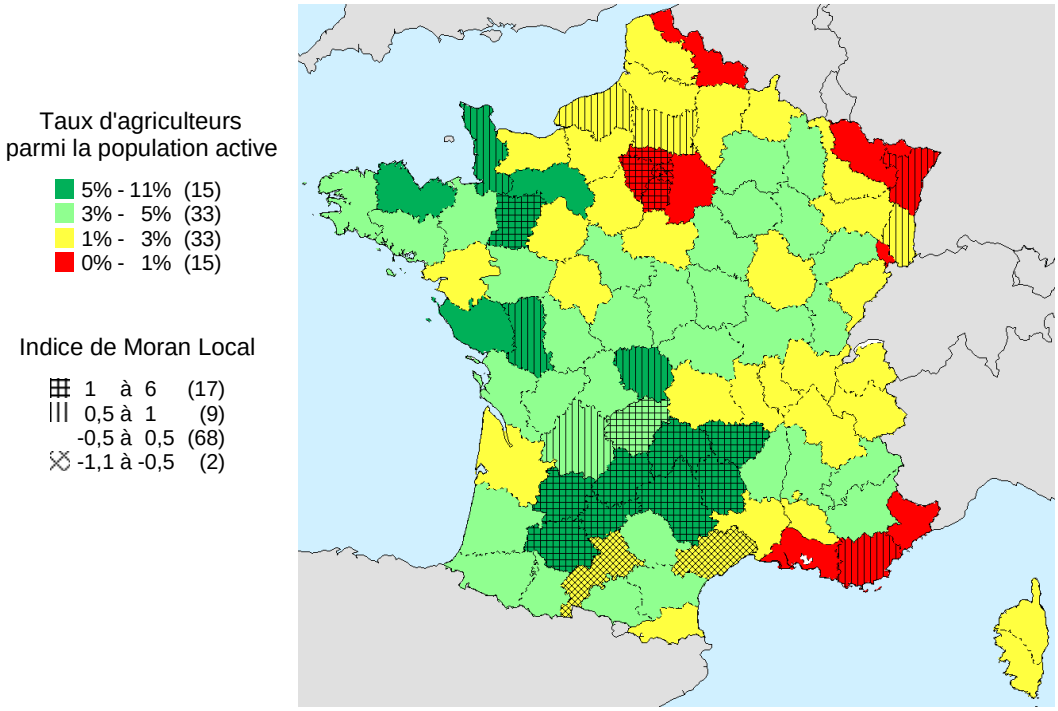


Fig. 125 –Carte des coefficients de Moran Locaux pour le taux d'agriculteur par départements

On observe qu'il y a davantage de départements qui ressemblent à leurs voisins (26 départements dont le coefficient est compris supérieur à 0,5) que de départements opposés à leurs voisins (2 départements dont le coefficient est inférieur à -0,5). On peut faire l'analyse suivante :

- Parmi les départements ressemblants à leurs voisins (en hachures verticales et horizontales), nous avons :
  - Dans les zones rurales : principalement le Massif Central et une partie du Sud-Ouest.
  - Dans les zones urbaines : principalement l'île de France et l'Alsace
- Parmi les départements opposés à leurs voisins (hachures en diagonales), nous avons surtout les départements contenant les grandes agglomérations du Sud-Ouest (Toulouse et Montpellier) qui sont entourés de départements ruraux.

Nous allons maintenant voir les améliorations apportées aux coefficients locaux et la réinterprétation des coefficients globaux.

2.2 Amélioration des coefficients

Nous verrons en premier l'amélioration des coefficients locaux car elle est la plus intéressante. Nous en déduirons une amélioration des coefficients globaux. Puis, nous détaillerons le nouveau coefficient de Geary symétrique qui est plus facile à cartographier. Finalement, nous verrons l'utilité de ces coefficients sur des exemples.

2.2.1 Amélioration des coefficients locaux

Le coefficient de *Moran Local* souffre de plusieurs défauts :

- Sa valeur est proportionnelle au nombre de voisins. Ainsi, il est difficile de comparer deux valeurs de ce coefficient pour deux objets ayant un nombre de voisins différents.
- De plus, pour que la valeur puisse être comparable à un coefficient de *Moran Global* afin d'utiliser les valeurs de référence -1 et 1, la valeur est corrigé par le nombre de voisins moyen $L_{moy} = \frac{L}{N}$.

Ainsi, on se rend compte qu'il y a un problème au niveau de la prise en compte du nombre de voisins. Cela nous a amené à proposer un nouveau coefficient de *Moran Local* que nous appelons coefficient de *Moran Local Amélioré*.

Nous avions :

$$MoranLocal(X,O_i) = \frac{1}{V(X)\times L_{moy}} \times \sum_{j=1,j\neq i}^{N} l_{ij} \times (X_i - \overline{X}) \times (X_j - \overline{X})$$

d'où $$MoranGlobal(X) = \frac{1}{N}\sum_{i=1}^{N} MoranLocal(X,O_i)$$

Nous le remplaçons par :

$$MoranLocalAmélioré(X,O_i) = \frac{1}{V(X)\times L_i} \times \sum_{j=1,j\neq i}^{N} l_{ij} \times (X_i - \overline{X}) \times (X_j - \overline{X})$$

Avec Le nombre de liens entre l'objet O_i et ses voisins : $L_i = \sum_{j=1,i\neq j}^{N} l_{ij}$

Le fait de remplacer L_{moy} (le nombre moyen de liens) par L_i (le nombre réels de liens entre l'objet O_i et ses voisins) peut sembler mineur mais il rétablit une certaine logique dans le calcul du coefficient local. En effet, maintenant, notre coefficient de *Moran Local Amélioré* est indépendant du nombre de ses voisins et on peut toujours considérer 1 et -1 comme des valeurs de référence en remarquant que le coefficient de *Moran Local Amélioré* est toujours du même ordre que le coefficient de Moran global. Nous avons en effet :

$$MoranGlobal(X) = \frac{\sum_{i=1}^{N} \left(L_i \times MoranLocalAmélioré(X,O_i)\right)}{\sum_{i=1}^{N} L_i}$$

Cela signifie que le coefficient de *Moran Global* est la moyenne pondérée des coefficients de *Moran Locaux Améliorés* et que donc, on peut interpréter le coefficient de *Moran Local Amélioré* à peu près de la même manière que le coefficient de *Moran Global*.

À partir de notre expérience avec le coefficient de *Moran Global*, nous avons décrit un coefficient de *Geary Local Amélioré*. Nous avons :

$$GearyGlobal(X) = \frac{1}{V(X)\times 2\times L} \times \sum_{i=1}^{N} \sum_{j=1,j\neq i}^{N} l_{ij} \times (X_i - X_j)^2$$

Nous partons de la contrainte suivante :

$$GearyGlobal(X) = \frac{\sum_{i=1}^{N}(L_i \times GearyLocalAmélioré(X, O_i))}{\sum_{i=1}^{N} L_i}$$

De là, nous déduisons la formule du coefficient de *Geary Local Amélioré* :

$$GearyLocalAmélioré(X, O_i) = \frac{1}{V(X) \times 2 \times L_i} \times \sum_{j=1, j \neq i}^{N} l_{ij} \times (X_i - X_j)^2$$

Les coefficients de *Geary Local Amélioré* et de *Moran Local Amélioré* permettent des analyses complémentaires comme nous le verrons ultérieurement. Mais nous allons d'abord voir l'amélioration des coefficients globaux.

2.2.2 Amélioration des coefficients globaux

Nous avons remarqué que les coefficients globaux sont en fait les moyennes pondérées des coefficients locaux. La pondération se fait par le nombre de lien avec le voisinage. Ainsi, le coefficient d'un objet situé à la bordure compte en moyenne deux fois moins que celui d'un objet situé au centre car il a généralement deux fois moins de voisins. De plus, la pondération par le nombre de voisins n'a pas vraiment de justification réelle. Pour cette raison, nous créons des nouveaux indices globaux qui sont la moyenne des indices locaux, ainsi, il n'y aura plus de discrimination entre les objets situés à la bordure et ceux situés au centre.

Nous posons :

$$\begin{aligned} MoranGlobalAmélioré(X) &= \frac{1}{N} \times \sum_{i=1}^{N} MoranLocalAmélioré(X, O_i) \\ &= \frac{1}{N \times V(X)} \times \sum_{i=1}^{N} \left(\frac{1}{L_i} \times \sum_{j=1, j \neq i}^{N} l_{ij} \times (X_i - \overline{X}) \times (X_j - \overline{X}) \right) \end{aligned}$$

Et

$$\begin{aligned} GearyGlobalAmélioré(X) &= \frac{1}{N} \times \sum_{i=1}^{N} GearyLocalAmélioré(X, O_i) \\ &= \frac{1}{N \times V(X) \times 2} \times \sum_{i=1}^{N} \left(\frac{1}{L_i} \times \sum_{j=1, j \neq i}^{N} l_{ij} \times (X_i - X_j)^2 \right) \end{aligned}$$

Dans les faits, on constate peu de différence entre les coefficients globaux originaux et ceux améliorés.

Nous allons maintenant voir une amélioration des coefficients de Geary permettant une meilleure interprétation.

2.2.3 Amélioration des coefficients de Geary

La cartographie du coefficient de *Geary Local Amélioré* est plus problématique que celle du coefficient de *Moran Local Amélioré*. En effet, le coefficient de Geary n'est pas symétrique car il

varie entre 0 et l'infini avec 1 comme valeur pivot tandis que le coefficient de Moran est symétrique car il a comme « références » -1 et 1 avec 0 comme valeur pivot. Notre but est donc de construire un coefficient de *Geary Global Symétrique* et un coefficient de *Geary Local Symétrique*.

Tout d'abord nous réinterprétons le coefficient de Geary original. En effet, on peut considérer le coefficient de Geary comme le ratio de deux variances :

- La variance classique de la variable *X* : $V(X) = \frac{1}{N}\sum_{i=0}^{N}(X_i - \overline{X})^2$
- La « variance de voisinage » globale de la variable *X* :

$$Vvois(X) = \frac{1}{2 \times L} \times \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} l_{ij} \times (X_i - X_j)^2$$

Le coefficient de *Geary Global* s'écrit donc :

$$GearyGlobal(X) = \frac{Vvois(X)}{V(X)}$$

Les valeurs de références de l'indice de Geary se réinterprète donc de la façon suivante :

- *GearyGlobal(X)* tend vers 0 quand *Vvois(X)* est très petit par rapport à *V(X)*
- *GearyGlobal(X)* tend vers l'infini quand *Vvois(X)* est très grand par rapport à *V(X)*
- *GearyGlobal(X)* est égal à 1 quand *Vvois(X)* est identique à *V(X)*

Pour créer notre indice symétrique, nous décidons d'utiliser le rapport entre *Vvois(X)* et *V(X)* de la manière suivante :

- Si $Vvois(X) \leq V(X)$ alors

$$\begin{aligned} GearyGlobalSymetrique(X) &= 1 - \frac{Vvois(X)}{V(X)} \\ &= 1 - GearyGlobal(X) \end{aligned}$$

- Si $Vvois(X) \geq V(X)$ alors

$$\begin{aligned} GearyGlobalSymetrique(X) &= 1 - \frac{V(X)}{Vvois(X)} \\ &= 1 - \frac{1}{GearyGlobal(X)} \end{aligned}$$

De cette manière, nous avons :

- *GearyGlobalSymetrique(X)* tend vers 1 quand *Vvois(X)* est très petit par rapport à *V(X)*
- *GearyGlobalSymetrique(X)* tend vers -1 quand *Vvois(X)* est très grand par rapport à *V(X)*
- *GearyGlobalSymetrique (X)* est égal à 0 quand *Vvois(X)* est identique à *V(X)*

L'exemple suivant montre la symétrie du coefficient de *Geary Global Symétrique*. En effet, supposons les cas de figure suivant :

- Avec $Vvois(X)$=5 et $V(X)$=1, nous avons :
  - $GearyGlobal(X) = \frac{5}{1} = 5$

- $GearyGlobalSymetrique(X) = \frac{1}{5} - 1 = -\frac{4}{5} = -0.80$

- Dans le cas « contraire », $Vvois(X)$=1 et $V(X)$=5, nous avons :
  - $GearyGlobal(X) = \frac{1}{5} = 0.25$
  - $GearyGlobalSymetrique(X) = 1 - \frac{1}{5} = \frac{4}{5} = 0.80$

Cependant, l'aspect symétrique n'a pas beaucoup d'intérêt dans le cas du coefficient global car il n'y a pas de problématique de cartographie. La création d'un coefficient de *Geary Local Symétrique* est donc plus importante. Nous nous inspirons alors du coefficient de *Geary Global Symétrique* pour définir le coefficient de *Geary Local Symétrique* de l'objet O_i pour la variable X:

- Si $VvoisLocale(X, O_i) \leq V(X)$ alors

$$\begin{aligned} GearyLocalSymetrique(X, O_i) &= 1 - \frac{VvoisLocale(X, O_i)}{V(X)} \\ &= 1 - GearyLocalAmélioré(X, O_i) \end{aligned}$$

- Si $VvoisLocale(X, O_i) \geq V(X)$ alors

$$\begin{aligned} GearyLocalSymetrique(X, O_i) &= 1 - \frac{V(X)}{VvoisLocale(X, O_i)} \\ &= 1 - \frac{1}{GearyLocalAmélioré(X, O_i)} \end{aligned}$$

Avec la « variance de voisinage » locale de l'objet O_i pour la variable *X* :

$$VvoisLocale(X, O_i) = \frac{1}{2 \times L_i} \times \sum_{j=1, j \neq i}^{N} l_{ij} \times (X_i - X_j)^2$$

L'interprétation de ce coefficient est la même que pour le coefficient global. De plus, le coefficient de *Geary Local Symétrique* est beaucoup plus facile à cartographier et à analyser que le coefficient de *Geary Local Amélioré*, justement en raison de son aspect symétrique. On remarquera toutefois qu'il n'y a pas de relation simple entre le coefficient de *Geary Local Symétrique* et coefficient de *Geary Global Symétrique*.

Nous allons maintenant voir des exemples d'utilisation de ces coefficients.

2.2.4 Exemple

À l'aide des nouveaux coefficients, nous allons analyser la carte des taux d'agriculteurs par départements.

Nous commençons par le coefficient de *Moran Local Amélioré*. L'analyse est facilitée car nous nous sommes servis des valeurs de références 0, -1 et 1 pour déterminer les plages de valeurs sans avoir eu besoin de les chercher de façons empiriques comme c'était le cas avec le coefficient de *Moran Local* original. Dans cette nouvelle analyse, nous avons séparé les objets plutôt semblables à leur voisins (entre 0,5 et 1) et les objets très semblables à leur voisinage (valeurs

supérieures à 1). L'analyse reste cependant globalement la même que celle faite avec le coefficient de *Moran Local*.

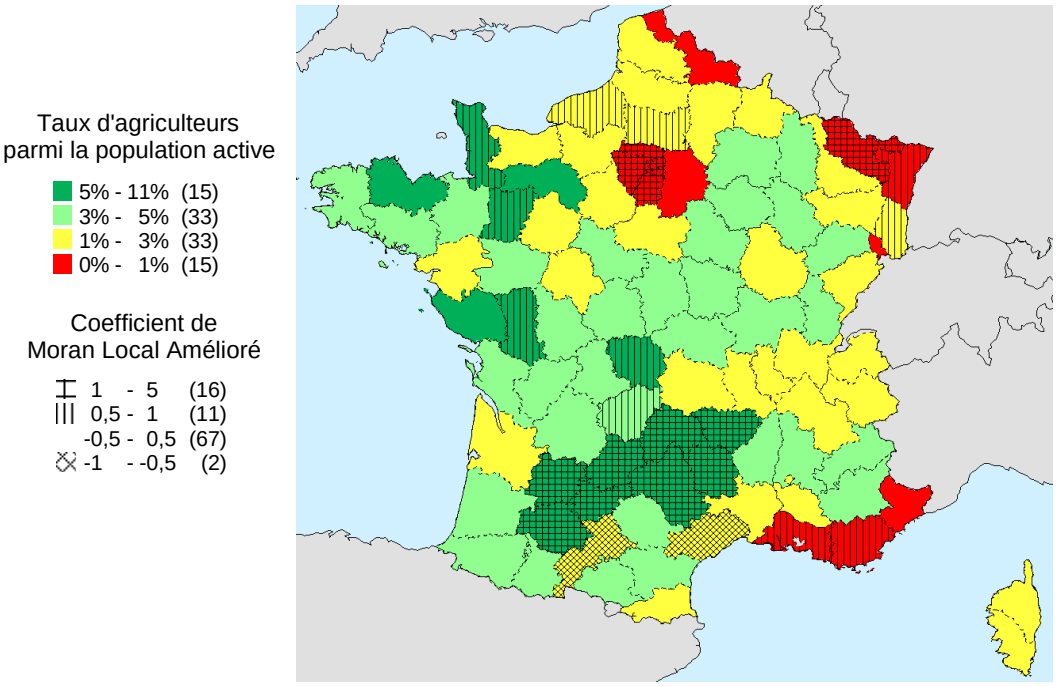


Fig. 126 –Carte des coefficients de Moran Locaux Améliorés

On constate toutefois des différences significatives qui sont mises en évidence par le tableau suivant. Il donne les différences les plus importantes entre le coefficient de *Moran Local* et le coefficient de *Moran Local Améliorée*.

Nom du département	Agriculteur	Coefficient de Moran Local	Coefficient de Moran Local Amélioré	Différence	Changement de plage de valeurs
JURA	3,23 %	0,00	0,00	0,00	Non
HAUTES-ALPES	3,24 %	0,00	0,00	0,00	Non
CHER	3,35 %	0,01	0,01	0,00	Non
DROME	3,44 %	-0,02	-0,02	0,00	Non
…					
LOT-ET-GARONNE	*5,81 %*	*1,77*	*1,32*	*-0,45*	Non
AVEYRON	*9,32 %*	*4,40*	*3,94*	*-0,46*	Non
MOSELLE	**0,74 %**	**0,49**	**1,10**	**+0,61**	**Oui (+2)**
SEINE-SAINT-DENIS	0,02 %	1,34	2,00	+0,66	Non
PARIS	0,02 %	1,36	2,03	+0,67	Non
LOZERE	*9,00 %*	*3,51*	*2,62*	*-0,89*	Non
LOT	*7,06 %*	*3,72*	*2,78*	*-0,94*	Non
CANTAL	*9,98 %*	*5,76*	*4,30*	*-1,46*	Non

Fig. 127 –Tableau ordonné selon les différences entre le coefficient de Moran local et sa version améliorée

On constate que la différence la plus significative (vis-à-vis des plages de valeurs) est pour le département de la Moselle qui, étant situé en bordure était pénalisé à cause du faible nombre de ses voisins et avait donc un score assez faible.

Le coefficient de *Geary Local Symétrique* apporte quant à lui une tout autre vision. En effet, il permet de distinguer les zones de fractures (valeurs comprises entre -1 et -0.5) des zones de continuité (valeurs comprises entre 0.5 et 1) en passant par les zones où les différences avec le voisinage sont dans la « moyenne » (valeurs comprises entre -0.5 et 0.5). Ainsi nous pouvons observer sur la carte suivante que l'on peut tracer une diagonale allant du Nord-Ouest au Sud-Ouest comprenant :

- Du côté au Nord et à l'Est une zone de continuité où les départements sont faiblement ruraux ou plutôt urbains.
- Du côté au Sud et à l'ouest une zone de discontinuité particulièrement marquée sur le Massif Central et une partie du Sud-Ouest.

Dans cette dernière zone, les causes de discontinuité sont de deux natures différentes :

- Les départements plutôt urbains (Toulouse et Montpellier) entourées de départements plutôt ruraux.
- Les départements très ruraux (Creuse, Gers, Lozère) entourés de départements moins ruraux.

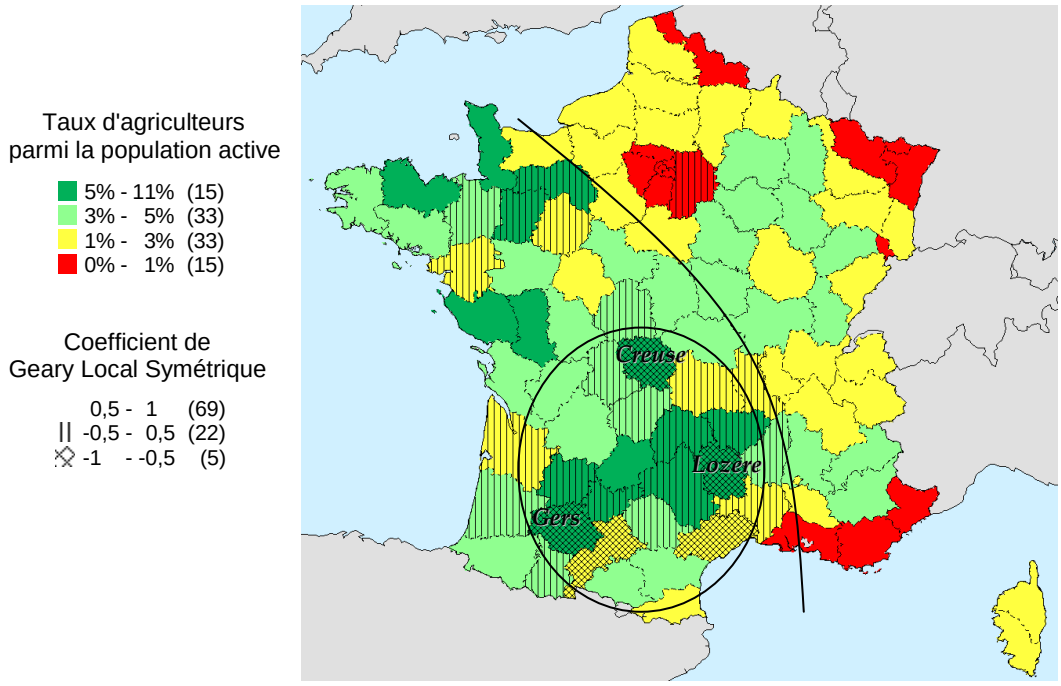


Fig. 128 –Carte des coefficients de Geary Locaux Symétriques

Les coefficients globaux sont les suivants

- Coefficient de *Moran Global Amélioré* = 0,47
- Coefficient de *Geary Global Amélioré* = 0,52
- Coefficient de *Geary Global Symétrique* = 0,42

Pour rappel, nous avions :

- Coefficient de *Moran Global* = 0,49

- Coefficient de *Geary Global* = 0,57

Les nouveaux indicateurs globaux donnent des résultats similaires aux anciens indicateurs. Ils indiquent toujours que les objets ressemblent d'avantage à leurs voisins que des objets pris au hasard.

2.3 Conclusion

Nous avons amélioré les coefficients d'autocorrélation globaux et locaux. Les améliorations les plus intéressantes concernent les coefficients de *Moran Local Amélioré* qui évite le biais lié au nombre de voisins tout en conservant les mêmes valeurs de référence que le coefficient de *Moran Global* (c'est-à-dire -1, 0 et 1). Dans le même esprit, nous avons défini un coefficient de *Geary Local Symétrique* qui utilise aussi -1, 0 et 1 comme valeurs de référence. Les valeurs de références sont importantes car elles permettent de définir automatiquement des plages de valeurs pour la cartographie de ces coefficients locaux.

Nous avons aussi constaté que l'amélioration des coefficients globaux n'est pas essentielle. En effet, dans le cas des coefficients globaux traditionnels, le biais induit par le nombre de voisins est souvent plus faible que pour les coefficients locaux car les différences locales ont tendance à se compenser.

3 Contributions à l'amélioration de la Modélisation des Flux

La modélisation des flux a pour but la prédiction des flux. Il s'agit d'un problème largement traité en Géographie. Nous allons d'abord voir les différentes familles de modélisation des flux en nous intéressant plus particulièrement au fonctionnement du modèle de Tobler, à ses avantages et à ses limites.

Dans la seconde partie, nous allons proposer une amélioration de ce modèle afin d'éviter d'obtenir des flux négatifs, ce qui est impossible dans la réalité car les flux représentent des quantité forcément positive ou nulle.

Finalement, nous mettrons en œuvre le modèle amélioré dans le cadre de la modélisation des migrations d'entreprises au sein du département de la Loire-Atlantique. Ce travail a été effectué dans le cadre de la phase Recherche du projet R&D de GÉOBS, en collaboration avec Florent Charron durant sa thèse en Géographie à GÉOBS [*Char05*]. Nous verrons que les corrections effectuées sont minimes et ne perturbent pas la modélisation des flux. Nous comparerons finalement les flux réels et les flux prédits par le modèle afin de mettre en évidence des connections particulières entre certaines zones.

3.1 Etat de l'art

Dans un premier temps, nous allons voir comment s'organisent les différentes familles de modélisation des flux. Puis nous nous intéresserons au cas des modèles dits « à double contrainte » qui sont très utiles dans les cas où on dispose des informations suivantes pour chaque zone : la somme des départs et la somme des arrivées. Nous verrons finalement comment fonctionne le modèle de Tobler qui est un modèle additif à double contrainte.

3.1.1 Présentation générale

La modélisation des flux a fait l'objet de nombreuses recherches [*PS01*, *Gras98*]. Il s'agit de prédire (ou de retrouver) les flux entre les zones en s'aidant de certaines informations telles que :

- la distance entre les zones
- le « stock » dans chaque zone
- la somme des départs de chaque zone
- la somme des arrivées dans chaque zone

Tous les modèles utilisent la distance car elle est un facteur déterminant. Cependant, on dénombre au moins trois catégories de modèles :

- Les modèles les plus simples dits « sans contraintes » se contentent de respecter la somme des flux.
- Les modèles plus évolués dits à « simple contrainte » respectent soit la somme des départs de chaque zone, soit la somme des arrivées dans chaque zone.
- Enfin, les modèles dits à « double contrainte » respectent la somme des départs et la somme des arrivées dans chaque zone.

Les modèles les plus utilisés sont les modèles à « simple contrainte » car la plupart du temps, on ne connaît qu'un seul paramètre : la somme des départs ou bien la somme des arrivées. Cependant, dans le cadre de notre recherche, nous nous sommes plutôt intéressés aux modèles à double contrainte car nous disposions de toutes les informations (la somme des départs et la somme des arrivées dans chaque zone) et nous pouvions donc utiliser ce type de modèle.

Les modèles les plus utilisés sont les modèles multiplicatifs et sont considérés comme les modèles de base. Cependant, dans le cadre des modèles à doubles contraintes, il existe aussi les modèles additifs tels le modèle de G. Dorigo et W. Tobler [*DT83*]. Ce modèle nous a intéressé car il modélise chaque flux comme la différence entre un facteur d'*Attraction* et un facteur de *Rejet* tandis que les modèles multiplicatifs modélisent chaque flux comme la multiplication entre un facteur d'*Attraction* et un facteur de *Rejet*.

3.1.2 Formalisation classique des modèles à double contrainte

Nous avons *n* objets géographiques. Dans les modèle à double contrainte, le flux théorique entre deux objets géographiques O_i et O_j, est défini comme suit :

- Pour les modèles gravitaires (ou multiplicatifs) dont le modèle de A. Wilson [*Wils67*]:

$$F_{ij}^{*} = \left(R_i \times A_j\right) \times Param_{ij}$$

- Pour les modèle additifs dont le modèle de Tobler:

$$F_{ij}^{*} = \left(R_i + A_j\right) \times Param_{ij}$$

Les termes intervenants dans les formules sont :

- R_i le facteur de Rejet de O_i (qui est doit être déterminé)
- A_j le facteur d'Attraction de O_j (qui doit être déterminé)
- $Param_{ij}$ un paramètre spécifié concernant le couple d'objets géographiques O_i et O_j, et faisant souvent intervenir la distance. Ce paramètre est généralement symétrique $Param_{ij} = Param_{ij}$

Les deux types de contraintes sont les suivantes :

- la contrainte sur la somme des flux entrants :

$$\sum_{i=1}^{n} F_{ij}^{*} = E_j \qquad \text{avec } E_j \text{ le total des sorties de l'objet } O_j$$

- la contrainte sur la somme des flux sortants :

$$\sum_{j=1}^{n} F_{ij}^{*} = S_i \qquad \text{avec } S_i \text{ le total des sorties de l'objet } O_i$$

Dans le cadre du modèle de Tobler, nous avons deux formulation pour $Param_{ij}$:

- la formulation simple ne dépendant que de la distance :

$$Param_{ij} = \frac{1}{distance\left(O_i, O_j\right)}$$

- la formulation évoluée prenant en compte le « stock » d'un point de vue multiplicatif :

$$Param_{ij} = \frac{stock\left(O_i\right) \times stock\left(O_j\right)}{distance\left(O_i, O_j\right)}$$

3.1.3 Estimation des facteurs dans le cadre du modèle de Tobler

L'estimation des facteurs de *Rejet* et d'*Attraction* dans le cadre du modèle de Tobler, se fait par la résolution du système d'équation linéaire. En effet, les contraintes se traduisent de la façon suivante :

$$E_j = \sum_{i=1}^{n} F_{ij}^{*}$$
$$= \sum_{i=1}^{n} \left(\left(R_i + A_j \right) \times Param_{ij} \right)$$
$$= \sum_{i=1}^{n} \left(R_i \times Param_{ij} \right) + A_j \times \sum_{i=1}^{n} Param_{ij}$$

Et $S_i = \sum_{j=1}^{n} F_{ij}^{*}$

$$= \sum_{j=1}^{n} \left(R_i + A_j \right) \times Param_{ij}$$
$$= R_i \times \sum_{j=1}^{n} Param_{ij} + \sum_{j=1}^{n} \left(A_j \times Param_{ij} \right)$$

Ce système est un système linéaire à $2 \times n$ inconnues (il y a *n* inconnues R_j, $1 \le j \le n$, et *n* inconnues A_i, $1 \le i \le n$) comportant $2 \times n$ équations linéaires. Cependant, les équations sont liées car $\sum_{j=1}^{n} E_j = \sum_{j=1}^{n} \sum_{i=1}^{n} F_{ij}^{*} = \sum_{i=1}^{n} S_i$. Ainsi, nous supprimons arbitrairement une variable en fixant sa valeur (par exemple : $E_1 = 0$) et en supprimant une équation. Nous aboutissons alors à un système linéaire à $2 \times n - 1$ inconnues et comportant $2 \times n - 1$ équations linéaires indépendantes. Ce système est alors résolu informatiquement par un solveur d'équations.

Nous pouvons alors en déduire les facteurs de *Rejet* et d'*Attraction* et ensuite les flux théoriques entre les objets.

3.2 Amélioration

Le principal inconvénient de la modélisation de Tobler est qu'il est possible d'obtenir des flux négatifs, ce qui n'est pas conforme à la réalité. De plus, même si les flux négatifs sont très faibles et proches de 0, ils peuvent être en nombre assez important. Dans leur article, G. Dorigo et W. Tobler propose une méthode itérative « excluant » systématiquement les liaisons fautives (qui deviennent alors nulles). Cependant, cette méthode ne fonctionne pas dans tous les cas car elle peut aboutir à l'exclusion de toutes les liaisons d'un objet géographique avec les autres objets géographique. Le système d'équations linéaires n'admet alors plus de solution. Pour cette raison, nous avons cherché une autre méthode. La méthode que nous avons expérimentée est est la suivante : les flux négatifs sont ramenés à zéro et « le déficit » induit est reporté sur les flux positifs existants. Ainsi le flux théorique corrigé F_{ij}^{**} entre deux objets géographiques O_i et O_j est le suivant :

$F_{ij}^{**} = 0$ si $F_{ij}^{*} < 0$

$F_{ij}^{**} = C_{ij} \times F_{ij}^{*}$ sinon

Avec C_{ij} un facteur correctif dépendant de O_i et O_j et qui est compris entre 0 et 1.

La correction peut être faite de deux façons qui s'excluent mutuellement : on corrige en respectant soit la contrainte sur la somme des flux entrants, soit la contrainte sur la somme des flux sortants. Nous avons donc décidé de faire un compromis entre les deux qui permet de respecter les contraintes pour les objets géographiques ayant un « stock élevé » au détriment des objets géographiques ayant des stocks plus faibles. Le facteur de correction est donc :

$$C_{ij} = \frac{stock(O_i) \times CS_i + stock(O_j) \times CE_j}{stock(O_i) + stock(O_j)}$$

Avec *CSi*, le facteur correctif des flux théoriques sortant de O_i :

$$CS_i = \frac{S_i}{S_i^+} \text{ avec } S_i^+ = \sum_{j, F_{ij}^* > 0} F_{ij}^*$$

et *CEj*, le facteur correctif des flux théoriques entrant dans O_j :

$$CE_j = \frac{E_j}{E_j^+} \text{ avec } E_j^+ = \sum_{i, F_{ij}^* > 0} F_{ij}^*$$

Nous allons maintenant voir l'application pour la modélisation des migrations des entreprises à l'intérieur du département de la Loire Atlantique.

3.3 Application pour l'analyse des migrations des entreprises à l'intérieur du département de la Loire-Atlantique

Nous disposons pour chaque commune de la Loire-Atlantique du nombre d'arrivées et du nombre de départs d'entreprises par commune entre 1992 et 2000. Nous avons aussi pour chaque commune, le nombre moyen d'entreprises présentes (le stock).

Une première analyse est le solde relatif par rapport au stock (nombre d'entreprises présentes en moyenne).

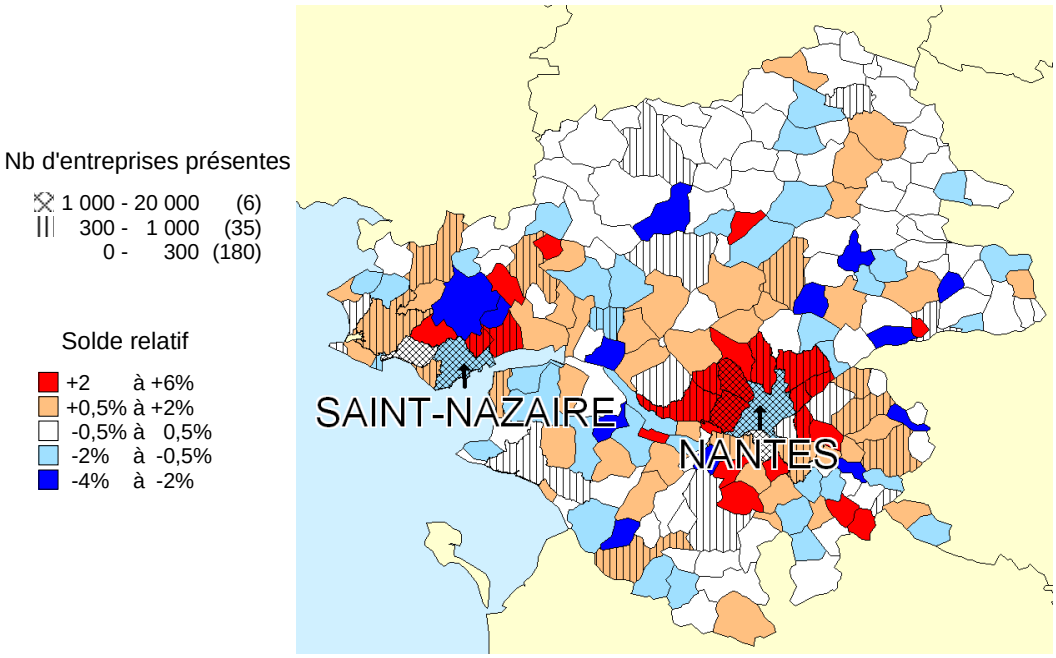


Fig. 129 –Carte des soldes relatifs dans le cadre des migrations d'entreprises au sein du département de la Loire-Atlantique

Nous constatons que les plus importants transferts tant en valeurs relatives (pourcentage) que en valeurs réelles (nombre d'entreprises) concernent les villes de Nantes et de St-Nazaire. Il s'agit a priori d'une migration des entreprises partant des centres-villes pour aller dans la périphérie immédiate.

Si nous nous focalisons sur la zone de Nantes, nous observons les flux migratoires réels suivants (allant de 5 à 100 entreprises déplacées) :

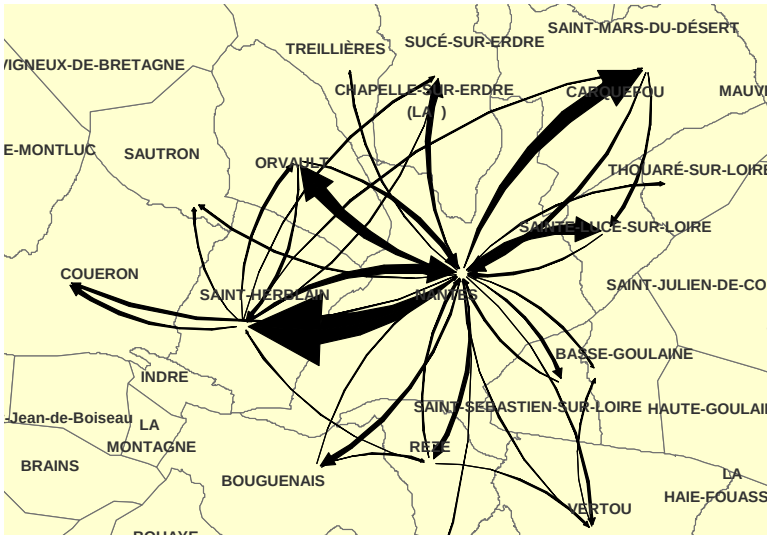


Fig. 130 –Flux migratoires réels des d'entreprises pour Nantes et sa périphérie

Cette carte montre que les destinations privilégiées des entreprises situées à Nantes sont les communes de St-Herblain, Carquefou, Orvault et Ste-Luce.

Nous allons maintenant voir les flux prédits par la modélisation. Tout d'abord, en calculant les flux théoriques, sans amélioration, nous obtenons seulement quelques flux très légèrement négatifs. En utilisant l'amélioration, nous supprimons ces flux négatifs et le report du déficit sur les autres flux entraîne des corrections très faibles de l'ordre de 2% dans le pire des cas. Le tableau suivant donne un extrait des valeurs calculées pour les flux théoriques.

Départ	Arrivée	Flux	Flux Corrigé	Différence
AIGREFEUILLE-SUR-MAINE	AIGREFEUILLE-SUR-MAINE	0	0	0
ANCENIS	ANCENIS	0	0	0
ANCENIS	MACHECOUL	0,15	0,15	0
ANETZ	ANETZ	0	0	0
…				
CHATEAU-THEBAUD	CHAPELLE-BASSE-MER (LA)	-0,02	0	0,02
VALLET	ROUGE	-0,02	0	0,02
BOUGUENAIS	NANTES	31,57	31,56	0,02
LOROUX-BOTTEREAU (LE)	SAFFRE	-0,02	0	0,02
…				
VALLET	REZE	-0,33	0	0,33
SAINT-SEBASTIEN-SUR-LOIRE	NANTES	32,35	32,01	0,34
SAINT-JULIEN-DE-CONCELLES	SAINT-SEBASTIEN-SUR-LOIRE	-0,34	0	0,34
NANTES	SAINT-SEBASTIEN-SUR-LOIRE	38,00	37,51	0,49
BASSE-GOULAINE	SAINT-SEBASTIEN-SUR-LOIRE	-0,55	0	0,55
REZE	NANTES	71,55	70,96	0,59
NANTES	REZE	78,42	77,49	0,92
REZE	SAINT-SEBASTIEN-SUR-LOIRE	-2,36	0	2,36
SAINT-SEBASTIEN-SUR-LOIRE	REZE	-2,49	0	2,49

Fig. 131 –Flux migratoires théoriques ordonnés selon la différence croissante entre la valeur théorique « brute » et la valeur théorique corrigée

Si nous nous focalisons sur la zone de Nantes, nous observons que les flux migratoires théoriques sont très similaires des flux réels comme le montre la carte suivante :

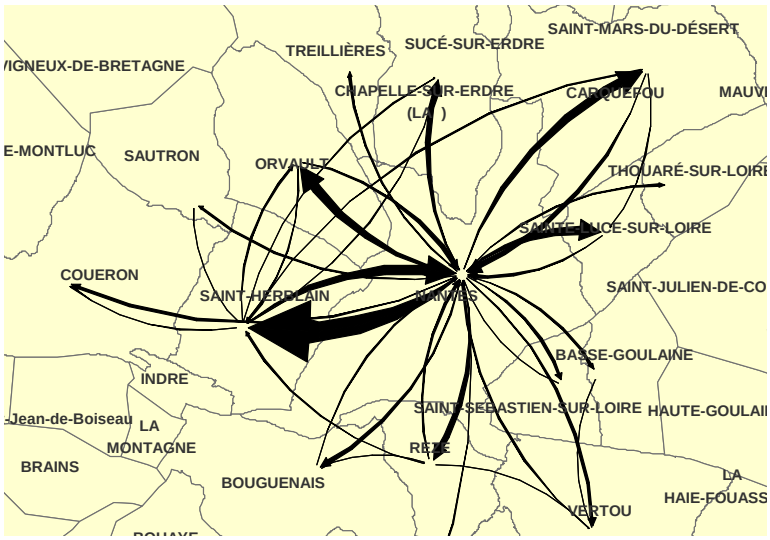


Fig. 132 –Flux migratoires théoriques des d'entreprises pour Nantes et sa périphérie

Nous avons aussi cartographié la carte des différences entre les flux réels et les flux théoriques. Il s'agit surtout de sous-estimations (allant de 1 à 20 entreprises), c'est-à-dire que les flux réels sont plus importants que les flux théoriques.

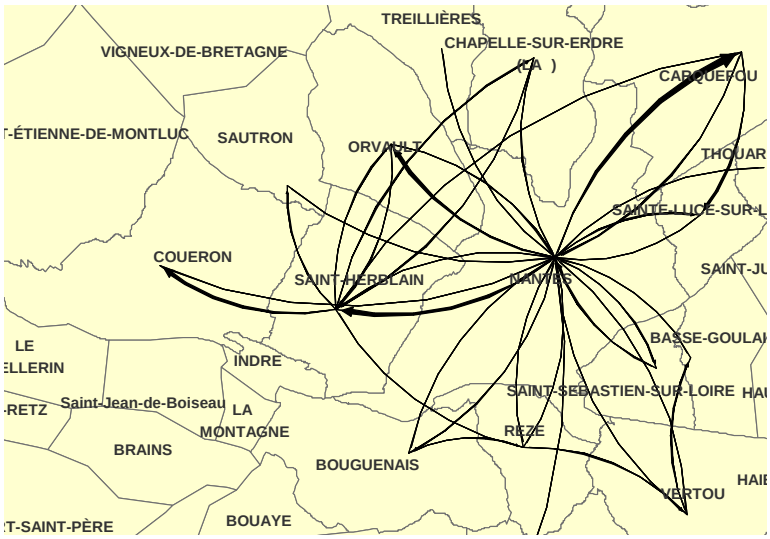


Fig. 133 – Carte des sous-estimations des flux pour Nantes et sa périphérie

Les flux les plus sous-estimés sont de Nantes vers St-Herblain, Carquefou et Orvault, d'une part, et de St-Herblain vers Coueron et de Carquefou vers Ste Luce sur Loire, d'autre part. On en déduit que :

- St-Herblain, Carquefou, Orvault sont beaucoup plus attractives que prévues pour les entreprises Nantaise.
- Coueron et Ste Luce sur Loire sont des zones attractives pour les entreprises déjà situées en périphérie.

Ce dernier résultat concernant la périphérie n'était pas mis en évidence par l'analyse des flux réels et c'est la comparaison avec les flux théoriques qui a mis en avant l'existence d'un

deuxième mouvement de migration de la périphérie immédiate (St-Herblain et Carquefou) pour une périphérie plus éloignée encore (Coueron et St-Luce) par rapport au centre de l'agglomération Nantaise.

3.4 Conclusion

Nous avons amélioré le modèle de Tobler afin d'éviter qu'il engendre des flux négatifs. Nous avons vu que les corrections effectuées sont minimes et ne perturbent pas la modélisation des flux. De plus, dans le cadre de la modélisation des migrations d'entreprises au sein du département de la Loire-Atlantique, la comparaison entre les flux réels et les flux prédits par le modèle a permis de mettre en évidence la forte attraction des entreprises nantaises pour la périphérie immédiate et dans un deuxième temps, l'attraction des entreprises déjà situées en périphérie pour des communes encore plus éloignées de Nantes.

BIBLIOGRAPHIE

Introduction

AYK00 Aufaure M. A., Yeh L., and Zeitouni K. Fouille de données spatiales. In R. Jeansoulin et C. Garbay eds. H. Prade, editor, *Le temps, l'espace et l'évolutif en Sciences du Traitement de l'Information*, pages 319-328. CEPADUES, 2000.
[http://www-rocq.inria.fr/~aufaure/Chap-FDS-I3.pdf]
Keywords: Méthodes d'analyse de données géographiques

Chad97 Chadule (groupe). *Initiation aux Pratiques Statistiques en Géographie*. Armand Colin, 1997. ISBN 2-200-01534-8.
Keywords: Histogramme des fréquences, Boites et moustaches, Courbe de concentration, Régression linéaire multiple, Matrice de corrélations, Analyse en Composantes Principales (ACP), Nuées dynamiques

Char89 Charre J. *Statistique et territoire*. GIP Reclus, 1989. ISBN 2-86912-060-2.

Sand89 Sanders L. *L'analyse statistique des données en géographie*. GIP Reclus, 1989. ISBN 2-86912-028-0.
Keywords: Analyse en Composantes Principales (ACP), Analyse Factorielle des Correspondances (AFC), Classification Ascendante Hiérarchique (CAH)

Zeit99 Zeitouni K. Etat de l'art sur l'extension du data mining aux bases de données géographiques. laboratoire PRiSM, Université de Versailles Saint-Quentin, 1999.
[http://www.prism.uvsq.fr/rapports/1999/document_1999_10.ps]
Keywords: Méthodes d'analyse de données géographiques

Classification de Données

AGGR98 Agrawal R., Gehrke J., Gunopulos D., and Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications.

In *ACM SIGMOD International Conference on Management of Data*, pages 94-105, 1998.
[http://citeseer.ist.psu.edu/article/agrawal98automatic.html]
Keywords: CLIQUE, Classification par grille

ABKS99 Ankerst M., Breunig M., Kriegel, and Sander J. OPTICS: Ordering points to identify the clustering structure. *ACM SIGMOD International Conference on Management of Data*, 28(2):49-60, Juin 1999.
[http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/OPTICS.pdf]
Keywords: OPTICS, Classification par densité

Berk02 Berkhin P. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.
[http://citeseer.ist.psu.edu/berkhin02survey.html]
Keywords: Algorithmes de classification, état de l'art

BKKS01 Breunig M., Kriegel H., Kröger P., and Sander J. Data bubbles: quality preserving performance boosting for hierarchical clustering. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 30(2):79-90, 2001.
[http://citeseer.ist.psu.edu/breunig01data.html]
Keywords: Data bubbles, Classification incrémentale, résumé de données

CCFM97 Charikar M., Chekuri C., Feder T., and Motwani R. Incremental clustering and dynamic information retrieval. In *29th Symposium on Theory of Computing*, pages 626-635, 1997.
[http://citeseer.ist.psu.edu/charikar97incremental.html]
Keywords: Classification Ascendante Hiérarchique (CAH), Classification incrémentale

DLR77 Dempster A. P., Laird N. M., and Rubin D. B. *Maximum likelihood from incomplete data via the EM algorithm*. J. Royal Statist. Soc. B, 39:1--39, 1977.
Keywords: EM, Modèle de mélange

EKSX96 Ester M., Kriegel H. P., Sander J., and Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *2nd International Conference on Knowledge Discovery and Data Mining, KDD*, pages 226-231, 1996.
[http://web.cs.ualberta.ca/~joerg/papers/KDD-96_final.pdf]
Keywords: DBSCAN, Classification par densité

Fish87 Fisher D. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2:139-172, 1987.
Keywords: COBWEB, Classification incrémentale

Fish96 Fisher D. Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Intelligence Research*, 4:147-180, 1996.
[http://citeseer.ist.psu.edu/article/fisher96iterative.html]
Keywords: COBWEB, Classification incrémentale

GRS98 Guha S., Rastogi R., and Shim K. CURE: an efficient clustering algorithm

for large databases. In *ACM SIGMOD International Conference on Management of Data*, pages 73-84, 1998.
[http://citeseer.ist.psu.edu/guha98cure.html]
Keywords: CURE, Classification hiérarchique par agglomération, échantilonnage

HK98 Hinneburg A. and Keim D. An efficient approach to clustering in large multimedia databases with noise. In *Knowledge Discovery and Data Mining, KDD*, pages 58-65, 1998.
[http://citeseer.ist.psu.edu/hinneburg98efficient.html]
Keywords: DENCLUE, Classification par densité

JD88 Jain A. K. and Dubes R. C. *Algorithms for Clustering Data*. Prentice Hall, New Jersey, 1988.
[http://www.cse.msu.edu/~jain/Clustering_Jain_Dubes.pdf]
Keywords: Algorithmes de classification, état de l'art

JMF99 Jain A. K., Murty M. N., and Flynn P. J. Data clustering: a review. *ACM Computing Surveys*, 31(3):264-323, 1999.
[http://citeseer.ist.psu.edu/jain99data.html]
Keywords: Algorithmes de classification, état de l'art

KHK99 Karypis G., Han E., and Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer: Special Issue on Data Analysis and Mining*, 32(8):68-75, 1999.
[http://citeseer.ist.psu.edu/karypis99chameleon.html]
Keywords: CHAMELEON, Classification hiérarchique par agglomération, partition de graphe

KR90 Kaufman L. and Rousseeuw P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York, 1990.
Keywords: Algorithmes de classification, état de l'art, CLARA, DIANA, PAM, AGNES

Maha36 Mahalanobis P. C. On generalized distance in statistics. *Proceedings of the National Inst. Sci. (India)*, 12:49-55, 1936.
Keywords: Distance de Mahalanobis

MC02 Mazlack L. and Coppock S. Using soft computing techniques to integrate multiple kinds of attributes in data mining. In *5th International FLINS Conference, Computational Intelligent Systems for Applied Research*, pages 137-144, Septembre 2002.
[http://www.ececs.uc.edu/~mazlack/professional/FLINS.2002.pdf]
Keywords: Données mixtes, état de l'art

NH94 Ng R. T. and Han J. Efficient and effective clustering methods for spatial data mining. In *20th International Conference on Very Large Data Bases, VLDB*, pages 144-155, Septembre 1994.
[http://citeseer.ist.psu.edu/ng94efficient.html]
Keywords: CLARANS, k-médoïdes, échantilonnage, classification

RM02 Raschia G. And Mouaddib N. SaintEtiQ : a Fuzzy Approach to Database Summarization. In *International Journal of Fuzzy Sets and Systems*, 129(2):137-162, Juillet 2002
Keywords: SaintEtiQ, données floues, classification

vRij79 Van Rijsbergen C. *Information Retrieval, 2nd edition.* Dept. of Computer Science, University of Glasgow,1979.
[http://www.dcs.gla.ac.uk/Keith/Preface.html]
Keywords: coefficient de Jaccard, classification

SCZ98 Sheikholeslami G., Chatterjee S., and Zhang A. WaveCluster: A multi-resolution clustering approach for very large spatial databases. In *24th International Conference on Very Large Data Bases, VLDB*, pages 428-439, 1998.
[http://citeseer.ist.psu.edu/sheikholeslami98wavecluster.html]
Keywords: WaveCluster, Classification par grille

SHR96 Struyf A., Hubert M., and Rousseeuw P. Clustering in an object-oriented environment. *Journal of Statistical Software*, 1(4):1-30, 1996.
[http://www.jstatsoft.org/v01/i04/paper/clus.pdf]
Keywords: Algorithmes de classification, état de l'art, CLARA, DIANA, PAM, AGNES

WYM97 Wang W., Yang J., and Muntz R. STING: A statistical information grid approach to spatial data mining. In *23th International Conference on Very Large Data Bases, VLDB*, pages 186-195, Athens, Greece, 1997.
[http://citeseer.ist.psu.edu/article/wang97sting.html]
Keywords: STING, Classification par grille

ZRL96 Zhang T., Ramakrishnan R., and Livny M. BIRCH: an efficient data clustering method for very large databases. In *ACM SIGMOD International Conference on Management of Data*, pages 103-114, 1996.
[http://citeseer.ist.psu.edu/zhang96birch.html]
Keywords: BIRCH, Classification incrémentale, résumé de données

Visualisation de Classifications

BDGHJS02 Bar-Joseph Z., Demaine D., Gifford D., Hamel A., Jaakkola T., and Srebro. K-ary clustering with optimal leaf ordering for gene expression data. In *International Workshop on Algorithms in Bioinformatics, WABI*, 2002.
[http://www.bioinformatics.uwaterloo.ca/papers/02clustering.pdf]
Keywords: Linear Ordering Problem (LOP), classification hiérarchique

BD00 Bock H. H. and Diday E. *Analysis of Symbolic Data*. Springer, 2000. ISBN 3-540-66619-2.
Keywords: Analyse et visualisation de données symboliques

ESBB98 Eisen M. B., Spellman P. T., Brown P. O., and Botstein D. Cluster

analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA*, 95(25):14863-14868, Décembre 1998.
[http://rana.lbl.gov/papers/Eisen_PNAS_1998.pdf]
Keywords: Visualisation de la classification

ID90 Inselberg A. and Dimsdale B. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *IEEE Visualization*, pages 361-378, 1990.
[http://www.ifs.tuwien.ac.at/~mlanzenberger/informatica-feminale03/se188095/auth/00146402.pdf]
Keywords: Visualisation en coordonnées parallèles

Jamb99 Jambu M. *Méthodes de base de l'analyse de données*. Eyrolles, 1999. ISBN 2-212-05256-1.
Keywords: Analyse des données

Scot92 Scott D. W. *Multivariate Density Estimation. Theory, Practice, and Visualization*. John Wiley & Sons, Inc., 1992. (Chapter 3).
Keywords: Visualisation multivariée, histogrammes

Siir00 Siirtola H. Direct manipulation of parallel coordinates. In *Conference on Human Factors in Computing Systems,ACM CHI*, volume 2 of *Interactive posters*, pages 119-120, 2000.
[http://infoviz.cs.uta.fi/papers/hsiirtola.pdf]
Keywords: Visualisation en coordonnées parallèles

Tuke77 Tukey J. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 1977.
Keywords: Boites et moustaches (box-and-whisker plot)

YWR03 Yang J., Ward M. O., and Rundensteiner E. A. Interactive hierarchical displays: a general framework for visualization and exploration of large multivariate data sets. *Computers and Graphics*, 27(2):265-283, Avril 2003.
[http://davis.wpi.edu/~xmdv/docs/CandG.pdf]
Keywords: Visualisation en coordonnées parallèles

Lissage Spatial

Aure91 Aurenhammer F. *Voronoi Diagrams - A Survey Of A Fundamental Geometric Data Structure*, ACM Computing Surveys, 23, 345- 405, 1991.

BG95 Bailey T. and Gatrell A. *Interactive Spatial Data Analysis*. Longman, London, 1995.
Keywords: spatial smoothing, kernel density

Bezd93 Bezdek J. Fuzzy Systems - What Are They, and Why. *IEEE Transactions on*

Fuzzy Systems, 1(1):1-5, 1993.
Keywords: fuzzy logic

CM02 Comanicu D. and Meer P. *Mean Shift: A Robust Approach Toward Feature Space Analysis*. IEEE Trans. Pattern Anal. Mach. Intell. 24, 5, 603-619, 2002.

Davi86 Davis J. C. *Statistics and Data Analysis in Geology*. Wiley, New York, 1986.
Keywords: spatial smoothing, trend surface analysis, fourier transform

Fort86 Fortune S. J. A sweepline algorithm for Voronoi diagrams. In *Proc. 2nd ACM Symp. Computational Geometry*, pages 313-322, Juin 1986. ISBN 0-89791-194-6.
[http://portal.acm.org/citation.cfm?id=10549&coll=portal&dl=ACM]

FN91 Franke R. and Nielson G. M. Scattered data interpolation and applications: A tutorial and survey. In H. Hagen and D. Roller, editors, *Geometric Modelling: Methods and Their Application*, pages 131-160, Berlin: Springer-Verlag, 1991.
Keywords: interpolation, état de l'art

Hard71 Hardy R. Multiquadratic equations of topography and other irregular surfaces. *J Geophysical Research*, 76:1905-1915, 1971.
Keywords: interpolation, multiquadric interpolation, radial basis functions

Krig51 Krige D. G. *A statistical approach to some mine valuation and allied problems on the Witwatersrand*. Master's thesis, University of the Witwatersrand, 1951.
Keywords: interpolation, Kriging

Math76 Mather P. M. *Computational Methods of Multivariate Analysis in Physical Geography*. Wiley, New York, 1976.
Keywords: spatial smoothing, trend surface analysis

Sain94 Sain S. R. *Adaptive kernel density estimation*. Thèse, Rice University, Houston, Texas, 1994.
Keywords: spatial smoothing, adaptive kernel density

Shep68 Shepard D. S. A two-dimensional interpolation function for irregularly spaced data. In *Proceedings of the 1968 ACM National Conference, New York*, pages 517-524, 1968.
[http://portal.acm.org/citation.cfm?id=810616]
Keywords: interpolation, Inverse Distance Weighting

Sibs81 Sibson R. A brief description of the natural neighbour interpolant. In V. Barnett, editor, *Interpreting Multivariate Data*, pages 21-36. Wiley, New York, 1981.
Keywords: interpolation, natural neighbour, triangulation

Silv86 Silverman B. W. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.
Keywords: spatial smoothing, kernel density

Wats92 Watson D. F. *Contouring: A Guide to the Analysis and Display of Spatial*

Data. Pergamon Press, Oxford, 1992.
Keywords: interpolation, état de l'art

Sectorisation

AK95 Alpert C. J. and Kahng A. B. Recent directions in netlist partitioning: A survey. *Integration: The VLSI J.*, 19:1-81, 1995.
Keywords: Rééquilibrage de partitions, mise en oeuvre des transferts

FM82 Fiduccia C. M. and Mattheyses R. M. A linear-time heuristic for improving network partitions. In *DAC '82: Proceedings of the 19th conference on Design automation*, pages 175-181, Piscataway, NJ, USA, 1982. IEEE Press. ISBN 0-89791-020-6.
Keywords: Rééquilibrage de partitions, calcul des transferts

HL95 Hendrickson B. and Leland R. A multi-level algorithm for partitioning graphs. In *Proceedings of Supercomputing'95*, San Diego, CA, Décembre 1995. ACM/IEEE.
[http://www.chg.ru/SC95PROC/509_BHEN/SC95.HTM]
Keywords: Partitionnement de graphe, random matching

HB95 Hu Y.F. and Blake R.J. An optimal dynamic load balancing algorithm. Technical report, Daresbury Laboratory, Warrington, UK, 1995.
[http://citeseer.ist.psu.edu/hu95optimal.html]
Keywords: Rééquilibrage de partitions, calcul des transferts

KK99 Karypis G. and Kumar V. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359-392, 1999.
[http://www-users.cs.umn.edu/~karypis/publications/Papers/PDF/mlevel_serial.pdf]
Keywords: Partitionnement de graphe, Partitionnement multiniveaux

KK98 Karypis G. and Kumar V. Multilevel k-way partioning scheme for irregular graphs. *Journal of Parallel and Distributed Computing*, 48:96-129, 1998.
[http://www-users.cs.umn.edu/~karypis/publications/Papers/PDF/mlevel_kway.pdf]
Keywords: Partitionnement de graphe, Partitionnement multiniveaux k-way

KSK03 Karypis G., Schloegel K., and Kumar V. *ParMeTis: Parallel Graph Partitioning and Sparse Matrix Ordering Library, Version 3.1*. University of Minnesota, Dept. of Computer Science, Août 2003.
[http://www-users.cs.umn.edu/~karypis/metis/parmetis/files/manual.pdf]
Keywords: Partitionnement de graphe, manuel d'utilisation, bibliothèque

Kris84 Krishnamurthy B. An improved min-cut algorithm for partitioning VLSI networks. *IEEE Trans. Computers*, 33(5):438-446, 1984.
Keywords: Rééquilibrage de partitions, calcul des transferts

OR94 Ou C.W. and Ranka S. Parallel incremental graph partitioning using linear programming. In *Supercomputing '94: Proceedings of the 1994 conference on Supercomputing*, pages 458-467, Los Alamitos, CA, USA, 1994. IEEE Computer Society Press. ISBN 0-8186-6605-6.
[http://portal.acm.org/citation.cfm?id=198492#]
Keywords: Rééquilibrage de partitions, calcul des transferts

PSL90 Pothen A., Simon H., and Liou K. Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. and Appl.*, 11:430-452, 1990.
Keywords: Partitionnement de graphe, bissection spectrale

SKK97 Schloegel K., Karypis G., and Kumar V. Multilevel diffusion schemes for repartitioning of adaptive meshes. *J. Parallel Distrib. Comput*, 47(2):109-124, 1997.
[http://www-users.cs.umn.edu/~karypis/publications/Papers/PDF/mlevel_diffuse_serial.pdf]
Keywords: Rééquilibrage de partitions, état de l'art

SKK03 Schloegel K., Karypis G., and Kumar V. Graph partitioning for high-performance scientific simulations. In Jack Dongarra, Ian Foster, Geoffrey Fox, William Gropp, Ken Kennedy, Linda Torczon, and Andy White, editors, *Sourcebook of Parallel Computing*, pages 491-541. Morgan Kaufman, San Francisco, 2003. Chap. 18.
[http://www-users.cs.umn.edu/~karypis/publications/Papers/PDF/gpchapter.pdf]
Keywords: Partitionnement de graphe, état de l'art

WC00 Walshaw and Cross. Mesh partitioning: A multilevel balancing and refinement algorithm. *SIAM Journal on Scientific Computing*, 22(1):63-80, Janvier 2000.
[http://citeseer.ist.psu.edu/article/walshaw98mesh.html]

Arbre de Décision

BFOS84 Breiman L., Friedman J. H., Olshen R. A., and Stone C. J. Classification and regression trees. Technical report, Wadsworth International, Monterey, CA, 1984.
Keywords: Arbre de Décision, CART

BU95 Brodley C. E. and Utgoff P. E. Multivariate decision trees. *Machine Learning*, 19(1):45-77, 1995.
[http://citeseer.ist.psu.edu/brodley92multivariate.html]
Keywords: Arbre de Décision, LMDT, arbre multivarié

CHH99 Coppersmith D., Hong S. J., and Hosking J. R. M. Partitioning nominal attributes in decision trees. *Data Mining and Knowledge Discovery*, 3(2):197-217, 1999.

[http://citeseer.ist.psu.edu/coppersmith99partitioning.html]
Keywords: Arbre de Décision, partitionnement binaire quasi-optimal pour une variable qualitative

HMS66 Hunt E. B., Marin J., and Stone P. T. *Experiments in Induction*. Academic Press, New York, 1966.
Keywords: Arbre de Décision, CLS

Kass80 Kass G. V. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2):119-127, 1980.
Keywords: Arbre de Décision, CHAID

LS97 Loh W. Y. and Shih Y. S. Split selection methods for classification trees. *Statistica Sinica*, 1997.
[http://citeseer.ist.psu.edu/loh97split.html]
Keywords: Arbre de Décision, fragmentation

MAR96 Mehta M., Agrawal R., and Rissanen J. SLIQ: A fast scalable classifier for data mining. *Lecture Notes in Computer Science*, 1057, 1996.
[http://citeseer.ist.psu.edu/mehta96sliq.html]
Keywords: Arbre de Décision, partitionnement binaire quasi-optimal pour une variable qualitative

MS63 Morgan J. N. and Sonquist J. A. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58:415-434, 1963.
Keywords: Arbre de Décision, AID

Murt98 Murthy S. K. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2(4):345-389, 1998.
[http://citeseer.ist.psu.edu/murthy97automatic.html]
Keywords: Arbre de Décision, état de l'art

Quin93 Quinlan J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.
Keywords: Arbre de Décision, ID3, C4.5

Autocorrélation Spatiale

Anse95 Anselin L. Local Indicators of Spatial Association: LISA. *Geographical Analysis*, 27(2):93-115, 1995.
Keywords: Autocorrélation spatiale locale, coefficient de Moran local

Flah01 Flahaut B. L'autocorrélation spatiale comme outil géostatistique d'identification des concentrations spatiales des accidents de la route. Février 2001.
[http://www.cybergeo.presse.fr/modelis/flahaut/flahaut.pdf]
Keywords: Autocorrélation spatiale,coefficient de Moran local, Application

Gear50 Geary R. C. The contiguity ratio and statistical mapping. *The Incorporated statistician*, 5:115-145, 1950.
Keywords: Autocorrélation spatiale, coefficient de Geary

JG03 Jacquez G. and Greiling D. Local clustering in breast, lung and colorectal cancer in Long Island, New York. *International Journal of Health Geographics*, 2(3), 2003.
[http://www.ij-healthgeographics.com/content/2/1/3]
Keywords: Autocorrélation spatiale locale, coefficient de Moran local, Application

Mora50 Moran P. Notes on continuous stochastic phenomena. *Biometrika*, 37:17-23, 1950.
Keywords: Autocorrélation spatiale, coefficient de Moran

PS97 Pumain D. and Saint-Julien T. *Localisations dans l'espace*, volume 1 of *L'analyse spatiale*. Armand Colin, 1997. ISBN 2-200-01897-5.
Keywords: Courbe de concentration, Autocorrélaton spatiale, méthode des quadrats

Modélisation des Flux

Char05 Charron F. *Les dynamiques de localisation des activités économiques et les motivations des chefs d'entreprise dans leurs choix d'implantation, l'exemple de la Loire-Atlantique*. Thèse CIFRE « GÉOBS – CNRS, Université de Nantes », 19 janvier 2005.
Keywords: Etude des flux d'entreprises au sein de la Loire-Atlantique

DT83 Dorigo G. and Tobler W. Push pull migration laws. *Annals of the Association of American Geographers*, 73(1):1-17, Mars 1983.
[http://www.geog.ucsb.edu/~tobler/publications/pdf_docs/movement/migration/PushPull.pdf]
Keywords: Modélisation des flux, modèle additif à double contrainte

Gras98 Grasland C. Interaction spatiale et interaction territoriale. Octobre 1998.
[http://www.grasland.cicrp.jussieu.fr/grasland/anspa/inter/Ateg98_4.htm]
Keywords: Modélisation des flux

PS01 Pumain D. and Saint-Julien T. *Les interactions spatiales*. Armand Colin, 2001. ISBN 2-200-26146-2.
Keywords: Modélisation des flux, Fonctions d'interaction spatiale

Wils67 Wilson A. A statistical theory of spatial distribution models. *Transportation Research*, 1:253-269, 1967.
Keywords: Modélisation des flux, modèle multiplicatif à double contrainte

LISTE DES PUBLICATIONS

Conférences

Candillier C., Couronné P., Gelgon M., and Mouaddib N. MSAT: A Multiscale Spatial Analysis Tool. In *Proceedings of the 19th International Conference: Informatics for Environmental Protection (EnviroInfo 2005)*, Brno, Czech Republic, 2005. Hrebicek, Racek (Eds.), Masaryk University. ISBN : 80-210-3780-6
Keywords: Méthodes d'analyse spatiale multi-échelles

Candillier C. and Mouaddib N. Outil de classification et de visualisation de grands volumes de données mixtes. In *5èmes journées d'Extraction et de Gestion des Connaissances (EGC 2005)*, Paris, France, 2005. Editions Cépaduès. ISBN : 2-85428-677-4
Keywords: Classification de données, visualisation de classifications

Candillier C. La sectorisation à partir de pôles - Application à la détermination de bassins d'emplois. In *Journée des Doctorants*, Ecole Polytechnique, Nantes, France, 2004.
Keywords: sectorisation, pôles, partition de graphes, rééquilibrage

Rapport technique

Candillier C. and Mouaddib N. A clustering and display software for large databases with mixed attributes. (soumis à *the 5th IEEE International Conference on Data Mining*, New Orleans, Louisiana, USA, 2005).
Keywords: Classification de données, visualisation de classifications

LISTE DES FIGURES

TABLE DES MATIÈRES

Méthodes d'Extraction de Connaissances à partir de Données (ECD) appliquées aux Systèmes d'Information Géographiques (SIG)

Christophe CANDILLIER

Résumé

Le travail effectué durant cette thèse concerne l'étude des méthodes d'Extraction de Connaissances à partir de Données (ECD) dans le cadre des Systèmes d'Information Géographiques (SIG). Nous avons non seulement mis en œuvre et amélioré des méthodes d'ECD classique (*Classification de Données, Visualisation de Classifications)* mais aussi des méthodes d'ECD spatiales liées à des méthodes d'analyse spatiale (*Lissage Spatial, Détermination de Pôles*, *Sectorisation*). Nous avons effectué notre travail de recherche au sein de la société GÉOBS spécialisée dans l'analyse des données géographiques (spatiales), et nous avons donc expérimenté, appliqué et vérifié ces méthodes sur des jeux de données fournis par GÉOBS et liés à des problématiques de Développement Économique, de Géomarketing, d'Analyse de Risque, d'Environnement, de Santé, etc. Ce mémoire offre une vision globale concernant un ensemble de problématiques et de méthodes d'analyse. Il met ainsi en avant la complémentarité des méthodes utilisées qui sont souvent connectées entre elles soit du point de vue technique soit du point de vue de leur utilisation. Finalement, ce fut un travail très enrichissant car il a touché à de nombreuses problématiques et à d'aussi nombreuses méthodes d'extraction de connaissances.

Mots-clés : Fouille de données, Extraction de Connaissances à partir de Données (ECD), Systèmes d'Information Géographiques (SIG), Classification de Données, Visualisation de Classifications, Arbres de Décision, Lissage Spatial, Sectorisation, Autocorrélation Spatiale, Modélisation des Flux

Abstract

During this PhD thesis, we have studied methods for Knowledge Discovery in Databases (KDD) applied to Geographic Information Systems (GIS). We have improved both classical KDD methods (*Data Clustering, Cluster Visualization*) and spatial KDD methods linked with spatial analysis methods (*Spatial Smoothing, Hot Spot Extraction*, *Spatial Partitionning*). We have worked in GÉOBS, a company expert in spatial data analysis. So our KDD methods have been implemented and tested with data sets provided by GÉOBS in relation with Economic Development, Risk Analysis, Environnement, Health, etc. This report gives a wide point of view on a range of analysis methods and their related problems. It points up the complementarity between theses methods which can be connected either in a technical way or in a user way. Eventually, this work was very enriching because it has concerned many problems and as many KDD tools.

Key-words: Data Mining, Knowledge Discovery in Databases (KDD), Geographic Information Systems (GIS), Data Clustering, Cluster Visualization, Decision Tree, Spatial Smoothing, Spatial Partitionning, Spatial Autocorrelation, Flow Modeling