Développement et mise en place d'une méthode de classification multi-blocs : application aux données de l'OQAI.

Mory Ouattara

CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS | le cnam

École Doctorale Informatique, Télécommunications et Électronique de Paris

CEDRIC

THÈSE DE DOCTORAT

présentée par : **Mory OUATTARA**

soutenue le : **18 mars 2014**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Spécialité : **Informatique**

Développement et mise en place d'une méthode de classification multi-bloc

Application aux données de l'OQAI

THÈSE DIRIGÉE PAR

M. BADRAN Fouad	*Professeur, CNAM*
Mme. NIANG Ndèye	*Maître de Conférence, CNAM*
Mme. MANDIN Corinne	*Ingénieure, CSTB*

RAPPORTEURS

M. PALUMBO Francesco	*Professeur, Université Federico II Naples*
M. QANNARI El Mostafa	*Professeur, Oniris Nantes*

EXAMINATEURS

M. NADIF Mohamed	*Professeur, Université Paris V*
M. BISSON Gilles	*Chargé de recherche, CNRS Grenoble*
Mme. DORIZZI Bernadette	*Professeur à TéléCOM SudParis, INT*

INVITE

M. DEROUBAIX Pierre	*Ingénieur, ADEME*

A mon père
A ma mère
A mon épouse Mariame
A ma fille Djuma
A toute ma famille

Remerciements

Je voudrais tout d'abord à remercier le professeur Fouad BADRAN qui a accepté de diriger cette thèse et qui m'a encadré pendant ces 3 années de thèse avec rigueur et pertinence.

Je tiens particulièrement à remercier Madame Ndèye NIANG, qui durant cette période a toujours été présente et a su me guider dans mon travail. Elle a toujours su apporter le regard critique nécessaire sur mes travaux et me proposer des voies de recherche pertinentes. Elle a su me donner l'envie et la motivation de mener ces travaux à bien, et pour cela je lui dois beaucoup.

Je remercie Messieurs les professeurs Francesco PALUMBO et El Mostafa QANNARI pour avoir accepté d'être rapporteurs de ma thèse et pour leur regard pertinent sur mon travail.

Je tiens à remercier également Madame le professeur Bernadette DORIZZI, Monsieur Gilles BISSON et Monsieur le professeur Mohamed NADIF pour avoir accepté d'examiner ma thèse.

Je tiens à remercier également Madame Corinne MANDIN et toute l'équipe de l'OQAI qui m'ont fourni un cadre de travail idéal pendant ma thèse et qui m'ont permis de développer mes idées. Ils ont toujours su apporter leurs expertises et un regard critique nécessaire sur les aspects techniques et applicatifs de mon travail.

Je tiens à remercier l'ADEME représentée par Monsieur Pierre DEROUBAIX et le CSTB pour le financement qu'ils m'ont attribué pour 3 ans et qui m'a permis de mener à bien les travaux présentés dans ce mémoire.

Merci à Madame Sylvie THIRIA pour sa disponibilité et sans qui cette thèse n'aurait pas débutée.

Merci à Giorgio RUSSOLILLO pour sa disponibilité, son aide et ses idées lors nos discussions. Nos échanges m'ont toujours permis de prendre du recul pour mieux cerner les difficultés rencontrées.

Je remercie également mon oncle Anzoumana GBANE et toute sa famille qui m'ont été d'un grand soutien pendant tout mon cursus. Je remercie infiniment mes parents pour tout ce qu'ils m'ont donné.

Merci à mon épouse Mariame pour sa patience et son soutien pendant ces années de thèse, merci également à ma fille Djuma pour son calme et nos moments de bonheur.

Enfin, merci à tous mes ami(e)s : Cindie, Nacera, Manu, Mohammed, Remi, Assia.

Résumé

La multiplication des sources d'information et le développement de nouvelles technologies engendrent des bases de données complexes. Dans les études environnementales sur la pollution de l'air intérieur par exemple, la collecte des informations sur les bâtiments se fait au regard de plusieurs thématiques comme les concentrations en polluants, les caractéristiques techniques, etc., engendrant ainsi des données de grande dimension avec une structure multi-blocs définie par les thématiques. L'objectif de ce travail visait à développer des méthodes de classification adaptées à ces jeux de données de grande dimension et structurées en blocs de variables.

La première partie de ce travail présente un état de l'art des méthodes de classification en général et dans le cas de la grande dimension en particulier. Dans la deuxième partie, trois nouvelles approches de classification d'individus décrits par des variables structurées en blocs ont été proposées. La méthode 2S-SOM (Soft Subspace Self Organizing Map) est une approche de type subspace clustering basée sur une modification de la fonction de coût de l'algorithme des cartes topologiques à travers un double système de poids adaptatifs défini sur les blocs et sur les variables. Pour surmonter la dépendance des résultats de la méthode SOM par rapport aux paramètres d'initialisation, nous proposons deux approches de recherche de consensus de SOM, CSOM (Consensus SOM) et Rv-CSOM,qui prennent en compte la structuration en bloc des variables. Dans l'approche CSOM, l'objectif est de favoriser les cartes les meilleures au sens d'un certain critère de validation, alors que Rv-CSOM va plutôt privilégier les cartes similaires par évaluation de la liaison entre cartes. Enfin, la troisième partie présente une application de ces méthodes sur le jeu de données de la campagne nationale « logements » menée par l'Observatoire de la Qualité de l'Air Intérieur (OQAI) afin de définir une typologie des logements français au regard des thématiques : qualité de l'air intérieur, caractéristiques constructives du bâtiment, composition des ménages et habitudes des occupants.

Mots clés : classification, multi-blocs, subspace clustering, consensus, SOM

Abstract

The multiplication of information sources and the development of new technologies generate complex databases, often characterized by a relatively high number of variables compared to individuals. In particular, in the environmental studies on the indoor air quality, the information is collected according to several aspects (technical characteristics, pollutant indoor concentrations, etc.), yielding column partitioned or multi-block data set. However, in case of high dimensional data, classical clustering algorithms are not efficient to find clusters which may exist in subspaces of the original space. The goal of this work was to develop clustering algorithms adapted to high dimensional data sets with multi-block structure.

The first part of the work presents the state of the art on clustering methods. In the second part, three new methods of clustering are developed : the subspace clustering method 2S-SOM (Soft Subspace-Self Organizing Map) is based on a modified cost function of the Self Organizing Maps method across a double system of weights on the blocks and the variables. Then we propose two approaches to find the consensus of self-organized maps : CSOM (Consensus SOM) and RV-CSOM based on weights determined from initial partitions.The last part presents an application of these methods to the French Observatory of indoor air quality database on housing. A nationwide survey was carried out to determine typologies of dwellings relatively to indoor air quality, building structure, household characteristics and habits of the inhabitants.

Keywords : clustering, multi-block, subspace clustering, cluster ensemble, SOM

Table des matières

Liste des tableaux

Table des figures

Introduction

Nous vivons principalement dans des espaces clos, qu'il s'agisse de lieux accueillant du public (transport, écoles, hôpitaux etc.), des bâtiments professionnels (bureaux et commerces) ou d'espaces privés (logements individuels ou collectifs). Or des signaux forts, en particulier le scandale de l'amiante en France, ont mis en évidence les conséquences potentiellement très graves liées à la présence de substances toxiques dans les bâtiments. Nous sommes tous exposés aux polluants présents dans l'atmosphère des environnements clos. Et les problèmes de santé dus à cette exposition sont multiples avec des manifestations cliniques très diverses qui pour la plupart ne sont pas spécifiques des polluants [Kirchner *et al.* 2011].

Dans ce contexte, des préoccupations de santé environnementale émergent en réclamant des données fiables et indépendantes pour prévenir et agir. L'Observatoire de la Qualité de l'Air Intérieur (OQAI) créé en 2001 dont l'opérateur scientifique et technique est le Centre Scientifique et Technique du Bâtiment (CSTB) en lien avec les ministères en charge de la Santé et du Logement, l'Agence de l'Environnement et de la Maîtrise de l'Energie (ADEME), l'Agence nationale de l'habitat (ANAH), puis l'AFFSET (Agence française de sécurité sanitaire de l'environnement et du travail, aujourd'hui ANSES) organise plusieurs campagnes de mesure des polluants de l'air intérieur dans les environnements clos. L'enjeu numéro un de ces campagnes est de développer la connaissance de la pollution intérieure, de ses origines et de ses effets sur la santé. Ainsi, il s'agit de rassembler des connaissances en vue d'identifier les situations préoccupantes pour la population et d'élaborer des recommandations dans le domaine des bâtiments, de leur conception et leur mise en œuvre à leur utilisation par les occupants.
Ainsi, suite à la campagne nationale «logements» (CNL) réalisée par l'OQAI, entre 2003 et 2005 dans 567 logements représentatifs du parc des 24 millions de résidences principales de la France continentale métropolitaine, l'OQAI a engagé une campagne nationale avec

une collecte des données sur les bâtiments à usage de bureaux (qualité de l'air intérieur, performance énergétique) et sur les occupants (confort perçu). Cette campagne vise d'une part à dresser une typologie descriptive des immeubles de bureaux au regard des informations collectées sur les aspects qualité de l'air intérieur et performance énergétique et sur les aspects santé et confort perçus par les occupants, d'autre part elle cherche à identifier des facteurs prédictifs de la qualité de l'air intérieur.

Ma thèse s'effectue dans le cadre d'une convention de recherche entre l'ADEME, l'OQAI et le Centre d'Etude et de Recherche en Informatique et Communication (CEDRIC) du CNAM.

L'objectif général est de proposer une méthode de recherche d'une unique typologie d'un vaste ensemble de données d'enquêtes (questionnaires) et de mesure structurées en blocs thématiques pouvant contenir des observations atypiques et manquantes. En rapport avec l'objectif premier de la campagne nationale «bureaux», nos travaux se situent dans le domaine de la classification automatique ou apprentissage non-supervisée.

L'apprentissage non-supervisé vise à découvrir la structure intrinsèque d'un ensemble d'observations en formant des classes. Ces classes sont constituées d'individus qui partagent les mêmes caractéristiques. La complexité de cette tâche s'est fortement accrue ces deux dernières décennies lorsque les masses de données disponibles ont vu leur volume accroître. La taille des données, mesurée selon deux dimensions (le nombre de variables et le nombre d'observations) peut être grande. De plus, la façon de collecter les données peut engendrer des problèmes lors de l'exploration et de l'analyse de ces données. Pour cela, il est fondamental de mettre en place des outils de traitement de données permettant une meilleure détection des connaissances disponibles dans ces données, facilitant la visualisation et la compréhension des données et identifiant les facteurs pertinents.

Dans cette thèse, nous nous intéressons aux cartes topologiques. Les cartes topologiques permettent, à l'aide d'un algorithme d'auto-organisation (Self Organizing Map, SOM), de quantifier d'une part de grandes quantités de données en formant les classes d'observations similaires et d'autre part de projeter les classes obtenues de façon non-linéaire sur une grille. Cette grille permet ensuite de visualiser la structure des données dans un espace constitué en général de deux dimensions tout en respectant, dans le même temps, la to-

pologie initiale des données. Cette famille d'algorithmes est généralement développée pour un jeu de données, or on rencontre dans la pratique de plus en plus de données multi-vues ou multi-blocs (un bloc ou une vue étant la collecte des variables selon une thématique précise) comme c'est généralement le cas des données de l'OQAI. Il apparaît donc bénéfique de développer un module de traitement des données multi-blocs dans les algorithmes d'auto-organisation avec comme objectif de prendre en compte la structure multi-blocs des données et de supprimer ou du moins atténuer l'influence de toute information non-pertinente dans le processus de quantification.

Ce mémoire est organisé en 6 chapitres.
Le chapitre 1 est consacré à la présentation du contexte et des objectifs de cette thèse. Après une brève présentation des notions de pollution de l'air intérieur, de santé et de confort perçu par les occupants et de performances énergétiques, nous présentons la campagnes nationale «bureaux» objet de cette thèse.

Le chapitre 2 est consacré à la présentation des modèles d'apprentissage non supervisé et principalement les méthodes de partitionnement basées sur la méthode des cartes auto-organisées SOM. Il présente le contexte de l'étude de la classification, en expliquant en détail différentes approches de classification non supervisée. Un autre problème important discuté dans le cadre de ce chapitre est l'évaluation des résultats de la classification.

La problématique et la structure multi-blocs des données de l'OQAI nous ont naturellement conduit à un état de l'art des méthodes de traitement de grandes bases de données. L'apprentissage de grandes bases de données impose aux algorithmes de classification des contraintes liées à la nature des variables, à la distribution associée ou aux relations entre les variables. Ces difficultés ont conduit à l'émergence de nouvelles méthodes de classification.

Le chapitre 3 n'a pas la prétention d'être exhaustif, nous nous sommes limités à proposer une vue ou une représentation assez synthétique des méthodes de partitionnement adaptées aux données de grande dimension. Nous les avons regroupées en différentes catégories. Il s'agit des méthodes de réduction de dimension, des méthodes de sélection locale et globale des variables, des méthodes de bi-partitionnement et des méthodes de recherche d'un consensus de plusieurs partitions. Dans le cadre de cette thèse, nous nous sommes

intéressés plus particulièrement aux algorithmes à base de pondération et aux méthodes de recherche de consensus dans le cadre des cartes auto-organisées.

Le chapitre 4 présente la première contribution de cette thèse à savoir, la méthode Soft Subspace clustering basée sur les cartes auto-organisatrices (2S-SOM) dédiée aux données multi-blocs. En effet, dans le but de rechercher une typologie d'observations décrites par des variables structurées préalablement en blocs, nous avons développé un modèle des cartes topologiques qui est une extension des travaux de Jing *et al.* [2007] et de Chen *et al.* [2012] proposés pour la méthode des K-moyennes. 2S-SOM repose sur un double système de poids recherchés par modification de la fonction de coût de SOM. Ces poids évaluent la contribution relative des blocs et des variables dans la classification. Ce nouvel algorithme d'apprentissage non supervisé utilisant l'algorithme d'apprentissage topologique a une portée pratique et nous montrons sur des exemples l'amélioration qu'il apporte aussi bien en classification qu'en visualisation.

La deuxième contribution de cette thèse fait l'objet du chapitre 5, nous y proposons deux approches de recherche de consensus entre partitions : Consensus SOM (CSOM et R_V-CSOM). Ces méthodes servent d'une part à étudier la robustesse des méthodes des cartes topologiques et d'autre part à proposer un consensus de cartes topologiques auto-organisées obtenues sur divers blocs de variables. Le principe de la méthode CSOM consiste à prendre en compte la qualité spécifique de chaque carte servant à définir le consensus à travers des poids définis localement pour chaque carte. R_V-CSOM tient compte de l'information commune aux cartes dont on cherche le consensus.

Le chapitre 6 est consacré à l'application des approches proposées aux données de l'OQAI.

Première partie

État de l'art

Chapitre 1

Contexte et Objectifs

1.1 Introduction

Les bâtiments de bureaux sont des environnements d'intérêt en termes de santé publique puisque ce sont les seconds lieux de vie après l'habitat pour la population adulte active qui y travaille. De manière générale, l'air à l'intérieur de ces bâtiments peut être contaminé par de nombreux polluants provenant de l'air extérieur, du bâtiment et de ses équipements ou des occupants et de leurs activités. Ces polluants de l'air intérieur peuvent être regroupés en trois grandes catégories : les polluants chimiques, biologiques ou physiques. Qu'ils soient chimiques, biologiques ou physiques, les pathologies qui y sont associées sont multiples et entraînent, dans la plupart des cas, des maux bénins qui peuvent néanmoins être désagréables (maux de tête, nausées, étourdissements, fatigue, allergies, irritations, etc). Certaines substances telles que le radon ou l'amiante sont liés à l'apparition de maladies plus graves comme les cancers.

De plus, certains paramètres participant directement au niveau du confort des occupants comme le renouvellement de l'air, l'humidité et la température sont également impliqués dans l'apparition de la pollution (moisissures, acariens) [De Baudouin 2006; Mosqueron et Nedellec 2001] et dans l'augmentation du taux d'émission de Composés Organiques Volatils (COV) dans l'air. En effet, le taux d'humidité relative qui représente la quantité de vapeurs d'eau présente dans l'air (exprimée en pourcentage de la saturation) est une fonction de la température. Un taux d'humidité élevé, d'une part favorise l'infestation par les acariens, et d'autre part entraîne des condensations sur les zones froides des surfaces. Cette humidité de surfaces (et de matériaux), à son tour, favorise la prolifération fongique. Un fort taux d'humidité dans les matériaux pourrait également contribuer à ac-

célérer leur dégradation, entraînant l'émission de composés chimiques [Sabine 2005]. Par ailleurs, Van der Wal *et al.* [1997] montrent à travers une étude expérimentale que l'augmentation de la température entraîne une augmentation significative du taux d'émission de COV pour des matériaux d'aménagement intérieurs tels que les moquettes et les PVC. Plus particulièrement, le dégagement de formaldéhyde varie en fonction des conditions de température et d'humidité.
Ces paramètres influent sur la perception de l'environnement et sur la qualité de l'air intérieur par les occupants, et sont donc à prendre en considération.

La qualité de l'air des environnements de bureaux est actuellement mal connue en France. En outre, la connaissance du parc de bâtiments en termes de nombre, répartition géographique et typologie est très limitée. Il existe donc un réel manque de données sur l'exposition des travailleurs dans ces bâtiments, ce qui empêche de mettre en place des politiques de gestion ou des recommandations dédiées. Par ailleurs, à l'heure des politiques d'économie d'énergie, impulsées notamment par le Grenelle de l'environnement, la connaissance des performances énergétiques du parc d'immeubles de bureaux apparaît comme étant essentielle. En préalable à une présentation plus détaillée, dans la section 1.5 de la campagne nationale «bureaux» menées par l'OQAI, il est nécessaire de définir les notions de qualité de l'air intérieur (QAI), de santé et de confort perçu par les occupants et de performances énergétiques (PE) des immeubles de bureaux.

1.2 La pollution de l'air intérieur

La contamination de l'air intérieur est souvent due à la présence nombreux polluants chimiques, biologiques ou physiques [Kirchner *et al.* 2011; De Baudouin 2006] provenant de l'air extérieur, du bâtiment et de ses équipements ou des occupants et de leurs activités.

1.2.1 Les polluants chimiques

Les polluants chimiques de l'air intérieur, de loin les plus nombreux, sont retrouvés sous la forme gazeuse, d'aérosols, ou sous forme particulaire. Leur toxicité tient principalement de leur formule chimique. Les principaux polluants sont :

- Le monoxyde de carbone (CO), gaz très toxique, qui provient de la combustion incomplète de matériaux carbonés charbon, pétrole, essence, fioul, gaz, bois ou éthanol.

Il s'agit de la première cause de mortalité accidentelle par émanations toxiques en France. En cause le plus souvent, les installations de chauffage mal réglées ou mal entretenues.

- Les Composés organiques volatils (COV) recouvrent une grande variété de substances chimiques ayant pour point commun d'être composés de l'élément carbone et d'autres éléments tels que l'hydrogène, les halogènes, l'oxygène, ... et d'être volatils à température ambiante. Regroupés au sein de grandes familles définies en fonction de leur formule chimique, ils sont utilisés dans la fabrication de nombreux produits et matériaux (peinture, vernis, colles, moquette, carrelage, nettoyants, tissus neufs, etc.). Les COV englobent des familles très variées et présentent, selon les substances et les niveaux d'exposition, des effets divers sur la santé comme des irritations de la peau, des muqueuses et du système pulmonaire, des nausées, maux de tête et vomissements.
- Les oxydes d'azote (NO_x), gaz formés d'azote et d'oxygène, comprenant le monoxyde d'azote (NO) et le dioxyde d'azote (NO_2) ; ils sont émis lors de combustions à haute température ; la pollution intérieure provient essentiellement des appareils de chauffage ou de production d'eau chaude, des gazinières, du tabagisme ou de la circulation automobile (transfert de la pollution extérieure à l'intérieur des bâtiments) ; leur effet sanitaire concerne principalement les phénomènes d'irritation de l'appareil respiratoire (poumons).

1.2.2 Les polluants biologiques

Les polluants microbiologiques ne sont pas moins nocifs : moisissures, bactéries, virus, acariens, pollens etc. Les polluants biologiques peuvent se rencontrer un peu partout dans les espaces clos : moquettes, revêtements muraux, matériaux d'isolation, installations sanitaires, circuits de distribution d'eau, systèmes de climatisation et de ventilation, etc.
La présence de cette contamination peut poser notamment des problèmes de santé pour les personnes fragiles, en particulier les enfants, les asthmatiques et les personnes âgées. Elle peut provoquer des allergies voire des infections respiratoires (rhinites, par exemple) et pulmonaires (asthme, légionellose, et contribuent au développement de l'asthme. La chaleur, l'humidité en lien avec les dégâts des eaux, les remontées capillaires, des locaux mal entretenus, une mauvaise maintenance des installations d'eau chaude et de climatisation ou même la présence d'un malade (grippe) favorisent la présence, voire la prolifération de nombreux agents microbiologiques, augmentant les risques de leur diffusion dans l'air intérieur et l'eau.

1.2.3 Les polluants physiques

Les pollutions physiques, comme les particules fines et ultrafines ou encore le radon, peuvent aussi avoir de sérieuses conséquences sanitaires. Ainsi, ce gaz radioactif cancérogène qu'est le radon est présent à l'état naturel dans certaines régions et peut pénétrer à l'intérieur des bâtiments par les défauts d'étanchéité. Les particules en suspension provenant de diverses sources extérieures (trafic, industries, etc.) et d'activités domestiques (cuisson, chauffage, tabagisme, etc.) peuvent provoquer des atteintes respiratoires (asthme, broncho-pneumopathie chronique obstructive (BPCO) etc.) et cardio-vasculaires.

1.2.4 Impact de la qualité de l'air intérieur sur la santé et sur la productivité des occupants

Depuis les années 70, de nombreuses études ont mis en évidence l'impact de la qualité de l'air intérieur sur la santé et le confort des occupants [Mosqueron et Nedellec 2001, 2004; Roulet 2004; Brightman *et al.* 2008; Billionnet 2012]. En particulier, dans l'étude américaine de référence BASE [1] sur les immeubles de bureaux, 28% des participants ont rapporté 1 ou plusieurs jours d'arrêt maladie [Apte *et al.* 2000]. Les symptômes associés à ces maladies sont multiples et variables avec des causes souvent aspécifiques et multifactorielles [Merlo 2002; EPA 1995]. Ils peuvent être respiratoires (rhinorrhée, obstruction nasale, sécheresse nasale, toux, etc), oculaires (sécheresse des yeux, larmoiement, etc), cutanés (sécheresse de la peau, éruptions cutanées), sensoriels (impression de mauvaises odeurs, mauvais goûts, éblouissements, etc) et généraux (fatigue, difficulté de concentration, perte de la mémoire). Ces symptômes sont souvent présents sur les lieux de travail et disparaissent plus ou moins rapidement en dehors des bâtiments. Leur présence peut avoir un impact fort sur la productivité des occupants [Tuomainen *et al.* 2002; Fisk *et al.* 2011]. Ainsi, dans un environnement ayant une mauvaise qualité de l'air intérieur, en plus des arrêts maladies, les difficultés de concentration sont un réel frein pour la productivité des travailleurs. D'après l'étude BASE 40% des personnes interrogées pensent avoir une productivité réduite en lien avec une qualité de leur environnement intérieur de travail dégradée [Apte et Daisey 1999; Apte *et al.* 2000; Brightman *et al.* 2008].

1. Building Assessment Survey and Evaluation, réalisée entre 1994 et 1998 sur 100 immeubles de bureaux, EPA [1995]

1.3 Confort

La sensation de bien-être est une notion subjective dépendant de la propre évaluation d'une personne d'un environnement donné. La notion de confort peut toutefois être interprétée objectivement à partir des mécanismes physiologiques du corps humain. On distingue le confort thermique, visuel et acoustique.

Confort thermique

Les conditions de confort thermique dépendent de quatre paramètres physiques caractérisant l'environnement et de deux paramètres caractéristiques de l'individu :

- la température de l'air ;
- la température moyenne de rayonnement ;
- la vitesse de l'air ;
- l'humidité de l'air ;
- le niveau d'activité de l'individu ;
- la résistance thermique de son habillement.

A partir de ces 6 paramètres, il est alors possible de définir des indices de confort thermique permettant de caractériser le niveau de confort d'un individu placé dans une ambiance donnée [Parat *et al.* 2009].

Confort visuel

Le confort visuel est une impression subjective liée à la quantité, à la qualité et à la distribution de la lumière. Le confort visuel dépend d'une combinaison de paramètres physiques : l'éclairement, la luminance, le contraste, l'éblouissement et le spectre lumineux, et de facteurs physiologiques, psychologiques liés à l'individu tels que son âge et son acuité visuelle [Bremond-Gignac *et al.* 2002; Dubois 2006].

Confort acoustique

Le confort acoustique est lié à la qualité et à la quantité des évènements appréhendés par un auditeur. Les attentes des occupants concernant le confort acoustique consistent généralement à vouloir concilier deux besoins [CSTB 2005] :

- d'une part, ne pas être dérangés ou perturbés dans leurs activités quotidiennes par des bruits aériens (provenant d'autres locaux voisins), des bruits de chocs ou d'équipements (provenant des différentes parties du bâtiment) et par les bruits de l'espace

extérieur (transports, passants, chantier, etc.) ;
- mais, d'autre part, intérieur (logement, salle de classe, bureau) et extérieur en percevant les signaux qui leur sont utiles ou qu'ils jugent intéressants.

1.4 Performances Energétiques (PE)

De nombreuses composantes rendent compte de la performance énergétique d'un bâtiment. Ainsi, la consommation d'énergie d'un bâtiment, bien qu'indicateur pertinent de la performance énergétique globale, n'est pas suffisante pour qualifier la performance énergétique de l'enveloppe du bâtiment et celle des équipements techniques de chauffage, de refroidissement, de ventilation et d'éclairage. Par ailleurs, un bâtiment à haute performance énergétique d'enveloppe et/ou d'équipements techniques peut présenter une performance énergétique globale faible à cause de son usage, de sa gestion ou du comportement des occupants [Parat *et al.* 2009]. Pour ces différents raisons, il est essentiel de rendre compte de la performance globale des bâtiments. De plus, à taille et usage égaux, certains bâtiments consomment 3 à 4 fois moins d'énergie que d'autres et les connaissances actuelles montrent que l'on peut améliorer la qualité de l'environnement intérieur tout en réduisant très fortement la consommation d'énergie [Roulet *et al.* 2006; De Baudouin 2006].
La section suivante, présente la campagne nationale «bureaux» (CNB).

1.5 Campagne Nationale «Bureaux» (CNB)

Pour remédier au manque de connaissances sur le parc des bâtiments à usage de bureaux, et sur l'exposition des travailleurs dans ces bâtiments, l'Observatoire de la Qualité de l'Air Intérieur (OQAI) organise une campagne nationale de mesure de la qualité de l'air intérieur et d'évaluation des aspects santé et confort perçu ainsi que les performances énergétiques dans les immeubles de bureaux. Le protocole de cette Campagne Nationale «Bureaux» (CNB) est basé sur le programme européen HOPE[2] [Bluyssen *et al.* 2003] validé par de nombreux experts européens. Une adaptation et un affinement du protocole initial de HOPE ont néanmoins été nécessaires pour l'appliquer aux bâtiments à usage de bureaux français.

2. HOPE Health Optimisation Protocol for Energy-effeicient buildings est un programme de recherche multidisciplinaire européen impliquant 9 pays a été mis sur pied pour évaluer le confort et la santé dans les bâtiments administratifs et résidentiels et leur relation avec la consommation d'énergie [Bluyssen *et al.* 2003].

1.5.1 Objectifs

Le but de la CNB est dans un premier temps d'élaborer un état descriptif du parc des immeubles de bureaux de France métropolitaine en termes de qualité de l'air intérieur, de confort perçu par les occupants et de performances énergétiques afin d'établir des typologies. Ensuite, dans une seconde étape, il s'agira de construire à partir des typologies précédentes des indicateurs de classement des immeubles de bureaux.

La CNB vise donc à établir des regroupements d'immeubles en classes homogènes, par rapport à la qualité de l'air intérieur, à la santé et au confort perçu par les occupants et par rapport aux performances énergétiques, permettant d'expliquer les facteurs de dégradation de la qualité de l'air intérieur, l'effet de la présence de certains polluants sur la santé et l'appréciation du confort par les occupants et les performances énergétiques des immeubles. Puis, à travers ces groupements, il s'agira de définir des indices de classement localement selon les thématiques QAI, santé et confort perçu par les occupants et performances énergétiques des immeubles de bureaux et un indice global de description des immeubles de bureaux. Ainsi, les méthodes statistiques utilisées pour atteindre ces objectifs seront d'une part non supervisées pour établir les typologies descriptives et d'autre part supervisées pour établir un classement des immeubles de bureaux. Les principales contributions de cette thèse concernent la partie non-supervisée. Une revue bibliographique et une présentation des méthodes de classification non-supervisées seront effectuées dans le chapitre 2. La section suivante est dédiée à la présentation des données de la CNB.

1.5.2 Organisation

La campagne nationale «bureaux» se déroule en deux phases :

Phase 1

La phase 1 concerne un premier échantillon d'immeubles répartis en France métropolitaine dont le choix est détaillé dans la section 1.5.4. Elle sera effectuée pendant une journée par trois techniciens enquêteurs formés préalablement (l'un d'eux aura exclusivement en charge le volet énergie). Cette phase est dédiée au recueil des données sur le bâtiment, les aspects santé et confort perçu par les occupants et la consommation énergétique. Dans cette phase, la collecte des données est réalisée grâce à des mesurages environnementaux dans cinq bureaux par immeuble et à travers quatre questionnaires :

– les mesurages environnementaux sont réalisés dans cinq bureaux (ou au niveau des

postes de travail occupés, pour le cas d'espaces ouverts). Les bureaux seront choisis pour être représentatifs des bureaux de l'immeuble : bureaux individuels, open space, bureau aveugle s'il y a lieu. Ils seront autant que possible répartis sur les différentes façades (différentes orientations). Seront mesurés sur la journée les paramètres suivants : concentrations en composés organiques volatils (COV) et aldéhydes, particules ultrafines (comptage en nombre), température, humidité relative et concentrations en CO_2. Un point extérieur servira de référence (mesure de COV).

- un questionnaire d'accompagnement de la mesure décrivant les conditions de mesures des concentrations des polluants.
- un questionnaire descriptif des éléments techniques du bâtiment. Les caractéristiques de l'immeuble et les usages susceptibles d'influencer la qualité de l'air intérieur seront renseignés précisément : descriptif général, descriptif des environnements intérieur et extérieur, entretien, activités dans l'immeuble, etc
- un questionnaire relatif à l'énergie : relevés des consommations énergétiques et descriptifs des ouvrants, parois, planchers, conditions de chauffage, équipements et installations, etc
- un auto-questionnaire, devant être rempli individuellement par les occupants du bâtiment, sur leur bien-être et leur perception du confort du bâtiment (ce questionnaire sera distribué aléatoirement aux occupants de l'immeuble avec un minimum de 50 occupants pour permettre le calcul d'indicateurs tel que BSI [3] (Building Synptom Index). Ce questionnaire est divisé en 4 parties :
  - informations générales ;
  - bien-être et symptômes ;
  - confort perçu par les occupants ;
  - autres aspects liés au bureau ;

Les données recueillies lors de cette phase permettront de mener des études pour obtenir une classification des immeubles de bureaux selon les critères de santé, de confort perçu par les occupants et de performances énergétiques. Cette thèse intervient dans le cas de cette phase.

Phase 2

La phase 2 porte sur un sous-échantillon représentatif des typologies établies à partir de l'échantillon de la phase 1. L'objectif est d'approfondir les données obtenues et par

3. Le Building Symptom Index (BSI) équivaut au nombre moyen, par occupant, de symptômes attribuables au bâtiment, c'est-à-dire disparaissant en dehors de celui-ci

conséquent, consolider les résultats de la phase 1. Pour cela, des mesures approfondies seront réalisées dans 5 bureaux (ou postes de travail) par immeuble, pendant une semaine complète afin d'affiner les typologies. Il s'agit :

- les Composés Organiques Volatils (COV), les aldéhydes, et l'ozone ;
- PM_{10}, $PM_{2,5}$ et particules ultra fines.
- fibres minérales artificielles et amiante pour le cas où l'enquête en phase 1 a permis d'identifier la présence d'amiante ;
- allergènes d'acariens, de chat et de chien ;
- endotoxines, flore fongique et flore bactérienne ;
- paramètres relatifs à la ventilation (concentration en CO_2, débits d'air extrait) ;
- confort thermique, acoustique et éclairage

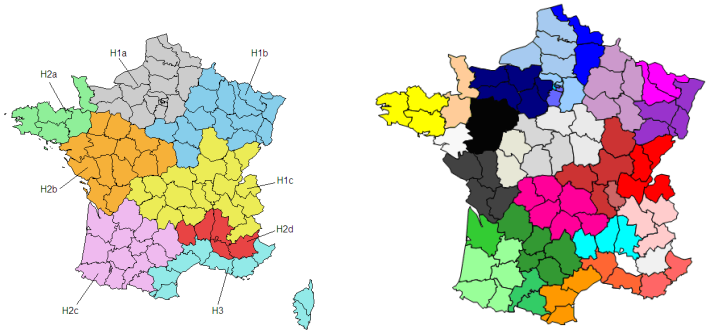
1.5.3 Échantillon et structure des données

Pour pouvoir extrapoler les résultats de la CNB à l'ensemble du parc de bureaux du territoire français, l'échantillon d'immeubles de la CNB doit permettre d'estimer les paramètres sans biais et avec une précision acceptable. Par ailleurs, compte tenu de l'inégale répartition des immeubles de bureaux par rapport à la démographie des départements, plusieurs niveaux d'échantillonnage seront nécessaires à la constitution de l'échantillon afin d'assurer à chaque immeuble de bureaux du territoire Français la même probabilité d'être choisi.

Un plan de sondage est réalisé sur deux niveaux. D'une part, il est lié aux 8 zones climatiques (ZC) de la réglementation thermique française dont le but est de fixer une limite maximale à la consommation énergétique des bâtiments neufs pour le chauffage, la ventilation, la climatisation, la production d'eau chaude sanitaire et l'éclairage (cf. Figure 1). D'autre part, le plan de sondage est lié à la définition de zones dites d'enquêtes (ZE). Les ZE sont composées de 1 à 6 départements contigus au sein d'une zone climatique (ZC). Les 4 villes les plus importantes en terme de nombre d'immeubles de bureaux sont directement considérées comme des zones d'enquête à part entière. Le sondage à degrés est le suivant :

- premier degré : pour chacune des 8 zones climatiques de la réglementation thermique française, des tirages équiprobables des zones d'enquêtes sont effectués ;
- deuxième degré : dans chacune des zones d'enquête tirées au sort, un nombre d'immeubles de bureaux proportionnel au nombre total d'immeubles dans la zone ou la ville est tiré au sort, pour atteindre un total de 300 immeubles. Le tableau 1.1 et les

figures 1.1(a) et 1.1(b) présentent les 8 zones climatiques (ZC) ainsi que les zones d'enquêtes (ZE) associées.



(a) Zones climatiques ; H1a, H1b, H1c, H2a, H2b, H2c, H2d, H3 sont les 8 ZC

(b) Zones d'enquêtes définies (32) ; Une même couleur indique qu'ils définissent une zone d'enquête

FIGURE 1.1 – Les deux niveaux de sondage ayant servi à définir l'échantillon d'immeubles de «bureaux» de la CNB

TABLE 1.1 – Répartition des départements par zone climatique et zone d'enquête

ZC	ZE	Dept	ZC	ZE	Dept
H1a	Z1	59, 02	H2a	Z1	29, 22, 56
H1a	Z2	62, 80, 76, 60	H2a	Z2	35, 50
H1a	Z3	14, 61, 27, 28, 78	H2b	Z1	44
H1a	Z4	77, 93, 95	H2b	Z2	85, 79, 17, 16
H1a	Z5	91, 94	H2b	Z3	41, 18, 36
H1a	Z6	92	H2b	Z4	86, 37
H1a-Paris	Paris	75			
H1b	Z1	45, 89, 58	H2b	Z5	53, 49, 72
H1b	Z2	08, 51, 10, 52, 55	H2c	Z1	33
H1b	Z3	54, 57	H2c	Z2	40, 64, 32, 65, 47
H1b	Z4	67	H2c	Z3	31, 09
H2c-Toulouse	Toulouse	31			
H1b	Z5	88, 68, 90, 70	H2c	Z4	24, 46, 82, 81, 12
H1c	Z1	69	H2d	Z1	84, 04
H1c	Z2	38, 73, 05	H2d	Z2	48, 07, 26
H1c	Z3	25, 39, 01, 74	H3	Z1	66, 11, 34
H1c	Z4	21, 71, 03, 42	H3	Z2	30, 13, 20
H1c	Z5	87, 23, 19, 63 43, 15	H3	Z3	83, 06
H1c-Lyon	Lyon	69			
H3-Marseille	Marseille	13			

1.5.4 Complexité liée aux données

Caractéristiques des données de la CNB

Les attributs décrivant les observations sont situés à 3 niveaux. Le niveau 1 est constitué des attributs décrivant l'immeuble dans son ensemble. Le niveau 2 se compose d'attributs décrivant les conditions de réalisation des 5 mesurages (5 bureaux par immeuble) des contaminants dans l'immeuble. Le niveau 3 est composé d'attributs décrivant les aspects santé et confort perçu par les occupants (au moins 50 occupants) par immeuble. L'échantillon de la CNB est composé de 1758 attributs de type qualitatif ou quantitatif structurés en blocs. Il comporte des variables catégorielles nominales ou ordinales et des variables quantitatives discrètes ou continues. La diversité des types de variables combinée à la grande dimension des données et à la présence d'observations présentant des comportements particuliers par rapport au reste de l'échantillon (atypiques) ou des données manquantes limitent assez rapidement les méthodes statistiques classiques de classification dédiées au traitement des données.

Plusieurs méthodes ont été proposées par différents auteurs pour étudier ces données dites mixtes dans le cadre de l'analyse factorielle. Elles reposent généralement soit sur un codage optimal des modalités des variables qualitatives pour ensuite les traiter comme des variables numériques à travers une analyse en composantes principales (ACP), soit sur une transformation adéquate des variables quantitatives en qualitatives. Dans le cadre de la classification d'individus, lorsque ces derniers sont décrits par des variables qualitatives, une méthode usuelle consiste à réaliser une ACM au préalable et à effectuer la classification sur les coordonnées factorielles alors quantitatives. Ces processus de discrétisation ou de transformation des variables conduisent souvent à des pertes d'information d'où la nécessité de définir une distance prenant en compte la structure mixte des données facilitant par ailleurs l'interprétation des résultats.

Structuration en blocs

La CNB est organisée au travers d'enquêtes menées dans les immeubles de bureaux et les données collectées sont structurées en quatre grands thèmes : les caractéristiques de l'immeuble (matériaux de construction, environnement extérieur, les ouvrants, etc), la qualité de l'air intérieur (les mesures des concentrations des contaminants chimiques, physiques, biologiques, etc), la santé et la perception du confort (la température, la luminosité, la ventilation, etc) et les performances énergétiques (consommation d'énergie, etc).

Une structure simplifiée des données est présentée dans la figure 1.2. L'étude descriptive des données de la CNB est étroitement liée à la compréhension du mode de fonctionnement, par thématique, de l'ensemble des immeubles de bureaux. Il est alors naturel d'une part, de procéder à l'étude des blocs relatifs aux thématiques qualité de l'air, santé et confort et performances énergétiques des immeubles de bureaux, d'autre part, de déterminer par thématique les variables initiales les plus informatives en identifiant les ressemblances et les dissemblances entre variables. Par ailleurs, du fait de la spécificité de chaque thématique, il apparaît important de mesurer à la fois l'importance des blocs et des variables dans une étude plus globale afin d'étudier les éventuelles relations entre les thématiques. La démarche adoptée pour le traitement des données de la CNB doit donc prendre en compte les particularités des données et les objectifs de traitement associés.

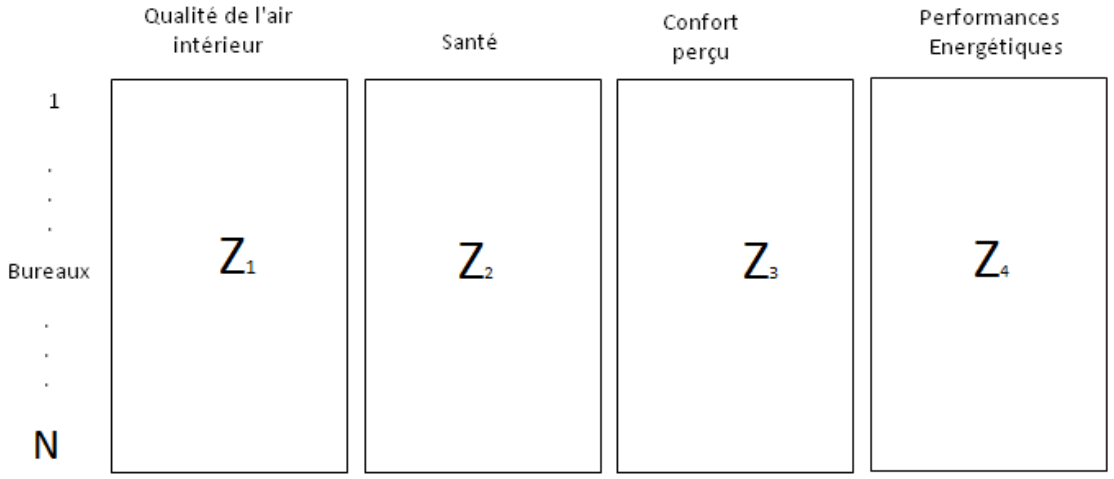


FIGURE 1.2 – Exemple de structuration en blocs

Grande dimension des données

Pour Agrawal *et al.* [1998] les données de grande dimension sont associées à des tables ayant des centaines d'attributs et sont souvent d'une grande variabilité. Or, selon Berchtold *et al.* [1997], la recherche de groupes dans un échantillon de grande dimension n'est pas toujours significative, car la densité moyenne des points, quelle que soit la région de l'espace considérée, peut être faible. De plus, assez souvent dans les données de grande dimension, on rencontre des variables ayant une distribution uniforme et considérées comme du bruit. Par conséquent, la notion de distance définie sur toute la dimension des données devient inefficace pour la classification. Les classes peuvent donc être portées par différents

sous espaces composés de combinaisons d'attributs de l'espace initial des données. Dans la littérature statistique, 4 grandes familles d'approches sont dédiées à la classification des données de grande dimension : les méthodes de réduction de dimension, les méthodes de sélection globale et locale des variables, les méthodes de type subpace clustering et les méthodes de recherche de consensus, etc. Ces méthodes sont présentées dans le chapitre 3.

1.6 Conclusion

La base de 300 immeubles de bureaux et les informations collectées constituent une base d'informations riche pour les thématiques qualité de l'air intérieur, santé et confort perçu par les occupants et les performances énergétiques des immeubles de bureaux. L'objectif final de la CNB étant la définition d'un indicateur global de classement, il est alors indispensable de définir une typologie globale du parc des bâtiments à usage de bureaux. La recherche de ces typologies induit plusieurs problèmes statistiques majeurs. Le principal problème concerne la détermination de groupes d'observations et d'attributs initiaux résumant l'information à l'intérieur de chaque thématique.

Face à cette problématique de classification qui tient compte de plusieurs critères (QAI, santé et confort perçu et performances énergétiques) objet de cette thèse, il existe de nombreuses méthodes basées sur de solides fondements théoriques qui permettent d'établir les classes a priori recherchées. Historiquement, plusieurs méthodes ont été proposées par différents auteurs pour étudier les données ayant une structure mixte [Escofier 1979; Saporta 2006]. De même, de nombreuses approches ont été proposées pour prendre en compte les aspects grande dimension et structuration multi-blocs des données en classification [Kriegel *et al.* 2009; Agrawal *et al.* 1998; Vega-Pons et Ruiz-Shucloper 2011].

Les méthodes d'aide aux décisions multicritères qui sont spécifiques à la nature et aux objectifs du problème rencontré dans cette thèse proposent des solutions pouvant prendre diverses formes en fonction de la problématique (de choix, d'affectation ou de classement) [Maystre *et al.* 1994; Ben Mena 2000]. Il s'agit généralement de trouver la solution la plus adaptée parmi un ensemble de solutions. Par ailleurs, ces méthodes nécessitent aussi une forte interaction entre les différentes thématiques. Rapporté au cas de la CNB, il est indispensable de connaître les avantages et inconvénients de chacun des paramètres (variables). De plus ce type de méthode est initialement développé pour des échantillons de faible dimension, d'où la nécessité de procéder à un élagage des variables dans chaque thématique à travers la sélection des variables et la classification des individus.

Compte tenu des données en cours d'acquisition de la campagne dans les immeubles de bureaux, les méthodes développées ont été testées et consolidées à partir d'un autre jeu de données de l'OQAI, à savoir celui de la campagne nationale dans les «Logements» (CNL). Le but de la CNL menée entre 2003-2005 était dans un premier temps d'élaborer un état descriptif de la qualité de l'air dans les logements en tenant compte des différentes situations (bâtiments, occupant) et établir un premier bilan des paramètres déterminant la pollution intérieure (source type d'habitat, ventilation, comportements, saisons, situation géographiques, etc), puis d'identifier les situations à risque, en estimant l'exposition des populations concernées et élaborer des recommandations et conseils pour l'amélioration de la qualité de l'air intérieur dans les logements (limitation des émissions des produits, réglementation technique, sensibilisation des professionnels ou des usagers, etc.).

Le chapitre 2, sans être exhaustif, pose les bases de la classification permettant de répondre aux interrogations suivantes :

- Qu'est ce qu'une classe ?
- Quels sont les attributs à utiliser ?
- Les observations contiennent-elles des objets atypiques ?
- Les variables doivent-elles être normalisées ?
- Quelles mesures de dissimilarité utiliser entre deux objets ?
- Comment optimiser la prise en compte des différentes thématiques ?
- Quelles méthodes de classification doit-on utiliser sur ces données ?
- Les données contiennent-elles des groupes homogènes ?
- Quel est le nombre exact de classes ?
- Les groupes découverts sont-ils valides ?

Chapitre 2

Classification

2.1 Introduction

Le but de la classification est de découvrir des groupes d'observations dans un ensemble de données non-étiquetées. Les groupes recherchés, communément appelés classes, forment des ensembles homogènes d'observations qui partagent des propriétés communes à travers des variables ou attributs. Les techniques de classification font partie de la statistique exploratoire multidimensionnelle. L'objectif est de regrouper les lignes ou les colonnes d'un tableau afin de découvrir et d'expliciter une structuration des données, il s'agit d'un problème de typologie ou de taxinomie ("clustering" chez les anglo-saxons), de classification non supervisée ou encore d'apprentissage sans professeur. Les classes inconnues à l'avance sont déterminées par la méthode de classification. Elles sont à distinguer des méthodes de classement dont l'objectif est de classer au mieux de nouvelles observations dans des classes connues ou déterminées a priori. Il s'agit alors d'un problème de discrimination, de classement, de classification supervisée (les anglo-saxons parlent simplement de "classification") ou d'apprentissage avec professeur.

On distingue deux grandes familles de méthodes de classification : celles dites "model-based" ou modèles de mélanges qui reposent sur des hypothèses probabilistes sur les distributions des observations [Banfield et Raftery 1993] et celles appelés "distance-based" qui utilisent des notions géométriques de similarité [Berkhin 2004]. Dans ce mémoire, on se limitera à ces dernières.

On peut aussi distinguer deux grandes familles de techniques de classification (Cf. figure 2.1) : les méthodes hiérarchiques qui produisent des classes de moins en moins fines

par regroupement des observations (méthodes ascendantes) ou des classes de plus en plus fines par division (méthodes descendantes) et les méthodes non-hiérarchiques ou de partitionnement direct qui produisent directement un regroupement de l'ensemble des éléments en un nombre K de classes fixé a priori. Cette famille comprend les méthodes des cartes auto-organisées. Nous choisissons de les décrire plus en détail dans la section 2.3.2.2 car elles sont à la base des méthodes que nous proposons dans cette thèse.

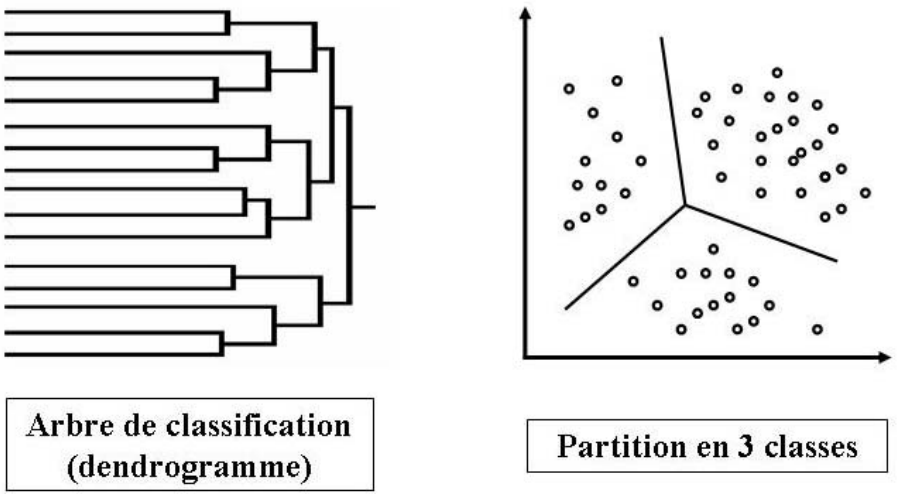


FIGURE 2.1 – Le dendrogramme d'une classification hiérarchique ascendante
textc

Pour atteindre l'objectif de regroupement des observations en classes homogènes, on fait appel à la notion de similarité qui permet d'évaluer la proximité entre deux observations. Dans la littérature, il existe de nombreux algorithmes basés sur des formalismes différents, sur plusieurs mesures de similarité et sur diverses stratégies d'agrégation des observations pouvant alors donner sur le même jeu de données, des classifications différentes [Berkhin 2004; Jain *et al.* 1999a]. La figure 2.2 illustre cette diversité des résultats dans le cas d'une classification ascendante hiérarchique.
Après quelques rappels sur les concepts généraux, les principales approches usuelles de classification sont présentées dans la section 2.3 ; plus particulièrement la méthode des cartes topologiques auto-organisées sur laquelle reposent les principales contributions de cette thèse. Enfin, les critères permettant d'évaluer la pertinence des résultats d'une classification sont présentés dans la section 2.4.

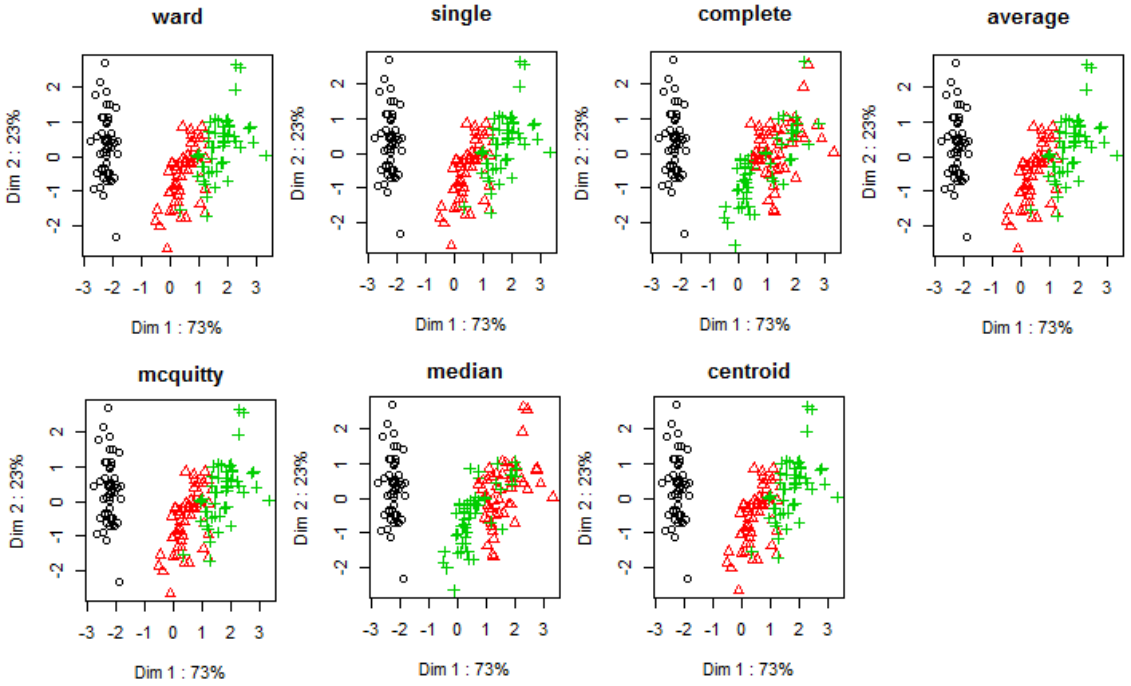


FIGURE 2.2 – Projection des classes fournies par l'algorithme de classification ascendante hiérarchique avec différents critères sur la base Iris de Fisher dans le premier plan factoriel d'une ACP. Les trois classes de cette base sont représentées par les signes $o, +$ et Δ

.

2.2 Rappels de notions générales sur la classification

L'objectif de la classification est de construire une typologie pour caractériser un ensemble de N observations décrites par p variables. Pour atteindre cet objectif, on émet l'hypothèse que les observations collectées ne sont pas toutes issues de la même population homogène, mais plutôt de K sous populations. Cette section définit les concepts nécessaires à la compréhension et à la définition d'une classification sur un ensemble de données.

On note $\mathcal{Z}$ l'ensemble des observations, $z_i = (z_i^1, \ldots, z_i^j, \ldots, z_i^p)$ où $i = 1, \ldots, N$. z_i^j désigne la valeur prise par l'observation z_i pour la variable z^j. L'ensemble des variables est noté $\mathcal{V} = \{z^1, \ldots, z^p\}$. La matrice des données de dimension $(N \times p)$ est notée Z. Dans toute la suite de ce mémoire de thèse, les termes objet et individu seront utilisés de manière équivalente pour désigner une observation.

2.2.1 Notations et définitions

Classe

Une classe c est un sous-ensemble de l'ensemble $\mathcal{Z}$ des observations à classer. En classification non-supervisée, la description des classes se fait soit par énumération de la liste des objets appartenant à la classe (description par extension), soit par l'énumération des propriétés caractéristiques des classes (description par intention). Dans les faits, les méthodes de classification fournissent généralement des descriptions par extension alors que l'utilisateur final est intéressé par la description en intention. Par rapport à l'ensemble $\mathcal{Z}$, la notion de classe induit la notion de partition de l'ensemble $\mathcal{Z}$.

Partition

Définition 1 (Partition) *Soit $\mathcal{Z}$ un ensemble d'observations, une partition P de $\mathcal{Z}$ en K classes est définie par $P = \{c_1, \ldots, c_k, \ldots, c_K\}$ avec $c_k \in \mathcal{P}$ (ensemble des parties de $\mathcal{Z}$) vérifiant :*
$(P_1) : c_k \neq \emptyset$ *pour* $k = 1, \ldots, K$
$(P_2) : \cup_{k=1}^{K} c_k = \mathcal{Z}$
$(P_3) : c_l \cap c_k = \emptyset \ \forall \ l \neq k.$

Les contraintes (P_2) et (P_3) peuvent souvent être relaxées. Si on relaxe la contrainte (P_3) que tout objet doit appartenir à une et une seule classe on obtient des classes empiétantes. La relaxation de (P_3) se rapproche alors plus de la réalité car il n'est pas rare d'obtenir des partitions pour lesquelles un objet peut appartenir à une ou plusieurs classes ; par exemple un ouvrage de statistique médicale peut être classé soit au rayon médical soit au rayon statistique.

Hiérarchie

Une hiérarchie est un ensemble de partitions emboitées. On la représente en général par un arbre hiérarchique appelé dendrogramme où les objets sont à la base et l'ensemble tout entier au sommet. La définition formelle d'une hiérarchie est la suivante :

Définition 2 (Hiérarchie) *$(\mathcal{H}, g)$ est une hiérarchie indicée si $\mathcal{H} \subseteq \mathcal{P}$ (ensemble des parties de $\mathcal{Z}$) et si g est une fonction à valeurs dans $\mathbb{R}^+$, telle que :*
$(h_0) : \emptyset \in \mathcal{H}$,
$(h_1) : \mathcal{Z} \in \mathcal{H}$,

$(h_2) : \forall z_i \in \mathcal{Z}, \{z_i\} \in \mathcal{H},$
$(h_3) : \forall h, h' \in \mathcal{H}, h \cap h' \in \{\emptyset, h, h'\},$
$(h_4) : h \subset h' \Longrightarrow g(h) < g(h')$
$(h_5) : g(h) = 0 \Longleftrightarrow \exists z_i \in \mathcal{Z}$ *tel que* $h = \{z_i\}.$

Les axiomes (h_0-h_3) définissent la hiérarchie et les axiomes h_4 et h_5 définissent respectivement la cohérence entre l'indice de hauteur g et la relation d'inclusion entre les classes. Les hiérarchies se lisent souvent plus simplement comme des arbres dont les nœuds terminaux et intermédiaires sont des classes.
Le regroupement des observations en classes dans les structures partition et hiérarchie se fait, en l'absence d'hypothèses concernant la distribution des données, sur des considérations géométriques. Les observations proches les unes des autres sont alors regroupées ensemble. La notion de regroupement induit la nécessité de définir une mesure de proximité entre observations.

Définition 3 (Mesure de dissimilarité) *On appelle mesure ou indice de dissimilarité sur un ensemble $\mathcal{Z}$, une application $d : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}^+$ tel que :*
$(a_0) : \forall z_i, z_j \in \mathcal{Z}, d(z_i, z_j) = 0 \Longleftrightarrow z_i = z_j,$
$(a_1) : \forall z_i, z_j \in \mathcal{Z}, d(z_i, z_j) = d(z_j, z_i),$

Définition 4 (Distance) *Une distance est une application $d : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}^+$ tel que :*
$(a_0) : \forall z_i, z_j \in \mathcal{Z}, d(z_i, z_j) = 0 \Longleftrightarrow z_i = z_j,$
$(a_1) : \forall z_i, z_j \in \mathcal{Z}, d(z_i, z_j) = d(z_j, z_i),$
$(a_2) : \forall z_i, z_j, z_k \in \mathcal{Z}, d(z_i, z_j) \leq d(z_i, z_k) + d(z_k, z_j),$

Définition 5 (Distance ultramétrique) *Une distance ultramétrique est une application $d : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}^+$ tel que :*
$(a_0) : \forall z_i, z_j \in \mathcal{Z}, d(z_i, z_j) = 0 \Longleftrightarrow z_i = z_j,$
$(a_1) : \forall z_i, z_j \in \mathcal{Z}, d(z_i, z_j) = d(z_j, z_i),$
$(a_3) : \forall z_i, z_j, z_k \in \mathcal{Z}, d(z_i, z_j) \leq max(d(z_i, z_k), d(z_k, z_j)),$

La notion de distance ultra-métrique constitue une base de la classification hiérarchique que nous présenterons dans la section 2.3.1.

En posant :

$$s(z_i, z_j) = \underset{z_i, z_j \in \mathcal{Z}}{max}(d(z_i, z_j)) - d(z_i, z_j)$$

on définit une mesure de similarité associée à la distance d. Remarquons que s est symétrique et vérifie $s(z_i, z_i) \geq 0 \ \forall \ z_i, \neq z_j$.

2.2.2 Inertie intraclasse et interclasse

Soit P une partition en K groupes c_k de l'ensemble $\mathcal{Z}$, on définira les quantités suivantes : w est le centre de gravité du nuage de points de l'ensemble $\mathcal{Z}$, $w_1, \ldots, w_K$ les centres de gravité des K sous nuages de points associés aux classes c_k, et $\mu_k = \frac{n_k}{N}$ est le poids de la classe c_k. L'inertie d'un nuage de points désigne la moyenne des carrés des distances au centre de gravité. L'inertie totale de l'ensemble $\mathcal{Z}$ est définie par :

$$\begin{aligned} I_T &= \frac{1}{N} \sum_{k=1}^{K} \sum_{z_i \in c_k} ||z_i - w||^2 \\ &= \underbrace{\sum_{k=1}^{K} \frac{1}{N} \sum_{z_i \in c_k} ||z_i - w_k||^2}_{Inertie-intraclasse} + \underbrace{\sum_{k=1}^{K} \mu_k ||w_k - w||^2}_{Inertie-interclasse} \end{aligned}$$

L'inertie intra-classe prend de faibles valeurs si les classes sont homogènes. L'inertie interclasse mesure à quel point les centres de classe w_k sont loin de w. Ainsi, lorsqu'il existe une structure en K classes dans l'ensemble $\mathcal{Z}$, un critère usuel de classification consiste à chercher la partition telle que l'inertie intra-classe soit minimale en optimisant :

$$P^* = \underset{P \in \mathcal{P}_K}{argmin} \left(\underbrace{\sum_{k=1}^{K} \frac{1}{N} \sum_{z_i \in c_k} ||z_i - w_k||^2}_{Inertie\ Intra} \right) \tag{2.1}$$

où $\mathcal{P}_k$ est l'ensemble des partitions en K classes possible de $\mathcal{Z}$. Remarquons que ce critère ne s'applique qu'à un nombre de classes fixé.

La connaissance de toutes les partitions possible de l'ensemble $\mathcal{Z}$ permet la détermination de la partition optimale au sens de la minimisation de l'inertie intra-classes. Cependant, dans la pratique sa résolution est impossible à cause de la cardinalité élevée de l'ensemble $\mathcal{P}_k$. En effet, pour $N = 19$ et $k = 4$ le nombre de partitions de l'ensemble $\mathcal{P}_k$ est supérieur

10^{10}. De nombreuses heuristiques proposent une solution approchée de la partition optimale en visitant un faible nombre de partitions.

2.2.3 Mesures de similarité pour données quantitatives et qualitatives

2.2.3.1 Données quantitatives

La première étape dans la grande majorité des algorithmes de classification est l'évaluation de la similarité entre les observations. Il s'agit de définir une matrice de taille $N \times N$ dont les entrées représentent la similarité entre deux observations. Le choix de la distance est primordial pour les méthodes de classification et doit être réalisé en tenant compte du type de données (numérique ou catégorielle). Le tableau (2.1) ci-dessous présente les principales distances usuelles utilisées dans la littérature pour les données numériques. Pour plus de détails sur les distances on pourra se référer aux revues présentées par Cha [2007] et Choi *et al.* [2010].

Euclidienne L_2	$d_{eucl} = \sqrt{\sum_j^p (z_i^j - z_{i'}^j)^2}$
City block L_1	$d_{cb} = \sum_j^p \lvert z_i^j - z_{i'}^j \rvert$
Minkowski L_q	$d_{mk} = \sqrt[p]{\sum_j^p (z_i^j - z_{i'}^j)^q}$
Chebychev L_∞	$d_{Cheb} = \max \lvert z_i^j - z_{i'}^j \rvert$

TABLE 2.1 – Quelques distances usuelles pour des données numériques

2.2.3.2 Données qualitatives

De nombreux exemples montrent que la distance euclidienne n'est pas adaptée pour les données catégorielles et binaires. Ainsi, comme dans le cas des données quantitatives, il existe plusieurs notions de dissimilarité spécifiques aux données catégorielles (Cf. tableau 2.2). Une connaissance a priori des données pourrait guider le choix de l'une ou l'autre des mesures. Chacune de ces mesures de similarité possède ses propres propriétés qui influencent les résultats de la classification. Lorsque les objets sont décrits par un ensemble de variables qualitatives, il est d'usage de représenter les p variables qualitatives initiales contenant chacune $m_1, \ldots, m_p$ modalités par un tableau d'indicatrices des modalités appelé tableau disjonctif complet noté X. On est alors en présence de données binaires x_i^j ($x_i^j \in \{0, 1\}$). Les indices de similarité présentés dans le tableau 2.2 sont ensuite proposés en combinant

de diverses manières les quatre nombres suivants associés à un couple i, i' d'individus :

- N_{11}, le nombre de caractéristiques communes (accords positifs) ;
- N_{10}, le nombre de caractéristiques possédées par i et pas par i' ;
- N_{01}, le nombre de caractéristiques possédées par i' et pas par i ;
- N_{00}, le nombre de caractéristiques que ne possèdent ni i, ni i' ;

Concordances simple (Sokal, Michener)	$\frac{N_{11}+N_{00}}{N_{11}+N_{01}+N_{10}}$
Jaccard	$\frac{N_{11}}{N_{11}+N_{01}+N_{10}+N_{00}}$
Russel-Rao	$\frac{N_{11}}{N_{11}+N_{01}+N_{10}}$
Ochiai	$\frac{N_{11}}{\sqrt{(N_{11}+N_{01})(N_{10}+N_{00})}}$
Ochiai II	$\frac{N_{11}\times N_{00}}{\sqrt{(N_{11}+N_{00})(N_{11}+N_{10})(N_{00}+N_{01})(N_{00}+N_{10})}}$
Dice	$\frac{N_{11}}{2N_{11}+N_{01}+N_{10}}$
Rogers-Tanimoto	$\frac{N_{11}+N_{00}}{N_{11}+2(N_{01}+N_{10})+N_{00}}$
Kulzinsky	$\frac{N_{11}}{N_{01}+N_{10}}$

TABLE 2.2 – Quelques distances usuelles pour des données catégorielles

Citons aussi la distance de Hamming

$$d_h(x_i^j, x_{i'}^j) = \sum_{j=1}^{p} \sigma(x_i^j, x_{i'}^j) \tag{2.2}$$

où

$$\sigma(x_i^j, x_{i'}^j) = \begin{cases} 0 \; si \; (x_i^j = x_{i'}^j) \\ 1 \; si \; (x_i^j \neq x_{i'}^j) \end{cases}$$

Il est aussi possible d'utiliser pour deux individus i et i' la distance du χ^2 :

$$d_{\chi^2}(i, i') = \sum_j \frac{N}{n_{.j}} (\frac{x_i^j - x_{i'}^j}{p})^2 \tag{2.3}$$

où $n_{.j}$ est le nombre d'individus ayant la modalité j. C'est par exemple le cas en analyse des correspondances multiples (ACM).

La distance d_{χ^2} montre que la similarité dépend du nombre de modalités possédées en commun par i et i' et de leur fréquence, ce qui revient à dire que deux individus ayant en commun une modalité rare sont plus proches que deux individus ayant en commun une modalité fréquente.

Une méthode classique pour traiter les données qualitatives consiste à effectuer une transformation des variables qualitatives via une analyse des correspondances multiples en variables quantitatives au préalable. Ce qui revient à coder tous les individus par les coordonnées factorielles résultant de cette transformation. Une fois les individus représentés

par des variables numériques, la classification est obtenue en utilisant un algorithme dédié aux variables quantitatives. Cependant, on perd la représentation initiale des données (les modalités). Cela engendre donc des difficultés d'interprétation.

2.2.3.3 Mesure de similarité pour données mixtes

Dans le cas où les individus sont décrits par des variables quantitatives et des variables qualitatives Huang [1998] et Lebbah *et al.* [2005] montrent que les formalismes habituels des problèmes de classification restent valables sous condition d'utilisation d'une distance adaptée. Une méthode classique consiste à combiner la distance euclidienne à celle de Hamming :

$$d_{mix} = d_{eucl} + d_h \tag{2.4}$$

Cependant dans cette combinaison, il est important de prendre en considération l'influence de la partie quantitative des données par rapport à la partie qualitative et inversement. Ainsi, certains auteurs ont proposé d'introduire un paramètre de pondération de cette influence. Cela sera discuté plus en détail dans le chapitre 4.
Nous avons présenté dans cette section quelques notations et définitions liées à la classification. La section suivante présente quelques méthodes usuelles de classification qui serviront de base aux approches proposées dans cette thèse.

2.3 Les méthodes classiques de classification

Les méthodes hiérarchiques fournissent, un arbre appelé dendrogramme, mettant en évidence une succession de partitions emboîtées de l'ensemble des observations alors que la famille des méthodes non-hiérarchiques ou de partitionnement produisent une partition directe de l'ensemble des éléments en un nombre K de classes fixé a priori. Cette dernière famille comprend les méthodes des cartes auto-organisées. Nous choisissons de les décrire plus en détail dans la section 2.3.2.2 car ils sont à la base des méthodes que nous proposons dans cette thèse.

2.3.1 Les méthodes de classification hiérarchique

Deux grands types de méthodes hiérarchiques existent : une descendante, dite divisive, et une ascendante, dite agglomérative. La première, moins utilisée, consiste à partir de la classe grossière contenant tous les objets, à partager celle-ci en deux puis, cette opération

est répétée à chaque itération jusqu'à ce que toutes les classes soient réduites à des singletons. La classification ascendante hiérarchique qui est la plus couramment utilisée, consiste à construire une succession de partitions emboîtées par regroupement successifs des observations en classes de moins en moins fines, jusqu'à l'obtention d'une seule classe contenant tous les objets. En considérant l'ensemble $\mathcal{Z}$ des observations z_i à classer, la CAH repose sur la définition d'une $d(z_i, z_{i'})$ distance entre les observations z_i et $z_{i'}$ et sur le choix d'une stratégie d'agrégation $d_C(c_l, c_k)$ déterminant la distance entre les classes c_l et c_k. Selon le mode de calcul de la distance d_C entre les classes, on retrouve dans la littérature différentes méthodes de classification hiérarchique. Ces méthodes sont résumées à l'aide de la formule de Lance et Williams [Gordon 1987] :

$$d_C(c_i, c_k) = \alpha_p d_C(c_l, c_i) + \alpha_q d_C(c_q, c_i) + \beta_p d_C(c_l, c_q) + \gamma \mid d_C(c_l, c_i) - d_C(c_q, c_i) \mid \quad (2.5)$$

où en fonction de la stratégie d'agrégation de deux classes, on définit les coefficients α, β et γ dans le tableau 2.3.

	α_p	α_q	β	γ
Lien simple	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Lien complet	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Moyenne	$\frac{n_p}{n_k}$	$\frac{n_q}{n_k}$	0	0
Centroide	$\frac{n_p}{n_k}$	$\frac{n_q}{n_k}$	$\frac{n_p n_q}{n_k^2}$	0
Médiane	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	0
Ward	$\frac{n_p+n_i}{n_k+n_i}$	$\frac{n_q+n_i}{n_k+n_i}$	$\frac{-n_i}{n_k+n_i}$	0

TABLE 2.3 – Les coefficients pour chaque stratégie d'agrégation des classes

L'algorithme de la classification ascendante hiérarchique se déroule comme suit :

Algorithme : Classification Ascendante Hiérarchique

Entrée : $\mathcal{Z}$

1. Initialiser les N classes c_k formées chacune d'une observation : $c_i = \{z_i\}$ et poser $d_C(c_i, c_{i'}) = d(z_i, z_{i'})$;
2. Fusionner les deux classes c_l et c_q les plus proches pour former une nouvelle classe $c_k = c_l \cup c_q$ tels que $d_C(c_l, c_q) = \min_{i,i'}(d_C(c_i, c_{i'}))$;
3. Calculer la distance entre la nouvelle classe c_k et les autres : $d_C(c_k, c_i)$ pour $i \neq l, q$;
4. Itérer : répéter n-1 fois les étapes 2 et 3 jusqu'à l'obtention d'une seule classe regroupant tous les objets.

Sortie : Un dendrogramme représentant les étapes de fusion des classes

La complexité de cet algorithme est quadratique. Ce qui contraint son application aux tableaux de taille raisonnable. Dans le contexte de la classification des données de grande dimension cet algorithme est appliqué sur une synthèse des données initiales qui peut être obtenue à travers une méthode de partitionnement directe telle que celles décrites ci-dessous.

2.3.2 Les méthodes de partitionnement direct

2.3.2.1 Méthode des K-moyennes

Le plus connu des algorithmes de partitionnement direct est celui des centres mobiles. Il consiste à chercher directement une partition des éléments en un certain nombre K de classes fixé a priori en minimisant le critère d'inertie intra-classe suivant :

$$J(\mathcal{C}) = \frac{1}{N} \sum_{k=1}^{K} \sum_{z_i \in c_k} \|z_i - w_k\|^2 = \frac{1}{N} \sum_{k=1}^{K} \frac{1}{n_k} J_{c_k} \tag{2.6}$$

où n_k est le nombre d'observations de la classe c_k, w_k désigne le vecteur moyen des observations de c_k et $J_{c_k} = \sum_{z_i \in c_k} \|z_i - w_k\|^2$ désigne l'erreur quadratique entre le vecteur w_k et les observations de c_k.

La minimisation de 2.6 se déroule en trois phases.

Algorithme : K-moyennes

Entrée : l'ensemble $\mathcal{Z}$ des observations, le nombre K de classes fixé a priori, le nombre T d'itérations

1. Initialiser les K centres de classe : on choisit arbitrairement K centres de classe ;
2. Itération de base (t>1)
   Étant donnée les centres de classe obtenus à l'étape (t-1) $\{w_k^{(t-1)}, k = 1, \ldots, K\}$, définir la partition $\pi^t = \{c_1^t, \ldots, c_K^t\}$ en affectant chaque observation z_i à la classe c_k^t dont il est le plus proche suivant le critère suivant : $z_i \in c_k^t$ si $d(z_i, w_k^t) = \min_{k=1,\ldots,K} d(z_i, w_k^{(t-1)})$
3. Calculer les nouveaux centres de classe :
   Pour chaque classe $c_k^{(t)}$ formée, associer un nouveau centre de classe w_k^t comme étant le vecteur moyen des éléments lui appartenant.

Répéter les étapes 2 et 3 jusqu'à la stabilisation des centres classes où jusqu'a atteindre le nombre T d'itérations fixé
Sortie : une partition π des observations en K classes

Ceci est la version de l'algorithme des K-moyennes proposée par Forgy [1965]. MacQueen [1967] propose une autre version dans laquelle le calcul des nouveaux centres est effectué après chaque réaffectation d'un individu. Cet algorithme est adapté au tableau de grandes tailles, sa complexité étant linéaire.

Une méthode de classification plus générale appelée nuées dynamiques a été proposée par [Diday 1971] qui contrairement à la méthode des K-moyennes permet une représentation plus générale des classes à travers un noyau formé de q points (plutôt que le seul centre de gravité), une loi de probabilité, un axe factoriel.
Dans la méthode des K-moyennes, chaque classe est représentée par son point moyen, ce qui a un sens géométrique et statistique pour des données numériques et non pour des données catégorielles (qualitatives). La méthode des K-moyennes a donc été étendue, à travers l'algorithme K-modes, aux données catégorielles. L'idée est la même que celle de K-moyennes et la structure géométrique ne change pas. La différence avec la méthode des K-moyennes réside dans le choix de la distance utilisée pour évaluer la proximité entre deux objets. La distance de Hamming définie par la relation 2.2 est couramment utilisée dans ce cas comme mesure de dissimilarité entre objet [Huang et Ng 1999]. Les centres de classes sont alors définis par le mode [1] de chaque classe.

1. Le mode d'une classe est la valeur la plus fréquente dans la classe

Pour les données mixtes, l'idée de l'extension des K-moyennes consiste à utiliser conjointement les algorithmes K-modes et K-moyennes sur chaque portion des données et d'utiliser la version suivante de la relation 2.4 :

$$d_{mix} = d_{eucl} + \gamma d_h \tag{2.7}$$

où d_{eucl} et d_h sont les distances utilisées respectivement pour les variables quantitatives et les variables qualitatives. γ est un paramètre de calibration de l'influence de la partie quantitative des données par rapport à la partie qualitative [Huang 1998].

La figure 2.3 montre un déroulement de l'algorithme des k-moyennes.

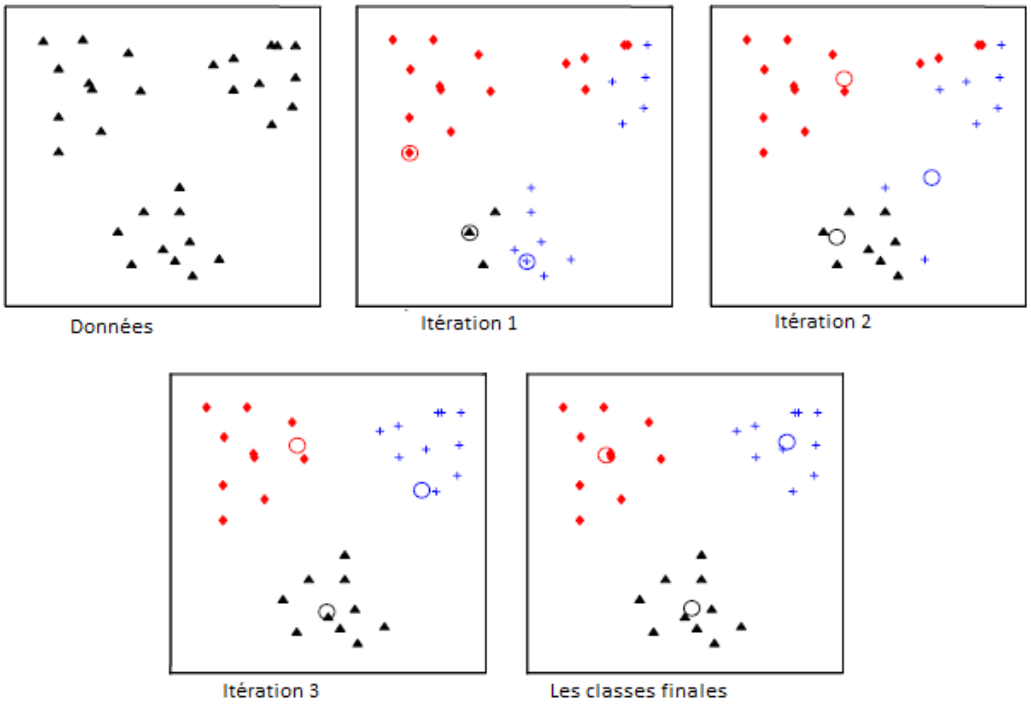


FIGURE 2.3 – Illustration de l'algorithme des K-moyennes sur un jeu de données défini dans $\mathbb{R}^2$ contenant 3 classes. L'étape Itération 1 correspond à l'étape d'initialisation, les étapes itérations 2 et 3 définissent l'évolution des centres de classe et l'étape des classes finales présente le résultat de l'algorithme après stabilisation des centres de classe [Jain 2010]

Malgré sa popularité, l'algorithme des K-moyennes présente plusieurs inconvénients majeurs en particulier la nécessité de connaître le nombre K a priori ; et par ailleurs, elle tend à trouver des classes à structure sphérique de taille relativement identiques. Ainsi, en présence d'une classe de très petite taille, ou d'une classe prédominante, elle aura tendance

à "vider " la plus grosse classe au profit de la petite. La partition finale ne reflétera donc pas correctement la structure des données en classes. De plus, cette catégorie d'algorithme de classification dépend fortement des paramètres initiaux et les résultats obtenus peuvent fortement différer en convergeant vers des minima locaux. Dans ce cas, pour y remédier, l'utilisateur peut effectuer différents choix initiaux arbitraires ou aléatoires et comparer les différents résultats obtenus afin de choisir celle fournissant l'optimum du critère. Cependant, Meilă [2006] montre que les K-moyennes convergent souvent vers des minima globaux si les classes des données sont bien séparées. Enfin, cet algorithme est inefficace en présence des données bruitées ou d'outliers.

Ce dernier problème est surmonté dans la méthode des K-médoïdes qui n'est pas limitée par le type de variable (catégorielle, quantitative ou mixte). Il s'agit d'un algorithme d'optimisation itérative combinant la réaffectation des objets aux classes avec une intervention des centres de classe et des autres points. Les médoïdes (centres de classe) sont les objets les plus appropriés pour représenter les classes dont le choix est dicté par la localisation d'un nombre prédominant de points à l'intérieur d'une classe [Kaufman et Rousseeuw 1990].

On montrera dans la section 2.3.2.2 suivante que la méthode des cartes topologiques qui est une version régularisée des K-moyennes est plus adaptée à la détection des classes à structure non-sphérique et à la présence d'outliers.
Les méthodes hiérarchiques dont la lecture de l'arbre hiérarchique qu'elles fournissent permet de déterminer le nombre optimal de classes, ont l'inconvénient majeur d'être coûteux en temps de calcul. À l'opposé les méthodes non-hiérarchiques ont l'avantage d'être adaptées aux données volumineux mais imposent la connaissance a priori du nombre exact de classes qui n'est en réalité pas connu en apprentissage non-supervisée. En présence d'un grand nombre d'observations $N > 10^3$ dans les jeux à traités, les limitations des approches hiérarchiques et des approches non-hiérarchiques sont compensées par la combinaison de ces deux familles de méthodes. Il s'agit pour pallier au coût élevé en temps de calcul des méthodes hiérarchiques d'appliquer une méthode non-hiérarchique de type K-moyennes ou les cartes SOM avec un nombre K de classes ou de cellules suffisamment élevé (K=50 par exemple), puis d'appliquer une CAH pour déterminer judicieusement le nombre de classes. Il est aussi usuel d'ajouter une troisième étape dite de consolidation dans laquelle la méthode des K-moyennes est appliquée sur les classes obtenues après coupure de l'arbre hiérarchique. C'est ce qu'on appelle la classification mixte [Saporta 2006].

2.3.2.2 La méthode des cartes topologiques

Les cartes topologiques ou auto-organisatrices qui appartiennent à la famille des méthodes neuronales, ont été introduites par Kohonen [1998]. Le but de ces cartes est de représenter des observations multidimensionnelles ($\mathcal{Z} \subset \mathbb{R}^p$) sur un espace discret de faible dimension (en général 1D ou 2D) qui est communément appelé la carte topologique. Dans ces méthodes, chaque classe est représentée par un neurone qui est caractérisé par un vecteur référent. La présentation suivante permet de comprendre les méthodes des cartes topologiques et de les positionner comme une extension de l'algorithme des K-moyennes.

Données, Référents et Cartes Topologiques

Les cartes topologiques font partie de la famille des méthodes de quantification vectorielle, au même titre que la méthode des K-moyennes, qui cherche une partition $\mathcal{C}$ de l'ensemble $\mathcal{Z}$ en K sous ensembles. Chaque sous-ensemble, noté $c \in \mathcal{C}$, est associé à un vecteur dit référent ou représentant w_c défini dans le même espace que les données de l'ensemble $\mathcal{Z}$. Soit $\mathcal{W} = \{w_c; c = 1, \ldots, K\}$ l'ensemble des vecteurs référents. Dans le cas des méthodes de quantification vectorielle, la partition $\mathcal{C}$ est souvent définie par une fonction d'affectation $\mathcal{X}$ permettant de définir les sous-ensembles c de la partition $\mathcal{C}$ tel que $c = \{z_i \in \mathcal{Z}/\mathcal{X}(z_i) = c\}$. L'ensemble $\mathcal{C}$ est constitué d'un ensemble de neurones interconnectés et le lien entre les neurones se fait par l'intermédiaire d'une structure de graphe non-orienté (cf. Figure 2.4). La structure de graphe sous-jacente à la carte $\mathcal{C}$ est induite par une distance souvent discrète σ sur $\mathcal{C}$ définie comme étant la longueur du plus court chemin. Pour chaque neurone $c \in \mathcal{C}$, la distance σ permet de définir la notion de voisinage d'ordre d de c :

$$V_c(d) = \{r \in \mathcal{C}, \sigma(c, r) \leq d\}$$

Plus précisément, le lien entre deux neurones r et c de la carte $\mathcal{C}$ est introduit par une fonction noyau $\mathcal{K}$ positive et symétrique telle que $\lim_{|x| \to \infty} \mathcal{K}(x) = 0$. Cette fonction noyau définit une zone d'influence autour de chaque neurone c de la carte : $\{r; \mathcal{K}(\sigma(c, r)) < \alpha\}$ où α est le seuil d'activation d'un neurone comme faisant partir du voisinage de c.

Dans la littérature, il existe plusieurs manières de définir la fonction $\mathcal{K}$. Nous noterons dans toute la suite de ce mémoire $\mathcal{K}^T = \mathcal{K}(\frac{\sigma}{T})$, avec T désignant un paramètre communément appelé température du modèle, la famille de définition générée par le choix de $\mathcal{K}$. Nous présentons ci-après les fonctions noyaux utilisées dans la littérature :

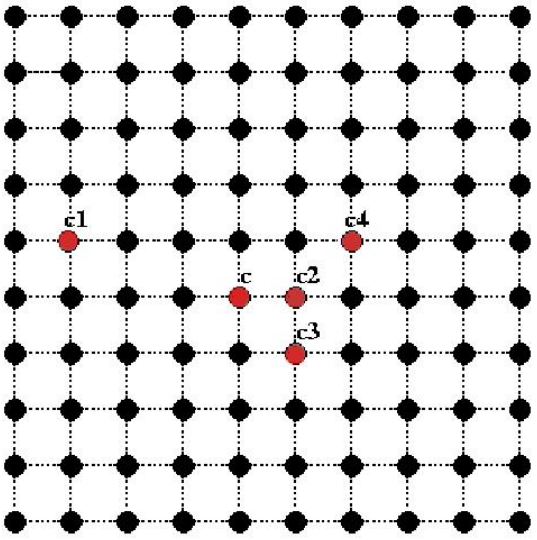


FIGURE 2.4 – Carte topologique, quelques distances entre les neurones : $\sigma(c, c_1) = 4$, $\sigma(c, c_2) = 1$, $\sigma(c, c_3) = 2$, $\sigma(c, c_4) = 3$.

- la fonction indicatrice :
  $\mathcal{K}(\sigma) = 1$ si $\sigma < 1$ et 0 sinon, ainsi $\mathcal{K}^T(\sigma) = 1$ si $\sigma < T$ et 0 sinon
- la fonction exponentielle :
  $\mathcal{K}(\sigma) = \exp(-|\sigma|)$ d'où $\mathcal{K}^T(\sigma) = \exp(-\frac{|\sigma|}{T})$
- la fonction "gaussienne" ; l'influence entre deux neurones dépend de la distance entre ces neurones
  $\mathcal{K}(\sigma) = \exp(-\sigma^2)$ d'où $\mathcal{K}^T(\sigma) = \exp(-\frac{\sigma^2}{T^2})$

Ainsi, dans ces relations, on remarque que T contrôle la taille du voisinage d'influence d'un neurone c. Plus T est petit plus ce voisinage est réduit et pour T suffisamment petit le voisinage est réduit à c.

L'algorithme associé aux cartes topologiques minimise la fonction de coût généralisée suivante :

$$\mathcal{G}^T_{SOM}(\mathcal{X}, \mathcal{W}) = \sum_{z_i \in \mathcal{Z}} \sum_{c \in \mathcal{C}} \mathcal{K}^T(\sigma(c, \mathcal{X}(z_i)))||z_i - w_c||^2 \tag{2.8}$$

La fonction $\mathcal{G}^T_{SOM}$ est une extension de la fonction de coût des K-moyennes dans laquelle la distance euclidienne entre une observation z_i et son référent $w_{\mathcal{X}(z_i)}$ est remplacée par une distance pondérée d_T représentant la somme pondérée des distances euclidiennes de z_i à chacun des vecteurs référents w_c du voisinage d'influence du neurone $\mathcal{X}(z)$

$$d_T(z_i, w_{\mathcal{X}(z_i)}) = \sum_{c \in C} \mathcal{K}^T(\sigma(c, \mathcal{X}(z_i)))||z_i - w_c||^2 \tag{2.9}$$

Le paragraphe suivant présente les deux algorithmes utilisés pour minimiser $\mathcal{G}^T_{som}$.

La version non-adaptive de l'algorithme de Kohonen

Dans cette version, la minimisation de $\mathcal{G}^T_{SOM}$ est semblable à celle des K-moyennes. Elle se réalise par itérations successives, chacune se décompose en deux phases :

- La phase d'affectation : elle consiste à créer une partition à l'aide de la fonction d'affection $\mathcal{X}$ en supposant que l'ensemble des référents $\mathcal{W}$ est fixé. Pour toute observation z_i, on a alors :

$$\mathcal{X}^T(z_i) = \underset{c \in \mathcal{C}}{argmin}\ d^T(z_i, w_c) \tag{2.10}$$

- La phase de minimisation : il s'agit de minimiser la quantité $\mathcal{G}^T_{SOM}$ par rapport à l'ensemble $\mathcal{W}$ des référents en fixant cette fois la fonction d'affectation $\mathcal{X}^T$ à sa valeur courante et $\mathcal{G}^T_{SOM}$ devient alors une fonction quadratique par rapport à $\mathcal{W}$ qui atteint un minimum pour $\frac{\partial \mathcal{G}^T_{SOM}}{\partial \mathcal{W}} = 0$. Les vecteurs référents sont alors actualisés grâce à la formule suivante :

$$w_c^T = \frac{\sum_{r \in \mathcal{C}} \sum_{z_i \in r} \mathcal{K}^T(\sigma(c, r)) z_i}{\sum_{r \in \mathcal{C}} \mathcal{K}^T(\sigma(c, r)) n_r} \tag{2.11}$$

  w_c^T représente le barycentre des vecteurs moyens des observations des cellules r et que chaque barycentre est pondéré par $\mathcal{K}^T(\sigma(c, r))n_r$ où n_r est le nombre d'observations captées par le référent r.

Sur le plan algorithmique, la version non-adaptative des cartes topologiques pour une valeur de T fixée se résume de la manière suivante :

Algorithme : SOM version non adaptative

Entrée : l'ensemble $\mathcal{Z}$ des observations,T fixé, choisir les référents initiaux (en général de manière aléatoire), la structure et la taille de la carte $\mathcal{C}$, le nombre d'itération N_i

1. Etape itérative t : l'ensemble des référents $\mathcal{W}^{t-1}$ de l'étape $t-1$ est connu,
   (a) Phase d'affectation : affectation des observations z_i aux référents en utilisant l'expression 2.10
   (b) Phase de minimisation : appliquer 2.11 pour déterminer l'ensemble des nouveaux référents $\mathcal{W}^t$.
2. Répéter l'étape itérative jusqu'à ce que l'on atteigne N_i ou jusqu'à la stabilisation de $\mathcal{G}^T_{SOM}$.

Sortie : une carte $\mathcal{C}$ représentant une partition des observations.

Version adaptative de l'algorithme SOM

Comme l'algorithme des K-moyennes, il existe une version stochastique de l'algorithme des cartes topologiques. En remarquant que dans la phase de minimisation il n'est pas obligatoire de trouver un minimum global de la fonction $\mathcal{G}^T_{SOM}$ pour la fonction d'affectation $\mathcal{X}$ fixée, il suffit de faire décroitre sa valeur en remplaçant 2.11 par une méthode de gradient, à l'itération t et pour un neurone c, on a :

$$w_c^t = w_c^{t-1} - \mu^t \frac{\partial \mathcal{G}^T_{som}}{\partial w_c^{t-1}} \tag{2.12}$$

où μ^t est le pas du gradient à l'itération t et

$$\frac{\partial \mathcal{G}^T_{som}}{\partial w_c} = 2 \sum_{z_i \in \mathcal{Z}} K^T(\sigma(c, \mathcal{X}(z_i)))(z_i - w_c) \tag{2.13}$$

Contrairement à la version batch, dans cette version à chaque itération une seule observation est soumise au modèle. De plus la fonction d'affectation $\mathcal{X}$ est identique à celle de l'algorithme des K-moyennes :

$$\mathcal{X}(z_i) = \underset{c}{argmin} ||z_i - w_c||^2 \tag{2.14}$$

À chaque présentation d'une observation z_i, les nouveaux référents sont alors calculés pour tous les neurones de la carte $\mathcal{C}$ en fonction du neurone sélectionné :

$$w_c^t = w_c^{t-1} - \mu^t \mathcal{K}^T(\sigma(c, \mathcal{X}^T(z_i)))(z_i - w_c) \tag{2.15}$$

Algorithme : SOM version adaptative

Entrée : l'ensemble $\mathcal{Z}$ des observations.

1. Phase d'initialisation : choisir les référents initiaux (en générale de manière aléatoire), la structure et la taille de la carte $\mathcal{C}$; fixer les valeurs de T_{max}, T_{min} et le nombre d'itérations N_i ;
2. Etape itérative t. L'ensemble des référents $\mathcal{W}^{t-1}$ de l'étape $t-1$ est connu,
   (a) Choisir une observation z_i
   (b) Calculer la nouvelle valeur de la température T en appliquant la formule :
   $$T = T_{max} * (\frac{T_{min}}{T_{max}})^{\frac{1}{N_i - 1}} \tag{2.16}$$
   pour cette valeur du paramètre, effectuer les deux phases suivantes :
   (c) Phase d'affectation : on suppose que $\mathcal{W}^{t-1}$ connu ; on affecte l'observation z_i au neurone $\mathcal{X}^T(z_i)$ défini à partir de la fonction d'affectation 2.14
   (d) Phase de minimisation : calcul de l'ensemble des nouveaux référents $\mathcal{W}^t$; les vecteurs référents sont modifiés selon la formule 2.13 en fonction de leur distance au neurone sélectionnée à l'étape d'affectation.
3. Répéter l'étape itérative en faisant décroître la valeur de T jusqu'à ce que l'on atteigne t=N_i.

Propriétés de SOM

Ordre topologique et visualisation

La décomposition de la fonction objectif $\mathcal{G}^T_{SOM}$ (2.8) dépendant de T permet de mettre son expression sous la forme [Yacoub *et al.* 2001] :

$$\begin{aligned}\mathcal{G}^T_{SOM}(\mathcal{X},\mathcal{W}) &= (\sum_c \sum_{r\neq c} \sum_{\mathcal{X}(z_i)=r} K^T(\sigma(c,r))||z_i - w_r||^2) + K^T(\sigma(c,c)) \sum_c \sum_{z_i \in c} ||z_i - w_c||^2 \\ &= \frac{1}{2} \sum_c \sum_{r\neq c} K^T(\sigma(c,r))(\sum_{\mathcal{X}(z_i)=r} ||z_i - w_c||^2 + \sum_{\mathcal{X}(z_i)=c} ||z_i - w_r||^2) \\ &\quad + K^T(\sigma(c,c)) \sum_c \sum_{z_i \in c} \mathcal{I}_c \end{aligned} \tag{2.17}$$

On remarque que suivant la valeur de T, chacun des deux termes de (2.17) a une importance relative dans la minimisation. Ainsi :

- le premier terme introduit la contrainte de conservation de la topologie des observations. En effet, si deux neurones c et r sont proches sur la carte, l'expression $\mathcal{K}^T(\sigma(c,r))$ est grande ; la minimisation de ce terme rapproche les deux sous-ensembles c et r liés aux cellules c et r. Leur proximité sur la carte entraîne alors une proximité des référents dans l'espace $\mathcal{Z}$ des observations. Plus T est grand plus ce terme a une importance relative dans la minimisation.
- le second terme correspond à la fonction objectif des K-moyennes pondérée par $\mathcal{K}^T(\sigma(c,c)) = \mathcal{K}(0)$. Son importance relative dépend du paramètre T. Pour des petites valeurs de T, ce terme est pris en considération dans l'apprentissage. Ce terme à tendance à fournir une partition de l'ensemble $\mathcal{Z}$ pour laquelle les classes sont compactes et dont le vecteur référent des classes est leur centre de gravité.

Ainsi, dans l'algorithme de Kohonen, les premières étapes correspondant aux grandes valeurs de T privilégient le premier terme de 2.17 et donc la préservation de l'ordre topologique. Le second terme prend de l'importance pour de petites valeurs de T et l'algorithme minimise une expression liée à l'inertie. Le choix idéal de la valeur de T permet de réaliser un compromis entre les deux termes de 2.17. L'ordre topologique ayant été obtenu pendant la première partie de l'algorithme, la minimisation s'emploie pour la suite à obtenir des sous-ensembles aussi compacts que possible. Il s'agit de la phase K-moyennes de l'algorithme qui consiste à s'adapter localement aux différentes densités des données. On peut donc résumer l'algorithme de Kohonen comme le calcul d'une solution des K-moyennes sous contrainte d'ordre topologique. Finalement, la structure de la carte formée est telle que chaque neurone et les neurones se situant dans son voisinage ont capté des observations similaires.

Gestion des données manquantes et des outliers

De manière générale, en présence de données manquantes, pour éviter de supprimer les observations, on peut remplacer une valeur manquante par la moyenne de la variable correspondante, mais cette moyenne peut être une très mauvaise approximation dans le cas où la variable présente une grande dispersion. Il est alors important de constater que l'algorithme de Kohonen supporte bien la présence des données manquantes, sans qu'il soit nécessaire de les estimer préalablement. En effet, lorsqu'on présente un objet z_i présentant des données manquantes, on détermine son neurone gagnant sur la base des dimensions renseignées. La fonction d'affectation 2.10 devient alors :

$$\mathcal{X}^T(z_i) = \underset{c\in\mathcal{C}}{argmin} \sum_{c\in C} \mathcal{K}^T(\sigma(c, \mathcal{X}(z_i)))||z_i - w_c||^2 \tag{2.18}$$

où la distance $||z_i - w_c||^2 = \sum_{j\in VM_i}(z_i^j - w_c^j)$, avec VM_i l'ensemble des indices des dimensions renseignées pour l'individu i. On peut utiliser les vecteurs avec données manquantes de deux façons :

- Si l'on souhaite les utiliser au moment de la construction des vecteurs référents, à chaque étape, une fois déterminé le neurone gagnant, la mise à jour des vecteurs référents ne porte que sur les composantes présentes dans le vecteur.
- Si l'on dispose de suffisamment de données pour pouvoir se passer des vecteurs incomplets pour construire la carte, on peut aussi se contenter de classer, après construction de la carte, les vecteurs incomplets en les affectant dans la classe dont le vecteur référent est le plus proche, au sens de la distance restreinte aux composantes présentes.

Cela donne de bons résultats, dans la mesure où une variable n'est pas complètement absente ou presque, et aussi dans la mesure où les variables sont corrélées. Par ailleurs, nous avons montré plus haut qu'en fin d'apprentissage, l'algorithme de Kohonen se comporte quasiment comme la méthode des K-moyennes. Cela permet alors d'estimer a posteriori les valeurs manquantes en les remplaçant par la valeur de la composante du vecteur référent correspondant à la donnée manquante. Il est clair que cette estimation est d'autant plus précise que les classes formées par l'algorithme sont homogènes et bien séparées les unes des autres. De nombreuses simulations ont montré tant dans le cas de données artificielles que de données réelles, qu'en présence de variables corrélées, la précision de ces estimations est remarquable [Cottrell *et al.* 2007].

La collecte des données peut engendrer des observations atypiques qui peuvent provenir des erreurs de saisie ou des pannes des instruments de mesures. D'un point de vue pratique, les *outliers* sont généralement éloignés des autres observations dans un plan factoriel de projection d'une ACP par exemple. Les *outliers* respectant la corrélation seront visibles sur les premiers plans factoriels alors que ceux détériorant la corrélation seront détectés sur les derniers plans. Il est souhaitable que les *outliers* n'affectent pas les résultats de la partition des observations. Dans le cas de l'algorithme SOM la contrainte de voisinage permet d'isoler les *outliers* dans des cellules de la carte qui sont elles mêmes isolées dans des régions de la carte [Kaski 1997].

Comme les méthodes de classification automatique qui se sont développées d'un point

de vue heuristique autour de l'optimisation d'un critère métrique, les cartes topologiques présentées ci-dessus reposent sur la notion de distance. Présentées ici dans le cas d'observations décrites par des variables quantitatives, elles peuvent être étendues à d'autres types de données à travers l'utilisation de distances adaptées (Cf. section 2.2.3).

De même que pour les méthodes de classification non-neuronales, plusieurs auteurs ont proposé des extensions des cartes topologiques utilisant des modèles de mélanges [Luttrell 1989, 1994; Anouar *et al.* 1998] pour formaliser l'idée intuitive de la notion de classe naturelle. Ces méthodes constituent une base pour une alternative aux méthodes proposées dans cette thèse. Nous présentons dans la suite le formalisme PRSOM des cartes topologiques probabilistes proposé par [Anouar *et al.* 1998].

2.3.3 Cartes topologiques et modèles probabilistes

Les ensembles à classifier sont généralement considérés comme des sous-ensembles d'une population plus grande, cependant les conclusions obtenues sur ces sous-ensembles sont souvent étendues à toute la population. Dans ce cas, il est nécessaire d'avoir recours à des modèles probabilistes. Ces modèles probabilistes formalisent l'hypothèse que les données $\mathcal{Z} = \{z_i, ..., z_N\}$ sont formées d'un mélange de différentes populations. Puis, ils définissent une classification des données en s'appuyant sur la distribution de probabilité des données [Govaert 2003]. L'algorithme appelé PRSOM (PRobalistic Self-Organizing Map) [Anouar *et al.* 1998] suppose que la distribution de probabilité $p(z/c)$ associée à chaque cellule c de la carte prend une forme analytique représentée par une loi gaussienne sphérique. Le mélange des cartes topologiques est défini par le formalisme bayésien introduit par Luttrell [1994]. Ce formalisme suppose que les observations $\mathcal{Z}$ sont générées selon le processus à trois étapes suivant :

- choix d'une cellule r de la carte $\mathcal{C}$ suivant la loi de probabilité a priori sur l'ensemble des cellules.
- la cellule r choisie permet de déterminer un voisinage avec une loi de probabilité conditionnelle $p(c/r)$.
- choix d'une observation z suivant la loi gaussienne $p(z/c)$ affectée à la cellule c.

Ce formalisme conduit à la détermination d'un générateur des données *p(z)* par un mélange de probabilités :

$$p(z) = \sum_{c,r \in \mathcal{C}} p(c,r,z) \tag{2.19}$$

$$= \sum_{c,r \in \mathcal{C}} p(c/z)p(c/r)p(r) \tag{2.20}$$

$$= \sum_{r \in \mathcal{C}} p(r)p_r(z) \tag{2.21}$$

avec

$$p_r(z) = p(z/r) = \sum_{c \in \mathcal{C}} p(c/r)p(z/c) \tag{2.22}$$

Les probabilités $p(r)$ a priori et $p_r(c)$ désignent respectivement les coefficients du mélange et les fonctions densités de chaque élément du mélange. Ainsi, on peut calculer $p(z)$ si l'on connaît pour chaque cellule c la fonction densité $f_c(z) = p(z/c)$ et la probabilité $p(c/r)$ de la cellule c connaissant r. Nous allons maintenant introduire la notion de voisinage dans le processus probabiliste. Supposons que chaque cellule c de la carte est très active si elle est proche de la cellule choisie r. Cette supposition permet de définir la probabilité $p(c/r)$ en fonction de la fonction d'évaluation du voisinage $\mathcal{K}^T$:

$$p(c/r) = \frac{\mathcal{K}^T(\sigma(c,r))}{\sum_{r \in \mathcal{C}} \mathcal{K}^T(\sigma(c,r))} \tag{2.23}$$

La définition complète de la quantité $p(z)$ est maintenant liée à la connaissance des quantités $p(r)$ et $p(z/c)$. Anouar *et al.* [1998] proposent d'utiliser le formalisme maximisant la vraisemblance du modèle pour estimer les quantités $p(r)$ et $f_c(z)$.

Les densités de probabilités a posteriori (relation 2.22) peuvent s'exprimer en fonction des distributions gaussiennes des différents neurones.

$$p_r(z) = \frac{\sum_{c \in \mathcal{C}} \mathcal{K}^T(\sigma(c,r)) f_c(z, w_c, \delta_c)}{\sum_{c \in \mathcal{C}} \mathcal{K}^T(\sigma(c,r))} \tag{2.24}$$

Ainsi, $p_r(z)$ apparaît comme un mélange local de densité gaussienne faisant intervenir principalement les neurones du voisinage de r sur la carte. L'ensemble des vecteurs référents $\mathcal{W}$ et des écarts types $\delta = \{\delta_c, c \in \mathcal{C}\}$ sont les paramètres à estimer. Sous hypothèse que les observations sont indépendantes, on obtient la vraisemblance classifiante suivante :

$$f(z_1, z_2, \ldots, z_N / \mathcal{W}, \delta, \mathcal{X}) = \prod_{i}^{N} p_{\mathcal{X}(z_i)}(z_i) \tag{2.25}$$

Cette expression est ensuite maximisée par rapport aux paramètres $\mathcal{W}$,δ et $\mathcal{X}$. D'une manière classique on réalise cet objectif en minimisant l'opposé de la log vraisemblance.

$$E(\mathcal{W},\delta,\mathcal{X}) = -\sum_{i=1}^{N} log(\sum_{r\in\mathcal{C}} \mathcal{K}^T(\sigma(\mathcal{X}(z_i),r))f(z_i,w_r,\delta_r)) \tag{2.26}$$

Comme pour l'algorithme classique de Kohonen, les deux phases d'affectation et de minimisation sont effectuées alternativement jusqu'à la convergence.

- Affectation : On suppose que l'ensemble des paramètres $\mathcal{W}$ et celui des écarts-types δ sont connus et fixés. La fonction d'affectation minimisant E est celle qui consiste à affecter chaque observation z_i au neurone le plus probable selon la densité p_r.

$$\mathcal{X}(z) = \underset{r}{argmax}\ p_r(z)$$

- Phase de minimisation. Au cours de cette phase, on suppose que la fonction d'affectation est constante et égale à la fonction d'affectation courante. On cherche alors à minimiser E par rapport à $\mathcal{W}$ et δ. On obtient alors les formules de mise à jour suivantes :

$$w_r^t = \frac{\sum_{i=1}^{N} K^T(\sigma(r,\mathcal{X}^{t-1}(z_i)) \frac{f_r(z_i,w_r^{t-1},\delta_r^{t-1})}{P_{\mathcal{X}^{t-1}(z_i)}(z_i)} z_i}{\sum_{i=1}^{N} K^T(\sigma(r,\mathcal{X}^{t-1}(z_i)) \frac{f_r(z_i,w_r^{t-1},\delta_r^{t-1})}{P_{\mathcal{X}^{t-1}(z_i)}(z_i)}}$$

$$(\delta_r^t)^2 = \frac{\sum_{i=1}^{N} K^T(\sigma(r,\mathcal{X}^{t-1}(z_i)) \frac{f_r(z_i,w_r^{t-1},\delta_r^{t-1})}{P_{\mathcal{X}^{t-1}(z_i)}(z_i)} ||w_r^{t-1} - z_i||^2}{p\sum_{i=1}^{N} K^T(\sigma(r,\mathcal{X}^{t-1}(z_i)) \frac{f_r(z_i,w_r^{t-1},\delta_r^{t-1})}{P_{\mathcal{X}^{t-1}(z_i)}(z_i)}}$$

Dans ces deux expressions, les paramètres à l'itération t s'expriment en fonction de ceux de l'itération t-1.

2.4 Les critères d'évaluation d'une classification

L'évaluation de la qualité d'une classification est un aspect très important pour valider les classes obtenues. Plusieurs critères de validation des classes sont présentés dans la littérature [Dubes et Jain 1976; Richard et Jain 1979; Halkidi *et al.* 2002; Jain 2008]. Ces critères peuvent être regroupés en 2 grandes familles. La famille des critères non-supervisées ou interne qui utilise uniquement les informations internes aux données telles que la distance entre les observations, pour quantifier l'adéquation entre les classes obtenues avec un algorithme de classification et l'idée que l'on se fait d'une bonne classification à savoir la séparabilité et la compacité des classes.

La famille des critères externes s'utilisant pour la validation d'un algorithme de classification fait appel à des connaissances externes à la classification. En effet, il est habituel d'utiliser des données étiquetées ; ces étiquettes ou labels peuvent être obtenus selon l'avis d'expert ou suite à l'application d'un algorithme de classification sur les données. On cherche alors, à évaluer la capacité d'un algorithme, à définir des classes dans lesquelles on retrouve des données ayant des labels identiques. Les critères de validation externe présentés dans la section 2.4.2 sont alors vus comme des mesures de comparaisons de partitions et de validation d'algorithme en classification.

2.4.1 Critères d'évaluation interne

Les critères de validation interne des classes sont basés sur la définition de mesures propres aux classes comme la distance entre les observations et leur centre de classe. Ils sont basés sur les propriétés voulant que :

- des individus d'une même classe partagent les mêmes propriétés (compacité).
- des individus appartenant à des classes différentes aient peu de propriétés en commun (séparabilité).

Pour évaluer le respect de ces deux notions, différentes mesures basées sur les distances entre les observations z_i et les centres de classe w_k ont été définies pour quantifier l'adéquation entre une partition et l'idée que l'on se fait d'une bonne classification. Cette section présente plus en détail les indices d'évaluation de la pertinence d'une classification.

Définition 6 (Somme des carrées) *La somme des carrées des erreurs (Mean Sqaure Error, MSE) permet d'évaluer la compacité des classes d'une classification. Elle vaut :*

$$MSE = \frac{1}{N} \sum_{k=1}^{K} \sum_{z_i \in c_k} ||z_i - w_k||^2 \tag{2.27}$$

où K est le nombre de classes. MSE correspond au critère à optimiser dans l'algorithme des K-moyennes.

Ce critère s'utilise pour la comparaison de partition de tailles identiques.

Définition 7 (Silhouette Value) *La silhouette value de Rousseeuw [1987] permet d'évaluer la compacité et la séparabilité des classes. Cet indice est défini pour chaque observation, pour chaque classe et pour la classification. Soit :*

- *a_i la moyenne des distances entre l'observation z_i et toutes les autres observations appartenant sa classe.*

- b_{ik} *la moyenne des distances entre l'observation* z_i *et les observations appartenant à la classe* k *avec* z_i *n'appartenant pas à la classe* k.
- b_i *le minimum des* $K-1$ *moyennes* b_{ik} *obtenues.*

Pour une observation z_i*, une classe* c_k *et une partition* $\mathcal{C}$ *la silhouette value est définie respectivement par :*

$$SV_z(z_i) = \frac{b_i - a_i}{max(b_i, a_i)} \tag{2.28}$$

$$SV_c(c_k) = \frac{1}{|c_k|} \sum_{z_i \in c_k} SV_z(z_i) \tag{2.29}$$

$$SV(C) = \frac{1}{K} \sum_{k=1}^{K} SV_c(c_k) \tag{2.30}$$

La quantité SV_z est comprise entre -1 et 1. Une valeur positive de SV_z et proche de 1 signifie que les observations appartenant à la même classe que z_i sont plus proches de cet objet que des autres observations des autres classes. Une valeur négative de SV_z et proche de -1 implique que z serait mieux classé dans une autre classe. Enfin si SV_z est proche de 0 cela implique que l'observation z se situe aux frontières de deux classes. SV_c évalue l'homogénéité de la classe k. Enfin, le coefficient SV_C varie également de -1 à 1, plus sa valeur est positive et grande plus ceci implique que les classes sont bien séparées et très compactes. Cet indice est une aide au choix de nombre de k de classes dans l'algorithme des K-moyennes. En effet, lorsque K n'est pas adéquat (trop petit ou trop grand), la valeur SV_c au niveau de certaines classes est très faible. Il faut alors calibrer K pour obtenir des quantités SV_c de même grandeur.

Définition 8 (Indice de Davies-Bouldin) *L'indice de Davies et Bouldin [1979] évalue la qualité d'une classification en mesurant la compacité et la séparabilité des classes à travers le calcul de la moyenne de la similarité entre les classes :*

$$DB(C) = \frac{1}{K} \sum_{l=1}^{K} \underset{k=1,\ldots,K, k \neq l}{max} \left(\frac{S_{db}(c_k) + S_{db}(c_l)}{d(w_k, w_l)} \right) \tag{2.31}$$

où $d(w_k, w_l)$ *est la distance entre les centres des classes* c_k *,* c_l *et* $S_{db}(c_k)$ *la moyenne des distances entre les observations de* c_k *et les centroïdes* w_k *de chaque classe.*

Pour des groupes compacts, la moyenne $S_{db}(c_k)$ de la distance au référent vecteur w_k est petite. Pour des groupes bien séparés, la distance $d(w_k, w_l)$ est grande. Une valeur faible de cet indice implique une classification de bonne qualité en termes de compacité et de séparabilité. Dans le cas de l'algorithme des K-moyennes, son application répétée sur un ensemble de données en faisant varier le nombre K permet par la suite de définir le nombre de classes K_I idéal comme celui minimisant cet indice.

Indices spécifiques aux cartes topologiques

Les cartes auto-organisées font partie des méthodes de quantification vectorielle qui ont des propriétés spécifiques, il semble donc naturel de les évaluer à l'aide de l'erreur de quantification moyenne que l'on définit ainsi :

Définition 9 (Erreur de quantification)

$$mqe = \frac{1}{N} \sum_{i=1}^{N} \|z_i - w_{c_i}\|^2 \tag{2.32}$$

où c_i est l'indice du prototype le plus proche de z_i

Le principe de conservation de la topologie des observations sur la carte implique d'évaluer la qualité de la topologie fournie par SOM. Le taux d'erreur topologique permet de quantifier la conservation de la topologie locale de l'espace des observations par la carte.

Définition 10 (Taux d'erreur topologique) *On considère qu'il y a une erreur topologique pour une observation si les deux neurones les plus proches de cette observation z_i en terme de distance ne sont pas voisins sur la carte. Le taux d'erreur topologique vaut alors :*

$$Tge = 1 - \frac{1}{N} \sum_{i=1}^{N} 1|_{N(c_i)} \left(\underset{c \neq \mathcal{X}(i)}{argmin}(||z_i - w_c||^2) \right) \tag{2.33}$$

où $1|_{N(c_i)}$ est la fonction indicatrice de l'ensemble des voisins du prototype le plus proche de l'observation z_i.

La mesure de quantification vectorielle et l'erreur topologique peuvent être contradictoires puisqu'elles évaluent des propriétés différentes de la carte topologique. La mesure de distorsion présentée ci-dessous crée un compromis entre ces deux mesures.

Définition 11 (Mesure de distorsion) *La mesure de distorsion prend en compte l'erreur de quantification vectorielle et la conservation de la topologie locale à travers l'introduction d'une pondération basée sur la fonction de voisinage définie dans SOM. Elle vaut*

l'erreur quadratique pondérée par la fonction de voisinage.

$$distorsion = \sum_{i=1}^{N} \sum_{c} K^T(c_i, c)||z - w_c||^2 \tag{2.34}$$

où $K^T(c_i, c)$ *est la fonction de voisinage.*

Remarquons que cette expression correspond à la valeur finale de la fonction objectif de SOM. Vesanto *et al.* [2003] décompose la relation 2.34 en trois termes correspondant à la variance des données dans le voisinage de chaque cellule, à la qualité de la topologie de la carte et au compromis entre la quantification vectorielle et la conservation de la topologie des observations.

2.4.2 Critères d'évaluation externe

La comparaison de deux partitions $\mathcal{C}$ et $\mathcal{C}'$ d'un même ensemble de données $\mathcal{Z}$ peut être réalisée à partir d'un tableau 2.4 de contingence $\mathcal{T} = (n_{kl})$, où n_{kl} désigne le nombre d'objets appartenant simultanément aux classes k et l des partitions respectifs $\mathcal{C}$ et $\mathcal{C}'$.

	1	...	l	...	K'	$n_{k.}$
1	n_{11}	...	n_{1l}	...	$n_{1K'}$	$\sum_l n_{1l}$
⋮						
k	n_{k1}	...	n_{kl}	...	$n_{kK'}$	$\sum_l n_{kl}$
⋮						
K	n_{K1}	...	n_{Kl}	...	$n_{KK'}$	$\sum_l n_{Kl}$
$n_{.l}$	$\sum_k n_{k1}$	...	$\sum_k n_{kl}$	...	$\sum_k n_{kK'}$	

TABLE 2.4 – Table de contingence entre deux partitions $\mathcal{C}$ et $\mathcal{C}'$ contenant respectivement K et K' classes ; n_{kl} est l'effectif d'observations appartenant simultanément à la classe k de la variable $\mathcal{C}$ et à la classe l de la variable $\mathcal{C}'$

Pour introduire les critères externes on redéfinit, à partir du tableau de contingence, les quantités N_{11}, N_{10}, N_{01} et N_{00} présentées dans la section 2.2.3.2 :

- N_{11}, le nombre de fois où deux observations sont dans la même classe dans $\mathcal{C}$ et dans $\mathcal{C}'$ (accords positifs)

$$N_{11} = \frac{1}{2}\left(\sum_{k=1}^{K}\sum_{l=1}^{K'} n_{kl}(n_{kl} - 1)\right) \tag{2.35}$$

- N_{10}, le nombre de fois où deux observations sont dans la même classe de $\mathcal{C}'$ et dans des classes différentes dans $\mathcal{C}$.

$$N_{10} = \frac{1}{2}\left(\sum_{k=1}^{K'} n_{k.}^2 - \sum_{k=1}^{K}\sum_{l=1}^{K'} n_{kl}^2\right) \tag{2.36}$$

– N_{01}, le nombre de fois où deux observations sont dans la même classe de $\mathcal{C}$ et des classes différentes $\mathcal{C}'$

$$N_{01} = \frac{1}{2}\left(\sum_{k=1}^{K'} n_{.l}^2 - \sum_{k=1}^{K}\sum_{l=1}^{K'} n_{kl}^2\right) \tag{2.37}$$

– N_{00}, le nombre de fois où deux observations sont dans des classes différentes de $\mathcal{C}$ et de $\mathcal{C}'$ (accords négatifs)

$$N_{00} = \frac{1}{2}\left(n^2 + \sum_{k=1}^{K}\sum_{l=1}^{K'} n_{kl}^2 - \left(\sum_{k=1}^{K} n_{k.}^2 + \sum_{k=1}^{K'} n_{.l}^2\right)\right) \tag{2.38}$$

2.4.2.1 Précision, rappel, F-mesure, Pureté

Les indices de précision et le coefficient de rappel sont des mesures asymétriques évaluant la similarité entre une partition $\mathcal{C}'$ fournit par un algorithme et les labels de référence $\mathcal{C}$.

Définition 12 (Indice de précision)
L'indice de précision indique la probabilité que deux objets soient regroupés dans la partition $\mathcal{C}'$ s'ils le sont dans la partition $\mathcal{C}$:

$$prec(\mathcal{C}, \mathcal{C}') = \frac{N_{11}}{N_{11} + N_{01}} \tag{2.39}$$

Définition 13 (Indice de rappel) *Le coefficient de rappel indique la probabilité que deux objets soient regroupés dans la partition $\mathcal{C}$ s'ils le sont dans la partition $\mathcal{C}'$:*

$$rapp(\mathcal{C}, \mathcal{C}') = \frac{N_{11}}{N_{11} + N_{10}} \tag{2.40}$$

Les critères de rappel et de précision prennent leurs valeurs dans l'intervalle $[0; 1]$, cependant une valeur proche de 1 n'implique pas que les partitions soient identiques. La combinaison de ces deux indices en prenant leur moyenne arithmétique, géométrique et harmonique, fournit respectivement le deuxième coefficient de Kulczynski, l'indice de Folkes et Mallows et la F-mesure. Ces trois critères sont symétriques et prennent leurs valeurs sur l'intervalle $[0, 1]$. Ils valent 1 si et seulement si les partitions $\mathcal{C}$ et $\mathcal{C}'$ sont identiques.

Définition 14 ($2^{ième}$ coefficient de Kulczynski)
Le deuxième coefficient de Kulczynski se définit comme la moyenne de l'indice de precision et du rappel :

$$\mathcal{K}(\mathcal{C}, \mathcal{C}') = \frac{1}{2}(prec(\mathcal{C}, \mathcal{C}') + rapp(\mathcal{C}, \mathcal{C}')) \tag{2.41}$$

Définition 15 (Indice de Folkes & Mallows) *L'indice de Folkes & Mallows est défini comme la moyenne géométrique de l'indice de précision et du coefficient de rappel*

$$\mathcal{FM}(\mathcal{C},\mathcal{C}') = \sqrt{prec(\mathcal{C},\mathcal{C}') \times rapp(\mathcal{C},\mathcal{C}')} \tag{2.42}$$

Définition 16 (F-mesure) *La F-mesure est la moyenne harmonique de l'indice de précision et du coefficient de rappel :*

$$\mathcal{F}(\mathcal{C},\mathcal{C}') = \frac{prec(\mathcal{C},\mathcal{C}') \times rapp(\mathcal{C},\mathcal{C}')}{prec(\mathcal{C},\mathcal{C}') + rapp(\mathcal{C},\mathcal{C}')} \tag{2.43}$$

En utilisant une moyenne harmonique pondérée, on définit la F_α-mesure comme suit :

$$\mathcal{F}_\alpha(\mathcal{C},\mathcal{C}') = \frac{(1-\alpha) \times prec(\mathcal{C},\mathcal{C}') \times rapp(\mathcal{C},\mathcal{C}')}{\alpha \times prec(\mathcal{C},\mathcal{C}') + rapp(\mathcal{C},\mathcal{C}')} \tag{2.44}$$

où α est un coefficient de pondération strictement positif. Notons que pour $\alpha > 1$ cet indice est négatif.

La pureté d'une partition

La pureté d'une partition s'évalue en quantifiant la cohérence d'une partition par rapport à une autre. La manière la plus simple d'évaluer la pureté est de rechercher le label majoritaire de chaque classe et de sommer le nombre d'observations ayant le label majoritaire par classe. La pureté se définit alors simplement par l'expression suivante :

$$Pur(C,C') = \frac{1}{N}\sum_{k=1}^{K} \underset{C_l}{argmax}(n_{kl}) \tag{2.45}$$

Cette valeur de la pureté est équivalente à l'estimation du pourcentage d'individus ayant le label majoritaire dans les classes de la partition C. Sa valeur est bornée dans $[0,\, 1]$; 1 implique que les individus formant les classes ont des labels identiques. Une formulation probabiliste de la pureté d'une partition consiste à calculer la probabilité qu'étant donnée une classe de la partition C, deux individus tirés au hasard sans remise aient le même label. En définissant par $\frac{n_{kl}}{n_k}$ la probabilité que le premier individu ait le label l et par $(\frac{n_{kl}}{n_k})^2$ la probabilité que le deuxième appartienne à la même classe. L'appartenance aux classes étant dure on évalue la pureté d'une classe par :

$$Pur_{prob}(c_k) = \sum_{l=1}^{K'} (\frac{n_{kl}}{n_k}) \tag{2.46}$$

Ce qui donne pour une partition :

$$Pur_{prob}(C) = \frac{1}{N}\sum_{k=1}^{K} n_k Pur_{prob}(c_k)^2 \tag{2.47}$$

Cette nouvelle mesure de pureté prend en compte la proportion des différents labels dans les classes en favorisant les classes ayant un nombre limité de labels.

2.4.2.2 Indice de Rand

L'indice de Rand indique la proportion de paires d'observations pour lesquelles deux partitions sont en accord. Il correspond à la mesure de similarité binaire simple correspondance qui prend ses valeurs dans l'intervalle $[0, 1]$ et est défini de la manière suivante :

$$\mathcal{R}(\mathcal{C}, \mathcal{C}') = \frac{N_{11} + N_{00}}{N_{11} + N_{01} + N_{10} + N_{00}} \tag{2.48}$$

Cependant, pour deux partitions définies aléatoirement la valeur de l'indice de Rand n'est pas nulle, d'autre part lorsque ces deux partitions ont des nombres de classes différents, l'indice de Rand peut être proche de 1. Hubert et Arabie [1985] et Chavent *et al.* [2001] proposent plusieurs variantes de l'indice initial de Rand pour surmonter ces problèmes. Le plus connu, l'indice de Rand ajusté $\mathcal{R}_a$ prend la valeur 0 lorsque les deux partitions sont définies aléatoirement et la valeur 1 lorsqu'elles sont identiques. L'indice de Rand est souvent plus élevé que sa version corrigée qui se définit en utilisant les notations de la section 2.4.2.

$$\mathcal{R}_a(\mathcal{C}, \mathcal{C}') = \frac{n^2 \sum_{i,j} n_{ij}^2 - \sum_i n_{i.}^2 \sum_j n_{.j}^2}{\frac{1}{2} n^2 (\sum_i n_{i.}^2 + \sum_j n_{.j}^2) - \sum_i n_{i.}^2 \sum_j n_{.j}^2} \tag{2.49}$$

2.4.2.3 Variation d'information

Le critère de comparaison Variation d'Information (VI) issu de la théorie de l'information, quantifie l'information apportée par la connaissance d'une partition $\mathcal{C}$ sur une partition $\mathcal{C}'$. Soit n_k le cardinal de la classe c_k. On définit par $P(k) = \frac{n_k}{N}$ la probabilité qu'une observation z_i choisie au hasard appartiennent à la classe k et par $P(k,l) = \frac{|c_k \cap c'_k|}{N} = \frac{n_{kl}}{N}$ la probabilité que des observations appartiennent aux classes $c_k \in C$ et $c'_k \in C'$.

Définition 17 (Entropie d'une partition)
L'entropie associée à une partition $\mathcal{C} = \{c_1, \ldots, c_K\}$ mesure l'incertitude de la variable aléatoire C dont la valeur est l'indice de la classe d'un objet prélevé aléatoirement dans

l'ensemble $\{1, \ldots, K\}$. *Elle est définie par :*

$$\mathcal{H}(\mathcal{C}) = -\sum_{k=1}^{K} P(k)log(P(k)) = -\sum_{k=1}^{K} \frac{n_{k.}}{N} log(\frac{n_{k.}}{N}) \tag{2.50}$$

L'entropie d'une partition est toujours positive et prend la valeur 0 lorsqu'il n'y a aucune incertitude quant à l'appartenance d'un objet à une classe. Autrement dit, lorsque $K = N$.

Définition 18 (Information mutuelle entre deux partitions)

$$\mathcal{I}(\mathcal{C}, \mathcal{C}') = \sum_{i=1}^{K}\sum_{i'=1}^{K'} P(k,l)log\left(\frac{P(k,l)}{P(k)P(l)}\right) = \sum_{i=1}^{K}\sum_{i'=1}^{K'} \frac{n_{kl}}{N} log\left(\frac{n_{kl}N}{n_{k.}n_{.l}}\right) \tag{2.51}$$

L'information mutuelle I est symétrique et positive. remarquons que lorsque deux partitions $\mathcal{C}$ et $\mathcal{C}'$ sont égales, on a : $I(\mathcal{C}, \mathcal{C}') = \mathcal{H}(\mathcal{C}) = \mathcal{H}(\mathcal{C}')$ L'information mutuelle I entre deux partitions $\mathcal{C}$ et $\mathcal{C}'$ quantifie l'information apportée par la variable aléatoire v associée à $\mathcal{C}$ sur la variable v' associée $\mathcal{C}'$ et réciproquement. Il est aussi possible de définir la version normalisée de l'information mutuelle NMI. L'indice NMI est indépendant du nombre de classes :

$$NMI(\mathcal{C}, \mathcal{C}') = \frac{\mathcal{I}(\mathcal{C}, \mathcal{C}')}{\sqrt{\mathcal{H}(\mathcal{C})\mathcal{H}(\mathcal{C}')}} \tag{2.52}$$

Définition 19 (Variation d'Information) *la variation d'information entre deux classifications* $\mathcal{C}$ *et* $\mathcal{C}'$ *est la somme de l'information sur* $\mathcal{C}$ *que l'on perd et de l'information sur* $\mathcal{C}'$ *que l'on gagne lorsqu'on passe de la partition* $\mathcal{C}$ *à la partition* $\mathcal{C}'$. *Formellement on a :*

$$VI(\mathcal{C}, \mathcal{C}') = [\mathcal{H}(\mathcal{C}) - \mathcal{I}(\mathcal{C}, \mathcal{C}')] + [\mathcal{H}(\mathcal{C}') - \mathcal{I}(\mathcal{C}, \mathcal{C}')] \tag{2.53}$$

$$= \mathcal{H}(\mathcal{C}) + \mathcal{H}(\mathcal{C}') - 2\mathcal{I}(\mathcal{C}, \mathcal{C}') \tag{2.54}$$

La figure 2.5 met en évidence la relation entre les quantités $\mathcal{H}(\mathcal{C})$, $\mathcal{I}(\mathcal{C}, \mathcal{C}')$ et $VI(\mathcal{C}, \mathcal{C}')$

Propriété 1 *La variation d'information est une distance sur l'ensemble des partitions* $\mathcal{C}$, $\mathcal{C}'$ *et* $\mathcal{C}''$, *elle présente les propriétés suivantes :*

1. *Positive : VI(*$\mathcal{C}$, $\mathcal{C}'$*) est toujours positif.*
2. *Séparabilité : VI(*$\mathcal{C}$, $\mathcal{C}'$*) s'annule si et seulement si les deux partitions sont égales*
3. *Symétrie : VI(*$\mathcal{C}$, $\mathcal{C}'$*)=VI(*$\mathcal{C}'$, $\mathcal{C}$*)*
4. *Inégalité triangulaire : VI(*$\mathcal{C}$, $\mathcal{C}'$*)+VI(*$\mathcal{C}'$, $\mathcal{C}''$*)>VI(*$\mathcal{C}$, $\mathcal{C}'$*)*

Nous présentons ci-dessous la version normalisée de la variation d'information :

$$NVI(\mathcal{C}, \mathcal{C}') = \frac{VI(\mathcal{C}, \mathcal{C}')}{\mathcal{I}(\mathcal{C}, \mathcal{C}')} \tag{2.55}$$

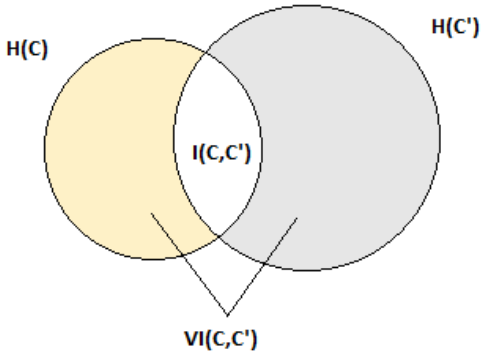


FIGURE 2.5 – La variation d'information (VI) en relation avec l'entropie $\mathcal{I}$

2.5 Conclusion

Dans ce chapitre, sans être exhaustif nous avons étudié les concepts de la classification et les principales méthodes classiques existantes. À savoir, les méthodes hiérarchiques dont la lecture de l'arbre hiérarchique qu'elles fournissent permet de déterminer le nombre optimal de classes. Cependant, ces méthodes ont l'inconvénient majeur d'être coûteux en temps de calcul. À l'opposé on retrouve les méthodes non-hiérarchiques qui ont l'avantage d'être adaptées aux données volumineux mais en imposant la connaissance à priori du nombre exact de classes qui n'est en réalité pas connu en apprentissage non-supervisée. La classification mixte permet une combinaison judicieuse des deux types de méthodes.
Nous avons ensuite présenté quelques mesures permettant d'évaluer la pertinence d'une classification et mis en évidence l'absence d'une mesure absolue de la performance. De nombreuses études bibliographiques ont montré qu'il existe un nombre important de méthodes de classification et qu'il est souvent difficile de choisir la méthode idoine pour un jeu de données. La section suivante présente les approches proposées dans la littérature pour traiter les données de grande dimension et rechercher un consensus entre les partitions.

Chapitre 3

Classification des données de grande dimension

3.1 Introduction

Dans les méthodes usuelles de classification, la similarité entre les observations est souvent déterminée par une distance prenant en compte toute la dimension des données [Jain *et al.* 1999b]. Les récentes avancées technologiques en termes de capacité de stockage d'une part, et la multiplication des sources d'information d'autre part, contribuent à la mise en place de bases de données complexes et de grande dimension. Ces données peuvent contenir des variables à forte variabilité ou à distribution uniforme. Dans ce cas, la similarité entre deux observations est souvent portée par un nombre limité de dimensions. Par conséquent, utiliser une mesure de similarité basée sur l'ensemble des variables peut s'avérer inefficace [Agrawal *et al.* 1998; Domeniconi *et al.* 2004; Kriegel *et al.* 2009].

Un autre problème plus connu sous le terme de "fléau de la dimension", implique la perte du pouvoir discriminant de la notion de distance au fur et à mesure que la dimension augmente. Cela se traduit par le fait que les observations sont pratiquement tout équidistantes les unes par rapport aux autres [Parsons *et al.* 2004]. Aggarwal *et al.* [2001] montrent théoriquement que dans un espace de dimension p, si l'on considère que les variables ont une distribution aléatoire d'observations et si l'on note d_{min} la distance entre les deux points les plus proches et d_{max} la distance entre les deux points les plus éloignés alors $\lim_{p\to\infty} \frac{d_{max}-d_{min}}{d_{min}} = 0$. Ce résultat implique que quand l'espace devient très grand les mesures classiques de distance deviennent inefficaces, car elles ne permettent pas de distinguer les points proches des points éloignés. Ce phénomène est accentué lorsque seules quelques di-

mensions sont importantes pour la classification, induisant ainsi l'intérêt de rechercher des classes dans des espaces de faible dimension.
Plusieurs approches ont été proposées pour la classification en grande dimension. Nous ne ferons pas une description exhaustive de tous les types d'approches, cependant le lecteur intéressé par plus détails pourra se référer aux revues présentées par Jain *et al.* [1999b]; Berkhin [2004]; Parsons *et al.* [2004]; Kriegel *et al.* [2009] et Vega-Pons et Ruiz-Shucloper [2011]. Néanmoins, ces méthodes peuvent être regroupées en quatre grandes familles. Les méthodes de réduction de dimension, les méthodes de sélection globale des variables, les méthodes de sélection locale des variables et les méthodes de recherche de consensus.
Les sections suivantes en font une présentation générale.

3.2 Réduction de la dimension

Les techniques de réduction des variables cherchent à résumer l'information des données initiales dans des espaces de faible dimension par rapport à l'espace initial. Le nouvel espace des variables est souvent défini à partir de combinaisons linéaires de toutes les variables initiales. Ces méthodes impliquent l'utilisation des approches factorielles telles que l'analyse en composantes principales ou la décomposition en valeurs singulières des matrices des données. A travers ces transformations, il s'agit de synthétiser, par combinaison linéaires des variables, l'information de l'espace initial. Classiquement, l'utilisation de la réduction des dimensions en classification est réalisée à travers une approche dite *Tandem*. Elle consiste à effectuer une classification sur les observations décrites par un nombre optimal de composantes factorielles issues de l'application préalable d'une méthode factorielle telle que l'Analyse en Composantes Principales (ACP) ou l'Analyse des Correspondances Multiples (ACM).
Le choix des composantes factorielles est un problème délicat et les résultats des approches de type tandem dépendent fortement de la structure de corrélation reliant les dimensions. Par ailleurs, les composantes factorielles maximisant l'inertie ne sont pas impérativement dédiées à la découverte des classes dans les données [Hubert et Arabie 1985; Vichi et Kiers 2001a]. Pour surmonter ces limitations, les méthodes Reduced-K-Means (RKM) [De Soete et Carroll 1994] et Factorial-K-Means (FKM) [Vichi et Kiers 2001b] recherchent simultanément la classification des individus et les composantes factorielles optimales pour la classification.

Ainsi, les méthodes FKM et RKM décomposent la matrice Z des observations en une

matrice binaire $U(N \times K)$ définissant les classes d'appartenance des observations, en une matrice $A(p \times Q)$ (Q étant le nombre de composantes factorielles retenues). Les colonnes de A indiquent la contribution des variables dans les classes de U et en une matrice F représentant les centroïdes des classes. Par exemple, pour un ensemble Z contenant 4 observations reparties en 2 classes et décrites par 5 variables, une partition des observations en deux classes donne :

$$U = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \qquad F = \begin{bmatrix} 1.2 & 1.3 \\ \text{-}0.1 & 0.2 \end{bmatrix} \qquad A = \begin{bmatrix} 0.7 & 0 \\ 0.7 & 0 \\ 0 & 0.7 \\ 0 & 0.7 \\ 0 & 0 \end{bmatrix}$$

La matrice U montre que les observations 1 et 2 sont dans la même classe et les observations 3 à 4 sont dans une autre classe. La matrice F contient les centroïdes des deux classes. La matrice A montre que les variables 1 et 2 sont associées au premier plan factoriel tandis que les variables 3 et 4 sont associées au second plan factoriel. La variable 5 n'intervient pas dans ce premier plan factoriel, elle correspond donc à du bruit dans les données.

Plus précisément, les fonctions objectifs associées à FKM et à RKM sont les suivantes :

$$F_{FKM}(U, F, A) = \| ZAA' - UFA' \|^2 = \| ZA - UF \|^2 \tag{3.1}$$

$$F_{RKM}(U, F, A) = \| Z - UFA' \|^2 \tag{3.2}$$

A travers la fonction objectif F_{FKM} (relation 3.1), FKM minimise la somme des carrées des distances entre les centroïdes dans l'espace de projection et les observations dans l'espace défini par les composantes factorielles. Ce qui correspond à l'inertie intra-classe des observations dans l'espace réduit aux Q composantes factorielles. Tandis que la fonction objectif F_{RKM} (relation 3.2) montre que RKM recherche la partition des observations minimisant la somme des carrées des distances entre les observations et les "pseudo centroïdes" définis dans des sous-espaces engendrés par les vecteurs colonnes de la matrice A. En ce sens, les méthodes FKM et RKM peuvent être vues comme des méthodes de sélection locale (Cf. section 3.3). Les résultats fournis par ce type d'approche sont particulièrement bons si les données présentent une structure globale de corrélation, en d'autres termes, si l'information contenue dans la plupart des variables peut être portée par un nombre restreint de composantes factorielles par exemple. Or, dans la pratique, peu de données de

grande dimension présentent une structure globale de corrélation. Par ailleurs, les composantes obtenues sont fonction de toutes les variables initiales, ce qui en grande dimension engendre des difficultés pour l'interprétation qui peut être longue et fastidieuse. Cependant, comme dans l'exemple précédent, il peut arriver que certaines variables initiales aient des coefficients faibles voir nuls dans la combinaison linéaire définissant une composante factorielle. Alors, en ne retenant que les plus forts coefficients on simplifie le problème, une alternative serait l'utilisation des méthodes factorielles "sparse" qui recherchent des combinaisons comportant un nombre important de coefficients nuls, la classification de variables ou les méthodes de sélection de variables. La sélection de variables va alors permettre de ne conserver qu'un nombre restreint de variables initiales. On distingue dans la suite les méthodes de sélection globale des méthodes "subspace clustering" et de bipartitionnement qui sélectionnent les variables localement au niveau des classes.

3.3 Sélection globale de variables

Les méthodes de sélection globale des variables cherchent à découvrir des variables pertinentes, au sens d'un certain critère, dans un jeu de données. Ces méthodes sont généralement basées sur des critères de sélection définis sur les variables ou sur des sous-espaces de variables [Mitra *et al.* 2002; I. Guyon 2003; Alelyani *et al.* 2013]. Cette famille d'approches nécessite généralement la définition des étapes essentielles d'évaluation de la pertinence d'une variable, de la procédure de recherche des variables pertinentes ainsi que d'un critère d'arrêt. Cette procédure, couramment utilisée en apprentissage supervisé, est guidée par la connaissance d'une variable cible (les étiquettes des données par exemple). Au sens de la classification, on dira qu'une variable ou qu'un groupe de variables est pertinent si sa suppression ne dégrade pas les performances de classification. On distingue en apprentissage non-supervisé trois types de méthodes de sélection globale des variables : les approches "Filtres", les approches "enveloppantes" ou "symbioses" et les approches "intégrées".

3.3.1 Approches "Filtres"

Ces méthodes sélectionnent les variables indépendamment de la classification. Elles se basent généralement sur des scores définis pour chaque variable à partir d'un certain critère qui repose généralement sur les propriétés des données. Les variables ayant des scores élevés sont ensuite utilisées pour la classification. De nombreux critères de sélection des variables sont proposés pour cette catégorie d'approches. Zhao et Liu [2007]; He *et al.* [2005]

définissent des scores sur les variables grâce à la laplacienne d'une matrice de similarité entre les individus. Ils proposent dans la méthode SPECtral feature selection (SPEC) d'estimer la pertinence des variables par décomposition spectral de la matrice de similarité S entre les individus dont les entrées sont définies à l'aide d'une fonction noyau $S(z_i, z_{i'}) = \exp(-\frac{||z_i - z_{i'}||^2}{2\sigma^2})$. Les auteurs construisent ensuite un graphe G sur la matrice S dont la matrice laplacienne L et sa version normalisée $\tilde{L}$ [1] servent ensuite de base aux calculs des poids des variables. Motivés par la théorie des graphes qui stipule que la structure d'un graphe est contenue dans son spectre, les auteurs définissent des poids sur les variables z^j à l'aide de la fonction : $\phi(z^j) = \frac{z^{jT}\tilde{L}z^j}{z^{jT}Dz^j} = \sum_{i=1}^{N-1} \alpha_i^2 \lambda_i$ où λ_i est un vecteur propre de la matrice $\tilde{L}$ et α_i le cosinus de l'angle formé par le vecteur propre associé à la valeur propre λ_i et la variable z^j. Les quantités $\phi(z^j)$ fournissent ainsi des scores sur les variables qui permettent de choisir les t variables pertinentes pour la classification.

Dash et Liu [2000] utilisent la notion d'entropie pour définir des scores sur les variables. Les auteurs quantifient la contribution des variables à l'entropie globale E de la matrice Z définie par :

$$E(z^1, \ldots, z^p) = -\sum_{z^1} \ldots \sum_{z^p} P(z^1, \ldots, z^p) log(P(z^1, \ldots, z^p))$$

avec $P(z^1, \ldots, z^p)$ désignant la probabilité jointe du point $(z^1, \ldots, z^p)$. Ils calculent pour chaque variable j, le score E_j de l'ensemble $\mathcal{V} - \{z^j\}$,

$$E_j(z^1, \ldots, z^p) = -\sum_{z^1} \ldots \sum_{z^p} P(z^1, \ldots, z^{j-1}, z^{j+1} \ldots, z^p) log(P(z^1, \ldots, z^{j-1}, z^{j+1} \ldots, z^p))$$

Les variables les moins pertinentes pour la classification qui sont celles ayant les scores E_j les plus faibles sont alors supprimées.
L'inconvénient principal de ces méthodes reste le choix du seuil pour les scores. Ce problème est surmonté par les méthodes "Symbioses".

3.3.2 Approches "Symbioses"

Contrairement aux approches filtres qui ignorent totalement l'influence des variables sélectionnées sur la performance de l'algorithme d'apprentissage, les approches "symbioses ou enveloppantes" utilisent l'algorithme d'apprentissage comme une fonction d'évaluation de la qualité des variables sélectionnées. Elles commencent par sélectionner un sous-espace des variables. Puis, elles évaluent les performances d'une méthode de classification sur

1. $L = D - W; \tilde{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$, où D la matrice diagonale des degrés du graphe et W la matrice d'adjacence du graphe G

l'ensemble sélectionné. Ce processus est répété plusieurs fois jusqu'à l'obtention du sous-ensemble de variables donnant la meilleure partition au sens d'un certain critère. Les approches symbioses classiques étant très coûteuses en temps, Dash et Liu [2000] proposent la version symbiose de leur approche filtre à travers un processus itératif en p itérations définies par application d'une méthode de classification sur l'ensemble $Z'_j(N \times j)$. Z'_j contient les j premières variables ayant les scores les plus élevés définis par le critère d'entropie. Puis, ils évaluent les performances de la classification à l'aide du critère d'inertie. Le processus s'arrête à l'étape t lorsque $\forall j$ tel que $1 < t < j \leq p$ les performances de classification sur la matrice Z'_j restent stables.
Pour surmonter l'indépendance des variables sélectionnées par rapport à la classification dans les approches filtres et diminuer le temps de calcul des approches symbioses les approches dites intégrées ont été développées.

3.3.3 Approches "Intégrées"

Ces méthodes exécutent la sélection de variables pendant le processus de l'apprentissage. Le sous-ensemble de variables ainsi sélectionnées sera choisi de façon à optimiser le critère d'apprentissage utilisé. Ainsi, Huang *et al.* [2005] proposent une extension de la méthode des K-moyennes dans laquelle, les poids définis sur chaque variable en tenant compte de sa dispersion dans la classe servent à sélectionner les variables pertinentes pour la classification.

Dans une approche plus directe, Witten et Tibshirani [2010] intègrent dans l'algorithme des K-moyennes un processus de sélection des variables. Les auteurs définissent une nouvelle fonction objectif (relation 3.3) des K-moyennes dont la maximisation est équivalente à la minimisation du critère d'inertie 2.6 présentée dans la section 2.3.2 :

$$\mathcal{J} = \sum_{j=1}^{p} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{i'=1}^{N} d(z_i^j, z_{i'}^j) - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in c_k} d(z_i^j, z_{i'}^j) \right) \tag{3.3}$$

où n_k est le nombre d'observations dans la classe k. Un vecteur poids $\beta = (\beta_1, \ldots, \beta_p)$ inclus dans la relation 3.3 définit des scores sur les variables. Le problème consiste donc à optimiser la relation suivante :

$$\underset{c_1,\ldots,c_K,\beta}{max} \sum_{j=1}^{p} \beta_j \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{i'=1}^{N} d(z_i^j, z_{i'}^j) - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in c_k} d(z_i^j, z_{i'}^j) \right) \tag{3.4}$$

où β est assujetti aux contraintes $||\beta||^2 \leq 1$, $||\beta||_1 \leq s$ et $\beta_j \geq 0$ $\forall j$. Le choix idéal du paramètre s conduit à un vecteur binaire pour β et les variables non-influentes ne sont pas prises en compte dans l'algorithme.
Remarquons que si les paramètres β_j sont tous identiques la relation 3.4 est identique au critère des K-moyennes classique. La solution du problème convexe 3.4 relativement aux poids β_i est donnée par :

$$\beta = \frac{\delta(\mathcal{J}_+, \Delta)}{||\delta(\mathcal{J}_+, \Delta)||} \tag{3.5}$$

où x_+ est la partie positive de x et $\Delta = 0$ si $||\beta|| < s$ et $\Delta > 0$ sinon, donc $||\beta||_1 = s$. δ est un opérateur de seuil défini par $\delta(x, c) = sign(x)(|x - c|)_+$. Algorithmiquement, l'approche de sélection des variables pour du "sparse clustering" inclut une étape supplémentaire de calcul des paramètres β dans l'algorithme des K-moyennes. Basée sur le même principe de définition des poids β, Witten et Tibshirani [2010] ont aussi proposé la méthode "sparse hierachical clustering" qui permet de faire de la sélection de variables en classification hiérarchique ascendante.

3.4 Classification des variables

Cette technique est généralement utilisée dans le cadre de la recherche de la multicolinéarités entre les variables, de la réduction du nombre de variables lorsqu'il est trop important ou de la transformation des variables en dimension indépendantes. Comme en classification d'individus, il existe des méthodes hiérarchiques de classification des variables et des méthodes de partitionnement des variables. Dans le cas de la classification hiérarchique, les stratégies utilisées pour la classification des variables sont les mêmes que pour la classification d'individus et plusieurs indices de similarité ont été proposés dans la littérature Nakache et Confais [2004].
La technique de classification de variables parmi celles les plus couramment utilisées est la procédure VARCLUS (SAS). La méthode VARCLUS de SAS est basée sur la division successive des variables. Elle commence par réaliser une analyse en composantes principales des variables. Les deux premières composantes factorielles associées aux deux plus grandes valeurs propres si la seconde est supérieure à 1 sont retenues. Ensuite, chaque variable est affectée à la composante principale qui lui est la plus corrélée. Les groupes obtenus sont à leur tour divisés en deux selon le même principe tant que la seconde valeur propre est supérieure à 1.
Dans une approche directe de partitionnement, Vigneau et Qannari [2003] recherchent une

partition en L classes des variables en maximisant un critère exprimant la colinéarité entre les variables d'une classe : $T = N \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} cov^2(z^j, u_k)$ sous la contrainte $u_k u_{k'} = 1$ où $\delta_{kj} = 1$ si la variable $j \in c_k$ et 0 sinon, et cov^2 est la covariance entre la variable z^j et la variable latente u_k représentant la classe k. Après initialisation des classes, la variable latente u_k pour chaque classe est le premier vecteur de la matrice dont les variables sont restreintes aux variables de la classe k. Puis, dans un processus itératif, une variable z^j est affectée à la classe qui maximise le carré de sa covariance avec la variable latente u_k.
Plasse [2006] remplace la mesure de dépendance entre deux variables quantitatives par une mesure de dépendance entre une variable numérique et une qualitative pour le traitement des variables qualitatives. Elle maximise le coefficient de corrélation linéaire entre les indicatrices des variables qualitatives z^j_{bin} et la variable latente u_k :

$$T = N \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} r^2 (\sum_{l=1}^{q} a_l z^{jl}_{bin}, u_k)$$

où a_l est la moyenne de la variable u_k dans le bloc l et r^2 désigne le coefficient de corrélation linéaire, et z^{jl}_{bin} l'indicatrice de la modalité l de la variable qualitative j.

3.5 Les méthodes de sélection locale de variables

Les méthodes de réduction des dimensions confrontent l'utilisateur au choix difficile du nombre nécessaire de composantes factorielles à retenir, tandis que les méthodes de sélection imposent le choix d'un certain nombre de variables. Pour atténuer la perte d'information engendrée par la suppression des variables ayant des influences faibles sans être négligeable, de nombreux auteurs proposent une alternative à la sélection brute des variables à travers des principes de pondérations des variables. Par ailleurs, malgré leurs succès dans bien des domaines d'applications, les algorithmes de sélection globale des variables sont peu efficaces en présence de classes définies dans différents sous-espaces.

Les concepts de subspace-clustering et de co-clustering ou de bipartionnement ont donc été développés pour surmonter les limites des méthodes de sélection globale et de réduction des variables [Parsons *et al.* 2004; Kriegel *et al.* 2009].

3.5.1 Subpace clustering

Nous illustrons les motivations du *subpace clustering* à travers un jeu de données simulées. Il s'agit de 400 observations décrites par 3 variables reparties en 4 classes contenant

chacune 100 observations. Les classes 1 et 2 sont définies par un mélange gaussien de variance 0.2 dont les moyennes sont 0.5 et -0.5 sur la variable 1, 0.5 et 0.5 sur la variable 2, sur la variable 3 elles sont caractérisées par une loi normale centrée réduite. Les classes 3 et 4 sont simulées selon le même principe. Les figures 3.1(a), 3.1(b) et 3.1(c) représentent la projection des variables par rapport aux rangs des observations (en abscisse), puis les figures 3.1(d), 3.1(e) et 3.1(f) représentent les observations dans les plans définis par les variables prises deux à deux. Ces figures montrent qu'aucune variable ne permet de séparer convenablement les quatre classes. Alternativement, la suppression d'une variable fournit les graphes 3.1(d), 3.1(e) et 3.1(f). Les classes 'o' et '$\triangle$' se distinguent des autres dans le plan composé des variables 1 et 2 (3.1(d)) alors que les classes '$+$' et '$\times$' se superposent dans ce même plan, car la suppression de la variable 3 supprime le bruit sur ces deux classes. Par ailleurs, l'algorithme classique des K-moyennes fournit des performances faibles à cause de la prise en compte de l'ensemble des variables dans la classification alors que chaque classe est définie par seulement deux variables. Les figures 3.2(a), 3.2(b) et 3.2(c) illustrent les limites de la méthode des K-moyennes sur ce jeu de données simulées.
Le subspace clustering consiste alors à retrouver chaque classe dans un sous-espace composé de variables pertinentes et une variable peut être pertinente pour une ou plusieurs classes. Des heuristiques plus sophistiquées pouvant être regroupées en deux catégories sont alors développées pour déterminer de manière optimale les sous-espaces associés aux classes de la classification.

3.5.1.1 Les approches "Top-Down"

Cette catégorie d'approches détermine une première classification à l'aide de l'ensemble des variables. Un poids associé à chaque variable est ensuite utilisé dans une nouvelle phase d'un processus itératif pour réaffecter les observations aux classes. La difficulté principale dans cette catégorie est la définition du nombre de classes et du nombre de variables formant le sous-espace associé à une classe. On rencontre dans la littérature, les algorithmes standards tels que PROCLUS, COSA.
PROCLUS (PROjected CLUStering, [Aggarwal *et al.* 1999]) fut le premier algorithme de type "top-down" proposé dans la littérature. PROCLUS se présente comme une extension de l'algorithme des K-médoïdes au subspace clustering en déterminant les représentants les plus centraux des classes. Son processus se déroule en trois étapes : initialisation, itération et raffinement. La première phase utilise les techniques "gloutonnes" pour échantillonner et sélectionner les K-médoïdes potentiels. L'idée consiste ensuite à définir autour des médoïdes des groupes homogènes d'observations. Dans la phase d'itération, l'algorithme

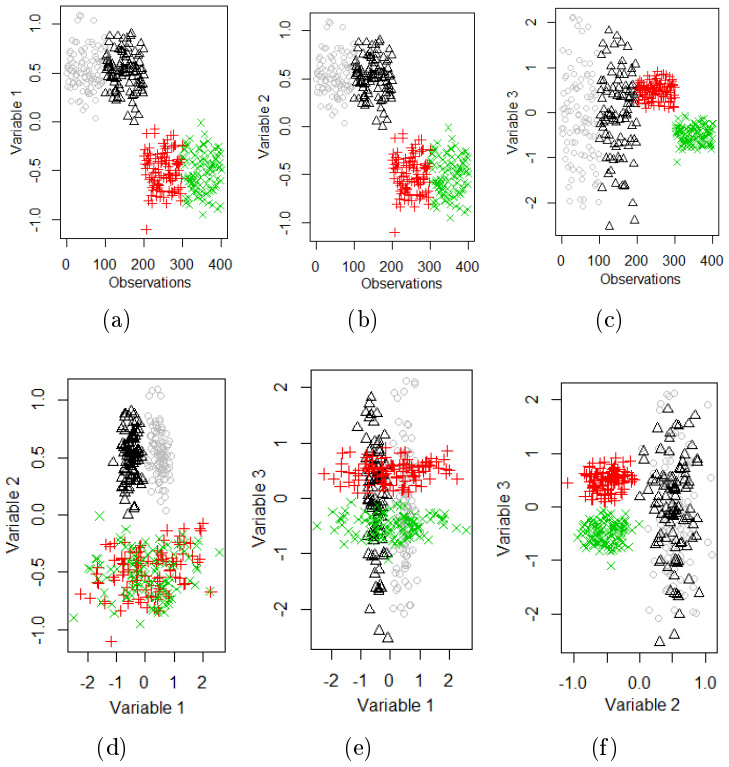


FIGURE 3.1 – Représentation des classes. Les figures 3.1(a), 3.1(b) et 3.1(c) représentent les classes par rapport à chaque dimension uniquement. Les figures 3.1(d), 3.1(e) et 3.1(f) représentent les classes dans les plans composés des variables prises deux à deux

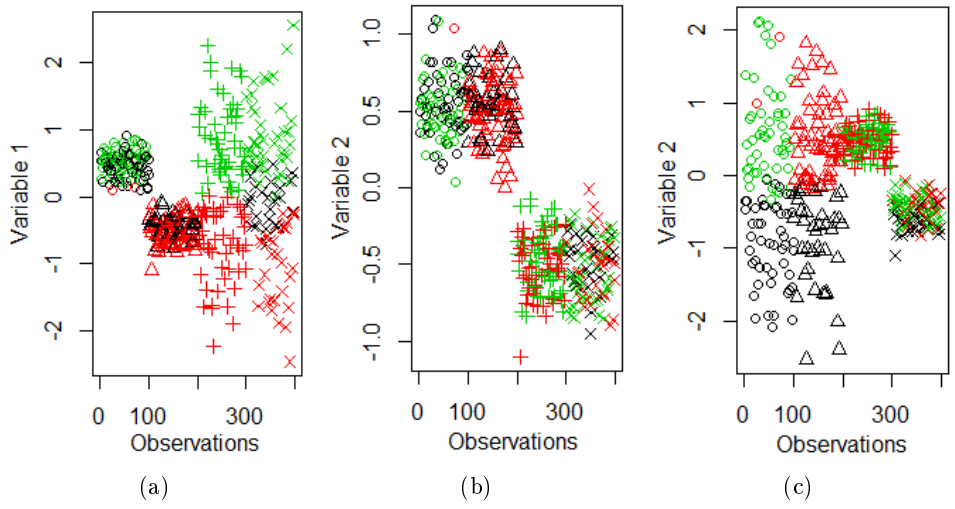


FIGURE 3.2 – La structure des classes fournies par application de l'algorithme des K-moyennes

calcule la pertinence des médoïdes à travers la moyenne des distances entre les observations d'une classe et le médoïde correspondant. Les médoïdes les moins pertinents sont ensuite remplacés par des nouveaux médoïdes choisis aléatoirement. Dans la phase de raffinement, pour chaque classe formée, un ensemble de variables de taille $l < p$ fixée a priori est choisi tel qu'il minimise la distance des observations au médoïde. Enfin PROCLUS réaffecte les observations aux nouveaux médoïdes. La procédure s'arrête lorsque la qualité d'un résultat ne change pas après un certain nombre de changements des médoïdes. Comme l'algorithme des K-moyennes, PROCLUS se spécialise dans la détermination des classes à structure sphérique qui sont cette fois représentées par un ensemble de médoïdes et de sous-espace correspondants.

ORCLUS étend la méthode PROCLUS à la recherche de classes dans des sous-espaces formés par des combinaisons linéaires des variables initiales. Cet algorithme contient les phases d'affectation des observations, de détermination des sous-espaces et de regroupement des classes proches dans l'espace. La phase d'affectation est identique à PROCLUS. Pour chaque classe, un sous-espace de vecteurs orthonormés est formé de vecteurs propres de la matrice de variance-covariance des observations qu'elle contient.
COSA (Clustering Objects on Subsets of Attributes, [Friedman et Meulman 2004]) est un algorithme itératif de type "top-down" qui associe à chaque observation des poids sur les dimensions. L'algorithme commence par affecter un poids identique à l'ensemble des variables, puis il détermine les k-plus proches voisins (knn) de chaque observation. Les poids forts, obtenus par optimisation d'un critère objectif, sont affectés aux variables dont la dispersion est faible dans les groupes de l'algorithme knn. Ce processus est alors répété jusqu'à la stabilisation des poids.

3.5.1.2 Les approches "Bottom-Up"

Les approches de type "Bottom-Up" utilisent les méthodes de classification basées sur un maillage de l'espace des observations en définissant pour chaque dimension un histogramme. Puis, les intervalles ayant une densité d'observations supérieure à un seuil fixé a priori définissent des classes pour chaque variable. Les auteurs font l'hypothèse que si un espace de q dimensions présente une forte densité d'observations alors tout espace composé de $q-1$ dimensions de cet espace est aussi dense. Ce principe conduit à des classes qui se chevauchent. Cette famille de méthodes comprend entre autres les méthodes : CLIQUE,

ENCLUS . L'algorithme CLIQUE (CLustering In QUest, [Agrawal *et al.* 1998]) est fondé sur la notion de densité, il recherche automatiquement des sous-espaces de plus grande dimensionnalité contenant des classes de forte densité. CLIQUE partitionne l'espace des observations en unités rectangulaires denses. Les unités denses voisines sur la grille sont regroupées pour former des classes. Étant donné un espace initial à p dimensions, CLIQUE procède comme suit :

- Découpage de chaque dimension de l'espace en I intervalles de même largeur.
- Détermination de l'ensemble des cellules denses dans l'ensemble des sous-espaces de l'espace des données.
- Détermination des classes comme un ensemble maximal de cellules denses contigües.

L'utilisation des régions à forte densité d'observations pour déterminer les classes permet à CLIQUE d'être indépendant du nombre K de classes fixé a priori dans la plupart des algorithmes de classification et de retrouver des classes de forme quelconque. De nombreuses variantes de CLIQUE ont été proposées. ENCLUS (ENtropy-based CLUStering) qui est une approche semblable à CLIQUE, utilise un critère basé sur la notion d'entropie pour sélectionner des sous-espaces de données denses d'observations. Comme CLIQUE, ENCLUS utilise des unités rectangulaires fixées à l'avance. Cependant, ENCLUS recherche directement des sous-espaces contenant potentiellement une ou plusieurs classes d'observations contrairement à CLIQUE.

3.5.1.3 Les approches de pondération des variables

Huang *et al.* [2005] dans W-K-Means puis Jing *et al.* [2007] dans Entropy weighting K-Means (EWKM) proposent de définir un système de pondération par modification de la fonction de coût associée à l'algorithme des K-Moyennes en y introduisant des poids. Dans la méthode EWKM, les auteurs minimisent simultanément, l'inertie intra-classe et maximise un terme d'entropie négatif dans le processus d'apprentissage. EWKM calcule pour chaque variable des poids inversement proportionnels à leur variance dans chaque classe. Le sous-espace de variables pertinents pour chaque classe est défini en se basant sur ces poids, facilitant ainsi l'interprétation des classes. Chen *et al.* [2012] étendent dans la méthode Feature Group K-means (FGKM), la méthode EWKM à la classification d'individus décrits par un grand nombre de variables structurées en blocs. Un second terme d'entropie négative défini au niveau des blocs permet d'établir des scores sur ces derniers et ainsi faire

de la sélection des blocs. La fonction objectif à minimiser associée à cet algorithme est :

$$\mathcal{J}(U,W,\alpha,\beta) = \sum_{k=1}^{K}(\sum_{i=1}^{N}\sum_{b=1}^{B}\sum_{z^j\in\mathcal{Z}^b} u_{ik}\alpha_{kb}\beta_{kbj}d(z_i^j,w_k^j) + \lambda\sum_{b=1}^{B}\alpha_{kb}log(\alpha_{kb}) + \eta\sum_{j=1}^{p_b}\beta_{kbj}log(\beta_{kbj})) \quad (3.6)$$

sous les contraintes : $\begin{cases} \sum_{k=1}^{K} u_{ik} = 1,\ u_{ik}\in\{0,1\} \\ \sum_{b=1}^{B}\alpha_{kb} = 1,\ 0\leq\alpha_{kb}\leq 1,\ 1\leq b\leq B, \\ \sum_{z^j\in\mathcal{V}^b}\beta_{kbj} = 1,\ 0\leq\beta_{kbj}\leq 1,\ 1\leq\ k\leq K \end{cases}$

où U est la matrice définissant la partition, W la matrice des centres de classe, α l'ensemble des poids α_{kb} des blocs b dans les classes et β l'ensemble des poids β_{kbj} des variables j du bloc b dans les cellules de la carte. La solution du problème de minimisation de la relation 3.6 s'obtient à l'aide du multiplicateur de Lagrange. Cette méthode est à la base de l'approche que nous proposons dans le chapitre 4.

Les méthodes présentées dans cette section ont l'inconvénient majeur de ne pas prendre en compte la corrélation entre les variables et privilégient généralement la partition des observations par rapport aux variables. Dans la section suivante, nous présentons les méthodes effectuant une double partition des lignes et des colonnes d'un tableau.

3.5.2 Bi-partitionnement

Les méthodes de Bi-partitionnement ou de Co-clustering recherchent une partition $\mathcal{C}$ des observations z_i et une partition $\mathcal{C}'$ des variables z^j dont les classes caractérisent les classes de la partition $\mathcal{C}$. Cette famille de méthodes a suscité beaucoup d'intérêts dans les domaines telles que l'analyse de données textuelles et la génétique où l'objectif est de définir des classes de documents par des classes de mots ou de gènes. Elles visent à obtenir des blocs individus/variables ou lignes/colonnes les plus homogènes selon des critères métriques ou probabilistes comme cela est illustré sur la figure 3.3. Cette figure montre l'intérêt de la classification croisée (à droite) qui fournit des blocs nets et illustratifs.

Il existe plusieurs algorithmes de bi-partitionnement [Charrad et Ben Ahmed 2011] :

- Les méthodes basées sur des algorithmes de partitionnement simple : ces méthodes appliquent un algorithme de classification simple sur les lignes et sur les colonnes séparément. Les bi-classes sont construites à partir des classes obtenues sur les lignes et sur les colonnes. Cette famille inclut les méthodes Croeuc, Croki2, Crobin et Cromul présentées par Govaert [1983, 1984]

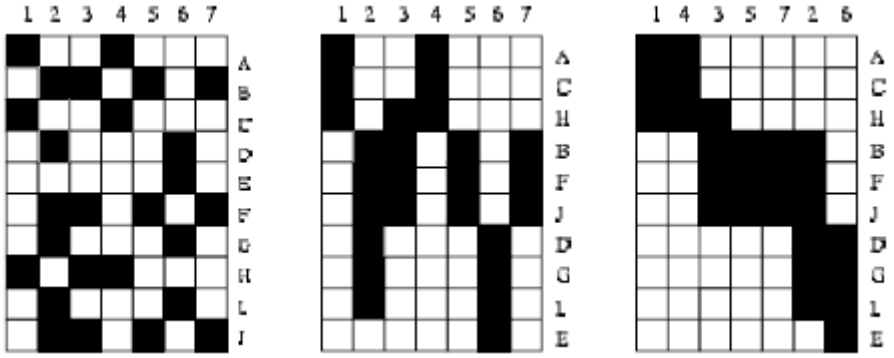


FIGURE 3.3 – Classification simple et classification croisée [Govaert 1983]

– Les méthodes divisives : ces méthodes procèdent par découpage itératif afin d'aboutir à des classes. Cette famille inclut les méthodes one-way splitting et two-way splitting de Hartigan [1975].

3.5.2.1 Méthodes de bi-partitionnement simple

Govaert [1983, 1984] définit les méthodes Croeuc, Croki2, Crobin et Cromul pour la classification croisée d'individus décrits respectivement par des variables continues, binaires, pour le découpage en bloc d'un tableau de contingence et les tableaux de variables qualitatives. Défini pour des données continues, l'algorithme Croeuc optimise le critère métrique suivant :

$$\mathcal{J}_{croeuc}(\mathcal{C},\mathcal{C}',W)=\sum_{k=1}^{K}\sum_{l=1}^{K'}\sum_{z_i\in c_k}\sum_{z^j\in c'_l}(z_i^j-w_k^l)^2 \tag{3.7}$$

où $c_k \in \mathcal{C}$, $c'_l \in \mathcal{C}'$, et W désignent respectivement une classe de la partition des lignes en K classes, une classe de la partition des colonnes en K' classes et W une matrice $K \times K'$ correspondant au résumé de l'information de Z ; w_k^l est le représentant d'un bloc dans Z. La recherche des partitions $\mathcal{C}$ et $\mathcal{C}'$ se fait grâce à un algorithme itératif utilisant deux fois l'algorithme des K-moyennes. Dans l'étape 2 de l'algorithme Croeuc ci-dessous, on applique l'algorithme de classification des K-moyennes sur les lignes en bloquant la partition sur les colonnes, puis sur les colonnes en bloquant la partition sur les lignes. Ce qui revient à optimiser alternativement les critères suivants (déduits de la relation 3.7 en fixant successivement les partitions $\mathcal{C}$ et $\mathcal{C}'$) :

$$\mathcal{J}_{croeuc1}(\mathcal{C},W/\mathcal{C}')=\sum_{k=1}^{K}\sum_{z_i\in c_k}\sum_{l=1}^{K'}\#c'_l(y_i^l-w_k^l)^2 \tag{3.8}$$

où $y_i^l = \frac{\sum_{z^j \in c'_l} z_i^j}{\#c'_l}$ et

$$\mathcal{J}_{croeuc2}(W, \mathcal{C}'/\mathcal{C}) = \sum_{l=1}^{K'} \sum_{z^j \in c'_l} \sum_{k=1}^{K} \#c_k (x_k^j - w_k^l)^2 \tag{3.9}$$

où $x_k^j = \frac{\sum_{z_i \in c_k} z_i^j}{\#c_k}$ et $\#x$ désigne le cardinal de l'ensemble x. Notons que cet algorithme se réduit à l'algorithme des K-moyennes classique sur les lignes (respectivement sur les colonnes) lorsque $K = N$ (respectivement $K' = p$).

Algorithme : Croeuc

Entrée : $\mathcal{Z}$, le nombre K de classes en lignes et le nombre K' de classes en colonnes.

1. Initialiser les partitions $\mathcal{C}^0$, $\mathcal{C}'^0$ et l'ensemble W^0.
2. A l'itération t+1, calculer les partitions et la matrice résumé $(\mathcal{C}^{t+1}, \mathcal{C}'^{t+1}, W^{t+1})$ à partir de $(\mathcal{C}^t, \mathcal{C}'^t, W^t)$;
   (a) Calculer $(\mathcal{C}^{t+1}, \mathcal{C}'^t, W')$ à partir de $(\mathcal{C}^t, \mathcal{C}'^t, W^t)$ en optimisant la relation 3.8 à l'aide de la méthode K-moyennes.
   (b) Calculer $(\mathcal{C}^{t+1}, \mathcal{C}'^{t+1}, W^{t+1})$ à partir de $(\mathcal{C}^{t+1}, \mathcal{C}'^t, W')$ en optimisant la relation 3.9 à l'aide de la méthode K-moyennes.
3. Recommencer l'étape 2 jusqu'à la convergence de l'algorithme.

Sortie : Les partitions $\mathcal{C}$ et $\mathcal{C}'$

Lorsque les données sont binaires ($z_i^j \in \{0, 1\}$), le critère à minimiser de la relation 3.7 devient :

$$\mathcal{J}_{Crobin}(\mathcal{C}, \mathcal{C}', A) = \sum_{k=1}^{K} \sum_{l=1}^{K'} \sum_{z_i \in c_k} \sum_{z^j \in c'_l} |z_i^j - a_k^l| \tag{3.10}$$

où $a_k^l \in \{0, 1\}$ et A la matrice binaire résumée de la matrice binaire Z.

L'algorithme associé est quasiment identique à celui de Croeuc. Dans les étapes 2(a) et 2(b) de l'algorithme Crobin, pour trouver les partitions optimales $\mathcal{C}^{t+1}$ et $\mathcal{C}'^{t+1}$, on utilise doublement l'algorithme des nuées dynamiques Diday [1971] avec la distance L_1 pour optimiser les critères 3.11 et 3.12 suivants déduits de la relation 3.10

$$\mathcal{J}_{Crobin1}(\mathcal{C}, W/\mathcal{C}') = \sum_{k=1}^{K} \sum_{z_i \in c_k} \sum_{l=1}^{K'} (y_i^l - \#c'_l a_k^l)^2. \tag{3.11}$$

où $y_i^l = \sum_{z^j \in c'_l} z_i^j$ et

$$\mathcal{J}_{Crobin2}(W, \mathcal{C}'/\mathcal{C}) = \sum_{l=1}^{K'} \sum_{z^j \in c'_l} \sum_{k=1}^{K} (x_k^j - \#c_k a_k^l)^2. \tag{3.12}$$

où $x_i^k = \sum_{z_i \in c_k} z_i^j$. À la convergence de cet algorithme, les centres de classes sont caractérisées par la valeur 0 ou 1 la plus fréquente dans les blocs.

Algorithme : Crobin

Entrée : $\mathcal{Z}$, le nombre K de classes en lignes et le nombre K' de classes en colonnes.

1. Initialiser les partitions $\mathcal{C}^0$, $\mathcal{C}'^0$ et la matrice A.
2. Calculer les partitions et la matrice résumé $(\mathcal{C}^{t+1}, \mathcal{C}'^{t+1}, A^{t+1})$ à partir de $(\mathcal{C}^t, \mathcal{C}'^t, A^t)$;
   (a) Calculer $(\mathcal{C}^{t+1}, \mathcal{C}'^t, A')$ à partir de $(\mathcal{C}^t, \mathcal{C}'^t, A^t)$ en optimisant la relation 3.11 à l'aide de l'algorithme des nuées dynamiques sur la matrice y_i^l $(N \times K')$, avec des noyaux de la forme $(\#c'_1 a_k^1, \ldots, \#c'_{K'} a_k^{K'})$
   (b) Calculer $(\mathcal{C}^{t+1}, \mathcal{C}'^{t+1}, A^{t+1})$ à partir de $(\mathcal{C}^{t+1}, \mathcal{C}'^t, A')$ en optimisant la relation 3.12 à l'aide de l'algorithme des nuées dynamiques sur la matrice x_i^k $(K \times p)$, et en recherchant les noyaux de la forme $(\#c_1 a_1^l, \ldots, \#c_K a_K^l)$
3. Recommencer l'étape 2 jusqu'à la convergence de l'algorithme.

Sortie : Une réorganisation des lignes et des colonnes fournissant les blocs homogènes de 1 ou 0.

Enfin, l'algorithme Croki2 traite le cas où Z est une matrice de contingence (Cf. section 2.4). Cet algorithme recherche dans un processus itératif alterné la partition des lignes de $\mathcal{Z}$ en K classes et la partition des colonnes de $\mathcal{Z}$ en K' à l'aide du critère de convergence suivant basé sur la métrique du KHI2 :

$$\chi^2(C, C') = \sum_{k=1}^{K} \sum_{l=1}^{K'} \frac{(f_{kl} - f_{k.} f_{.l})^2}{f_{k.} f_{.l}}$$

avec $f_{kl} = \frac{n_{kl}}{N}$. Fondé sur ce critère, l'usage de l'algorithme des nuées dynamiques pour optimiser un critère de convergence propre aux lignes et aux colonnes de la matrice de contingence fournit les partitions en lignes et en colonnes recherchées.

Optimisation en lignes. On bloque la partition en colonnes $\mathcal{C}'$ et on travaille que sur la partition en lignes $\mathcal{C}$. Un nouveau tableau de contingence U est défini à partir du tableau initial Z tel que

$$u_{kl} = \sum_{j \in c'_l} n_{kj}$$

Et on cherche donc à minimiser le critère suivant :

$$\mathcal{J}_{Croki1}(\mathcal{C}) = \sum_{k=1}^{K} \sum_{z_i \in c_k} u_{i.} \sum_{l=1}^{K'} \frac{(u_{kl} - g_{kl})^2}{u_{.l}} \tag{3.13}$$

Optimisation en colonnes. Contrairement à l'étape précédente, la partition en lignes $\mathcal{C}$ est fixée. Un nouveau tableau de contingence V est défini à partir du tableau initial tel que

$$v_{kl} = \sum_{i \in c_k} n_{kj}$$

Et on cherche donc à minimiser le critère suivant :

$$\mathcal{J}_{Croki2}(\mathcal{C}) = \sum_{l=1}^{K'} \sum_{j \in c'_k} v_{.j} \sum_{k=1}^{K} \frac{(v_{kl} - g_{kl})^2}{v_{k.}} \tag{3.14}$$

Cromul, une version modifiée de la méthode Croki2, est destinée à l'analyse de tableaux de variables qualitatives ou de questionnaires. Son principe consiste à appliquer Croki2 au tableau disjonctif complet associé aux questionnaires.
Malheureusement, les algorithmes Croeuc, Croki2, Crobin et Cromul requièrent la connaissance du nombre de classes en lignes et en colonnes. Ces méthodes proposées par Govaert [1984], reposent sur des critères métriques différents suivant le type des données. Nadif et Govaert [1993] présentent un formalise du problème de la classification croisée sous l'approche modèle de mélange pour mieux appréhender les résultats fournis par ces algorithmes et donner une justification théorique de ceux-ci. Nous ne présenterons pas ce formalisme dans cette thèse. Cependant, on pourra se référer aux extensions des algorithmes Croeuc et Crobin proposées par Jollois [2003] pour plus de détails.

3.5.2.2 Les méthodes divisives

Hartigan [1975] propose dans l'algorithme divisif one-way splitting un découpage en blocs homogènes des objets. Cet algorithme se concentre principalement sur la partition des objets, en essayant de construire des classes de telle manière que les variables aient une variance intra-classe inférieure à un certain seuil fixé. L'idée de base de l'algorithme est de n'utiliser que les variables ayant une variance supérieure au seuil dans une classe donnée pour découper cette classe. Lorsque les données sont directement comparables d'une variable à une autre, Hartigan [1975] propose un deuxième algorithme divisif, Two-way splitting, qui choisit à chaque étape entre une division de l'ensemble des objets et une

division de l'ensemble des attributs. Ce choix est basé sur la réduction au maximum de l'hétérogénéité du groupe d'objets ou de variables à diviser. Afin de respecter les contraintes hiérarchiques imposées pour cet algorithme, les divisions effectuées à une étape ne sont jamais remises en cause aux étapes suivantes. L'avantage de cet algorithme est qu'il ne nécessite pas de savoir à l'avance le nombre de blocs à obtenir.

3.6 Recherche de consensus en classification

Dans le chapitre 2 et dans la première partie de ce chapitre nous avons montré qu'il existe une variété de méthodes de classification : les méthodes des K-moyennes, les méthodes neuronales, les méthodes de classification hiérarchique, etc. Cependant, il est bien connu qu'il n'existe pas une méthode capable de retrouver la meilleure partition indépendamment du jeu de données.

Face à la grande diversité des méthodes de classification, l'application de différents algorithmes de partitionnement sur un même jeu de données peut fournir des partitions pouvant être totalement différentes. Par ailleurs, la plupart des méthodes de classification reposent sur des algorithmes itératifs sensibles à leurs paramètres d'initialisation tels que les centres initiaux pour les K-moyennes ou la taille du voisinage pour les cartes SOM. En apprentissage non-supervisé, le choix d'une de ces partitions dépend alors du critère de validation utilisé qui est lui-même dépendant de l'objectif à atteindre.

Dans ce contexte, s'inspirant du succès des méthodes d'agrégation de modèles en apprentissage supervisé, émerge l'idée de *cluster ensemble* ou de *clustering aggregation*.Cette idée reprend les concepts plus anciens de recherche d'un consensus de partitions proposés par Régnier [1983] et repris par Gordon et Vichi [1998].

Ces méthodes d'ensemble cluster consistent à combiner plusieurs partitions d'un même jeu de données pour améliorer les performances des classifications [Régnier 1983; Breiman 1996; Gordon et Vichi 1998; Strehl et Ghosh 2002]. Elles reposent sur l'idée que le processus de fusion permet de compenser les éventuelles erreurs d'un algorithme et qu'une décision d'un groupe ou ensemble est plus fiable qu'une décision individuelle. Autrement dit, les différentes partitions d'un même ensemble d'observations fournissent des informations complémentaires dont la synthèse améliorerait les performances globales de classification de l'ensemble. Topchy *et al.* [2004b] présentent des arguments théoriques montrant notamment

que la partition finale π^* converge vers la partition réelle sous-jacente aux données lorsque l'ensemble Π des partitions dont on cherche le consensus devient grand.

Soit $\Pi = \{\pi_1, \ldots, \pi_B\}$ l'ensemble de B partitions de l'ensemble $\mathcal{Z}$. Le but des méthodes d'agrégation de classifications est de trouver une partition π^* représentant au mieux la structure des données $\mathcal{Z}$ à partir de l'ensemble des informations disponibles dans les B partitions.
La partition π^* communément appelée *consensus*, s'obtient à travers l'optimisation d'une fonction $\mathcal{J}$ définie à partir des partitions π^b. La partition π^* peut alors être vue comme la partition centrale associée à toutes les partitions π^b choisies parmi l'ensemble Π^* des partitions de $\mathcal{Z}$ [Guénoche 2011].

La recherche exhaustive de π^* parmi l'ensemble $\Pi_K \subset \Pi^*$ des partitions en K classes des objets est impossible. Le nombre de partitions formant Π_K étant dissuasif : $m = \frac{1}{K!}\sum_{l=1}^{K}\binom{K}{l}(-1)^{K-l}k^n$ soit 171 798 901 possibilités de former 4 groupes à partir de 16 objets. On se contente alors de rechercher π^* comme une combinaison des partitions de Π qui doit être choisi de façon à obtenir la plus grande diversité entre les partitions π^b. Cela permet d'avoir de meilleures performances pour la partition consensus π^*.
Les méthodes de *cluster ensemble* se déroulent en deux principales étapes : génération ou diversification des partitions et agrégation des partitions permettant de fournir la partition consensus. La figure 3.4 illustre les deux étapes du processus de *cluster ensemble*.
Dans la littérature, plusieurs auteurs définissent des propriétés souhaitées pour la partition consensus que nous pouvons résumer ci-dessous [Jain *et al.* 1999b; Topchy *et al.* 2004a] :

1. **Robustesse**, la méthode de recherche de consensus adoptée doit avoir en moyenne des performances meilleures que celles des partitions de Π
2. **Consistance**, la partition consensus doit être proche de la plupart des partitions de l'étape de diversification.
3. **Stabilité**, la partition consensus doit être insensible à de petites variations des partitions de l'étape de diversification, au bruit et aux observations aberrantes.
4. **Nouveauté**, la partition consensus doit mettre en évidence des propriétés des données qui sont difficilement atteignables par chaque partition de l'ensemble Π.

Les méthodes de recherche de consensus de partitions se distinguent selon les stratégies de génération des partitions initiales et selon les fonctions objectif utilisées dans la phase d'agrégation.

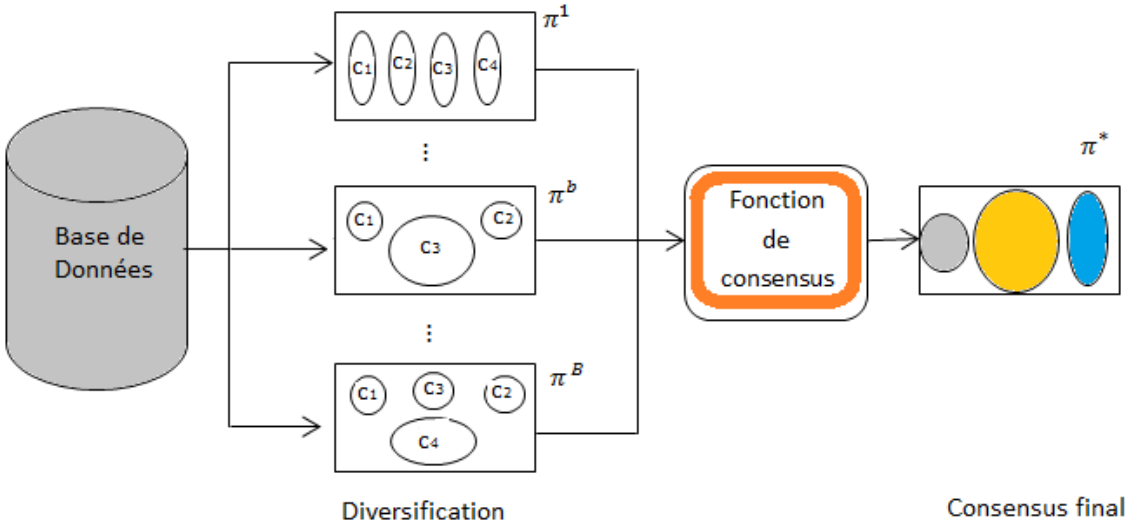


FIGURE 3.4 – Processus de recherche de consensus de partitions

3.6.1 L'ensemble des partitions

De manière générale, il n'existe pas de contrainte particulière sur le processus de génération des partitions initiales. Elles peuvent être obtenues par variation des paramètres d'un algorithme, par hybridation (méthodes multi stratégies) ou par ré-échantillonnage. Cependant, le consensus issu de l'agrégation est conditionné par ces partitions. Elles déterminent dans une grande mesure la fonction de consensus à utiliser notamment des partitions redondantes, trop fortement corrélées ou au contraire trop différentes peuvent nécessiter une prise en compte particulière. Les différentes procédures peuvent être regroupées comme ci-dessous :

- **Ensemble de partitions défini par variation des paramètres** : l'ensemble Π est constitué de partitions fournies par K applications d'un même algorithme sur l'ensemble $\mathcal{Z}$. Chaque application est associée à un unique système de paramètres. Dans ce contexte, différentes initialisations des centres de classe de l'algorithme des K-moyennes ou des K-modes sont couramment utilisées ([Strehl et Ghosh 2002; Fred et Jain 2003; Iam-On *et al.* 2008]). D'autres algorithmes de type K-médoïdes, tels que PAM (Partitioning Around Medoïds) qui contrairement à la méthode des K-moyennes définit les classes comme des sous-ensembles d'observations proches des medoïdes permettent aussi la diversification [Halkidi *et al.* 2001].
- **Ensemble de partitions défini par hybridation** : ce type d'ensemble s'obtient généralement par application de plusieurs algorithmes sur le même jeu de données

Strehl et Ghosh [2002]; Iam-On *et al.* [2008].

– **Ensemble de partitions défini sur des sous-ensembles** :
un algorithme de partitionnement est appliqué sur différents ensemble de variables. Ces derniers peuvent être des groupes de composantes factorielles (choix aléatoire de k composantes factorielles parmi p) ou des groupes ou blocs de variables initiales [Strehl et Ghosh 2002] décrivant les mêmes individus. La diversification s'obtient aussi selon le principe de ré-échantillonnage du bagging : un algorithme de partitionnement est appliqué sur différents échantillons bootstrap [Dudoit et Fridlyand 2003].

Ces différents mécanismes de diversification sont résumés dans la figure 3.5.

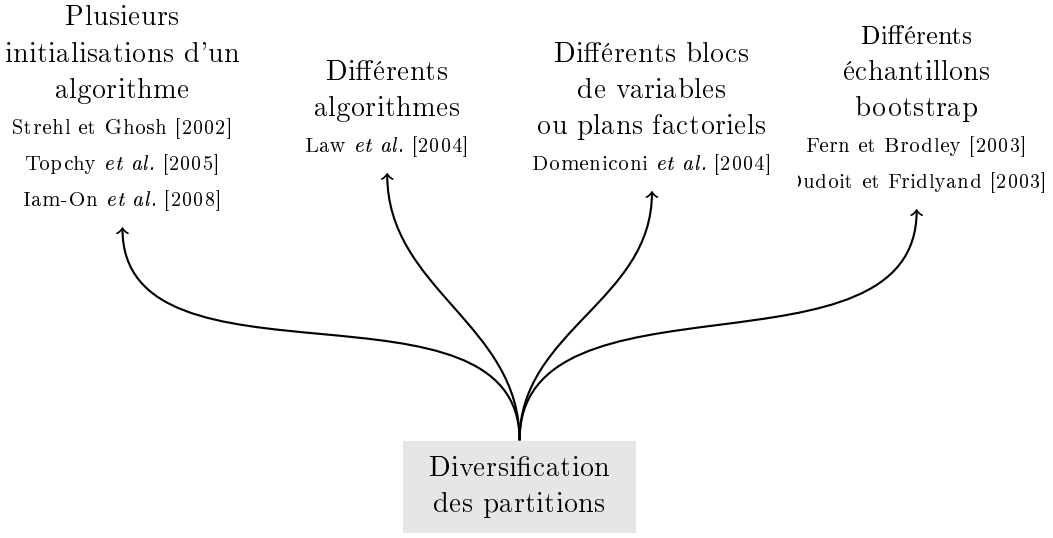


FIGURE 3.5 – Diagramme des principales approches de détermination de l'ensemble de diversification

La section suivante présente des approches d'agrégation permettant de déterminer la partition π^*.

3.6.2 Fonctions consensus

L'étape d'agrégation des partitions, commune à toutes les méthodes de *cluster ensemble*, repose sur une fonction consensus qui doit être définie de façon appropriée afin d'améliorer les performances de classification de la partition finale π^*.

Plusieurs représentations des partitions formant l'ensemble de diversification sont possibles : chaque partition peut être associée à une variable catégorielle dont les modalités sont les labels de classes, au tableau disjonctif contenant les indicatrices des modalités ou à

la matrice d'adjacence des partitions. Les méthodes diffèrent selon la représentation utilisée et la manière d'obtenir le consensus. On distingue les méthodes basées sur un principe de vote [Dudoit et Fridlyand 2003; Nguyen et Caruana 2007; Ayad et Kamel 2008], celle basée sur le partitionnement d'une méta-matrice de similarité Co définie à partir des matrices d'adjacence des partitions [Fern et Brodley 2003, 2004; Li *et al.* 2007] et celles qui obtiennent la partition centrale ou médiane introduite par Régnier [1983] en maximisant une mesure d'association telle que l'indice de Rand entre elle et les partitions initiales [Gordon et Vichi 1998; Krieger et Green 1999].

Dans la majorité des méthodes, les éventuelles différences de qualité des partitions initiales, ni leurs relations ne sont prises en compte explicitement. Les méthodes de consensus pondérés permettent de pallier à ces limites en associant aux partitions initiales des poids fixes ou adaptatifs.

Dans le cas particulier où l'ensemble de diversification est obtenu par application d'un algorithme de partitionnement topologique, plusieurs auteurs ont proposé des heuristiques de recherche de consensus de cartes topologiques dans l'objectif de prendre en compte la structure topologique des observations [Georgakis *et al.* 2005; Baruque *et al.* 2007].

3.6.2.1 Consensus fondé sur un principe de Vote

Lorsque les partitions sont considérées comme des variables catégorielles, un problème important dans les approches par vote est la nécessité de mettre en correspondance les différentes classes des partitions initiales. En effet, les partitions étant obtenues de manière indépendante, il n'existe aucune garantie quant à la correspondance entre la classe 1 d'une partition π^a avec la classe 1 d'une autre partition π^b par exemple. Ce problème est d'autant plus délicat que les nombres de classes des partitions diffèrent. Le ré-étiquetage consiste à fixer une partition π de base, puis les classes des autres partitions sont mises en correspondance en observant le recouvrement des classes par maximisation de $\sum_{i=1}^{N} 1_{(\tau^b(\pi^b(z_i))=\pi(z_i))}$ où $\tau^b \in \mathcal{P}$ est une permutation des labels de la partition π^b.
Un vote au sein des partitions "ré-étiquetées" produit le consensus. Ayad et Kamel [2008] présentent un algorithme de vote cumulatif ainsi qu'une méthode de vote pondéré qui permet de calculer une densité de probabilité résumant les partitions initiales. Dudoit et Fridlyand [2003] déterminent les classes des observations dans le consensus par le mode de l'ensemble suivant : $y_i = \{\tau^1(\pi^1(z_i)), \ldots \tau^B(\pi^B(z_i))\}$, où $\tau^b(\pi^b(z_i))$ représente le label

de la classe d'appartenance de l'observation z_i dans la partition de correspondance de π^b relativement à la partition de base. Les partitions étant obtenues par ré-échantillonnage bootstrap, les auteurs évaluent ensuite la confiance du consensus par la proportion d'éléments de l'ensemble Π définissant pour chaque observation sa classe d'appartenance. En partant de ce principe de vote, plutôt dur, [Nguyen et Caruana 2007] proposent "Iterative Voting Consensus" un algorithme itératif en deux phases dans lequel les centres de classe de la partition consensus sont actualisés. Les auteurs définissent l'ensemble $Y = \{y_1, \dots y_N\}$ avec $y_i = \{\tau^1(\pi^1(z_i)), \dots \tau^B(\pi^B(z_i))\}$, où $\tau^b(\pi^b(z_i))$ est le label de z_i après la permutation des labels de π^b, puis initialisent la partition centrale π^* en K classes. À chaque itération, on définit pour une classe c_k^* de π^*, l'ensemble $P_k^b = \pi_k^b(c_k^*)$ correspondant aux labels pris par les observations de la classe c_k^* dans la partition π^b et par $y_k^* = \{mode(P_k^1), \dots, mode(P_k^B)\}$. À chaque étape la classe d'appartenance d'une observation est définie par le mode de cette classe puis, les observations z_i sont de nouveau affectées aux classes de π^* par la relation $\pi^*(z_i) = \underset{k}{argmin} \sum_{b=1}^{B} 1_{(y_{kb}^* \neq y_{ib})}$. Le processus s'arrête lorsque π^* se stabilise.
La plupart de ces méthodes nécessitent une partition de base pour le ré-étiquetage des classes. Cependant, le choix de l'heuristique de sélection de la partition de base a d'importantes conséquences sur les résultats obtenus et détermine notamment le choix du nombre de classes. Pour résoudre ce problème, Fred [2001] proposent une méthode de vote fondée sur l'usage des matrices d'adjacence associées aux partitions :

$$Co_\pi(z_i, z_i') \equiv \begin{cases} 1, \ si \ \pi(z_i) = \pi(z_{i'}) \\ 0, \ si \ \pi(z_i) \neq \pi(z_{i'}) \end{cases} \tag{3.15}$$

La moyenne des B matrices de similarité fournit la matrice de connectivité globale Co des observations.

$$Co(z_i, z_{i'}) = \frac{1}{B} \sum_{b=1}^{B} Co_{\pi^b}(z_i, z_{i'}); \ i, i' = 1, \dots, N. \tag{3.16}$$

Remarquons que dans le cas particulier où la diversification est obtenue par partitionnement de plusieurs échantillons bootstrap, certaines observations peuvent se répéter dans les échantillons bootstrap compte tenu du tirage avec remise. Dudoit et Fridlyand [2003] redéfinissent les entrées de la matrice Co en tenant compte de la fréquence d'apparition des observations :

$$Co(i, j) = \frac{a_{ij}}{m_{ij}}$$

avec $a_{ii'} = \sum_{b=1}^{B} 1_{(z_i \in \pi^b, z_{i'} \in \pi^b, \pi^b(z_i) = \pi^b(z_{i'}))}$ et $m_{ii'} = \sum_{b=1}^{B} 1_{(z_i \in \pi^b, z_{i'} \in \pi^b)}$.

Fred [2001] fixe un seuil t=0.5 afin de définir les classes à partir de la matrice de connectivité. Les classes s'obtiennent en regroupant dans la même classe les observations pour lesquelles la connectivité est supérieure à 0.5.

3.6.2.2 Consensus fondé sur le partitionnement d'une méta-matrice de similarité

En considérant la matrice de connectivité Co comme une nouvelle matrice de similarité, il est possible d'appliquer un algorithme de classification pour construire le consensus π^*.

[Fred et Jain 2005] proposent d'appliquer une CAH avec la stratégie d'agrégation *lien simple* sur la matrice Co. La figure 3.6 montre que cette démarche permet notamment de retrouver les classes ayant des structures spirales et bien séparées (cf. figure 3.6). Cependant, l'utilisation d'une classification ascendante hiérarchique (CAH) sur la matrice des connectivités Co hérite de ses limites, a savoir sa complexité d'ordre 2 par rapport au nombre d'observations $O(N^2)$ et le problème du choix de la stratégie d'agrégation.

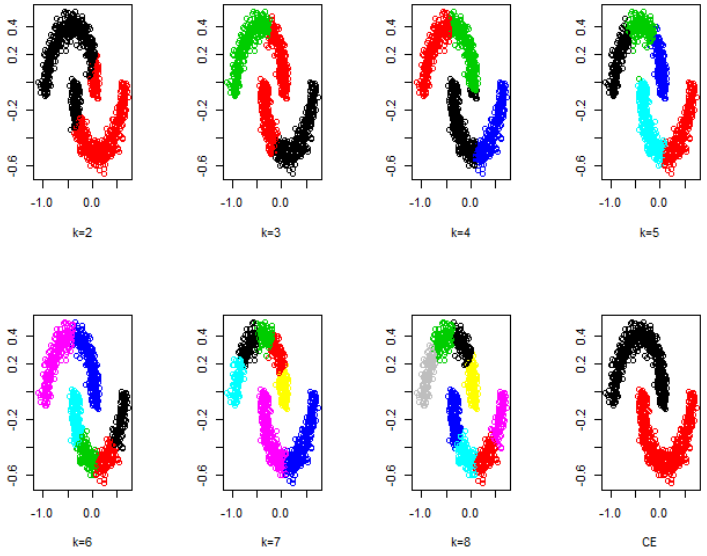


FIGURE 3.6 – Exemple de partition consensus en utilisant la matrice des co-associations sur 7 classifications obtenues en faisant varier le nombre de classes de 2 à 7 de la méthode des K-moyennes ; CE correspond à la classification consensus

Li *et al.* [2007] proposent une méthode de factorisation de la matrice non-négative Co pour déterminer le consensus. La notion de factorisation de matrice non-négative (Non-negative Matrix Factorization (NMF)) fait référence à la décomposition en valeurs singu-

lières d'une matrice M comme étant le produit deux matrices Q et S tel que $M \approx QSQ^T$ sous condition que les matrices Q et S soient non-négatives[2]. En définissant la distance entre deux partitions π^a et π^b par :

$$\Gamma(\pi^a, \pi^b) = \sum_{i,j}^{N} \gamma_{i,j}(\pi^a, \pi^b) \tag{3.17}$$

où $\gamma_{i,j}(\pi^a, \pi^b) = |Co_{\pi^a}(z_i, z_j) - Co_{\pi^b}(z_i, z_j)|$ la partition consensus est dans ce cas solution du problème d'optimisation suivant :

$$\pi^* = \underset{\pi \in \Pi^*}{argmin} \frac{1}{B} \sum_{b=1}^{B} \sum_{i,j}^{N} |Co_{\pi}(z_i, z_j) - Co_{\pi^b}(z_i, z_j)|$$

Ce qui revient à rechercher la matrice U telle que

$$Co_{\pi^*} = \underset{U}{min} ||Co - U||_F^2$$

avec $||.||_F$ désignant la norme de Frobenius. En recherchant U sous la forme QSQ^T et sous contrainte on obtient le problème :

$$Co_{\pi^*} = \underset{Q \geq 0,\ S \geq 0}{min} ||Co - QSQ^T||^2 \quad s.t. \quad Q^TQ = I \tag{3.18}$$

A chaque étape t du processus, les matrices Q et S valent :

$$Q_t = Q_{t-1} \sqrt{\frac{CoQ_{t-1}S_{t-1}}{Q_{t-1}Q_{t-1}^T CoQ_{t-1}S_{t-1}}} \quad et \quad S_t = Q_{t-1} \sqrt{\frac{Q_{t-1}^T CoQ_{t-1}}{Q_{t-1}^T Q_{t-1}S_{t-1}Q_{t-1}^T Q_{t-1}}}$$

Le consensus est fourni par la matrice Q. En modifiant légèrement la structure de la matrice Co, dans un second algorithme, Li et Ding [2008] pondèrent les différentes partitions afin d'éliminer les partitions redondantes à travers des poids α_b et telle que la matrice Co devient $\tilde{Co} = \sum_{b=1}^{B} \alpha_b Co_{\pi^b}$. Le consensus et les poids α s'obtiennent avec un processus itératif alternant la recherche de la matrice Q et la recherche des poids α.

En ce basant toujours sur la matrice Co , le problème de consensus peut être défini comme un problème de partitionnement de graphes ou d'hypergraphes. Dans ce cas, les partitions π^b sont représentées par des matrices disjonctives complètes $H_{(i)}$ dont les colonnes définissent les classes (cf. tableau 3.1).

2. Une matrice M est non-négative si $\forall\ i,\ j,\ M(i,j) \geq 0$

La concaténation horizontale des matrices $H_{(i)}$, correspondant aux partitions π^b de l'ensemble Π fournit la matrice d'adjacence H d'un hypergraphe à N sommets et $\sum_{b=1}^{B} K_{(b)}$ hyperarêtes. Les colonnes h_{ij} définissent les hyperarêtes où 1 indique que cette hyperarête contient le sommet i et 0 qu'elle ne contient pas le sommet i.

(a) 3 partitions

	Partitions		
	π^1	π^2	π^3
z_1	1	3	1
z_2	1	3	1
z_3	2	2	3
z_4	2	2	3
z_5	3	1	3
z_6	3	1	3

(b) Hypergraphe associé aux partitions

	H_1			H_2			H_3		
	H_{11}	H_{12}	H_{13}	H_{21}	H_{22}	H_{23}	H_{31}	H_{32}	H_{33}
z_1	1	0	0	0	0	1	1	0	0
z_2	1	0	0	0	0	1	1	0	0
z_3	0	1	0	0	1	0	0	0	1
z_4	0	1	0	0	1	0	0	0	1
z_5	0	0	1	1	0	0	0	0	1
z_6	0	0	1	1	0	0	0	0	1

TABLE 3.1 – Exemple de représentation d'un hypergraphe pour trois partitions π

En se basant sur le partitionnement de graphe, [Strehl et Ghosh 2002] proposent trois méthodes de recherche de consensus. Dans la méthode CSPA pour *Cluster-based Similarity Partioning Algorithm* les auteurs définissent un graphe $\mathcal{G}_1(O, Co)$ où les sommets sont les observations et les arêtes sont les entrées de la matrice de connectivité Co ($N \times N$). Un algorithme de partitionnement de graphe tel que METIS [Karypis *et al.* 1997] fournit la partition π^*.

Dans la méthode MCLA pour *Meta-CLustering Algorithm*, les auteurs définissent un graphe $\mathcal{G}_2(C, J)$ où les sommets sont les classes et les arêtes définissent la similarité entre deux classes en utilisant l'indice de Jaccard. Le graphe $\mathcal{G}(C, J)$ des classes est ensuite partitionné en utilisant la méthode de partitionnement d'hypergraphe HMETIS. Enfin, un vecteur d'association entre les instances et les classes est créé au sein de chaque méta-

groupe. Les instances sont ensuite classées dans le méta-cluster ayant le plus fort degré d'association.

Enfin dans la méthode HPGA pour *Hyper Graphe Partitioning Algorithm*, les auteurs définissent un hypergraphe $\mathcal{G}_2(O, A)$ où les sommets O sont les observations et les hyper-arêtes représentent les classes. Pour chaque classe de chaque partition π^b, une hyper-arête relie toutes les observations qu'elle contient. L'algorithme de partitionnement de graphe HMETIS, qui est une extension de METIS aux hyper-graphes, Karypis *et al.* [1997] partitionne ensuite l'hypergraphe $\mathcal{G}_2$ des partitions π_k en créant une classification consensus qui coupe le moins d'hyper-arêtes. Ces trois problèmes sont d'ordre NP-complet. Leur complexité valent respectivement $O(kN^2H)$, $O(kNH)$, $O(k^2NH^2)$.

3.6.2.3 Méthodes de recherche directe d'une partition centrale

Contrairement aux méthodes présentées dans la section 3.6.2.2, on cherche la partition π^* telle qu'elle maximise un critère d'association Γ avec les partitions $\pi^b \in \Pi$ Saporta [2006].

$$\pi^* = \underset{\pi \in \Pi^*}{argmax} \sum_{b=1}^{B} \Gamma(\pi, \pi^b) \tag{3.19}$$

où Π^* est l'ensemble des partitions de $\mathcal{Z}$.

L'une des méthodes les plus connus de cette famille est la méthode COBWEB [Fisher 1987] qui est une approche de classification conceptuelle reposant sur la notion de *partition utility*. La fonction d'utilité qui correspond au Γ de la relation 3.19 est définie par la relation suivante entre la partition centrale $\pi^* \in \Pi^*$ et les partitions de $\pi^b \in \Pi$:

$$\Gamma(\pi^*, \pi^b) = \sum_{k=1}^{K} p(\pi_k^*) \sum_{l=1}^{K_b} p(\pi_l^b/\pi_k^*)^2 - \sum_{l=1}^{K_b} p(\pi_l^b)^2 \tag{3.20}$$

avec $p(\pi_r^*) = \frac{\#\pi_r^*}{N}$, $p(\pi_l^b) = \frac{\#\pi_l^b}{N}$ et $p(\pi_l^b/\pi_r^*) = \frac{\#\pi_r^* \cap \pi_l^b}{\pi_r^*}$. Γ représentent l'accroissement du nombre attendu de classes de la partition π^b correctement retrouvé connaissant la partition π^* par rapport au nombre de classes sans la connaissance de la partition π^*.

Krieger et Green [1999] considèrent dans la méthode SEGWAY, l'indice de Rand ou sa version corrigée comme étant la fonction Γ de la relation 3.19. Ils maximisent, après

initialisation de la partition consensus π^*, l'expression

$$V(\pi^*) = \sum_{b}^{B} \alpha_b Rand(\pi^*, \pi^b)$$

où les α_b sont poids fixe positif vérifiant $\sum \alpha_b = 1$. Dans un processus itératif, pour chaque observation, on crée un ensemble de K-1 partitions par réaffectation de l'observation z_i aux K-1 classes (autres que sa classe d'appartenance) dans la partition π^*. La nouvelle partition π^* est celle parmi les $K-1$ partitions qui maximise $V(\pi^*)$. Le processus s'arrête lorsque la quantité $V(\pi^*)$ est stable.

Dans le domaine de la théorie de l'information, Strehl et Ghosh [2002]; Topchy *et al.* [2004a] définissent la partition centrale π^* en maximisant une version normalisée de l'information mutuelle :

$$NMI(\pi_a, pi^b) = \frac{-2\sum_{k=1}^{K_a}\sum_{l=1}^{K_b}\frac{n_{kl}}{N}log(\frac{n_{kl}N}{n_{k.}n_{.l}})}{\sum_{k=1}^{K_a} n_{k.}log(\frac{n_{l.}}{N}) + \sum_{l=1}^{K_b} n_{.l}log(\frac{n_{.l}}{N})} \tag{3.21}$$

L'information mutuelle normalisée est une mesure symétrique permettant de quantifier la quantité d'informations partagée par deux distributions et qui prend de grandes valeurs lorsque les partitions π^a et π^b sont identiques. La partition π^* est recherchée telle que :

$$\pi^* = \underset{\pi\in\Pi}{argmax} \sum_{b=1}^{B} NMI(\pi^*, \pi^b) \tag{3.22}$$

Dans le cadre des méthodes de recherche de la partition centrale il convient d'ajouter celles reposant sur des comparaisons par paires fondées sur l'analyse relationnelle Benhadda et Marcotorchino [2007]. La partition centrale est défine par sa matrice d'adjacence Co^* que l'on obtient par maximisation du critère de Condorcet (relation 3.23) fondé sur les matrices adjacences des partitions π^b.

$$C(Co^*) = B\sum_{i=1}^{N}\sum_{i'=1}^{N} Co(i,i')Co^*(i,i') + \overline{Co(i,i')}\ \overline{Co^*(i,i')} \tag{3.23}$$

avec $\overline{x} = 1 - x$ La formulation mathématique du problème relationnel consiste à trouver la partition centrale π^* telle que :

$$Co^* = \underset{\tilde{Co}}{max}\ C(\tilde{Co}) \tag{3.24}$$

Co^* vérifiant :

$$\begin{cases} Co^*(i,i) = 1, \forall i \\ Co^*(i,i') = Co^*(i',i), \forall i, i' \\ Co^*(i,i') + Co^*(i',i") - Co^*(i,i") < 1, \forall i, i', i" \end{cases} \tag{3.25}$$

La solution exacte de ce problème s'obtient par programmation linéaire dans le cas où le nombre d'individus à classifier serait relativement petit, mais dans la pratique Benhadda et Marcotorchino [2007] présente une heuristique permettant d'obtenir une solution approchée sans avoir à fixer au préalable le nombre de classes. Ceci constitue un avantage par rapport aux méthodes nécessitant la connaissance a priori du nombre de classes

3.6.2.4 Consensus de cartes topologiques

Dans l'algorithme d'apprentissage des cartes topologiques auto-organisées, chaque neurone ou cellule de la carte identifie des groupes particuliers d'observations, en tenant compte de l'information contenue dans les neurones de son voisinage, permettant donc de déterminer la structure topologique de la carte. En fonction des paramètres d'initialisations les cellules ou les régions de la carte peuvent varier en particulier celles présentant des observations atypiques. En terme de visualisation, il est possible d'améliorer les performances de la méthode SOM à travers une carte topologique consensus des cartes d'un ensemble de diversification [Baruque *et al.* 2007].

L'idée d'utiliser plusieurs cartes topologiques pour accroître les performances de classification est initiée par Blackmore et Miikkulainen [1995] dans l'algorithme Hierarchical SOM (HSOM). HSOM génère une structure pyramidale de SOM dans laquelle le premier niveau correspond à une carte topologique obtenue sur les données initiales et le niveau 2 correspond à de nouvelles cartes obtenues par application de SOM sur les cellules des cartes du niveau inférieur présentant une inertie intra-classe trop grande par rapport un seuil fixé et ainsi de suite jusqu'à ce que les inerties intra-classes des cellules soient tout inférieures au seuil. Contrairement à cette approche hiérarchique de SOM qui divise une carte SOM en plusieurs cartes SOM, les méthodes de fusion d'un ensemble de SOM recherchent une unique carte topologique résumant l'information contenue dans divers SOM.

Comme dans les méthodes classiques de recherche de consensus, la démarche des méthodes de *"Fusion de SOM"* se résume aussi en deux étapes : une étape de diversification par la création d'un ensemble de cartes topologiques et une étape d'agrégation des SOM.

L'ensemble $\Pi = \{\mathcal{C}^1, \ldots, \mathcal{C}^B\}$ de diversification des cartes topologiques peut être obtenu de diverses manières. Il peut s'agir : de résultats de l'application de différents algorithmes (SOM ou Neural Gas) de cartes topologiques sur le même ensemble de données, de résultats obtenus par application répétée d'un même algorithme neuronal avec différentes initialisations des paramètres. [Jiang et Zhou 2004; Georgakis *et al.* 2005; Saavedra *et al.* 2007].
Les procédures de fusion de SOM cherchent toutes à atteindre le même objectif à savoir : définir une carte topologique $\mathcal{C}^*$ telle que chacune de ses cellules résulte de l'agrégation de plusieurs cellules, appartenant aux cartes de l'ensemble Π de diversification, jugées "proches" au sens d'une certaine distance.
Les cellules de la carte fusionnée sont classiquement représentées par le vecteur moyen w_c^* des vecteurs référents des cellules à fusionner w_c^b.

$$w_c^* = \frac{1}{B} \sum_{b=1}^{B} w_c^b \tag{3.26}$$

En fonction du critère utilisé pour évaluer la similarité entre les cellules des cartes topologiques, on rencontre dans la littérature plusieurs procédures de fusion de SOM.

Remarquons qu'une cellule c d'une carte topologique peut être définie par un vecteur binaire de dimension $1 \times N$ correspondant à son indicatrice.

Georgakis *et al.* [2005] propose une méthode de programmation dynamique de fusion des SOM. Les auteurs initialisent la carte fusion $\mathcal{C}^*$ (choix d'une carte de l'ensemble Π), puis ils choisissent une carte $\mathcal{C}^b$. Chaque cellule $c \in \mathcal{C}^*$ de la carte est agrégée avec une cellule $c' \in \mathcal{C}^b$ tel que $c' = \underset{r \in \mathcal{C}^b}{argmin}(||w_c^* - w_r^b||^2)$. Une fois toutes les cellules agrégées, pour $B = 2$, la formule (3.26) permet de déterminer les nouveaux vecteurs référents de la carte fusion $\mathcal{C}^*$. Le processus est répété pour toutes les cartes de l'ensemble Π. Cette approche impose d'avoir des cartes $\mathcal{C}^*$ de même dimension. Ceci constitue une limitation majeure en particulier dans le cas multi-blocs où les dimensions des cartes sont induites par les dimensions des données. Cependant, la formule 3.26 ne prend pas explicitement en compte la notion de voisinage au niveau local des cartes $\mathcal{C}^b$. Par conséquent, deux neurones proches sur la carte fusion $\mathcal{C}^*$ ne sont pas nécessairement définis par des neurones proches au niveau local. Enfin dans l'algorithme présenté par les auteurs, la complexité est d'ordre $O((B-1)M^2)$ où M est le nombre de neurones des cartes.
Saavedra *et al.* [2007] considèrent la représentation sous la forme des référents pour déterminer la distance entre deux neurones. Chaque neurone peut être associé à une zone de

l'espace des observations appelé polygone de Voronoi [3] [Aurenhammer et Klein 2000]. Les neurones ayant des polygones de Voronoi similaires peuvent alors être considérés comme étant proches dans l'espace des observations. Ainsi, Saavedra *et al.* [2007] évaluent la similarité entre les neurones grâce à un vecteur binaire μ_c^k $(1 \times N)$ d'association des observations aux cellules c de la carte k :

$$\mu_c^b(z_i) \equiv \begin{cases} 1, & si\ (z_i) \in c \\ 0, & sinon \end{cases} \tag{3.27}$$

La fusion de deux vecteurs référents est basée sur la mesure de dissimilarité suivante :

$$ds(w_c, w_r) = \frac{\sum_{l=1}^{N} XOR(\mu_c^a, \mu_r^b)}{\sum_{l=1}^{N} OR(\mu_c^a, \mu_r^a)} \tag{3.28}$$

où a et b désignent deux cartes distinctes, XOR désigne le OU exclusif et prend la valeur 1 sur sa composante i si l'individu n'appartient pas simultanément aux cellules c et r, 0 sinon. OR désigne le OU inclusif et prend la valeur 1 si l'observation i appartient à au moins une des cellules c et r. À l'aide de la distance définie par 3.28, les auteurs agrègent par rapport à un seuil fixé les cellules les plus proches : $ds(w_c^k, w_r^l) < \theta_f$ et le référent correspondant.

Ces deux approches souffrent de deux problèmes. La taille de la carte fusion est identique à celle des cartes formant l'ensemble Π de diversification. Or, elle n'est en principe pas connue d'avance. De plus la notion de voisinage local au niveau des cartes topologiques formant Π n'est pas prise en compte dans la carte fusion.

3.7 Conclusion

Nous avons présenté dans ce chapitre les méthodes de classification permettant de surmonter les limites des méthodes classiques de classification face aux données de grande dimension. Les performances des méthodes présentées dans ce chapitre sont dépendantes de la nature des données et des objectifs à atteindre. Nous présentons dans la suite les deux principales contributions de cette thèse. Le chapitre 4 présente une nouvelle méthode de type subspace clustering pour le traitement des données de grande dimension structurées

3. On appelle région de Voronoï ou cellule de Voronoï associée à une cellule c, l'ensemble des points $z_i \in \mathcal{Z}$ qui sont plus proches de c que de tout autre cellule $r \in \mathcal{C}$. En d'autres termes

$$Vor(c) = \{z_i in \mathcal{Z} / \forall c \in \mathcal{C} || z_i - w_c || < || z_i - w_r ||\}$$

en blocs de variables et basée sur les carte topologiques et le chapitre 5 présentent deux approches dédiées à la fusion de SOM.

Deuxième partie

Approches proposées

Chapitre 4

Soft Subspace clustering SOM

4.1 Introduction

Dans la section 3.5.1 du chapitre 3, nous avons présenté les méthodes dédiées au subspace clustering. Nous proposons ici 2S-SOM, une méthode de soft subspace clustering, basée sur SOM, prenant en compte la structuration en blocs des variables et son extension aux données mixtes. Nous faisons l'hypothèse que les dimensions et les blocs de variables contribuent à différents niveaux à la détermination des classes. Ces contributions apportées par les variables et les blocs dans chaque classe sont alors mesurées par des poids.
La méthode est basée sur une version modifiée de la fonction de coût de SOM [Kohonen 1998] en introduisant des poids adaptatifs sur les blocs et sur les variables et un terme d'entropie négatif permettant de définir les contributions relatives des variables et des blocs inspirée des travaux de Huang et Ng [1999]; Jing *et al.* [2007] et de Chen *et al.* [2012]. L'idée de base consiste à rechercher itérativement une partition des observations et à déterminer pour chaque cellule des variables et des blocs spécifiques.

La méthode est présentée dans la section 4.2. La section 4.3 présente les propriétés de 2S-SOM. La méthode est illustrée sur des données réelles en section 4.4.

4.2 La méthode 2S-SOM

Nous rappelons quelques notations avant la présentation de l'approche 2S-SOM. On dispose de N observations z_i décrites par p variables divisées en B blocs. On note :

- Z la matrice de N observations $z_i \in \mathbb{R}^p$ avec $i = 1, \ldots, N$.
- $\mathcal{V} = \{z^j, j = 1, \ldots, p\}$ l'ensemble des variables divisé en B blocs de p_b variables tels que $p_1 + \ldots + p_b + \ldots + p_B = p$.

- α est une matrice $K \times B$ où K désigne le nombre de classes c dans Z, α_{cb} est le poids du bloc b dans la classe c.
- $\beta = [\beta_1, \ldots, \beta_B]$ est une matrice $K \times p$ où β_b est une matrice de dimension $K \times p_b$ définissant les poids $\beta_{cbj} (j = 1, \ \ldots, p_b)$ sur les variables du bloc b pour la cellule c.

4.2.1 2S-SOM

2S-SOM est une extension à SOM de l'algorithme de type subspace clustering FGKM qui est basé sur la méthode des K-moyennes [Chen *et al.* 2012]. Il repose sur une modification de la fonction de coût de SOM en introduisant un double système de poids α_{cb} $(b = 1, \ldots, B)$ et β_{cbj} $(j = 1, \ldots, p_b)$ définis respectivement sur les blocs et sur les variables pour chaque cellule c de la carte $\mathcal{C}$. La classification est donc obtenue par optimisation de la fonction objectif J_{2S-SOM} définie en (4.1).

$$\mathcal{J}^T_{2S-SOM}(\mathcal{X}, \mathcal{W}, \alpha, \beta) = \sum_{c \in C} \left(\sum_{b=1}^{B} \left(\sum_{z_i \in \mathcal{Z}} \alpha_{cb} \mathcal{K}^T(\sigma(\mathcal{X}(z_i), c)) d_{\beta_{cb}}(i) + J_{cb} \right) + I_c \right) \quad (4.1)$$

avec $d_{\beta_{cb}}(i) = \sum_{j=1}^{p_b} \beta_{cbj} (z^j_{ib} - \omega^j_{cb})^2$ et sous les contraintes :

$$\begin{cases} \sum_{j=1}^{p_b} \beta_{cbj} = 1, \ \beta_{cbj} \in [0,1], \forall \ c \in \mathcal{C}, \forall b \\ \\ \sum_{b=1}^{B} \alpha_{cb} = 1, \ \alpha_{cb} \in [0,1], \ \forall \ c \in \mathcal{C} \end{cases}$$

$I_c = \lambda \sum_{b=1}^{B} \alpha_{cb} log(\alpha_{cb})$ et $J_{cb} = \eta \sum_{j=1}^{p_b} \beta_{cbj} log(\beta_{cbj})$ représentent les entropies négatives pondérées et associées aux vecteurs poids relatifs aux blocs et aux vecteurs poids relatifs variables pour une cellule c. Ces termes permettent d'ajuster, selon les paramètres λ et η, les contributions relatives apportées par les variables et les blocs dans la classification. Cela sera détaillé dans la section 4.3.

L'optimisation de la fonction de coût $\mathcal{J}_{2S-SOM}$ s'effectue de façon alternée en quatre étapes : les deux premières phases d'affectation des observations aux classes et d'actualisation des vecteurs référents sont identiques à celles de SOM. Dans ces deux premières étapes, les valeurs des poids sont supposées connues et fixées à leur valeur courante. On a alors :

- Étape 1 : les référents $\mathcal{W}$ sont connus et fixés, les observations sont affectées aux

cellules en respectant l'équation (4.2) :

$$c_g(z_i) = \mathcal{X}(z_i) = \underset{c\in\mathcal{C}}{argmin} \left(\sum_{r\in\mathcal{C}} \mathcal{K}^T(\sigma(r,c)) \left(\sum_{b=1}^{B} \alpha_{cb} d_{\beta_{cb}}(i) \right) \right) \tag{4.2}$$

– Étape 2 : actualisation des centres de classe. Chaque cellule c de la carte est représentée par un vecteur référent qui représente au mieux les observations qu'elle a captées en minimisant l'équation (4.1) relativement à l'ensemble $\mathcal{W}$:

$$\omega_{c_g}^T = \frac{\sum_{z_i\in\mathcal{Z}} \mathcal{K}^T(\sigma(X(z_i), c_g))z_i}{\sum_{z_i\in\mathcal{Z}} \mathcal{K}^T(\sigma(\mathcal{X}(z_i), c_g))} \tag{4.3}$$

Les étapes 3 et 4 suivantes sont similaires à celles de FGKM proposée par [Chen *et al.* 2012]. Il s'agit de l'optimisation de la fonction objectif $\mathcal{J}_{2S-SOM}$ sous les contraintes

$$\sum_{j=1}^{p_b} \beta_{cbj} = 1,\ \beta_{cbj} \in [0,1], \forall\ c \in \mathcal{C}, \forall b$$

$$\sum_{b=1}^{B} \alpha_{cb} = 1,\ \alpha_{cb} \in [0,1],\ \forall\ c \in \mathcal{C}$$

en utilisant le multiplicateur de Lagrange.

– **Etape d'actualisation des poids** α_{cb} : Si les paramètres $\mathcal{X} = \widehat{\mathcal{X}}$, $\omega = \widehat{\omega}$ et $\beta = \widehat{\beta}$ sont connus et fixés à leurs valeurs courantes alors $\forall\ \lambda > 0$ on définit le lagrangien de la fonction $\mathcal{J}_{2S-SOM}$ qui dépend uniquement des paramètres α_{cb} par :

$$\begin{aligned}\mathcal{L}_{\alpha,\lambda} &= \mathcal{J}_{2S-SOM}(\widehat{\mathcal{X}}, \widehat{\omega}, \alpha, \widehat{\beta}) - \sum_{c\in\mathcal{C}} \mu_c (\sum_{b=1}^{B} \alpha_{cb} - 1) \\ &= \sum_{c\in\mathcal{C}} \left(\sum_{b=1}^{B} \left(\sum_{z_i\in Z} \alpha_{cb} \mathcal{K}^T(\sigma(\mathcal{X}(z_i), c)) d_{\beta_{cb}} + J_{cb} \right) + I_c - \mu_c (\sum_{b=1}^{B} \alpha_{cb} - 1) \right).\end{aligned}$$

où les termes μ_c sont les multiplicateurs de Lagrange associés aux contraintes. La solution $\widehat{\alpha}$ optimisant $\mathcal{J}_{2S-SOM}$ vérifie

$$\frac{\partial \mathcal{L}_{\alpha,\mu}}{\partial \alpha_{cb}} = \sum_{z_i\in\mathcal{Z}} \mathcal{K}^T(\sigma(\mathcal{X}(z_i), c)) d_{\beta_{cb}}(i) + \lambda(log(\alpha_{cb}) + 1) - \mu_c = 0 \tag{4.4}$$

et

$$\frac{\partial \mathcal{L}_{\alpha,\mu}}{\partial \mu_c} = \sum_{b=1}^{B} \alpha_{cb} - 1 = 0 \tag{4.5}$$

posons

$$\Psi_{cb} = \sum_{z_i\in\mathcal{Z}} \mathcal{K}^T(\sigma(\mathcal{X}(z_i), c)) d_{\beta_{cb}}(i)$$

$$(4.4) \Rightarrow \alpha_{cb} = \exp(\frac{-\lambda - \Psi_{cb}}{\lambda}) \exp(\frac{\mu_c}{\lambda})$$

et

$$(4.5) \Rightarrow \exp(\frac{\mu_c}{\lambda}) = \frac{1}{\sum_{b=1}^{B} \exp(\frac{-\lambda - \Psi_{cb}}{\lambda})}$$

$\mathcal{J}_{2S-SOM}^{T}(\widehat{\mathcal{X}}, \widehat{\omega}, \alpha, \widehat{\beta})$ atteint son minimum pour une cellule c et pour un bloc b en

$$\alpha_{cb} = \frac{exp(\frac{-\Psi_{cb}}{\lambda})}{\sum_{b=1}^{B} exp(\frac{-\Psi_{cb}}{\lambda})} \tag{4.6}$$

Remarquons que Ψ_{cb} se décompose dans (4.7) en la somme pondérée de la distance des observations z_i des cellules r du voisinage T de la cellule c et en la somme pondérée de la distance des observations $z_i \in c$ au centre de classe de c.

$$\Psi_{cb} = \sum_{z_i \in r, r \neq c} \mathcal{K}^T(r,c) d_{\beta_{cb}}(i) + \mathcal{K}^T(c,c) \sum_{z_i \in c} d_{\beta_{cb}}(i) \tag{4.7}$$

Le premier terme de (4.7) est proportionnel à la distance des observations appartenant aux cellules r du voisinage T de la cellule c par rapport au centre de classe de la cellule c et le second terme est proportionnel à l'inertie des observations z_i de la cellule c. Finalement, le poids d'un bloc sera donc d'autant plus important que ce bloc minimise simultanément l'inertie des observations appartenant à la classe et l'inertie des observations appartenant au voisinage T de la cellule c.

– **Etape d'actualisation des poids** β_{cbj} : de manière identique à l'étape précédente, si les paramètres $\mathcal{X} = \widehat{\mathcal{X}}$, $\omega = \widehat{\omega}$ et $\alpha = \widehat{\alpha}$ sont connus et fixés à leurs valeurs courantes alors $\forall\ \eta > 0$, on définit le lagrangien de la fonction $\mathcal{J}_{2S-SOM}$ qui dépend uniquement des paramètres par β_{cbj} par :

$$\begin{aligned}\mathcal{L}_{\beta,\eta} &= \mathcal{J}_{2S-SOM}((\widehat{\mathcal{X}}, \widehat{\omega}, \widehat{\alpha}, \beta) - \sum_{c \in \mathcal{C}} \mu_{cb} \left(\sum_{j=1}^{p_b} \beta_{cbj} - 1 \right) \\ &= \sum_{c \in \mathcal{C}} \left(\sum_{b=1}^{B} \left(\sum_{z_i \in Z} \alpha_{cb} \mathcal{K}^T(\sigma(\mathcal{X}(z_i), c)) d_{\beta_{cb}}(i) + J_{cb} - \mu_{cb}(\sum_{j=1}^{p_b} \beta_{cbj} - 1) \right) \right)\end{aligned}$$

où les variables μ_{cb} sont les multiplicateurs de Lagrange associés aux contraintes. La solution $\widehat{\beta}$ optimisant $\mathcal{J}_{2S-SOM}$ vérifie

$$\frac{\partial \mathcal{L}_{\beta,\mu_{ck}}}{\partial \beta_{ck}} = \Phi_{cbj} + \eta(log(\beta_{cbj}) + 1) - \mu_{cb} = 0 \tag{4.8}$$

et

$$\frac{\partial \mathcal{L}_{\beta,\mu_{cb}}}{\partial \mu_{cb}} = \sum_{j=1}^{p_b} \beta_{cbj} - 1 = 0 \tag{4.9}$$

où

$$\Phi_{cbj} = \sum_{z_i \in \mathcal{Z}} \alpha_{cb} \mathcal{K}^T(\mathcal{X}(z_i), c)(z_{ib}^j - \omega_{cb}^j)^2 \tag{4.10}$$

$$(4.8) \Rightarrow \beta_{cbj} = \exp(\frac{-\eta - \Phi_{cb}}{\eta}) \exp(\frac{-\mu_{cb}}{\eta})$$

et

$$(4.9) \Rightarrow \exp(\frac{-\mu_{cb}}{\eta}) = \frac{1}{\sum_{j=1}^{p_b} \exp(\frac{-\eta - \Phi_{cbj}}{\eta})}$$

$\mathcal{J}_{2S-SOM}^T(\widehat{\mathcal{X}}, \widehat{\omega}, \widehat{\alpha}, \beta)$ atteint son minimum pour une cellule c et pour un bloc b en :

$$\beta_{cbj} = \frac{\exp(\frac{-\Phi_{cbj}}{\eta})}{\sum_{j=1}^{p_b} \exp(\frac{-\Phi_{cbj}}{\eta})} \tag{4.11}$$

Remarquons que la fonction Φ_{cbj} (l'équation (4.10)) se décompose dans (4.12), pour la variable j, en la somme pondérée de la distance entre les observations z_i appartenant aux cellules r du voisinage T de c et en la somme pondérée de la distance des observations z_i au référent de la cellule c.

$$\Phi_{cbj} = \sum_{z_i \in r, r \neq c} \alpha_{cb} \mathcal{K}^T(r, c)(z_{ib}^j - \omega_{cb}^j)^2 + \mathcal{K}^T(c, c) \sum_{z_i \in c} \alpha_{cb}(z_{ib}^j - \omega_{cb}^j)^2 \tag{4.12}$$

Le poids d'une variable sera donc d'autant plus important qu'elle minimise simultanément la variance de la j$^{\text{ème}}$ composante des observations appartenant à la classe c et la somme des carrées des distances entre les j$^{\text{ème}}$ composantes des observations appartenant aux cellules r du voisinage T de la cellule c et le référent w_c de la cellule c.
Les coefficients de pondération α_{cb} et β_{cbj} définis par 2S-SOM indiquent respectivement l'importance relative des blocs et des variables dans les classes. Ainsi, plus le poids d'un bloc b ou d'une variable v_j est important, plus le bloc ou la variable contribue à la définition de la classe au sens où elle permet de réduire la variabilité des observations dans la cellule et dans son voisinage proche. Finalement, à la convergence, 2S-SOM fournit d'une part une carte topologique permettant de visualiser les données et d'autre part des systèmes de poids pour les classes de la classification.

En vue de faciliter l'interprétation des classes, si la taille de la carte conduit à un trop grand nombre de cellules, il est possible d'appliquer un algorithme de classification

ascendante hiérarchique (CAH) sous contrainte de voisinage sur la matrice composée des vecteurs référents pour réduire ce grand nombre de cellules en un nombre restreint de classes [Gordon 1996; Vesanto *et al.* 2000]. La contrainte de voisinage dans la CAH permet alors de conserver la topologie des observations fournie par la carte 2S-SOM. Dans le cas des évaluations présentées dans la section 4.4, les classes finales sont obtenues par application d'une CAH sous contraintes utilisant la stratégie d'agrégation de ward.

4.2.2 Version mixte de 2S-SOM

Initialement développée pour des données numériques, la méthode 2S-SOM s'étend aux données mixtes à travers l'adaptation de la distance euclidienne aux bases de données contenant des variables qualitatives et quantitatives. Pour prendre en compte les données mixtes, nous nous reportons aux travaux présentés par Lebbah *et al.* [2005] et par Chen et Marques [2005], qui présentent simultanément la version mixte quasi-identique de l'algorithme SOM. Les auteurs proposent une extension de la distance euclidienne classique pour la classification des données mixtes. Contrairement à l'algorithme NCSOM proposé par Chen et Marques [2005] dans lequel les auteurs combinent simplement la distance euclidienne classique et la distance de Hamming pour données binaires, Lebbah *et al.* [2005] propose d'utiliser dans l'algorithme MTM (Mixed Topological Map) un paramètre γ d'ajustement de l'influence de la partie quantitative des données par rapport à la partie qualitative dans l'évaluation de la similarité entre les observations.
Avant de présenter l'extension de 2S-SOM aux données mixtes rappellons la notion de codage binaire des variables qualitatives. Une variable qualitative à m modalités peut être codée par un vecteur de $\{0,1\}^m$ dont la forme dépend de la nature de la variable qualitative. Dans le cas d'une variable qualitative simple, le codage utilisé est dit disjonctif et équivaut à représenter sa modalité j par un vecteur formé de zéros sauf pour sa j ième composante qui est égale à 1. Dans le cas d'une variable qualitative ordinale (exemple : petit, moyen, grand, très grand) le codage dit additif équivaut à représenter sa modalité j par un vecteur formé de 1 pour ses j premières composantes et par des zéros pour les autres. Ici on considère une base de données mixtes, après codage, les observations $z_i \in \mathcal{Z} \subset \mathbb{R}^p \times \{0,1\}^q$, où p désigne le nombre de variables quantitatives et q le nombre total des variables binaires. Formellement, $z_i = (z_i^r, z_i^{bin})$ avec $z_i^r \in \mathbb{R}^p$ désigne la partie réelle et $z_i^{bin} \in \{0,1\}^q$ désigne la partie binaire des observations, q étant la somme des modalités de toutes les variables qualitatives.

La distance entre deux observations z_i et $z_{i'}$ définies dans $\mathbb{R}^p \times \{0,1\}^q$ vaut alors :

$$d_m(z_i, z_{i'}) = \sum_{j=1}^{p}(z_i^j - z_{i'}^j)^2 + \gamma \sum_{j=p+1}^{p+q} \delta(z_i^j, z_{i'}^j) \tag{4.13}$$

avec

$$\delta(z_i^j, z_{i'}^j) = \begin{cases} 0 \ si \ (z_i^j = z_{i'}^j) \\ 1 \ si \ (z_i^j \neq z_{i'}^j) \end{cases}$$

Le paramètre γ permet d'ajuster l'influence de la partie qualitative des variables par rapport à la partie quantitative et inversement. Huang [1997] estime après une évaluation expérimentale que $\gamma \in [\frac{\sigma}{3}, \frac{2\sigma}{3}]$ où σ est la moyenne des écart-types des variables de la partie continue des données. L'usage de 4.13 dans le processus d'apprentissage implique que l'affectation des observations aux cellules de la carte reste identique à celle de 2S-SOM. L'actualisation des centres de classe est spécifique à chaque portion des données :

- La formule 4.3 permet d'actualiser la partie quantitative des référents. Ainsi, pour chaque cellule c de la carte et en notant par w_{c^r} sa partie quantitative, on a :

$$\omega_{c^r}^T = \frac{\sum_{z_i \in \mathcal{Z}} \mathcal{K}^T(\sigma(X(z_i), c))z_i^r}{\sum_{z_i \in} \mathcal{K}^T(\sigma(\mathcal{X}(z_i), c))} \tag{4.14}$$

- On note par $w_{c^{bin}}$ la partie binaire du vecteur référent relatif à la cellule c. Afin de faire la mise à jour d'une composante l de ce vecteur on calcule la fréquence pondérée de 1 de la composante j notée F_1^T et la fréquence pondérée de 0 de la composante l notée F_0^T.

$$F_1^T(c, j) = \frac{\sum_{z_i \in \mathcal{Z}} \mathcal{K}^T(\mathcal{X}(z_i), c)z_{ij}^{bin}}{\sum_{z_i \in \mathcal{Z}} \mathcal{K}^T(\mathcal{X}(z_i), c)} \tag{4.15}$$

et

$$F_0^T(c, j) = \frac{\sum_{z_i \in \mathcal{Z}} \mathcal{K}^T(\mathcal{X}(z_i), c)(1 - z_{ij}^{bin})}{\sum_{z_i \in \mathcal{Z}} \mathcal{K}^T(\mathcal{X}(z_i), c)} \tag{4.16}$$

La modalité qui définit la cellule est celle dont la fréquence F est supérieure la somme des fréquences des autres modalités de la variable qualitative considérée :

$$\omega_{c^{binj}}^T = \begin{cases} 1 \ si \ \ F_1^T(c, j) > F_0^T(c, j) \\ 0 \ sinon \end{cases} \tag{4.17}$$

- Actualisation des poids α_{cb} et β_{cbj} : remarquons que lorsque les variables sont simplement disjonctif (absence d'ordre dans les modalités), l'inverse proportionnalité des poids rapport à inertie des observations dans les cellules de la carte implique que les modalités ayant des fréquences faibles auront naturellement des poids forts. Nous

proposons d'atténuer ces biais à travers une normalisation dans les quantités définies en (4.7) et (4.12)

$$\Psi_{cb}^{m} = \sum_{z_i \in \mathcal{Z}} \mathcal{K}^T(\sigma(\mathcal{X}(z_i), c)) d_{m\beta_{cb}}(i) \tag{4.18}$$

$$\Phi_{cbj}^{f_j} = \sum_{z_i \in \mathcal{Z}} \alpha_{cb} \mathcal{K}^T(\mathcal{X}(z_i), c)(\sum_{j=1}^{p_b} (z_{ibj} - \omega_{cbj})^2 + \gamma f_j \sum_{j=p_b+1}^{p_b+q_b} (z_{ibj} - \omega_{cbj})^2) \tag{4.19}$$

avec $d_{m\beta_{cb}}(i) = \sum_{j=1}^{p_b} \beta_{cbj}(z_{il} - w_{cbj})^2 + \gamma \sum_{j=p_b+1}^{p_b+q_b} \beta_{cbj} \delta(z_i^j, w_{cb}^j)$ et f_j est inversement proportionnelle à la fréquence de la modalité j de la variable associée. Les quantités p_b et q_b désignent le nombre de variables quantitatives et binaires du bloc b. Le remplacement dans les expressions définies en 4.11 et 4.6 des quantités (4.12) et (4.7) par 4.18 et 4.19, pour les données mixtes, permet d'obtenir une actualisation des poids sur les blocs et sur les variables.

4.3 Propriétés de 2S-SOM

Dans cette section, nous étudions les propriétés de conservation topologique de 2S-SOM et l'influence des poids α et β dans la classification en fonction des paramètres λ, η et du paramètre de voisinage T. En tenant compte des relations (4.6) et (4.11) si λ et η sont très grands alors $\alpha_{ck} \approx \frac{1}{B}$ et $\beta_{cbj} \approx \frac{1}{p_b}$. La fonction objectif $\mathcal{J}_{2S-SOM}^T$ peut se décomposer en (4.20) faisant apparaître les termes de conservation de la topologie des observations et de quantification vectorielle de 2S-SOM :

$$\begin{aligned}
\mathcal{J}_{2S-SOM}^T(\mathcal{X}, \mathcal{W}, \alpha, \beta) &= \sum_{c \in C} \left(\sum_{b=1}^{B} \left(\sum_{r \in \mathcal{C}} \sum_{z_i \in r} \alpha_{cb} \mathcal{K}^T(\sigma(r, c)) d_{\beta_{cb}} + J_{cb} \right) + I_c \right) \\
&= \mathcal{K}^T(\sigma(c, c)) \sum_{c \in C} \left(\sum_{b=1}^{B} \left(\sum_{z_i \in c} \alpha_{cb} d_{\beta_{cb}} + J_{cb} \right) + I_c \right) + \\
&\qquad \sum_{c \in C} \left(\sum_{b=1}^{B} \left(\sum_{r \neq c} \sum_{z_i \in r} \alpha_{cb} \mathcal{K}^T(\sigma(r, c)) d_{\beta_{cb}} \right) \right)
\end{aligned} \tag{4.20}$$

1. Le premier terme correspond à la fonction objectif proposée par [Chen *et al.* 2012] dans FGKM pondérée par $\mathcal{K}^T(\sigma(c, c)) = \mathcal{K}^T(0)$. Son importance relative dans 2S-SOM dépend alors de T ; plus T est petit plus ce terme prend de l'importance dans la minimisation, dans ce cas 2S-SOM est équivalent à FGKM. De plus, lorsque les

paramètres λ et η sont très grands, les poids α_{cb} et β_{cbj} deviennent constant alors, 2S-SOM est équivalent à l'algorithme des K-Moyennes.

2. Le deuxième terme introduit la contrainte de conservation topologique. Ce terme montre que si deux cellules sont proches sur la carte $\mathcal{C}$ alors $\mathcal{K}^T$ est grand car $\sigma(c,r)$ est petit ; la minimisation de ce terme rapproche les deux cellules r et c. Ainsi, la proximité sur la carte traduit donc une proximité dans l'espace des observations. De plus, lorsque $\lambda \to \infty$ et $\eta \to \infty$, les blocs sont équipondérés de même que les variables. Ainsi, 2S-SOM est équivalent à SOM.

Lorsque $\lambda \to \infty$ et η fixé les poids α_{cb} associés aux blocs étant tous égaux à $\frac{1}{B}$, alors, seuls les poids β_{cbj} des variables des blocs définissent les cellules. En considérant que le sous espace associé à la classe est donné par les variables ayant les plus forts poids, 2S-SOM peut être vu comme un algorithme de soft subspace clustering.
Lorsque $\eta \to \infty$ et λ fixé les poids des variables d'un bloc sont identiques à $\frac{1}{p_b}$. Dans ce cas seuls les blocs sont pénalisés selon leur capacité à définir les cellules. Dans ce contexte, 2S-SOM permet alors de déterminer pour chaque cellule les blocs qui lui sont spécifiques.

L'algorithme : 2S-SOM

Entrée : La matrice $\mathcal{Z}$

1. Initialisation : choisir la dimension de la carte, le voisinage initial T_o et final T_f des cellules, définir l'ensemble des centres de classe $\mathcal{W}^0$, initialiser les poids α_{cb}^0 sur les blocs et les poids β_{cbj}^0 sur les variables et le couple de paramètres λ et η.
2. Affectation : utiliser la formule (4.2) pour affecter chaque observation à sa cellule d'appartenance.
3. Actualisation des centres de classe : utiliser la formule (4.3) pour actualiser les référents des cellules.
4. Actualisation des poids sur les blocs : utiliser les formules (4.6 et 4.7) pour actualiser les poids sur les blocs.
5. Actualisation des poids sur les variables : utiliser les formules (4.12 et 4.11) pour actualiser les poids sur les variables.
6. Répéter les étapes 2 à 5 jusqu'à la convergence de l'algorithme vers un minimum.

Sortie : Une carte topologique, les poids α_{cb} sur les blocs et β_{cbj} sur les variables.

Finalement, l'algorithme 2S-SOM qui est une extension de la méthode SOM par ajout des deux étapes d'actualisation des poids sur les blocs et sur les variables hérite de la même complexité $O(N \times p \times K^2 \times N_A)$ où K est le nombre de cellulesde la carte et N_A le nombre d'apprentissages. De manière identique à la méthode SOM, 2S-SOM converge en général

vers un minimum local.

4.4 Évaluation

Dans cette section nous allons étudier d'une part, les performances de la méthode 2S-SOM, puis nous la comparons aux méthodes standards SOM et K-moyennes (KM), aux méthodes de type subspaces clustering : Entropy Weighted K-moyennes (EWKM), Feature Group K-Moyennes (FGKM). D'autre part, nous présenterons les propriétés de visualisation de 2S-SOM et sa capacité à identifier les variables de bruit.

4.4.1 Données

Les bases de données utilisées pour évaluer la méthode de fusion proposée sont extraites du répertoire de données de l'Université de Californie à Irvine (UCI) [Bache et Lichman 2013].

- Le jeu de données "Image Segmentation" (IS) contient 2310 observations et 19 variables décrivant les pixels de 7 images. Chaque observation représente un point d'une image décrite par deux blocs de 9 et 10 variables caractérisant le contraste de couleur de ce point sur l'image. Chaque observation possède une étiquette comprise entre 1 et 7.

- Le jeu de données "Cardiographie" (CT) contient 2126 cardiographies fœtales décrites par 21 variables regroupées en 3 blocs. Le bloc 1 contient 7 variables liées à la fréquence cardiaque d'un fœtus. Le bloc 2 contient 4 variables décrivant la variabilité de rythme cardiaque et le bloc 3 est composé de 10 variables définissant des histogrammes de la cardiographie du fœtus. Ces 2126 observations sont divisées en 10 classes.

- Le jeu de données "Dutch Maps Utility " (DMU) est composé de 2000 observations correspondant à l'écriture manuelle des 10 chiffres de la numération mathématique (0 à 9). La représentation sous forme d'image de ces 2000 observations est décrite par 649 variables structurées en 6 blocs contenant respectivement 76, 216, 64, 240, 47 et 6 variables.

Les données simulées D1 et D2 contiennent chacune 400 observations divisées en 4 classes de 100 observations décrites par 4 blocs de variables. La table D1 contient 100

variables réparties en 4 blocs de 25 variables et la table D2 contient 4 blocs de 5 variables. La table D3 contient 400 observations et 25 variables, elle est obtenue en rajoutant à la table D2 un bloc de 5 variables de bruit et 5% d'observations aberrantes. Le tableau 4.1 présente les caractéristiques des données en terme de structuration en blocs et de répartition des variables de bruit dans les blocs ainsi que les paramètres des cartes topologiques retenues. La figure 4.1 représente les classes des données simulées dans le plan factoriel d'une ACP sur les 3 tables.

	Structure en blocs		Structure de la carte			
Données	$\sharp blocs$	$\sharp VB$	Niter	Dim	$T_i \times T_f$	$(\lambda,\ \eta)$
IS	9-10		150	9×9	2×0.82	(3,31)
CT	7-4-10		150	10×10	2×0.1	(7,11)
DMU	76-216-64-240-47-6		150	10×7	3×0.2	(10, 20)
D1	25-25-25-25	9-18-10-7	150	10×9	2×0.1	(2, 3)
D2	5-5-5-5	2-2-4-4	150	10×9	3×0.2	(1, 5)
D3	5-5-5-5-5	2-2-4-4-5	150	10×10	3×0.2	(1, 5)

TABLE 4.1 – Caractéristiques des données et paramètres des cartes retenues pour les tables IS, CT, DMU, D1, D2 et D3 ; il s'agit des cartes minimisant simultanément l'erreur topologique et de quantification vectorielle . Les quantités $\sharp blocs$ et $\sharp VB$ correspondent respectivement à la dimension de chaque bloc et au nombre de variables de bruit par bloc. Les quantités Niter, Dim, $T_i \times T_f$ et $(\lambda,\ \eta)$ correspondent respectivement au nombre d'itérations, aux dimensions de la carte, à la taille du voisinage et aux paramètres λ et η d'ajustement des poids α et β, les meilleures

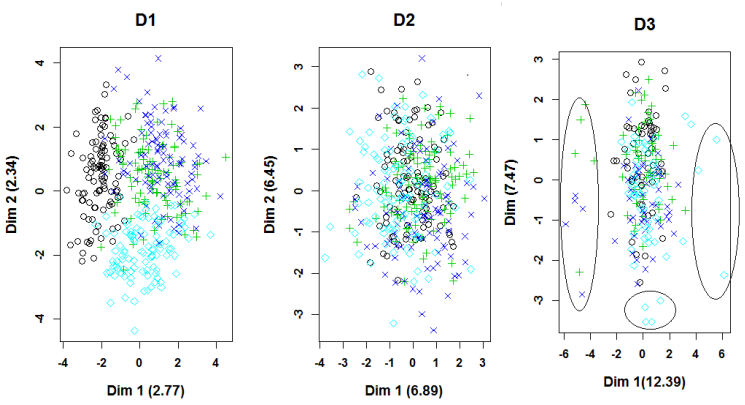


FIGURE 4.1 – Projection des classes des tables D1, D2 et D3 dans le premier plan factoriel d'une ACP ; les observations atypiques sont entourées.

Pour chaque jeu de données 2S-SOM a été appliqué à travers plusieurs initialisations des paramètres de l'algorithme : choix des centres initiaux, dimensions de la carte, taille du

voisinage, nombre d'itérations et les paramètres λ et η. Pour chaque couple de paramètres (λ, η) fixé la meilleure carte, celle minimisant simultanément l'erreur de quantification vectorielle et l'erreur topologique a été retenue pour chaque table (cf. TAB 4.1). D'autres indices internes de performances tels que la mesure de distorsion auraient pu être utilisés.

4.4.2 Comparaison des performances

Nous utiliserons ici les indices externes de précision, de rappel, de F-mesure et le coefficient de pureté d'une partition sur des données labellisées. L'algorithme 2S-SOM a fournit des cartes de tailles relativement grandes (TAB 4.1). Une classification ascendante hiérarchique sous contraintes utilisant la stratégie d'agrégation de ward a été appliquée sur les référents des cartes finales pour obtenir des partitions en un nombre de classes identique au nombre de labels dans les jeux de données initiaux. Ainsi, les cellules des cartes IS, DMU, CT10, D1, D2 et D3 ont été regroupées en respectivement 7, 10, 10, 4, 4 et 4 classes.

Données	Cl 1	Cl 2	Cl 3	Cl 4
D1	100	100	99	101
D2	100	95	107	98
D3	237	146	10	7

TABLE 4.2 – Répartition des observations dans les classes de la CAH

Les comparaisons des résultats (TAB 4.3) montrent de meilleures performances de 2S-SOM par rapport à FGKM sur tous les jeux de données et pour l'ensemble des indices de performance. Il en est de même pour 2S-SOM comparé à SOM et à EWKM à l'exception de la mesure de rappel pour la base CT pour SOM et pour la base DMU pour EWKM. Comparé à la méthode des K-Moyennes, 2S-SOM montre de meilleures performances sur les bases DMU et IS à l'exception du rappel pour la base IS. Pour la base CT, comparées aux classes des bases IS, DMU, D1, D2 et D3, les classes initiales sont déséquilibrées en termes de nombres d'observations, en particulier CT contient 4 classes de forts effectifs, que la stratégie d'agrégation de Ward utilisée ici ne permet pas de reconstituer convenablement. En effet, d'autres stratégies d'agrégation, notamment le lien minimum, permettent d'améliorer les performances en termes de précision (0.53), rappel (0.51), F mesure (0.50) et de pureté (0.50). Sur les données simulées D1, D2 et D3, la méthode 2S-SOM se révèle meilleure que l'ensemble des autres méthodes pour tous les indices. Les faibles performances des K-moyennes et SOM sont peut être dues à l'incapacité de ces méthodes à ignorer les variables de bruit ou les blocs de bruit et à l'absence d'une structure globale de corrélation entre les variables. Comparée aux méthodes de type subspace clustering FGKM et EWKM

la méthode proposée est meilleure en terme de performances pour tous les indices.

	Données	Indices	kM	EWKM	FGKM	SOM	2S-SOM
		Précision	0.38	0.66	0.60	0.63	**0.71**
		Rappel	**0.93**	0.70	0.63	0.67	0.74
	IS	F-mesure	0.50	0.64	0.59	0.59	**0.69**
		Pureté	0.41	0.59	0.63	0.61	**0.63**
		Precision	**0.50**	0.45	0.40	0.44	0.47
		Rappel	**0.53**	0.48	0.38	0.52	0.49
Réelles	CT	F-mesure	**0.48**	0.45	0.27	0.44	0.45
		Pureté	**0.47**	0.43	0.38	0.45	0.45
		Precision	0.59	**0.81**	0.60	0.75	0.80
		Rappel	0.61	**0.84**	0.80	0.78	0.82
	DMU	F-mesure	0.59	**0.80**	0.62	0.74	**0.80**
		Pureté	0.61	**0.77**	0.40	0.72	**0.77**
		Precision	0.37	0.98	0.90	0.31	**0.99**
		Rappel	0.35	0.65	0.60	0.28	**0.65**
	D1	F-mesure	0.36	0.77	0.77	0.29	**0.78**
		Pureté	0.47	0.72	0.72	0.38	**0.74**
		Précision	0.46	0.37	0.70	0.28	**0.87**
		Rappel	0.45	0.34	0.54	0.26	**0.61**
Simulées	D2	F-mesure	0.45	0.36	0.60	0.27	**0.70**
		Pureté	0.58	0.45	0.61	0.33	**0.71**
		Précision	0.33	0.35	0.75	0.35	**0.90**
		Rappel	0.28	0.30	0.48	0.27	**0.48**
	D3	F-mesure	0.31	0.36	0.61	0.29	**0.62**
		Pureté	0.37	0.47	0.49	0.35	**0.51**

TABLE 4.3 – Performances des classifications de 2S-SOM sur les données réelles et sur les bases D1, D2 et D3

4.4.3 Visualisation et détection des variables et des blocs de bruit

Données simulées

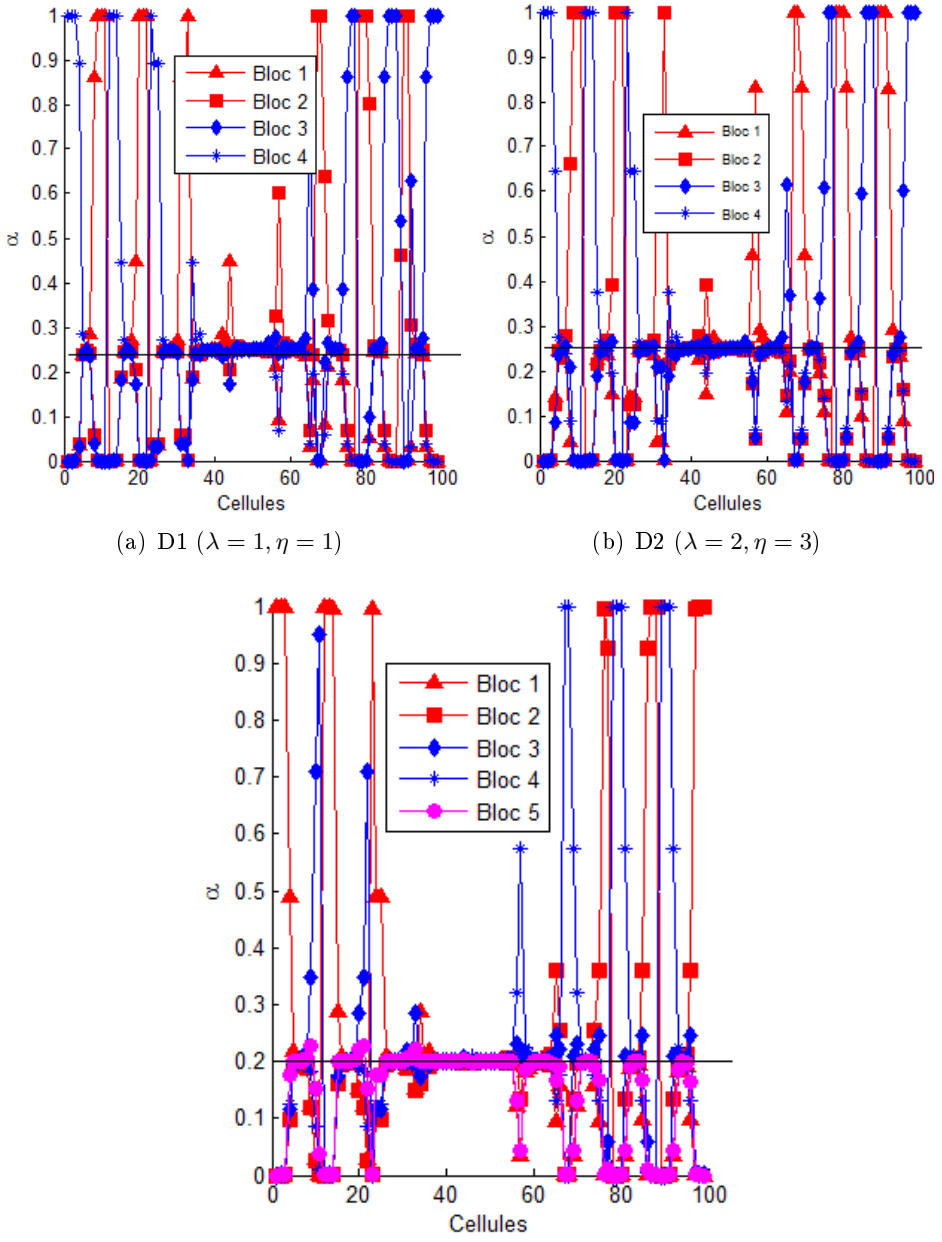
Les figures 4.2 représentent respectivement les poids α_{cb} attribués par 2S-SOM aux blocs dans les cellules des cartes topologiques associées aux tables D1, D2 et D3. Elles illustrent qu'à travers les poids forts, chaque bloc se spécialise dans la définition d'un certain nombre de cellules. Sur la figure 4.2(a) par exemple, les blocs 1 et 4 caractérisent mieux les observations appartenant à la première moitié des cellules de la carte. Les blocs 2 et 3 définissent les observations appartenant à la deuxième moitié des cellules de la carte.

Le constat est identique pour la table D2 avec les blocs 1 et 3 qui caractérisent les premières cellules et les blocs 2 et 4 les dernières cellules. Pour la table D3, on observe que les blocs 1, 2 contenant le moins de variables de bruit caractérisent un plus grand nombre de cellules parmi les premières et les dernières mais de façon exclusive. Les blocs 3 et 4 présentant plus de bruit donnent néanmoins des poids importants à un ensemble de cellules. Le bloc 5 porteur d'aucune information (il est composé uniquement de variables uniformes) n'est influent pour aucune cellule de la carte associée à la table D3 (Figure 4.2(c)). De manière générale, on observe que l'importance d'un bloc pour une cellule traduit la présence de variables fortement informatives pour la cellule par rapport aux autres blocs. En d'autres termes, la méthode proposée est robuste, par rapport à la présence de blocs contenant des variables de bruit.

Nous analysons dans la suite les poids des variables de la table D3 pour illustrer la capacité de l'algorithme à distinguer les variables de bruit et celles informatives. Dans les cellules ayant des poids forts pour le bloc 1 (figure 4.2(c)), les variables 1 et 5 (celles simulées avec une distribution uniforme) ont des poids faibles et ne sont donc pas prises en compte par 2S-SOM (figure 4.3(a)). Il en est de même pour les variables 1 et 3 du bloc 2 (figure 4.3(b)). Les blocs 3 et 4 mettent en évidence une seule variable pertinente (la seule à distribution non uniforme) pour les cellules influencées par ce bloc (figures 4.3(c) et 4.3(d)).
On constate aussi que dans le bloc 5 où toutes les variables ont une distribution uniforme, les poids sont quasi-identiques pour toutes les variables (Figure 4.3(e)).

Données réelles

Dans la suite, nous illustrons graphiquement la méthode uniquement sur la base IS, les résultats sont similaires pour les autres jeux de données.
La figure 4.4(a) donne une représentation graphique des poids α_{cb} définis sur les blocs par rapport aux cellules de la carte.
Pour les blocs, on observe que pour une majorité des cellules (plus de 67 %) les poids associés au bloc 2 sont nettement supérieurs à ceux du bloc 1. En d'autres termes, le bloc 2 apparaît donc plus pertinent pour déterminer la structure des classes de la carte. Une étude descriptive préalable de la table IS à travers une analyse en composantes principales permet, a priori, de supposer ce résultat. En effet, la visualisation de la répartition des individus dans les classes sur le premier plan factoriel d'une ACP sur toutes les variables (figure 4.4(b), IS) ou uniquement sur les variables de chaque bloc (figure 4.4(b), IS : bloc



(a) D1 ($\lambda = 1, \eta = 1$)

(b) D2 ($\lambda = 2, \eta = 3$)

(c) D3 ($\lambda = 1, \eta = 5$)

FIGURE 4.2 – Les poids α_{cb} associés aux blocs par rapport aux cellules des cartes associées aux tables D1, D2 et D3

1) et (figure 4.4(b) IS : bloc 2), permet de voir que le bloc 2 permet de mieux distinguer les classes.

Pour les variables, on formule, l'hypothèse qu'un poids β_{cbj} sur une variable v_j peut être considéré comme important s'il est supérieur au poids moyen $\frac{1}{p_b}$ de son bloc d'appar-

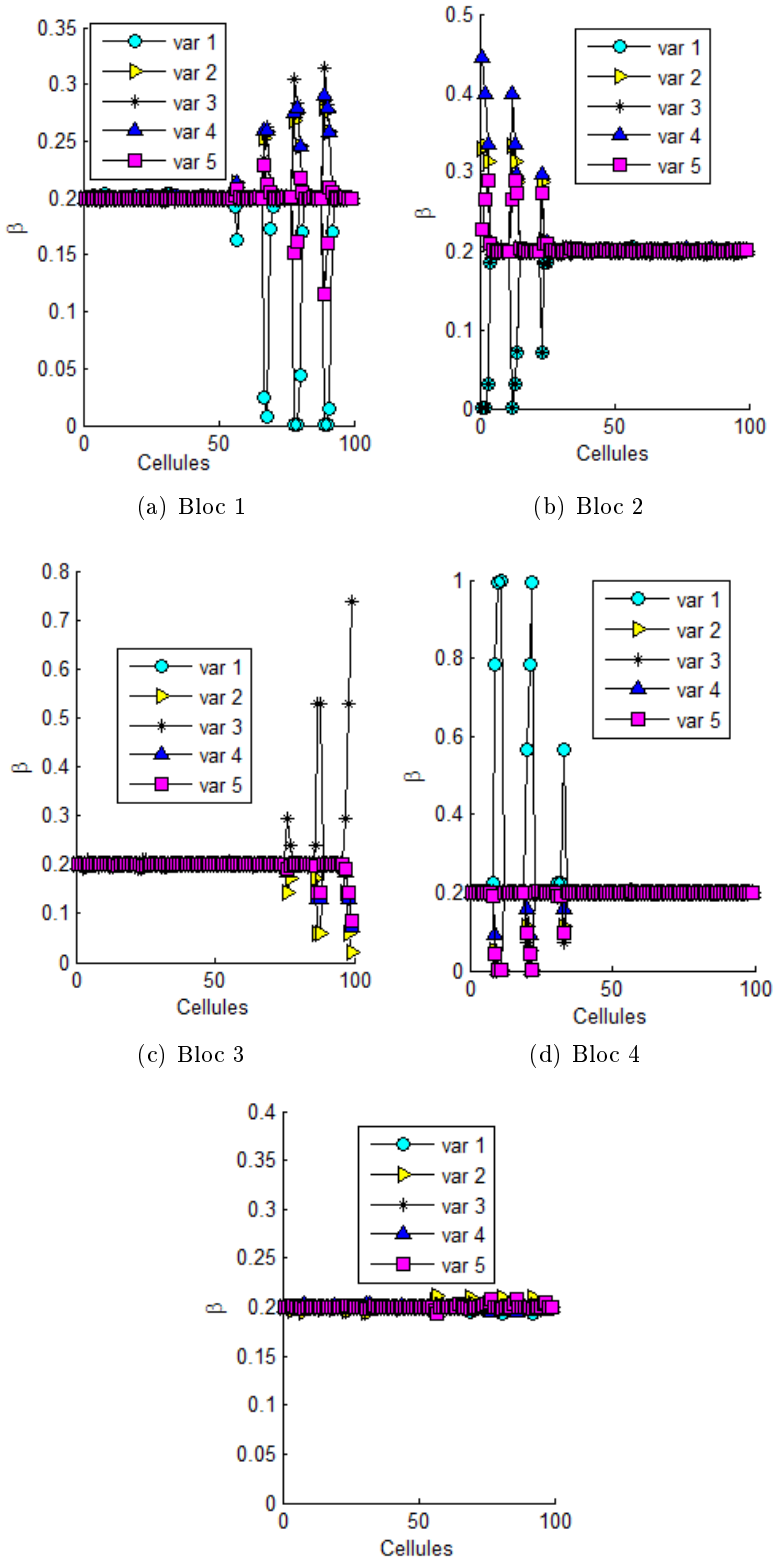
tenance. Les figures 4.5 à 4.6 représentent les poids des variables en fonction des cellules de la carte topologique fournie par 2S-SOM. On observe, dans le bloc 1, au seuil $\frac{1}{p_1} = 0.11$, que les variables 3, 7, 8 et 9 sont influentes dans la plupart des cellules de la carte. Les variables 1, 2, 4 sont très peu influentes dans la plupart des cellules 30 à 50 et les variables 5 et 6 définissent les cellules comprises entre 50 et 60. Dans le bloc 2, les variables 1, 2, 3, 4 et 8 sont fortement influentes dans la majorité des cellules comprises entre 40 et 55. A l'opposée, les variables 7 et 10 ne sont pas pertinentes pour ces cellules. Plus précisément, le sous-espace associé à la cellule 40 de la carte IS par exemple est constitué des variables 1, 2, 3, 4, 8 et 9 du bloc 2 uniquement relativement au seuil 0.10 fixé (Cf. les figures 4.6(a), 4.6(b), 4.6(c), 4.6(d), 4.6(e), 4.6(i) et 4.6(j)).

Nous illustrons maintenant les propriétés globales de 2S-SOM relativement aux paramètres λ et η. Les figures 4.7(a) et 4.7(b) représentent l'évolution de l'erreur de quantification vectorielle en fonction du couple λ et η. Lorsque les valeurs des paramètres λ et η augmentent, l'algorithme se stabilise au sens de l'erreur de quantification vectorielle ; leur influence sur cette dernière devient négligeable, 2S-SOM est alors semblable à SOM. Il apparaît deux valeurs particulièrement faibles assimilables à des "*outliers*" . Elle correspondent à un paramétrage d'apprentissage ($\lambda = 2$, $\eta = 16$, $\lambda = 11$, $\eta = 11$) qui dégrade la qualité topologique de la carte fournie par 2S-SOM.

Les figures 4.7(c) et 4.7(d) représentent l'évolution des moyennes des poids des cellules de la carte associée à chaque paramètre λ lorsque η est fixé. (Nous l'illustrons avec λ =3 (figure 4.7(e)) et $\eta = 3$ (figure 4.7(d))). On observe que lorsque $\lambda \to \infty$, les poids définis par l'algorithme 2S-SOM donnent les mêmes poids moyens aux blocs par cellule (cf. figure 4.7(c)). Tout se passe comme si on équilibrait l'influence des blocs, on ne prend en compte que les variables. Les poids moyens des variables définis par 2S-SOM pour l'ensemble des cellules de la carte permettent de regrouper les variables des blocs en groupes. Ainsi, dans le bloc 1 par exemple, les variables 3, 7 et 9 sont les plus pertinentes ($\beta_{cbj} > 0.11$) pour les cellules alors que les variables 4 et 5 sont moyennement pertinentes dans les cellules de la carte ($\beta_{cbj} \approx 0.11$) et les variables 1, 2, 6 et 8 aux poids inférieurs au seuil apportent de très faibles contributions dans les cellules de la carte. Il en est de même pour le bloc 2. (cf. figure 4.7(e)). Considérant chaque cellule de la carte lorsque $\lambda \to \infty$ pour η fixé, 2S-SOM permet de faire du subspace clustering en sélectionnant les variables ayant les plus forts poids pour la cellule.

Lorsque λ fixé et $\eta \to \infty$, on observe des situations inverses : l'algorithme 2S-SOM fournit des poids quasi identiques pour toutes les variables et privilégie donc les blocs. On peut alors déterminer les blocs importants pour chaque cellule de la carte, on fait ainsi de la

sélection de blocs plutôt que des variables séparément.
Au niveau de la visualisation, on observe une bonne conservation de la topologie des observations sur les cartes fournies par 2S-SOM sur chaque table. En effet, la méthode 2S SOM hérite des propriétés de SOM, en particulier les observations aberrantes sont isolées dans les classes 3 et 4 de la table D3 (Cf. figure 4.8).

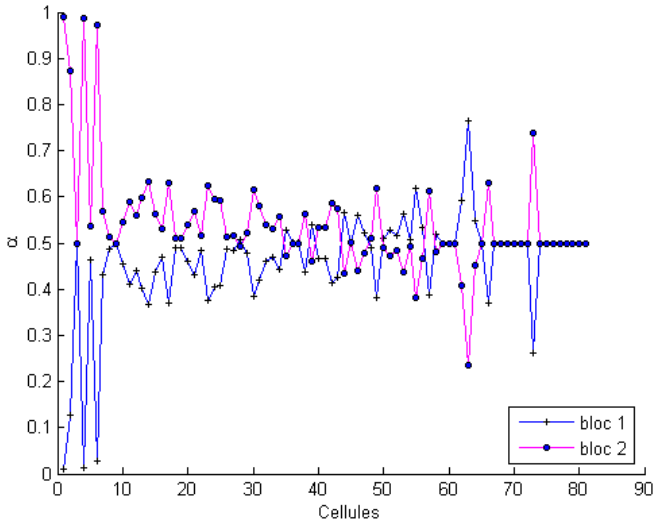


(a) Bloc 1 (b) Bloc 2

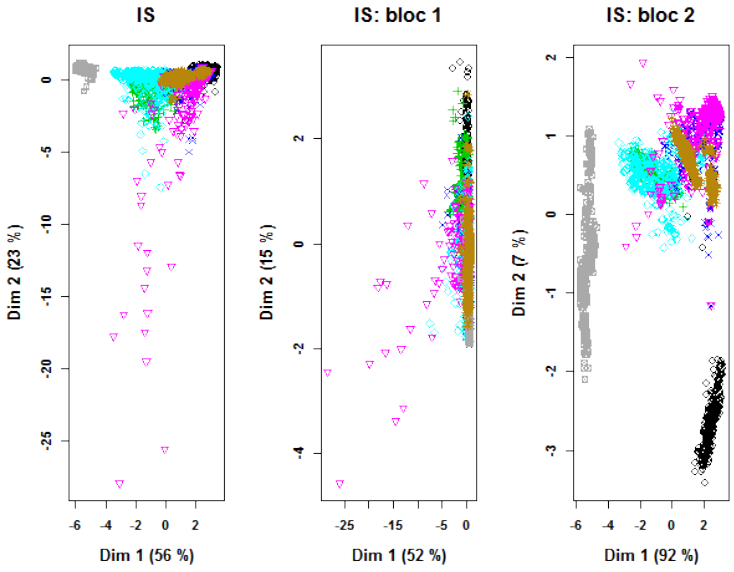
(c) Bloc 3 (d) Bloc 4

(e) Bloc 5

FIGURE 4.3 – Représentation des poids β des variables de la table D3 dans chaque bloc ($\lambda = 1$, $\eta = 5$)

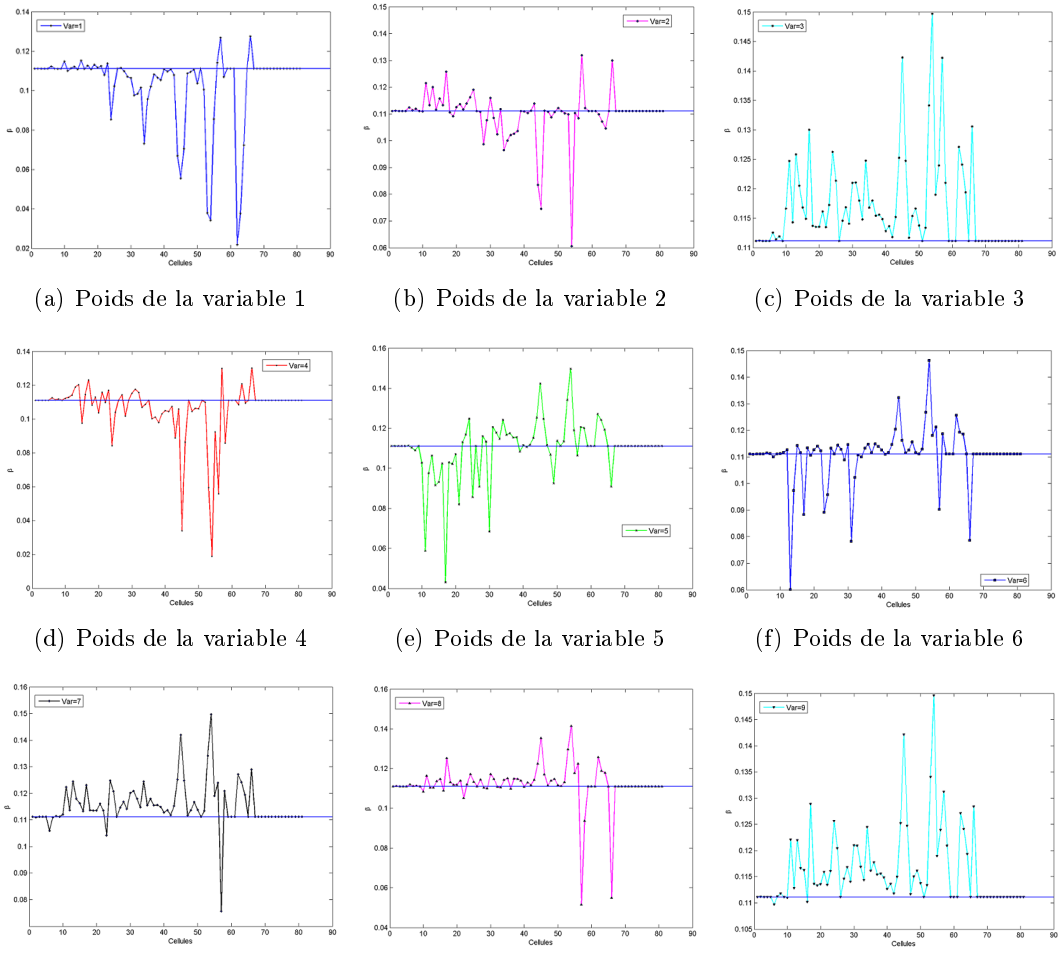


(a) Les poids α des blocs sur les cellules de la carte IS



(b) Projection des observations dans le premier plan factoriel défini par une analyse en composante principale sur la table IS

FIGURE 4.4 – Évaluation de la pertinence des blocs dans les cellules de la carte IS

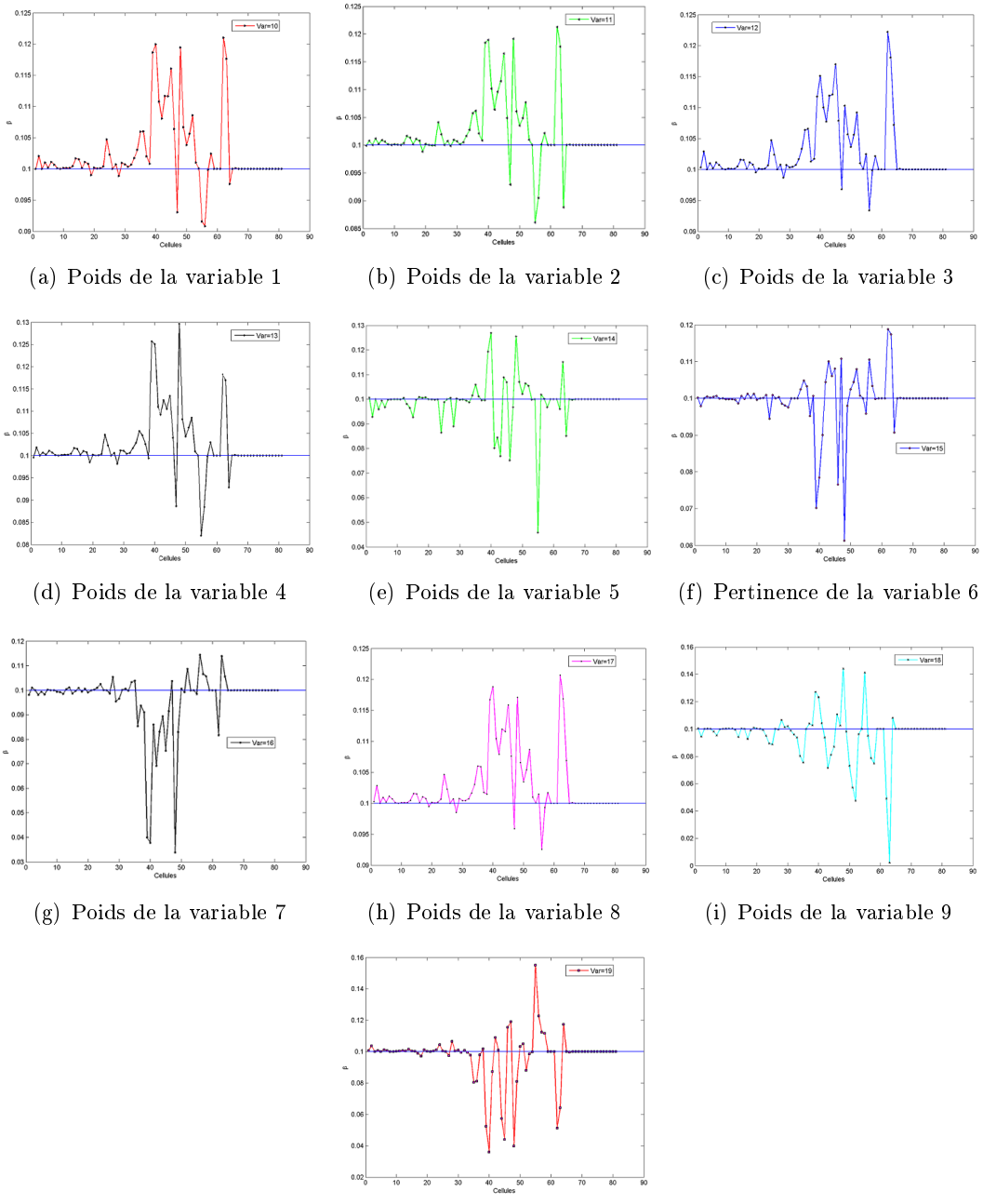


(a) Poids de la variable 1 (b) Poids de la variable 2 (c) Poids de la variable 3

(d) Poids de la variable 4 (e) Poids de la variable 5 (f) Poids de la variable 6

(g) Poids de la variable 7 (h) Poids de la variable 8 (i) Poids de la variable 9

FIGURE 4.5 – Représentation des poids β_{cbj} associés aux variables du bloc 1 par rapport au 81 cellules de la carte IS ; la ligne horizontale définit le seuil $\frac{1}{p_1} = 0.11$



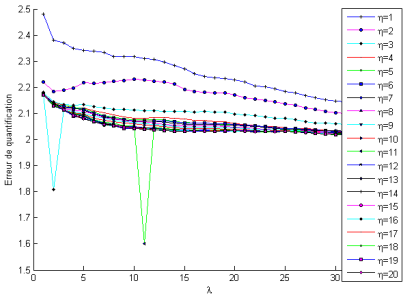
(a) Poids de la variable 1 (b) Poids de la variable 2 (c) Poids de la variable 3

(d) Poids de la variable 4 (e) Poids de la variable 5 (f) Pertinence de la variable 6

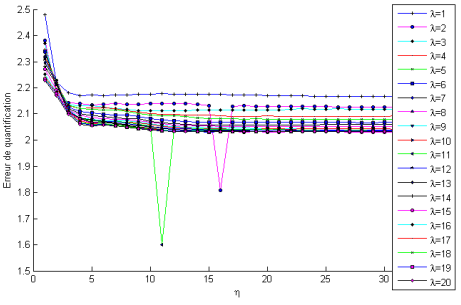
(g) Poids de la variable 7 (h) Poids de la variable 8 (i) Poids de la variable 9

(j) Poids de la variable 10

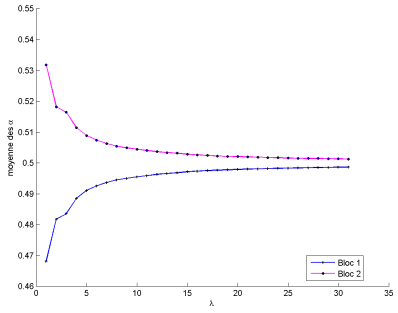
FIGURE 4.6 – Représentation des poids β_{cbj} associés aux variables du bloc 2 par rapport au 81 cellules de la carte IS ; la ligne horizontale définit le seuil $\frac{1}{p_2} = 0.10$



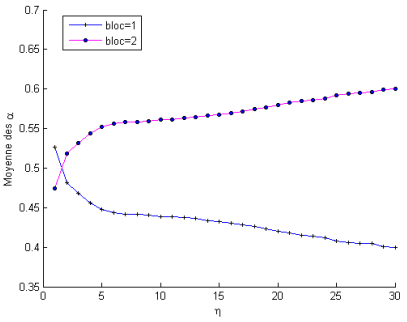
(a) L'évolution de l'erreur de quantification vectorielle par rapport au couple (λ, η), en abscisse les λ



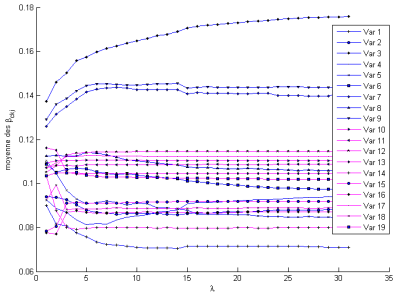
(b) L'évolution de l'erreur de quantification vectorielle par rapport au couple (λ, η), en abscisse les η



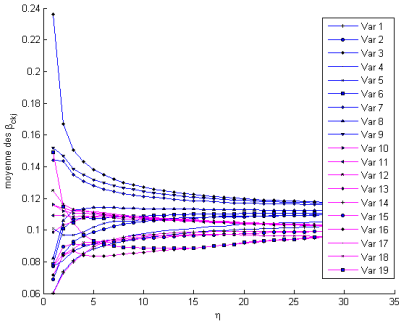
(c) L'évolution de la moyenne des poids des blocs sur les cellules pour les cartes obtenues avec λ variant et $\eta = 3$



(d) L'évolution de la moyenne des poids des blocs sur les cellules pour les cartes obtenues avec $\lambda = 3$ et η variant

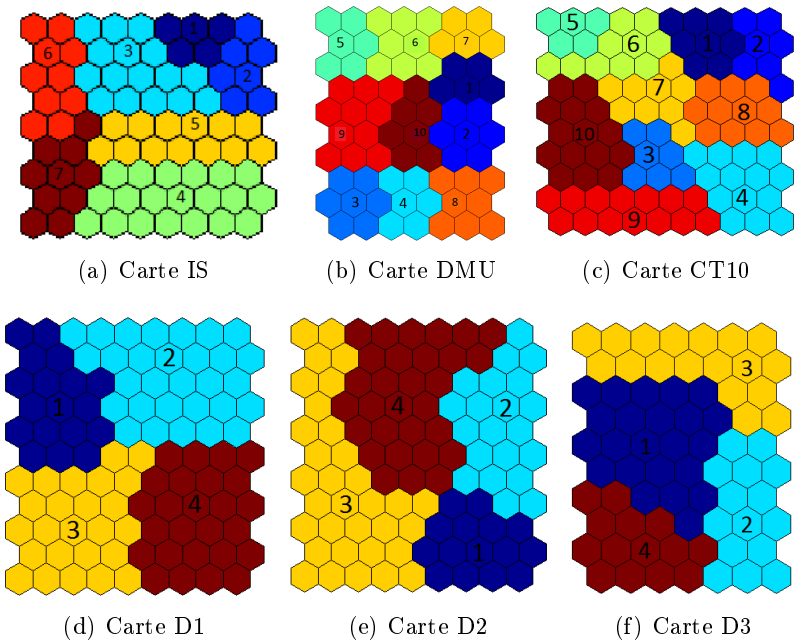


(e) L'évolution de la moyenne des poids des variables sur les cellules pour les cartes obtenues avec λ variant et $\eta = 3$



(f) L'évolution de la moyenne des poids des variables sur les cellules pour les cartes obtenues avec $\lambda = 3$ et η variant

FIGURE 4.7 – Les propriétés de 2S-SOM par rapport aux paramètres λ et η



(a) Carte IS (b) Carte DMU (c) Carte CT10

(d) Carte D1 (e) Carte D2 (f) Carte D3

FIGURE 4.8 – Visualisation des classes sur les cartes fournies par une CAH appliquée sur les cellules résultantes

4.5 Conclusion

La méthode 2S-SOM proposée dans ce chapitre permet de faire une classification d'un ensemble d'individus décrits par un grand ensemble de variables structurées en blocs et pouvant présenter des données aberrantes ou manquantes. Son application sur des données étiquetées fournit des partitions, globalement, plus en adéquation avec les partitions de référence que celles obtenues avec les méthodes SOM, K-Moyennes, EWKM et FGKM. Cependant, comme la méthode SOM, elle est sensible aux paramètres d'initialisation, ce qui pose un problème de stabilité des résultats.
La deuxième contribution de cette thèse, objet du chapitre suivant, permet de prendre en compte cet aspect. En effet, plutôt que de choisir la carte la meilleure, il est possible d'agréger plusieurs cartes afin d'améliorer les performances globales de la carte finale.

Chapitre 5

Fusion d'ensemble de SOM

5.1 Introduction

Ce chapitre s'intéresse au consensus de cartes topologiques pour des données multi-blocs. Lorsque les données sont de type multi-blocs, chaque bloc b étant de dimension p_b, les dimensions des cartes varient naturellement selon les blocs. Ainsi, les méthodes classiques de fusion de SOM qui imposent que la taille de la carte finale soit identique à celle des cartes formant l'ensemble de diversification sont inefficaces [Georgakis *et al.* 2005; Saavedra *et al.* 2007]. De plus, il est difficile d'évaluer la distance entre deux référents définis sur des blocs de tailles différentes.

Nous proposons dans ce chapitre, deux approches de fusion de SOM qui prennent en compte la structuration en bloc des variables que nous appellerons CSOM (Consensus SOM) et R_V-CSOM. L'idée du consensus dans ces deux approches est formalisée différemment : dans l'approche CSOM, l'objectif est de favoriser les cartes les meilleures au sens d'un certain critère de validation.

CSOM est donc une approche directe de type hiérarchique consistant en une double application de SOM. Une première application de SOM sur les blocs permet de réduire la variabilité des données en fournissant un résumé synthétique des données par des référents, ce qui peut être assimilé à l'ensemble de diversification obtenue par échantillonnage (les variables sont différentes par bloc). Au niveau 2, une première version de CSOM consiste à appliquer SOM sur la table contenant les résultats du niveau 1. Puis, dans une deuxième version dite pondérée, CSOM intègre dans la carte consensus finale, l'influence relative de chaque carte topologique formant l'ensemble de diversification à travers un mécanisme de

pondération locale s'appuyant sur la pertinence des cartes au sens de la mesure de distorsion.

La deuxième approche de recherche de consensus de SOM repose sur la prise en compte de l'information commune à toutes les partitions. Contrairement à CSOM pondéré, ici les poids reposent sur les relations entre les blocs, en ce sens, ils sont considérés comme plus globaux que les poids de CSOM qui sont propres à chaque carte. Cette approche va plutôt privilégier les partitions similaires. Son principe est basé sur la recherche d'une matrice compromis moyenne pondérée des matrices issues des partitions représentant au mieux la similarité entre les cartes topologiques. La fusion des cartes topologiques est alors obtenue à travers une classification de cette matrice compromis.
La section 5.2 présente le problème de fusion de SOM hiérarchique à deux niveaux basé sur une double application de l'algorithme SOM. La section 5.3 présente la méthode de fusion de SOM à travers le partitionnement d'une matrice consensus.

5.2 Approche directe de fusion de SOM (CSOM)

Comme présentée dans le chapitre 3.6.1, la démarche des méthodes de fusion de SOM se résume en une étape de diversification par la création d'un ensemble de cartes topologiques et en une étape agrégation de ces SOM.
Soit $\Pi = \{\mathcal{C}^1, \ldots, \mathcal{C}^b, \ldots, \mathcal{C}^B\}$ l'ensemble de diversification des cartes topologiques, $\mathcal{C}^b = (W^b, \pi^b)$, avec π^b et $W^b = \{w_1^b, \ldots, w_{N_b}^b\}$ définissant respectivement une partition des observations relativement au bloc b et la matrice des vecteurs référents $W^b(N_b \times p_b)$ de la carte $\mathcal{C}^b$; N_b est le nombre de cellules de la carte $\mathcal{C}^b$. Soit $\mathcal{C}^* = (W^*, \pi^*)$ la carte représentant la fusion des B cartes $\mathcal{C}^b$. Le problème revient alors à définir la partition π^* et la matrice des vecteurs référents $W^* = \{w_1^*, \ldots, w_{N_{cell}}^*\}$ de la carte fusion des cartes $\mathcal{C}^b$.

5.2.1 Principe de la méthode

La structure multi-blocs des données est induite par un regroupement de certaines variables selon des critères ou thématiques spécifiques. Les observations sont représentées dans la diversification par des vecteurs référents de même dimension dans chaque bloc. On définit par $\tilde{z}_i$ une nouvelle représentation de l'observation z_i telle que $\tilde{z}_i = (w_i^1, \ldots, w_i^b, \ldots, w_i^B)$ avec $w_i^b \in R^{p_b}$ dans le cas de données continues (ou à $R^{p_b} \times \{0,1\}^{q_b}$ dans le cas des don-

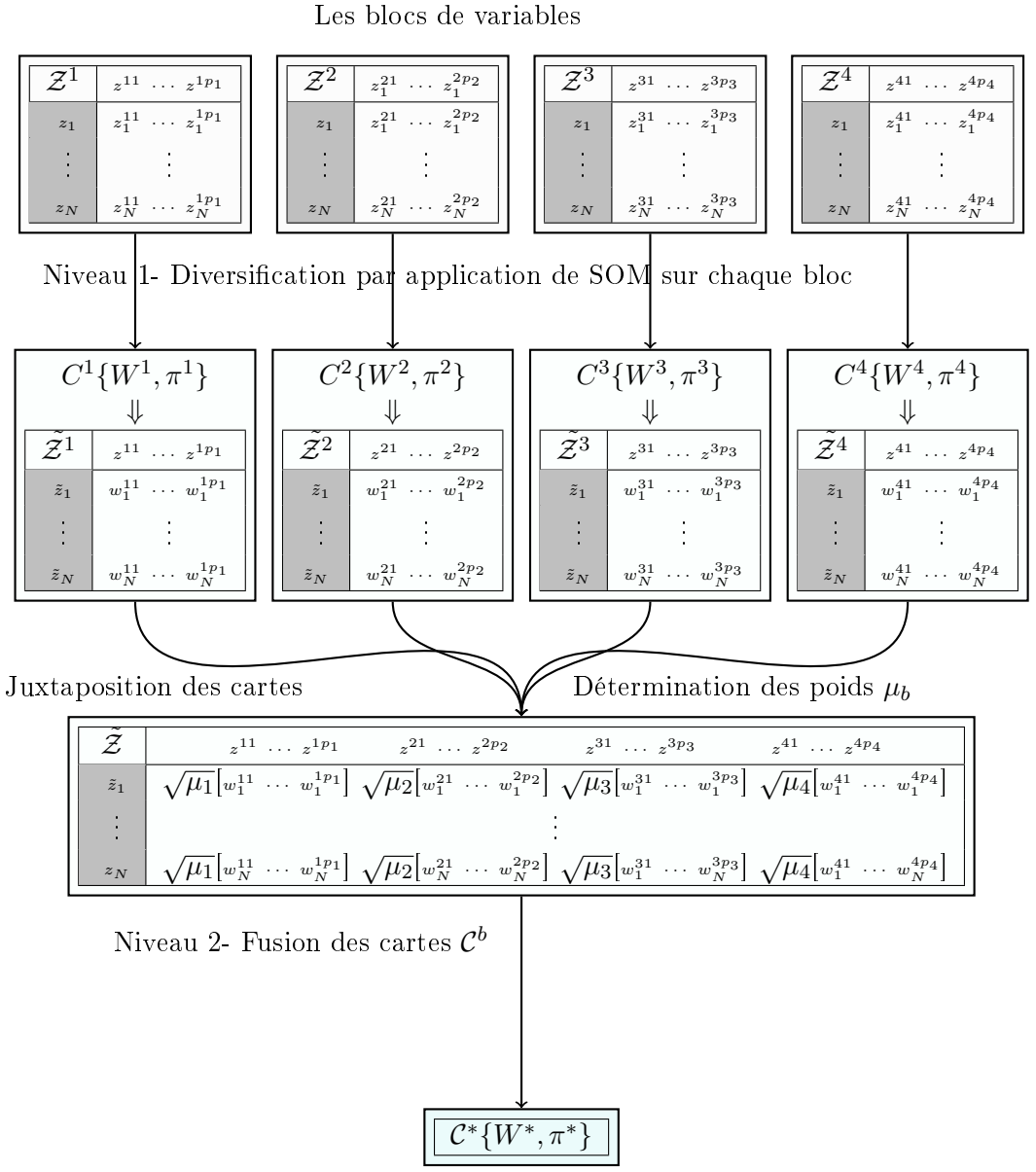


FIGURE 5.1 – Schéma des méthodes de fusion de SOM pour B=4 ; la diversification au niveau des blocs fournit les cartes $\mathcal{C}^1, \ldots, \mathcal{C}^4$ puis, la fusion donne la carte $\mathcal{C}^*$

nées mixtes). $\tilde{z}_i$ est donc la juxtaposition des vecteurs référents des cellules ayant capté l'observation z_i au niveau de chaque bloc. Soit $\tilde{\mathcal{Z}}$ $(N \times p)$ avec $p = p_1 + \ldots + p_B$ la matrice contenant les $\tilde{z}_i$. Chaque observation est remplacée par la concaténation de ces prototypes plutôt que par le label de sa classe dans la partition π^b permettant de conserver plus d'informations dans chaque bloc de variables. La figure 5.1 présente le schéma du mécanisme de fusion de SOM. La fusion finale se fait alors de deux manières : la première consiste à appliquer un algorithme de partitionnement topologique, en l'occurrence SOM, sur la matrice $\tilde{\mathcal{Z}}$. Cette approche sera notée CSOM$_S$. Ceci suppose un traitement équivalent des cartes formant l'ensemble de diversification dans la recherche du consensus. Or, les critères de validation d'une carte topologique présentés dans la section 2.4 montrent que les classifications obtenues au niveau de chaque bloc de variables peuvent être de qualités différentes. Il est donc possible d'établir, relativement à ces critères, une hiérarchie de qualité entre les cartes topologiques. Nous proposons donc de prendre en compte cette hiérarchie en introduisant des coefficients de pondération dans la fonction de coût de l'algorithme de fusion de SOM, (MTM, NCSOM pour les données mixtes). Cela consiste à remplacer, au niveau 2, la distance euclidienne classique utilisée dans l'algorithme SOM par la distance euclidienne pondérée :

$$d_f(\tilde{z}_i, \tilde{\omega}_c) = \sum_{b=1}^{B} \mu_b ||\tilde{z}_i^b - \tilde{w}_c^b||^2 \tag{5.1}$$

où μ_b est une quantité inversement proportionnelle à la valeur de l'indice de validation associée à chaque carte formant l'ensemble de diversification. Les vecteurs $\tilde{z}_i^b$ et $\tilde{w}_c^b$ sont associés respectivement aux observations et aux référents de la carte $\mathcal{C}^b$ du bloc b. Les poids μ_b vérifient $\sum_b \mu_b = 1$ et $0 \leq \mu_b \leq 1$. Dans le cas de la mesure de distorsion, on note par $Dist_b$ la valeur de l'indice de distorsion associée à la carte topologique du bloc b, divisée par la dimension p_b du bloc b. Les poids, dans l'étape de recherche de consensus, sont alors définis :

$$\mu_b = \frac{1}{Dist_b Dist_f} \tag{5.2}$$

où $Dist_f = \sum_{b=1}^{B} \frac{1}{Dist_b}$ est un terme de normalisation. L'algorithme de fusion CSOM accorde donc une importance particulière aux cartes de l'ensemble de diversification fournissant une meilleure quantification vectorielle et une meilleure conservation de la topologie des observations. Finalement, la fonction objectif de fusion de SOM est donnée par :

$$\mathcal{J}_{CSOM} = \sum_{\tilde{z}_i \in \tilde{\mathcal{Z}}} \sum_{c \in \mathcal{C}^*} \mathcal{K}^T(\sigma(\mathcal{X}(\tilde{z}_i), c)) \left(\sum_{b=1}^{B} \mu_b ||\tilde{z}_i^b - \tilde{w}_c^b||^2 \right) \tag{5.3}$$

où $\mathcal{C}^*$ est l'ensemble des cellules de la carte fusion. La relation 5.3 montre effectivement que cette nouvelle fonction objectif prend en compte la qualité relative des cartes de l'ensemble de diversification. L'optimisation de $\mathcal{J}_{CSOM}$, réalisée itérativement, est quasiment identique à celle de l'algorithme SOM :

- Affectation des observations :

$$\mathcal{X}(\tilde{z}_i) = \underset{c \in \mathcal{C}^*}{arg\ min} \sum_{c \in \mathcal{C}*} \mathcal{K}^T(\sigma(\mathcal{X}(\tilde{z}_i), c)) \sum_{b=1}^{B} \mu_b ||\tilde{z}_i^b - \tilde{w}_c^b||^2$$

- Actualisation des centres de classe

$$\tilde{w}_c^* = \frac{\sum_{\tilde{z}_i \in \tilde{\mathcal{Z}}} \mathcal{K}^T(\sigma(\mathcal{X}(\tilde{z}_i), c)) \tilde{y}_i}{\sum_{\tilde{z}_i \in \tilde{\mathcal{Z}}} \mathcal{K}^T(\sigma(\mathcal{X}(\tilde{z}_i), c))}$$

  où $y_i = [\sqrt{\mu_1}\tilde{z}_i^1, \ldots, \sqrt{\mu_B}\tilde{z}_i^B]$ et $\tilde{z}_i^b$ la portion de $\tilde{z}_i$ restreinte aux dimensions du bloc b.

Algorithme : Fusion de SOM

Entrée : L'ensemble des B blocs de variables

1. Diversifier : apprendre sur chaque bloc de variables une carte topologique.
2. Déterminer les poids μ_b à l'aide de la relation 5.2
3. Créer la matrice $\tilde{\mathcal{Z}}$ dont les entrées sont $\tilde{z}_i = y_i = [\sqrt{\mu_1}\tilde{z}_i^1, \ldots, \sqrt{\mu_B}\tilde{z}_i^B]$
4. Appliquer un algorithme de partitionnement topologique sur la matrice $\tilde{\mathcal{Z}}$ pour déterminer le consensus des cartes.

Sortie : la carte fusion des SOM

5.2.2 Évaluation de l'approche directe de fusion de SOM

La validation d'une approche en classification est d'utiliser des connaissances externes à la détermination des classes. Cela implique donc le calcul d'indices spécifiques ayant pour but de vérifier quelques-unes des propriétés de séparabilité, d'homogénéité, de compacité, etc. Le choix de l'indice dépend le plus souvent de l'objectif à atteindre. Cependant, en règle générale, une manière simple et cohérente d'évaluer la qualité des résultats d'un algorithme de classification consiste à utiliser les données étiquetés. On peut alors calculer la similitude entre les classes fournies par l'algorithme et les étiquettes préalablement définies sur les données. Par conséquent, on adopte une approche d'évaluation externe qui utilise des critères externes présentés dans la section 2.4.2. Il s'agit de la Pureté (Pur) pour sa facilité d'interprétation en terme de similarité, de l'Information Mutuelle Normalisée (NMI)

qui évalue la quantité d'information partagée par deux partitions et de la variation d'information (NVI) qui s'interprète comme une distance entre deux partitions, une propriété intéressante de ces deux derniers critères est leur insensibilité par rapport au nombre de classes. Afin de juger de la stabilité de l'approche proposée, nous présentons la moyenne et l'écart-type de ces indices pour différents apprentissages.

5.2.2.1 Les données

Nous utilisons les données du IS, CT, DMU présentées dans le chapitre 4 que nous complétons par un jeu de données simulé par les auteurs de la méthode FGKM contenant 600 observations décrites par 50 variables structurées en 3 blocs contenant 10, 10 et 30 variables. Ce jeu de données est disponible sur le CRAN du logiciel R.

5.2.2.2 Performance et visualisation

La méthode CSOM est comparée d'une part aux méthodes standards SOM et K-moyennes appliquées sur l'ensemble des données, d'autre part aux méthodes de cluster ensemble NMF et Weighted NMF proposées par Li *et al.* [2007] et à la méthode CSPA proposée par Strehl et Ghosh [2002]. Elle est aussi comparée à la même méthode simplifiée sans les poids CSOM_s.
L'ensemble de diversification constitué est formé des apprentissages obtenus par l'algorithme SOM sur chaque bloc de variables. Puis, on recherche le consensus des partitions obtenues au niveau de ces blocs de variables. Pour les méthodes basées sur SOM, lorsque le nombre de neurones de la carte finale est trop grand par rapport aux nombre de labels des données, une CAH sous contrainte de voisinage et consolidée permet ensuite de réduire le nombre de neurones sur les cartes. Le consensus est recherché 25 fois sur 25 ensembles de diversification. Le tableau 5.1 présente la moyenne des performances des algorithmes SOM, K-moyennes, NMF, WNMF, CSPA, CSOM et CSOM_s pour les 25 expériences.

On observe que la méthode CSOM a des performances en général meilleures que les méthodes standards SOM et K-moyennes et les méthodes de consensus de partitions NMF et WNMF pour les bases IS, CT3, FGKM. Sur la base DMU, comparée aux méthodes NMF, WNMF et CSPA, CSOM est meilleure cependant, on observe une dégradation de ses performances en terme de pureté par rapport à l'algorithme CSPA. De plus, les performances systématiquement bonnes de $CSOM$ par rapport à sa version sans poids CSOM_s montrent qu'il est important de prendre en compte à travers les poids μ_b la qualité relative

D	Indice	SOM	KM	NMF	WNMF	CSPA	CSOM	$CSOM_s$
IS	Pur	0.61(0.01)	0.61(0.02)	0.45(0.005)	0.58(0.001)	0.37(0.001)	**0.65(0.02)**	0.60(0.01)
	NMI	0.59(0.01)	0.62(0.01)	0.36(0.01)	0.51(0.01)	0.27(0.01)	**0.63(0.01)**	0.58(0.01)
	NVI	0.58(0.01)	0.56(0.01)	0.78(0.003)	0.65(0.003)	0.84(0.01)	**0.54(0.01)**	0.58(0.01)
CT	Pur	0.78(0.02)	0.78(0.02)	0.78(0.02)	0.78(0.02)	0.78(0.02)	0.78(0.02)	0.78(0.02)
	NMI	0.13(0.001)	0.15(0.001)	0.10(0.001)	0.10(0.001)	0.15(0.001)	**0.16(0.001)**	0.12(0.001)
	NVI	0.93(0.03)	0.92(0.03)	0.95(0.001)	0.96(0.001)	0.92(0.03)	**0.91(0.03)**	0.94(0.03)
DMU	Pur	0.78(0.02)	0.76(0.03)	0.84(0.001)	0.86(0.01)	**0.88(0.001)**	0.84(0.03)	0.75(0.03)
	NMI	0.75(0.02)	0.76(0.02)	0.79(0.005)	0.81(0.001)	0.83(0.001)	**0.85(0.03)**	0.76(0.02)
	NVI	0.39(0.001)	0.38(0.01)	0.34(0.001)	0.31(0.001)	0.28(0.001)	**0.26(0.001)**	0.39(0.01)
FGKM	Pur	0.55(0.02)	**0.61(0.03)**	0.53(0.001)	0.52(0.01)	0.51(0.001)	0.60(0.03)	0.59(0.03)
	NMI	0.18(0.02)	**0.27(0.02)**	0.12(0.005)	0.11(0.001)	0.12(0.001)	0.22(0.03)	0.22(0.02)
	NVI	0.89(0.001)	**0.84(0.01)**	0.94(0.001)	0.94(0.001)	0.93(0.001)	0.87(0.01)	**0.84(0.01)**

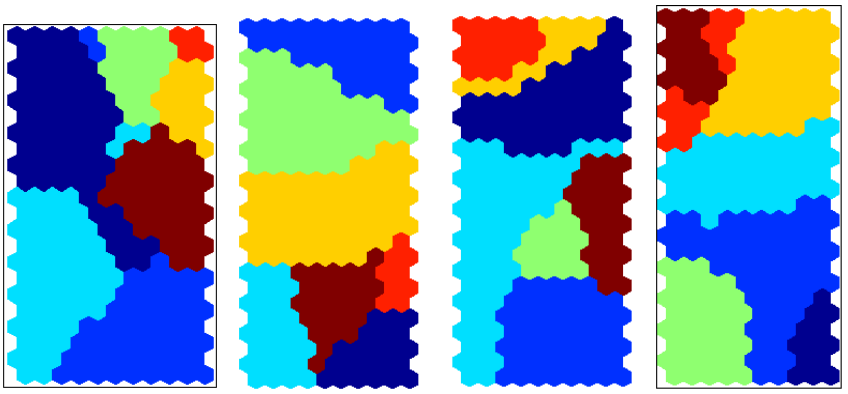
TABLE 5.1 – Performances des algorithmes SOM, K-moyennes (KM), NMF, WNMF, CSPA et Fusion de SOM (CSOM et $CSOM_s$) sur les bases IS, CT3, DMU et FGKM, les valeurs entre parenthèses sont les écarts-types des 25 apprentissages ; en gras les meilleurs algorithmes.

des partitions en terme de quantification et de conservation de la topologie des observations dans le consensus.

Les figures 5.2, 5.3, 5.5 et 5.4 présentent les classes sur les cartes topologiques fournies

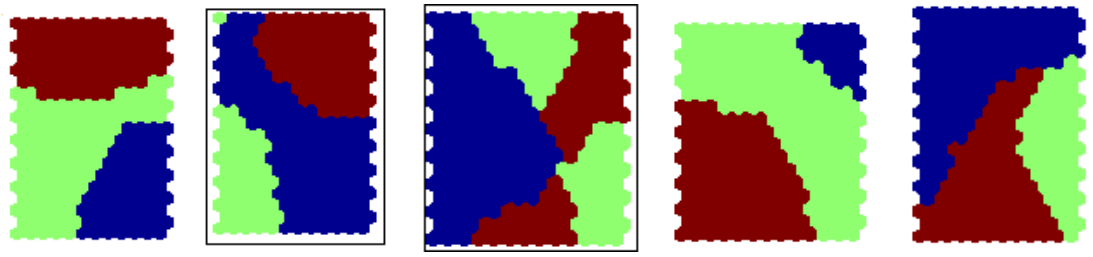
par SOM au niveau des blocs et sur les cartes finales des algorithmes CSOM et CSOM$_s$. Les valeurs entre parenthèses correspondent à la moyenne des 25 mesures de distorsion obtenues au niveau des blocs. Les figures 5.2(c), 5.3(d) et 5.4(d) obtenues par CSOM montrent une bonne conservation de la topologie des observations pour l'ensemble des tables. Pour la base IS par exemple, la carte du bloc 1 a une mauvaise structure topologique caractérisée par le fait que des neurones appartenant à une même classe se trouvent dans des regions éloignés sur la carte (figure 5.2(a)). La fusion tenant compte du bloc 2 ayant une meilleure conservation de la topologie permet d'atténuer ce phénomène 5.2(c).

Nous avons proposé dans cette section une approche directe de Fusion de SOM dédiée au traitement des données structurées en blocs de variables. L'utilisation de la mesure de distorsion pour pénaliser les partitions non-performantes, en termes de quantification et de conservation de la topologie des observations, permet d'améliorer significativement les performances de classification sur les jeux de données utilisés pour l'évaluation des performances de CSOM. L'utilisation d'autres critères tels que l'indice de Davie-Bouldin ou la "silhouette value" aurait pu permettre d'améliorer les performances en termes de séparabilité et d'homogénéité des données au détriment de la visualisation. L'application de l'algorithme CSOM sur la matrice $\tilde{\mathcal{Z}}$ permet à la carte C^* de prendre naturellement en compte la topologie des observations. Cependant, elle ne garantit pas une conservation de la topologie locale des observations au niveau des blocs. Malgré ces performances, les poids μ_b sont déterminés uniquement à partir des cartes SOM sans prendre en compte explicitement les similitudes entre les cartes SOM au niveau des blocs. Nous proposons, dans la deuxième approche, de tenir compte des liaisons entre les cartes topologiques.



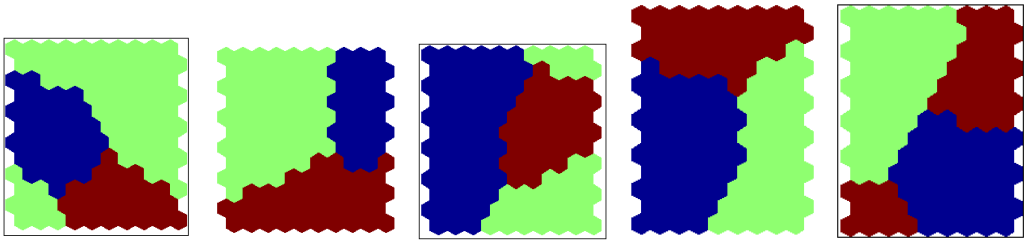
(a) Bloc 1 (0.35) (b) Bloc 2 (0.65) (c) CSOM (d) CSOM$_s$

FIGURE 5.2 – Visualisation des classes des cartes fournies par une CAH sous contrainte et consolidée par l'algorithme des K-moyennes sur les référents des cartes obtenues sur la base IS ; les cartes encadrées sont les moins performantes en terme de visualisation



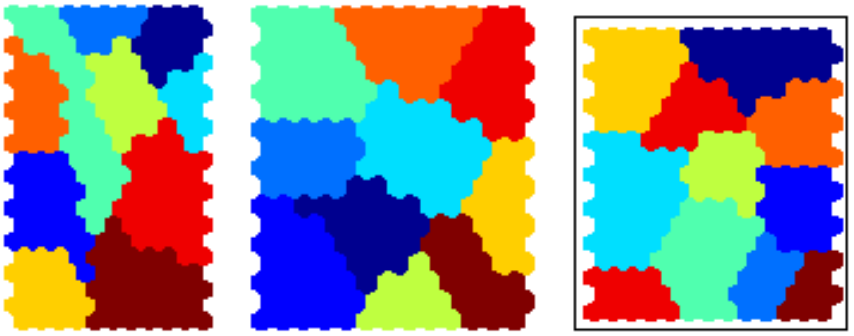
(a) Bloc 1 (0.23) (b) Bloc 2 (0.46) (c) Bloc 3 (0.31) (d) CSOM (e) CSOM$_s$

FIGURE 5.3 – Visualisation des classes des cartes fournies par une CAH sous contrainte et consolidée par l'algorithme des K-moyennes sur les référents des cartes obtenues sur la base CT ; les cartes encadrées sont les moins performantes en terme de visualisation

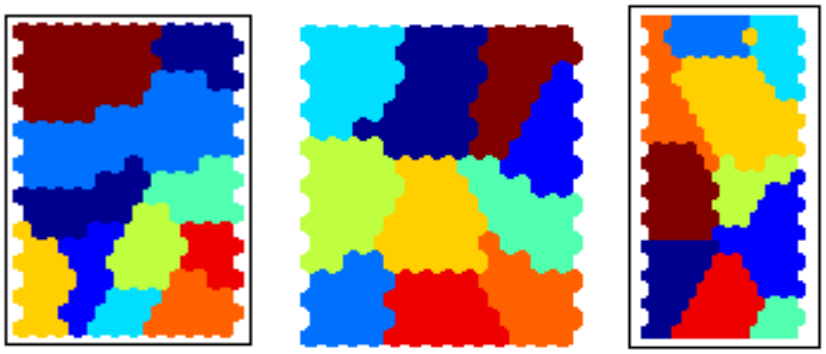


(a) Bloc 1 (0.35) (b) Bloc 2 (0.36) (c) Bloc 3 (0.29) (d) CSOM (e) CSOM$_s$

FIGURE 5.4 – Visualisation des classes des cartes fournies par une CAH sous contrainte et consolidée par l'algorithme des K-moyennes sur les référents des cartes obtenues sur la base FGKM ; les cartes encadrées sont les moins performantes en terme de visualisation



(a) Bloc 1 (0.07) (b) Bloc 2 (0.10) (c) Bloc 3 (0.06)



(d) Bloc 4 (0.08) (e) Bloc 5 (0.10) (f) Bloc 6 (0.58)

FIGURE 5.5 – Visualisation des classes des cartes fournies par une CAH sous contrainte et consolidée par l'algorithme des K-moyennes sur les référents des cartes obtenues sur la base DMU ; Les cartes encadrées sont les moins performantes en terme de visualisation

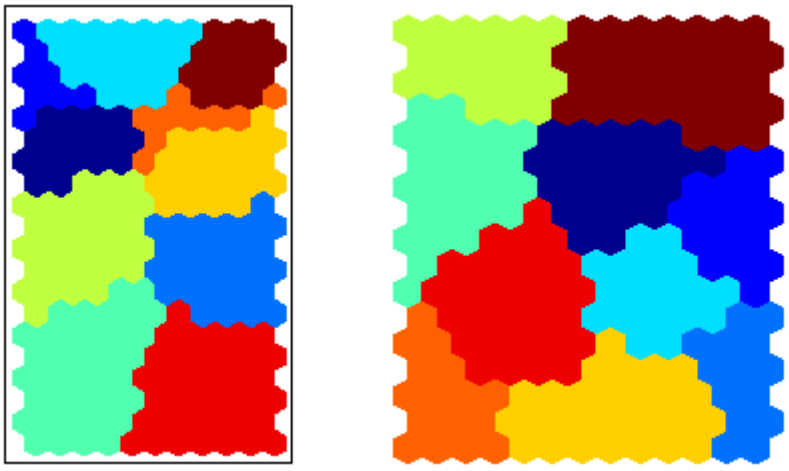


FIGURE 5.6 – Les cartes consensus fournies par les algorithmes CSOM (à gauche) et CSOM$_s$ (à droite)

5.3 Consensus fondé sur une matrice compromis, R_V-CSOM

Cette approche recherche le consensus en se basant sur le coefficient de corrélation vectorielle RV introduite par Robert et Escoufier [1976] pour déterminer un compromis entre les partitions de l'ensemble diversification.

5.3.1 Fusion de SOM fondée sur les matrices des référents $\tilde{Z}^b$

Nous proposons maintenant de fusionner les cartes topologiques en tenant compte de l'interrelation entre les blocs. On dispose des tableaux $\tilde{Z}^b$, $b = 1, \ldots, B$ de dimension $(N \times p_b)$. Dans chaque matrice $\tilde{Z}^b$, les individus sont décrits par les vecteurs référents des cellules les ayant capté sur la carte topologique correspondante au bloc b. On associe à chaque matrice $\tilde{Z}^b$, la matrice $X^b = \tilde{Z}^b \tilde{Z}^{b\prime}$ $(N \times N)$ des produits scalaires inter-individus, où $\tilde{Z}^{b\prime}$ est la matrice transposée de $\tilde{Z}^b$. C'est un objet représentatif du bloc $\tilde{Z}^b$. Il contient tous les liens inter-individus du point de vue de leur représentation par $\tilde{Z}^b$. Le problème ici consiste alors à rechercher la matrice de taille $N \times N$ qui réalise un bon compromis entre les liens inter-individus au regard de tous les blocs. On utilise le produit scalaire de Hilbert-Schmidt pour induire une distance entre les matrices X^b et ainsi les comparer. Ce produit scalaire, pour deux matrices X^a et X^b est défini par :

$$HS(X^a, X^b) = trace(DX^a DX^b)$$

où D est la matrice des poids associés aux individus, en général ces poids sont choisis uniformément égaux à 1/N. En remplaçant les matrices X^a et X^b par leur représentant normé $\frac{X}{||X||_{HS}}$ on obtient le coefficient de corrélation vectorielle R_V introduit par Robert et Escoufier [1976] entre les matrices :

$$R_V(X^a, X^b) = HS\left(\frac{X^a}{||X^a||_{HS}}, \frac{X^b}{||X^b||_{HS}}\right) = \frac{trace(DX^a DX^b)}{\sqrt{trace((DX^a)^2) trace((DX^b)^2)}} \qquad (5.4)$$

Le $R_V(X^a, X^b)$ correspond au cosinus de l'angle entre les points X^a, X^b et l'origine de l'espace de projection des représentants des matrices X^a et X^b relativement à la géométrie définie par le produit scalaire HS. Ainsi, plus le cosinus de l'angle des deux matrices X^a et X^b est grand, plus ces deux matrices se rapprochent et tendent à être colinéaires et plus les "liens" relatifs inter-individus représentés par ces deux matrices sont forts. Par conséquent, nous formulons le problème de consensus comme la recherche d'une matrice $\tilde{Co}$ de même nature que les matrices X^b et telle que $\tilde{Co}$ soit la plus corrélée possible au sens du produit scalaire HS avec les matrices X^b :

$$\tilde{Co} = \underset{C}{max} \sum_{b=1}^{B} R_V^2(C, X^b) \tag{5.5}$$

Nous recherchons une solution de l'équation (5.5) sous la forme d'une moyenne pondérée des matrices X^b :

$$\tilde{Co} = \sum_{b=1}^{B} \sigma_b X^b \tag{5.6}$$

où les quantités σ_b sont des paramètres de pondération liés aux interrelations entre les matrices X^b. Le problème est finalement équivalent à trouver la meilleure combinaison des matrices X^b au sens de la corrélation vectorielle. On démontre que les poids σ_b s'obtiennent de manière exacte comme les coordonnées (après normalisation) du premier vecteur propre associé à la plus grande valeur propre de la matrice des corrélations vectorielles R_V. La matrice R_V est symétrique, il est facile de vérifier que tous ces termes sont positifs. D'après le théorème de Peron-Frobenius [Gentle 2007], le sous-espace engendré par les vecteurs propres associés à la plus grande valeur est de dimension 1 ; de plus, les composantes du vecteur propre correspondant sont toutes de même signe que l'on peut choisir positif. Ainsi, le premier vecteur propre normalisé tel que $\sum \sigma_b = 1$, fournit la matrice $\tilde{Co}$ la plus corrélée avec toutes les matrices X^b au sens du coefficient R_V.
La matrice R_V est symétrique, il est facile de vérifier que tous ces termes sont positifs. D'après le théorème de Peron-Frobenius [Gentle 2007], le sous-espace engendré par les vecteurs propres associés à la plus grande valeur est de dimension 1 ; de plus, les composantes du vecteur propre correspondant sont toutes de même signe que l'on peut choisir positif. Ainsi, le premier vecteur propre normalisé tel que $\sum \sigma_b = 1$, fournit la matrice $\tilde{Co}$ la plus corrélée avec toutes les matrices X^b au sens du coefficient R_V.

5.3.2 Fusion de SOM fondée sur les partitions π_b

Dans le cas où l'on utilise les partitions π^b de SOM, on définit les matrices disjonctives complètes H^b associées aux B partitions de l'ensemble $\Pi = \{\pi^1, \ldots, \pi^B\}$. Dans ce cas, les matrices $X^b = H^b H^{b\prime}$ sont alors les matrices d'adjacence associées aux partitions π^b avec :

$$X^b(i,j) = \begin{cases} 1 & si \quad \pi^b(i) = \pi^b(j) \\ 0 & sinon \end{cases}$$

Dans la littérature, on utilise habituellement la matrice $Co = \frac{1}{B}\sum_{b=1}^{B} X^b$ dont les entrées sont égales à la fréquence avec laquelle deux individus regroupés ensemble dans une

classe pour déterminer le consensus [Strehl et Ghosh 2002; Li *et al.* 2007]. Mais cette méthode suppose, implicitement, que les différentes partitions ont la même importance. Or, les partitions formant l'ensemble Π n'ont généralement pas la même pertinence. Le formalisme présenté dans la section 5.3 va alors définir à travers un système de poids la matrice compromis $\tilde{Co}$ tel que plus il existe des matrices ayant des "directions" proches dans la géométrie définie par le produit scalaire précédent, plus elles vont avoir des pondérations importantes. Par contre, une matrice, dont la "direction" est très différente de la majorité des matrices, aura un coefficient de pondération faible.

La matrice $\tilde{Co}$ définie par la relation (5.5) peut alors être vue comme la matrice consensus des matrices d'adjacence X^b des partitions de l'ensemble Π. Puisqu'elle est de même nature que les matrices X^b qui sont des matrices de similarité, l'application d'une classification hiérarchique ascendante sur cette matrice $\tilde{Co}$ permet d'obtenir le consensus π^* des B partitions.

La matrix $\tilde{Co}$ définie dans cette approche est obtenue de manière similaire à la matrice compromis dans l'étude de l'intrastructure dans la méthode STATIS [Lavit *et al.* 1994]. Elle est basée sur le coefficient de corrélation vectorielle R_V introduit par Robert et Escoufier [1976]. Cependant, Smilde *et al.* [2009] montrent que ce coefficient est dépendant des dimensions des données initiales. Plus précisément, ils montrent que le coefficient R_V prend des valeurs artificiellement grandes pour de petites tailles d'échantillon, il décroit avec le nombre d'observations et croit avec le nombre de variables des blocs. Il serait alors souhaitable d'avoir des blocs de variables de faible dimension. Dans le cas contraire, le R_V peut être évalué à partir des matrices $\tilde{X^b} = X^b - diag(X^b)$ où $diag(X^b)$ désigne une matrice diagonale contenant les éléments diagonaux de la matrice X^b. Cependant, le R_V peut prendre dans ce cas des valeurs négatives, on perd alors la positivité des coefficients du premier vecteur propre et l'interprétation en termes de poids.

Dans toute la suite de ce chapitre on désignera par R_V-SOM$_1$ le résultat du consensus lorsque $X^b = H^b H^{b\prime}$.

5.3.3 Évaluation

Nous utiliserons de deux manières la méthode R_V-CSOM. Elle servira d'une part à déterminer un consensus de SOM pour données multi-blocs, d'autre part à étudier la robustesse de l'algorithme SOM par rapport aux paramètres d'initialisation.

Algorithme : R_V-CSOM

Entrée : L'ensemble de diversification est constitué de B cartes SOM.

1. Calculer les matrices X^b $(b = 1, \ldots, B)$ par la relation $X^b = \tilde{Z}^b \tilde{Z}^{b\prime}$ ou $H^b H^{b\prime}$
2. Calculer la matrice des corrélations vectorielles R_V entre les matrices X^b à l'aide l'équation 5.4
3. Diagonaliser cette matrice pour déterminer les coefficients σ_b et calculer la matrice des compromis à l'aide de l'équation 5.6
4. Selon l'objet considéré :
   - Appliquer une méthode partitionnement topologique sur la matrice $[\sigma_1 \tilde{Z}^1, \ldots, \sigma_B \tilde{Z}^B]$
   - Appliquer CAH sur la matrice $\tilde{Co}$ pour déterminer la carte consensus finale.

Sortie : La carte consensus

5.3.3.1 Recherche de consensus

Nous utilisons les bases de données IS, CT, DMU et FGKM. Pour chaque base de données, la diversification des cartes est apportée par les blocs. Les B cartes la formant s'obtiennent par application de SOM sur chacun des B blocs de variables de la base associée. Ainsi, on obtient autant de cartes topologiques qu'il y a de blocs de variables.

Pour un ensemble de diversification nous recherchons les poids permettant de définir la matrice consensus $\tilde{Co}$. Cette expérience est répétée 25 fois. Les tableaux 5.1(a) et 5.1(b) présentent respectivement la moyenne des poids définis par les algorithmes R_V-SOM sur les tables FGKM et DMU.

(a) Les poids fournis par R_V-CSOM et R_V-CSOM$_1$ au niveau des blocs pour la table FGKM

	Bloc 1	Bloc 2	Bloc 3
R_V-CSOM	0.32	0.34	0.34
R_V-CSOM$_1$	0.24	0.38	0.38

(b) Les poids fournis par R_V-CSOM et R_V-CSOM$_1$ au niveau des blocs pour la table DMU

	Bloc 1	Bloc 2	Bloc 3	Bloc 4	Bloc 5	Bloc 6
R_V-CSOM	0.13	0.19	0.17	0.18	0.16	0.16
R_V-CSOM$_1$	0.12	0.21	0.19	0.21	0.15	0.11

TABLE 5.2 – Les poids fournis par R_V-CSOM et R_V-CSOM$_1$ sur les blocs

Les méthodes R_V-CSOM et R_V-CSOM$_1$ traduisent globalement la même structure, il

n'y a pas de différence marquante entre les classements des blocs par ordre d'importance. On note cependant que les variations entre les blocs sont plus marquées au niveau de la méthode R_V-CSOM$_1$.

Afin de positionner R_V-CSOM par rapport aux méthodes de consensus présentées dans la littérature, nous appliquons les algorithmes de recherche de consensus basés sur la factorisation de matrice non-négative (NMF), Weighted NMF présentés par Ding *et al.* [2006] et Li et Ding [2008], la méthode d'ensemble cluster (CSPA) présentée par Strehl et Ghosh [2002] sur les données.

Le tableau 5.3 présente, par algorithme, la moyenne et l'écart type des indices de pureté (Pur), d'information mutuelle normalisée (NMI) et de variation d'information normalisée (NVI) pour les 25 expériences. Nous rappelons que les méthodes de fusion basées sur les matrices $\tilde{Z}^b$ et sur les partitions π^b seront notées respectivement R_V-CSOM et R_V-CSOM$_1$.

Comparée aux méthodes NMF, WNMF, CSPA et R_V-CSOM$_1$ la méthode R_V-CSOM est la plus performante sur les bases IS, CT, FGKM. Ce qui est du à l'utilisation des matrices $\tilde{Z}^b$ qui fournissent un bon compromis entre la représentation initiale des données et leur label dans les partitions π^b.

D	Indice	SOM	KM	NMF	WNMF	CSPA	R_V-CSOM	R_V-CSOM$_1$
IS	Pur	0.61(0.01)	0.61(0.02)	0.45(0.005)	**0.58(0.001)**	0.37(0.001)	0.55(0.001)	0.53(0.004)
	NMI	0.59(0.01)	0.62(0.01)	0.36(0.01)	0.51(0.01)	0.27(0.01)	**0.57(0.001)**	0.44 (0.005)
	NVI	0.58(0.01)	0.56(0.01)	0.78(0.003)	0.65(0.003)	0.84(0.01)	**0.61(0.001)**	0.71(0.003)
CT3	Pur	0.78(0.02)	0.78(0.02)	0.78(0.02)	0.78(0.02)	0.78(0.02)	**0.78(0.001)**	0.78(0.001)
	NMI	0.13(0.001)	0.15(0.001)	0.10(0.001)	0.10(0.001)	0.15(0.001)	**0.20(0.001)**	0.11(0.001)
	NVI	0.93(0.03)	0.92(0.03)	0.95(0.001)	0.96(0.001)	0.92(0.03)	**0.89(0.001)**	0.94(0.001)
DMU	Pur	0.76(0.02)	0.78(0.03)	0.85(0.001)	0.85(0.01)	**0.86(0.001)**	0.54(0.001)	0.82(0.001)
	NMI	0.76(0.02)	0.77(0.02)	0.79(0.005)	0.81(0.001)	**0.82(0.001)**	0.57(0.001)	0.79(0.004)
	NVI	0.39(0.001)	0.37(0.01)	0.34(0.001)	0.31(0.001)	**0.30(0.001)**	0.60(0.001)	0.35(0.001)
FGKM	Pur	0.55(0.02)	**0.61(0.03)**	0.53(0.001)	0.52(0.01)	0.51(0.001)	**0.59(0.001)**	0.53(0.002)
	NMI	0.18(0.02)	**0.27(0.02)**	0.12(0.005)	0.11(0.001)	0.12(0.001)	**0.23(0.001)**	0.14(0.003)
	NVI	0.89(0.001)	**0.84(0.01)**	0.94(0.001)	0.94(0.001)	0.93(0.001)	**0.87(0.0008)**	0.93(0.001)

TABLE 5.3 – Performances des algorithmes SOM, K-moyennes (KM), NMF, WNMF, CSPA, R_V-CSOM et R_V-CSOM$_1$ sur les bases IS, CT et DMU ;

5.3.3.2 Étude de la robustesse de SOM par rapport aux paramètres d'initialisation

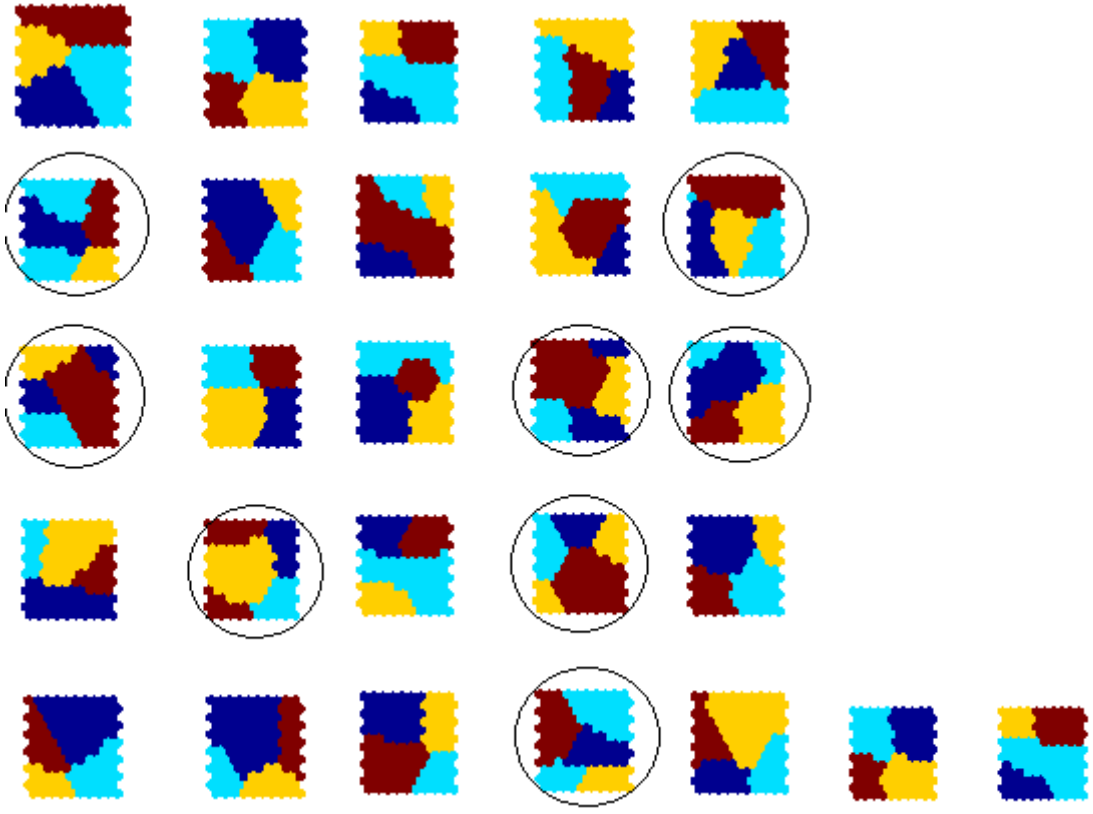
Dans cette section, l'objectif est de rendre robuste la méthode SOM par rapport aux paramètres d'initialisation. Pour cela les B cartes formant la diversification pour une base de données s'obtiennent par application répétée de l'algorithme SOM sur chaque table initiale constituée de l'ensemble des variables. Seuls les paramètres d'initialisation de l'algorithme SOM varient. Dans ce ensemble, B est arbitrairement fixé à 30 afin de comparer significativement les performances.

Les méthodes R_V-CSOM et R_V-CSOM$_1$ sont pratiquement meilleures en terme de performances pour toutes les tables sauf pour la table IS au niveau de la pureté 5.4. Ce résultat est dû à la forte corrélation entre les cartes de ce ensemble de diversification qui se traduit par des coefficients R_V forts entre les matrices X^b.

En terme de visualisation, les figures 5.7(a) et 5.8(a) présentent la structure des classes des cartes du deuxième ensemble de diversification. Les figures 5.7(b) et 5.8(b) présentent la structure des classes des cartes fusion obtenues par application de l'algorithme SOM sur la matrice $[\sqrt{\sigma_1}\tilde{Z}^1, \ldots, \sqrt{\sigma_B}\tilde{Z}^B]$. On observe à travers la proximité des cellules formant les classes, une bonne conservation de la topologie des observations lorsque σ_b est obtenue par R_V-CSOM.

D	Indice	SOM	KM	NMF	WNMF	CSPA	R_V-CSOM	R_V-CSOM$_1$
IS	Pur	0.61(0.01)	0.61(0.02)	0.60(0.005)	0.58(0.001)	**0.61(0.001)**	0.55(0.001)	0.53(0.004)
	NMI	0.59(0.01)	0.62(0.01)	0.36(0.01)	0.51(0.01)	0.27(0.01)	**0.62(0.001)**	0.56 (0.005)
	NVI	0.58(0.01)	0.56(0.01)	0.78(0.003)	0.65(0.003)	0.84(0.01)	**0.54(0.001)**	0.61(0.003)
CT3	Pur	0.78(0.02)	0.78(0.02)	0.78(0.02)	0.78(0.02)	0.78(0.02)	**0.78(0.001)**	0.78(0.001)
	NMI	0.13(0.001)	0.15(0.001)	0.10(0.001)	0.10(0.001)	0.15(0.001)	**0.19(0.001)**	0.15(0.001)
	NVI	0.93(0.03)	0.92(0.03)	0.95(0.001)	0.96(0.001)	0.92(0.03)	**0.89(0.001)**	0.94(0.001)
DMU	Pur	0.76(0.02)	0.78(0.03)	0.82(0.001)	0.79(0.01)	0.70(0.001)	0.62(0.001)	**0.85(0.001)**
	NMI	0.76(0.02)	0.77(0.02)	0.80(0.005)	0.78(0.001)	0.60(0.001)	0.81(0.001)	**0.83(0.004)**
	NVI	0.39(0.001)	0.37(0.01)	0.32(0.001)	0.35(0.001)	0.67(0.001)	0.51(0.001)	**0.29(0.001)**
FGKM	Pur	0.55(0.02)	0.61(0.03)	0.58(0.001)	0.59(0.001)	0.59(0.001)	**0.64(0.001)**	0.58(0.001)
	NMI	0.20(0.02)	0.27(0.02)	0.20(0.005)	0.20(0.001)	0.26(0.001)	**0.31(0.001)**	0.27(0.003)
	NVI	0.89(0.001)	0.84(0.01)	0.89(0.001)	0.88(0.001)	0.85(0.001)	**0.81(0.001)**	0.85(0.001)

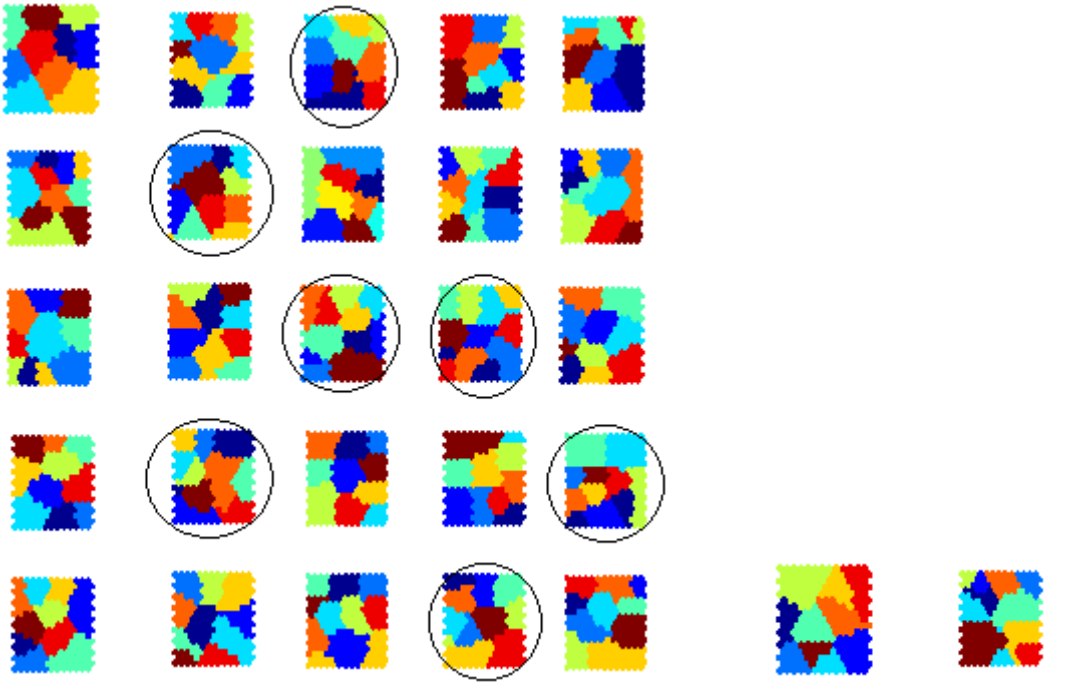
TABLE 5.4 – Performances des algorithmes SOM, K-moyennes (KM), NMF, WNMF, CSPA, R_V-CSOM et R_V-CSOM$_1$ sur les bases IS, CT et DMU ; la diversification est obtenue par variation des paramètres de l'algorithme SOM



(a) Les cartes des 25 apprentissages de l'ensemble de diversification pour la base FGKM ; les cartes encerclées présentent une performance topologique faible

(b) Les cartes consensus R_V-CSOM et R_V-CSOM$_1$ pour la base FGKM

FIGURE 5.7 – Visualisation de la structure des classes sur la carte topologique pour la base FGKM



(a) Les cartes des 25 apprentissages de l'ensemble de diversification pour la base DMU ; les cartes encerclées présentent une performance topologique faible

(b) Les cartes consensus R_V-CSOM et R_V-CSOM$_1$ pour la base DMU

FIGURE 5.8 – Visualisation de la structure des classes sur la carte topologique pour la base DMU

5.4 Conclusion

Dans ce chapitre, nous avons présenté deux méthodes de Fusion de SOM pour données multi-blocs. Ces méthodes montrent, l'intérêt de prendre en compte la qualité relative des cartes de l'ensemble de diversification et les relations les régissant.

Dans la première approche de type hiérarchique, l'usage de SOM au niveau 2 permet de prendre en compte la notion de topologie dans la carte finale. La seconde approche basée sur la prise en compte de la relation entre les cartes montre que les performances des méthodes de recherche consensus sont, en règle générale, dépendantes de la qualité de l'ensemble de diversifications. Cette approche peut aussi être utilisée pour améliorer la stabilité des résultats d'une classification.

Le chapitre suivant présente les applications des méthodes proposées aux données de l'OQAI.

Troisième partie

Application à la CNL

Chapitre 6

Application aux données de l'OQAI

6.1 Introduction

La collecte des données sur les immeubles à usage de bureaux étant en cours, les approches développement dans cette thèse sont appliquées aux données de la campagne nationale «Logements» (CNL).
Cette campagne vise à dresser un état de la pollution de l'air dans l'habitat afn de donner les éléments utiles pour l'estimation de l'exposition des populations, la quantification et la hiérarchisation des risques sanitaires associés, l'identifcation des facteurs prédictifs de la qualité de l'air intérieur.

Plus de 30 paramètres (chimiques, biologiques, physiques) de pollution ont été mesurés, sur une durée d'une semaine, à plusieurs emplacements à l'intérieur des logements, dans les garages attenants lorsqu'ils existent et à l'extérieur. Dans le même temps des informations détaillées ont été collectées sur les caractéristiques techniques des logements, sur leur environnement ainsi que sur les ménages et leurs activités au travers d'un questionnaire. Dans ce chapitre, nous recherchons une unique typologie des logements enquêtés en tenant compte de l'ensemble des paramètres relevés sur les logements : type, structure, taille, ancienneté, mais également revêtements intérieurs, aménagements, les habitudes des occupants, la structure des ménages et les polluants mesurés. La section 6.2 présente les données. La section 6.3 présente les applications de 2S-SOM à la base CNL.

6.2 Données

6.2.1 Un échantillon représentatif du parc français

567 résidences principales ont été désignées par tirage au sort dans 74 communes de 50 départements et de 19 régions. La méthode de sondage retenue a permis d'avoir un échantillon le plus représentatif possible du parc 24 millions de résidences principales de France continentale métropolitaine, en termes de diversité des structures, mais aussi de type de ménages. L'étude ne portant que sur des logements occupés, les résidences secondaires en ont été exclues. La figure 6.1 présente la répartition géographique des logements de CNL.

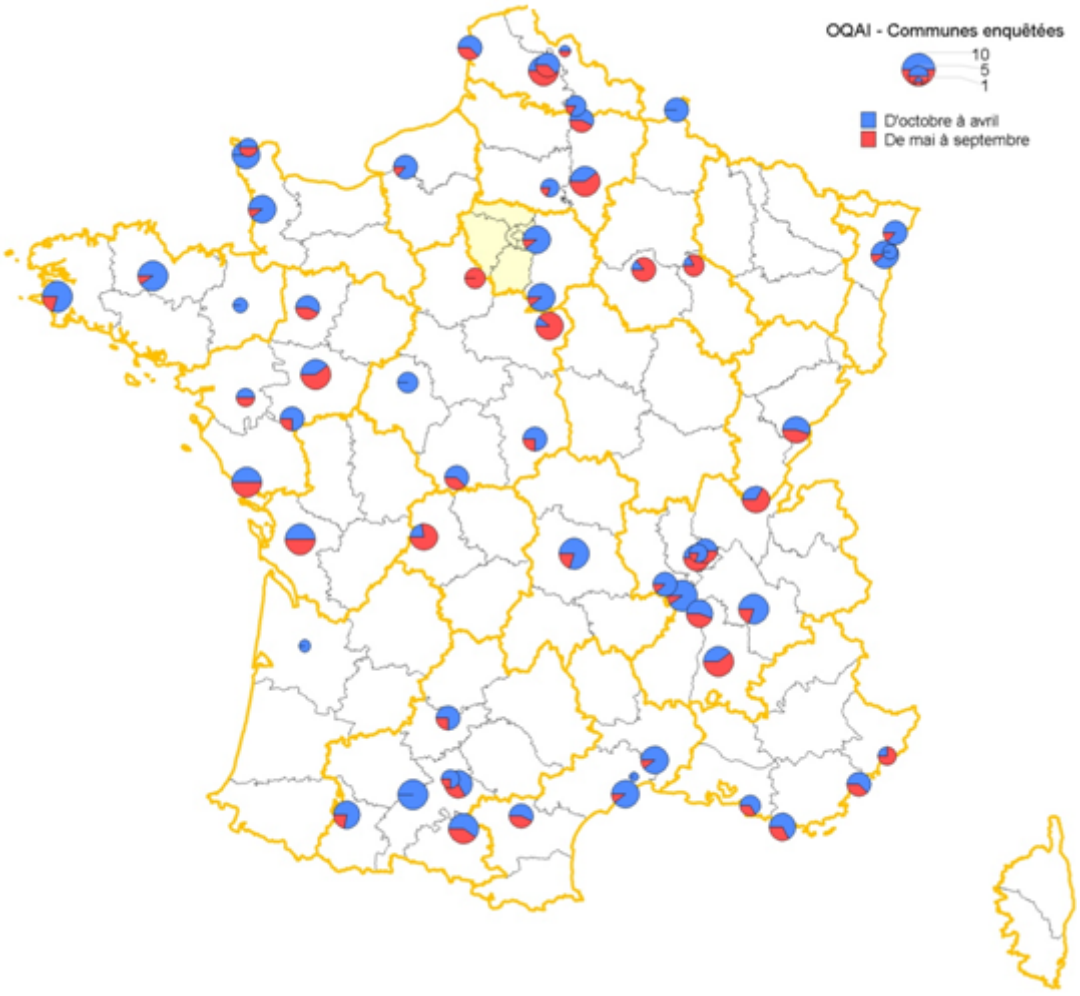


FIGURE 6.1 – Répartition géographique des logements enquêtés lors de la campagne nationale de l'OQAI Kirchner *et al.* [2011]

Plusieurs blocs d'informations ont été récoltés sur chacun des 567 logements de l'échantillon. Dans le cas de cette application, nous nous intéressons aux blocs relatifs aux polluants de l'air intérieur, aux caractéristiques des logements, à la structure des ménages et aux habitudes des occupants des logements.

6.2.2 Bloc Polluants (17 polluants)

Les polluants ou familles de polluants mesurés sont celles dont la dangerosité est connue et pouvant être présentes dans les logements. Ainsi, une trentaine de polluants chimiques, physiques et de biocontaminants ont été mesurés au cours de la CNL. Chaque polluant a été mesuré selon une stratégie d'échantillonnage spécifique : matériels, protocoles de pose, de prélèvement et d'analyse. Ainsi, les composés organiques volatils (COV) ont été mesurés pendant une semaine par diffusion sur cartouches dans la chambre des parents, dans le garage communicant et à l'extérieur. Les aldéhydes ont été prélevés par diffusion sur cartouches imprégnées de dinitrophénylhydrazine (DNPH) dans la chambre des parents et à l'extérieurs (figure 6.2(a)). Les cartouches ont ensuite été analysées en laboratoires.
La concentration en allergènes d'acariens est déterminée à partir de prélèvements par aspiration sur le matelas de la chambre investiguée (chambre des parents). Le sac de l'aspirateur est ensuite envoyé en laboratoire pour y être analysé par la méthode immuno-enzymatique ELISA (figure 6.2(b)).
Les particules sont prélevées de manière active par aspiration d'air, filtration et impaction sur un filtre, dans le séjour pendant une semaine (de 17h à 8h en semaine et toute la journée pendant le week-end) à l'aide d'un Minipartisol ($PM_{2,5}$ et PM_{10}) équipé d'un échantillonneur à 2 têtes (figure 6.2(c)).

(a) Support adsorbant solide utilisé pour les prélèvements de COV et aldéhydes

(b) Prélèvement par aspiration des acariens sur un matelas

(c) Appareils de prélèvement des particules

FIGURE 6.2 – Matériels de mesure des polluants de l'air

Dans cette application, nous avons retenu les polluants présentant un nombre raisonnable de valeurs manquantes ($NaN < 10\%$). Le tableau 6.1 présente les statistiques descriptives des polluants, et leurs effets sanitaires.

	Moy	Min	Max	Std	% NaN	Effets sanitaires
$PM_{2.5}$	37.07	1.2	567.7	51.86	49%*	Respiratoires et cardio-vasculaires
PM_{10}	53 .59	1.6	522.6	64.14	47%*	Respiratoires et cardio-vasculaires
DERP1	16.75	0.005	608	48.89	22%*	(allergies, asthmes)
DERF1	7.15	0.01	129	14.6	22%*	(allergies, asthmes).
Formaldéhyde	18.51	1.02	70.75	10.42	3%	Respiratoires
Acétaldéhyde	7.81	0.98	51.78	5.8	3%	Irritations yeux, tractus respiratoire
Acroléine	0.59	0	5.45	0.51	3%	Respiratoires
Hexaldéhyde	4.83	0.38	89	6.16	3%	
Benzène	0.83	0.00	7.13	0.77	5.29%	Neurologiques et immunologiques, leucémie
Méthoxy-propanol	1.27	0.00	45.94	3.07	5.29%	Testiculaires
Trichloroéthyléne	1 .96	0.42	0.00	75.66	5.29%	Neurologiques, Cancers
Toluène	5.87	0.42	109	9.11	5.29%	
Tétrachloroéthylèı	0.58	0.00	104.77	4.59	5.29%	Rénaux, Neurologiques, Cancers
Styrène	0.3	0.00	8.3	0.431	5.29%	Neurologiques
Butoxy-éthylène	0.57	0.00	12.3	1.07	5.29%	
Triméthylbenzène	1.35	0.00	22.38	1.92	5.29%	
Dichlorobenzène	8 .6	0.00	804.21	42.21	5.29%	Rénaux
N-décane	3.15	0.00	310.04	14.91	5.29%	

TABLE 6.1 – Description des substances ou des paramètres mesurés ; Moy, Min, Max, Std et % NaN désigne respectivement la moyenne, le minimum, le maximum, l'écart-type et la proportions de valeurs manquantes ; * correspond aux polluants non pris en compte dans les analyses

La mesure des différents paramètres de pollution ont ensuite été complétée par la recherche des sources et des déterminants de ces paramètres. Il s'agissait de découvrir la source des contaminants grâce à des informations précises et pertinentes. Des informations détaillées ont donc été collectées sur les caractéristiques techniques des logements et leur environnement ainsi que sur les ménages, leurs activités et le temps passé au contact de la pollution. En tout, plus de 600 questions par logement ont été renseignées. Ces questions ont été regroupées en plusieurs thématiques. Dans cette application on s'intéresse à

3 thématiques.

6.2.3 Bloc Logements (72 variables)

Les données descriptives des sources ou situations de pollution en lien direct avec les logements ou leur environnement sont recueillies d'entrée de jeu par un technicien enquêteur. Elles concernent la situation et les caractéristiques physiques du logement (proximité de sources de pollution extérieures, type et année de construction, nombre d'étages), la description intérieure du logement (taille et descriptif des pièces d'habitation et des dépendances, présence d'un garage communiquant, caractéristiques des systèmes de ventilation, de chauffage et de cuisson, équipements sanitaires, aération/ventilation, travaux de rénovation, etc.), les types de revêtement (sols, murs, plafonds), de menuiseries et d'équipements (ménagers, meubles en bois, tapis, rideaux, literie) et la qualité globale de l'environnement (présence d'humidité, sources potentielles de pollution extérieures), etc.

6.2.4 Bloc Ménage (11 variables)

Dans ce bloc, plus de 1612 individus ont été suivis dans les 567 logements de la CNL. Les informations collectées sur ces occupants sont liées d'une part à la structure des ménages (composition des ménages, sexe, statut d'occupant, activité professionnelle, niveau d'étude, revenu, enfants).

6.2.5 Bloc Habitudes des ménages (45 variables)

Dans ce bloc, les informations collectées sur ces occupants sont liées aux habitudes de vie (entretien du logement, présence de plantes et d'animaux domestiques, loisirs, pratiques de cuisson, utilisation de cosmétiques, de la voiture et/ou de désodorisants d'ambiance, tabagie).

Nous allons, dans une étude multi-blocs, rechercher une unique typologie des logements de la base CNL au regard des blocs des polluants, logements, ménage et habitudes des ménages qui contiennent respectivement 17 et 72, 45, 11 variables de types quantitatives ou qualitatives. L'annexe 6.4 présente les tableaux de variables pour chaque bloc.

6.3 Recherche d'une typologie globale de la base CNL

La recherche de la typologie se présente de la manière suivante :

1. Utiliser la méthode de Soft-Subspace Clustering 2S-SOM pour déterminer un ensemble Π de diversification constitué de cartes topologiques. Cet ensemble de diversification est obtenu en faisant varier les paramètres de 2S-SOM, notamment l'initialisation des centres de classes.

2. L'objectif étant de rendre robuste la typologie finale par rapport aux paramètres du modèle, sur Π nous appliquons les méthodes de fusion des cartes SOM du chapitre 5 pour obtenir la typologie finale.

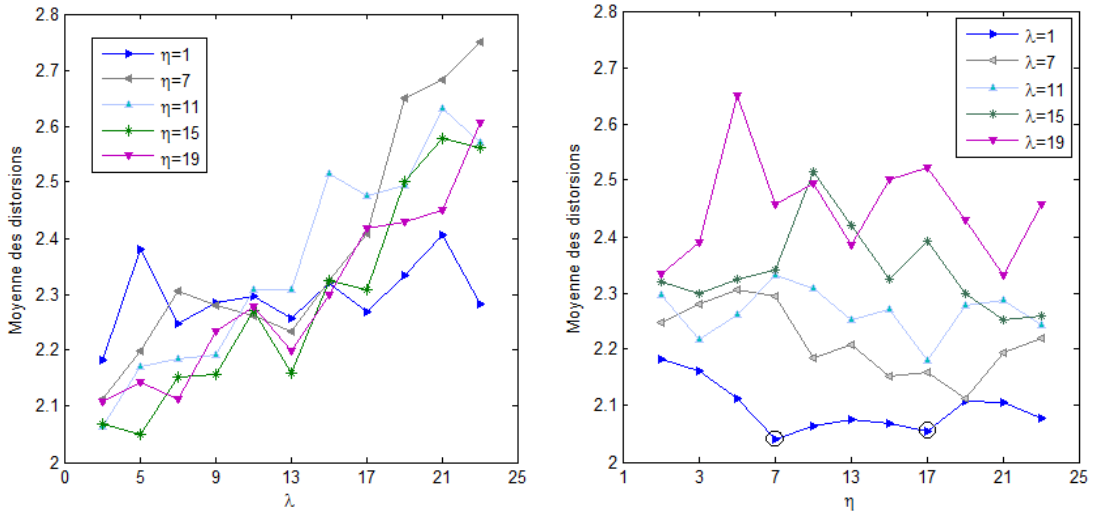
Dans la phase de pré-traitement des données, nous avons utilisé l'algorithme SOM classique pour estimer, selon le principe présenté dans la section 2.3.2.2 du chapitre 2, les valeurs manquantes des variables du bloc polluant présentant au plus 10% de valeurs manquantes. Ainsi, nous travaillons sur l'échantillon de 567 observations, les $PM_{2.5}$, PM_{10} et les allergènes ne sont pas pris en compte dans cette étude.

6.3.1 Application de 2S-SOM et de FSOM à la CNL

Plusieurs applications de 2S-SOM ont été réalisées en faisant varier ses paramètres. Au niveau de SOM, la taille de la carte est 12×10 pour 150 itérations, 3 et 1 désignent le voisinage initial et final des cellules. Les figures 6.3(a) et 6.3(b) montrent l'évolution moyenne de la mesure de distorsion pour les paramètres (λ, η) fixés. La figure 6.3(a) montre que l'étude directe qui consiste à considérer identiquement les blocs et les variables perturbe les performances de la mesure de distorsion (λ grand et η grand).

Sur la figure 6.3(b) les meilleures performances sont atteintes pour $\lambda = 1$, ainsi nous avons retenu sur les données CNL, $\lambda = 1$ et $\eta = 7$.

Pour le couple de paramètres $(\lambda = 1, \eta = 7)$, 10 apprentissages de cartes faisant varier les centres de classes ont été réalisés. Pour $\lambda = 1$, on affecte à priori des poids forts aux blocs minimisant l'inertie intra-classes dans les cellules. Chaque bloc se spécialise dans la caractérisation d'un certain nombre de cellules. Le bloc des polluants est celui qui permet à priori de déterminer le plus des groupes homogènes d'observations dans les cellules comme cela est illustré sur la figure 6.3.1 zone D. Les cellules dans les zones représentées par "A", "B" et "C" sont caractérisées respectivement par les blocs logement, ménage et habitudes.



(a) (b)

FIGURE 6.3 – Évolution de la moyenne de la mesure de distorsion pour les couples de paramètres λ et η

6.3.2 Analyse de la carte finale

La figure 6.5 présente les référents de la carte 2S-SOM, obtenue à partir des 567 logements. Chaque référent peut être assimilé à un logement type ; il possède une valeur pour chacune des variables prises en compte dans l'analyse. Les cartes suivantes représentent la projection en 2D de ces variables. Elles présentent un ordre topologique bien organisé : les variables corrélées sont regroupées dans le premier plan d'une analyse en composante principale appliquée à la table des référents afn de faciliter l'interprétation. On observe sur chacune des figures 6.5(a), 6.5(b), 6.5(c) et 6.5(d) la structure de la carte par rapport aux variables du bloc correspondant.

La matrice $\tilde{Z}$ présenté dans la figure 5.1 des observations sert à déterminer la partition consensus. Les données de la CNL n'étant pas étiquetées, une estimation de nombre K de classes dans la matrice $\tilde{Z}$ est effectuée l'aide de l'indice de Davies-Bouldin Davies et Bouldin [1979]. Le nombre K de classes fixé à priori est celui qui minimise l'indice de Davies-Bouldin. La figure 6.6 montre que pour K=5, algorithme des K-moyennes fournit le bon nombre de classes dans le partitionnement de la carte fusion finale.

Nous allons maintenant décrire les 5 groupes obtenus sur la base CNL par rapport aux variables d'intérêt. Une variable est dite d'intérêt pour une classe si son poids est significativement supérieur à un seuil fixé. Les figures 6.7, 6.8, 6.9 et 6.10 présentent les

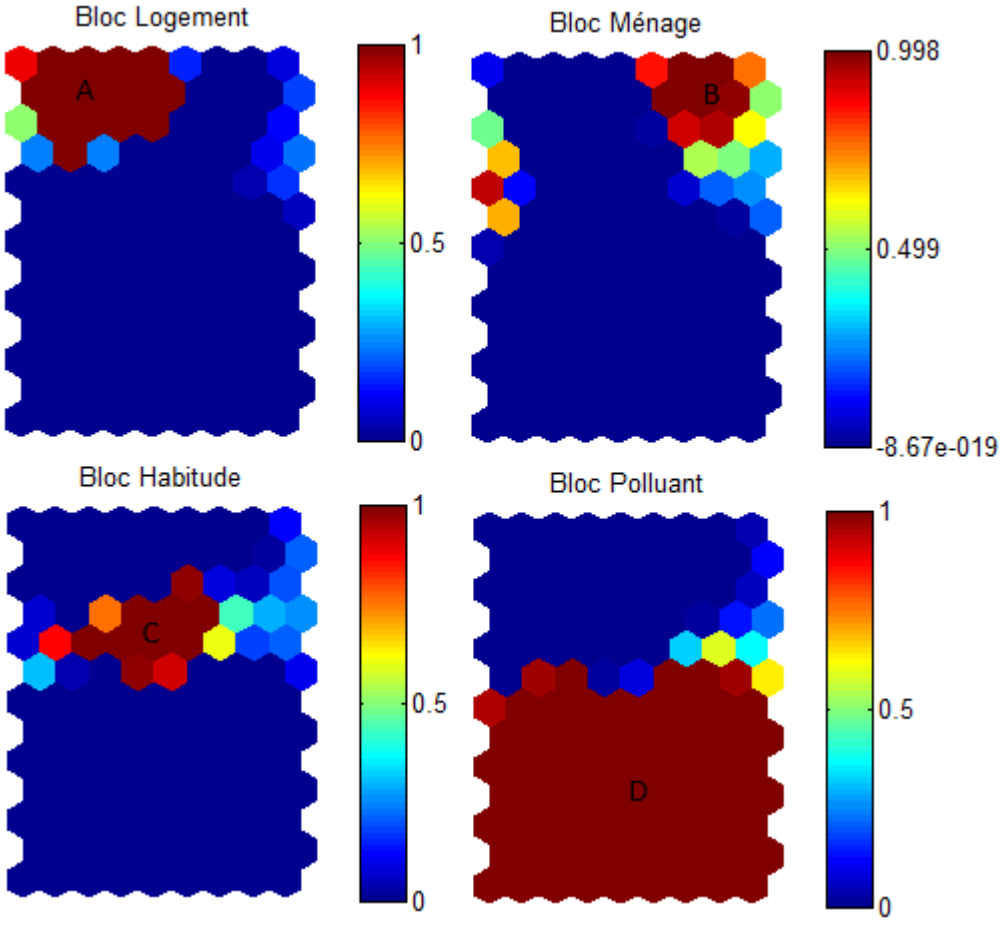
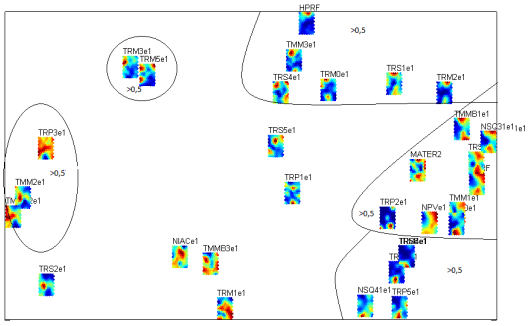


FIGURE 6.4 – Les poids des blocs sur chaque cellule de la carte ; Les codes couleurs correspondent aux poids des blocs dans les cellules

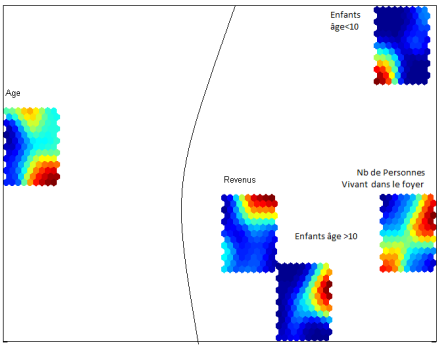
caractéristiques des classes ; en ordonnée le label de la classe et en abscisse la moyenne de la variable dans la classe accompagné d'un intervalle de confiance.

Groupe 1, 110 observations

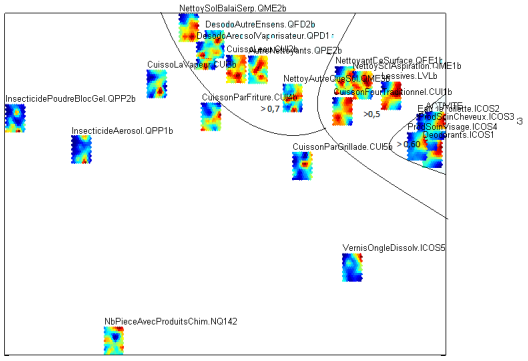
Ces ménages contiennent souvent des familles en couple (85 % versus 74 %). Ils comportent en moyenne 3 personnes par foyer. Ces ménages exercent une profession intermédiaire (73 % versus 19%). Par rapport au diplôme, on retrouve plus de personnes ayant suivi un enseignement technique court (40% versus 23%). Ils vivent avec des revenus inférieurs à la moyenne de l'échantillon.
Par rapport à la structure technique, ces logements ont une plus petite surface (97 m^2 versus 110m^2), sans endroit pour bricoler (80 % versus 36 %).



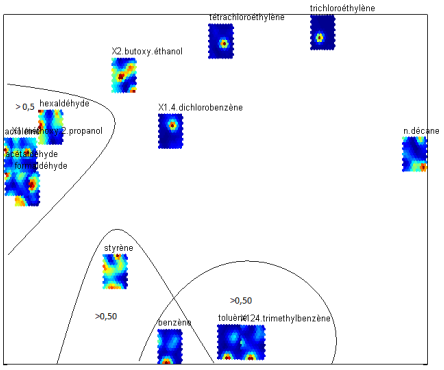
(a) Bloc logement (54% d'inertie)



(b) Bloc ménage (67% d'inertie)



(c) Bloc habitude (62% d'inertie)



(d) Bloc polluant (52% d'inertie)

FIGURE 6.5 – Liaison entre les variables continues descriptives ayant servies à l'apprentissage

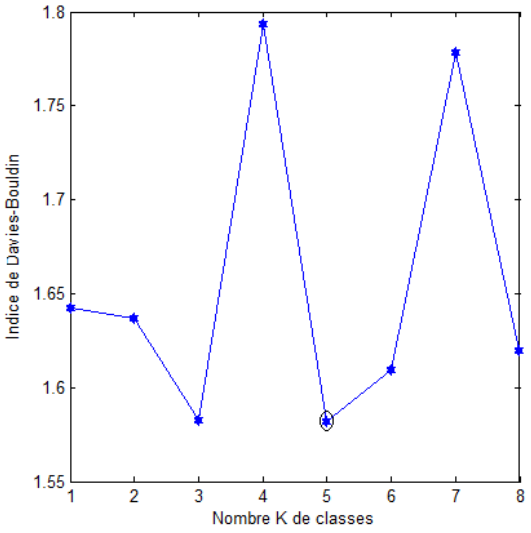


FIGURE 6.6 – Indice de Davies Bouldin pour K fixés

On retrouve des ménages ayant une activité moyenne dans leur logement, ils sont un peu plus nombreux à être fumeurs (55% versus 44 %) et à avoir des animaux domestiques. Ils bricolent beaucoup pour la plupart (72 % versus 51 %).
Par rapport aux polluants, les concentrations sont significativement élevées pour les aldéhydes (figure 6.10).

Groupe 2, 157 observations

Ces ménages contiennent plus de familles monoparentales (22 % versus 8 %) et sont moins souvent en couple que dans le reste de l'échantillon (61.3 % versus 74 %). Ils comportent en moyenne 3 personnes par foyer (3 personnes versus 2.83). Ces ménages sont souvent des employés (30 % versus 20 %) ou exercent une profession intermédiaire (53 % versus 19%). Par rapport au diplôme, on retrouve plus de personnes ayant un bac +3/4 (24% versus 11%).
Relativement à la structure technique, ces logements se distinguent significativement par les caractéristiques suivantes : ce sont des logements de type collectif (53 % versus 40 %), dont les occupants sont locataires. Ces logements ont une plus petite surface (93 m^2 versus 110m^2) et n'ont pas d'endroit pour bricoler (86 % versus 36 %).
Par rapport aux habitudes, on retrouve des ménages ayant une activité moyenne dans leur logement, ils sont un peu plus nombreux à être fumeurs et à posséder des animaux domestiques. Ils bricolent et jardinent peu pour la plupart.
Par rapport aux polluants, on retrouve des concentrations significativement faibles pour les aldéhydes et les hydrocarbures.

Groupe 3, 59 observations

Ces ménages sont jeunes (44 ans versus 49 ans) et relativement nombreux (3.50 versus 2.8). Ils sont tous en couple (100 % versus 74 %) et ont des revenus semblables à la moyenne de l'échantillon. Cette classe contient de nombreux artisans (22 % versus 6 %) et des personnes ayant des professions intermédiaires (54 % versus 22 %). À l'inverse, elle ne contient aucun cadre supérieur (0 % versus 25 %). Leur profil de diplôme est souvent proche de celui de l'échantillon sauf pour l'enseignement technique court et long (59 % versus 31 %).
Par rapport à l'échantillon complet, ces logements se distinguent significativement par les caractéristiques suivantes : ce sont des logements individuels (81 % versus 60 %), de grande surface (135m^2 versus 110m^2), disposant d'un jardin ou d'une cour privative (93 % versus

62 %). Ces logements disposent d'un endroit pour bricoler (94 % versus 64 %), d'un garage très fréquemment utilisé (70 % versus 52 %), plutôt attenant (41 % versus 25 %).
Ce groupe est constitué de ménages ayant une activité assez élevée dans leur logement, ils bricolent et jardinent peu. Ils ont peu de plantes et sont moins nombreux à avoir des animaux. Ils sont majoritairement non fumeurs.
Par rapport aux polluants, les concentrations sont semblables à celles de l'échantillon.

Groupe 4, 104 observations

Ces ménages vivent majoritairement en couple (85 % versus 74 %) avec en moyenne, un enfant de plus de 10 ans (0.78 enfants versus 0.59 enfants). Ils sont plus nombreux que dans l'échantillon (3.2 personnes versus 2.83).
C'est la classe la plus riche (3670 euros versus 2522 euros). Ils sont en majorité cadres supérieurs (87% versus 25 %) ayant un diplôme au moins bac +5 (46.5 % versus 11 %) ou d'enseignement technique court (29.4 % versus 23 %).
Par rapport à l'échantillon complet, ce type de logements correspond aux grandes surfaces (135 m^2 versus 110 m^2) de construction ancienne (1950 versus 1958), en majorité propriété des occupants (69 % versus 63 %).
Les occupants ont une activité assez moyenne dans leur logement. Majoritairement non-fumeurs, ils bricolent et jardinent peu. Ils ont peu de plantes et sont moins nombreux à avoir des animaux.
Par rapport au bloc des polluants, ces ménages présentent des concentrations assez élevées en formaldéhyde (20 ppb versus 18 ppb), en trichloroéthylène (8.43 ppb versus 2 ppb) et en tétrachloroéthylène (1.56 ppb versus 0.59 ppb).

Groupe 5, 137 observations

Cette classe est la plus âgée (69 ans versus 49), elle est constituée en majorité des retraités. On y trouve plus de personnes seules (35.5 % contre 17 %). En moyenne, ces ménages ont des revenus inférieurs à la moyenne de l'échantillon (2177 versus 2522). Ils ont souvent quitté l'école après le certificat d'études primaires (26.1 % contre 10 %) et tirent l'essentiel de leurs ressources de pensions (97.8 % versus 24 %) mais jamais d'une activité salariée (0 % contre 64 %).
Ils habitent généralement des logements collectifs (65% versus 60%) de taille moyenne équipés de meubles en bois massif. Ils entretiennent peu leur logement.
Ils présentent de faibles concentrations pour l'ensemble des aldéhydes. Au niveau des hy-

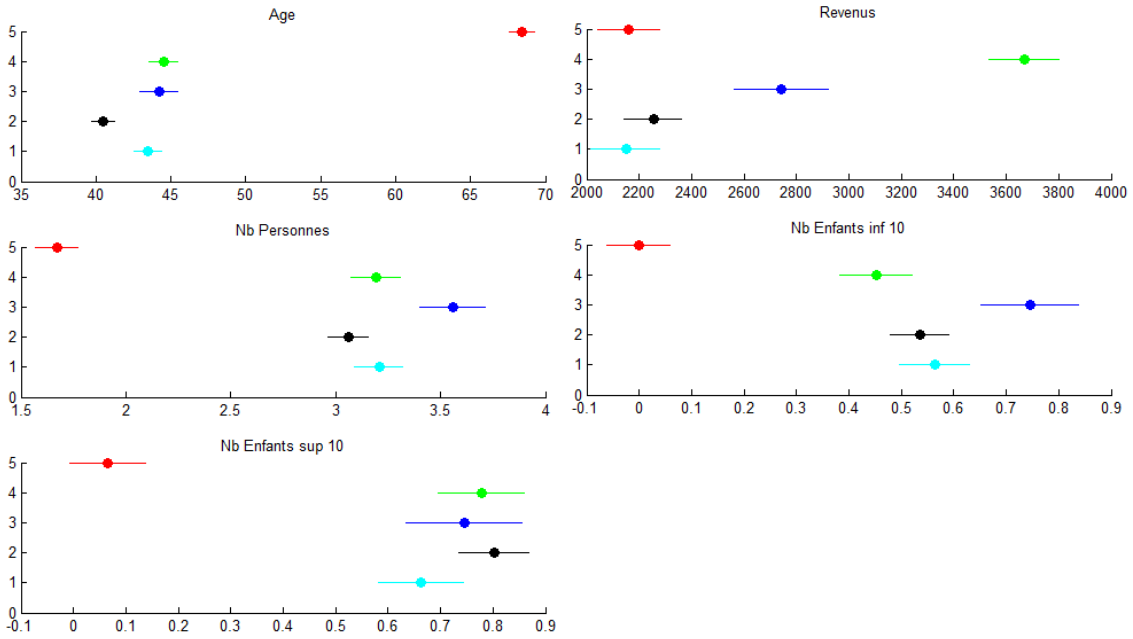


FIGURE 6.7 – Description des classes par rapport aux variables du bloc ménage

drocarbures, seuls les concentrations de toluène sont significativement élevées dans cette classe (6.88 ppb versus 5.87 ppb).

Typologies	Logement	Ménage	Habitude	Polluant	Gle	2S-SOM
Logement	1	0.05	0.05	0.03	0.33	0.35
Ménage	0.05	1	0.07	0.02	0.17	0.45
Habitude	0.05	0.07	1	0.03	0.13	0.30
Polluant	0.03	0.02	0.03	1	0.04	0.25
Gle	0.33	0.17	0.13	0.04	1	0.19
2S-SOM	0.35	0.45	0.30	0.25	0.19	1

TABLE 6.2 – L'information mutuelle normalisée entre les différentes typologies ; Logement, Ménage, Habitude, Polluant, Gle et 2S-SOM correspondent respectivement aux typologies obtenues sur les blocs logement, ménage, habitude et polluants, sur l'ensemble des données et par application de 2S-SOM

6.3.3 Comparaison de la partition consensus avec les partitions obtenues sur chaque bloc

L'AFSSET a réalisé dans le cadre de l'OQAI, une classification des logements au regard des niveaux des concentrations mesurées dans l'air intérieur pour les différents composés organiques volatils (COV). Cette classification des logements a conduit à proposer une typologie en 4 groupes de la pollution par les COV (typologie multi pollution). Par ailleurs,

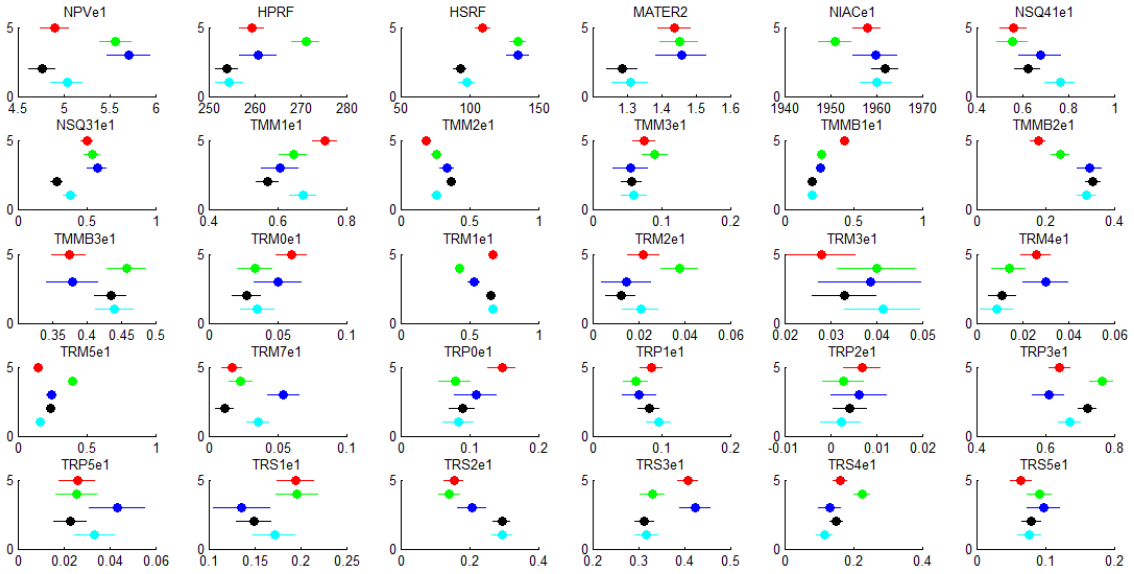


FIGURE 6.8 – Description des classes par rapport aux variables du bloc logement

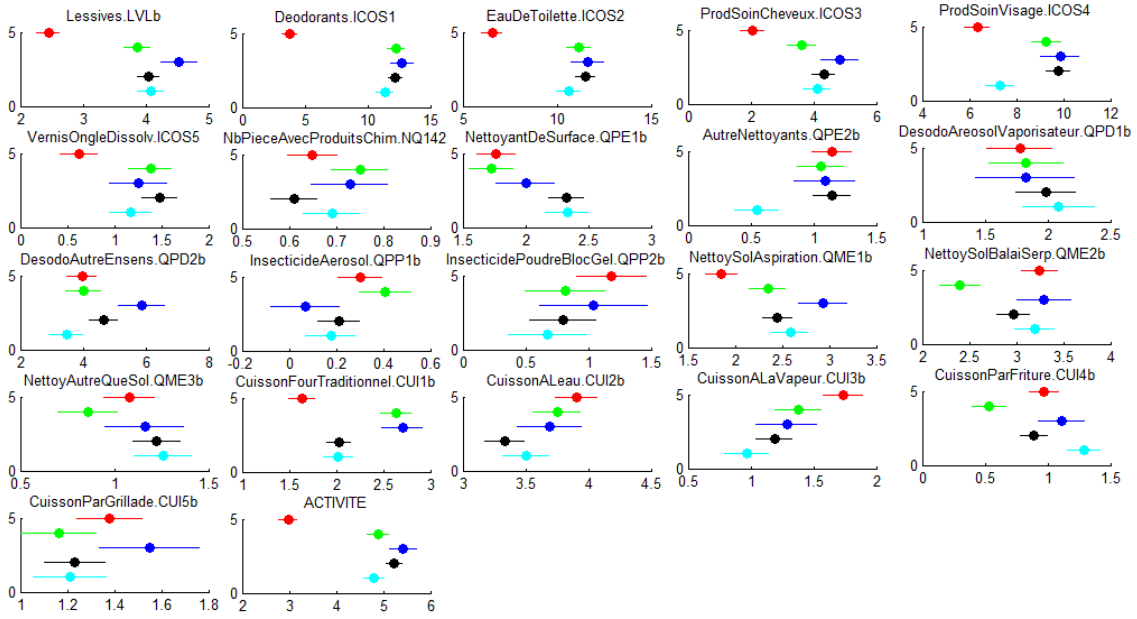


FIGURE 6.9 – Description des classes par rapport aux variables du bloc habitude

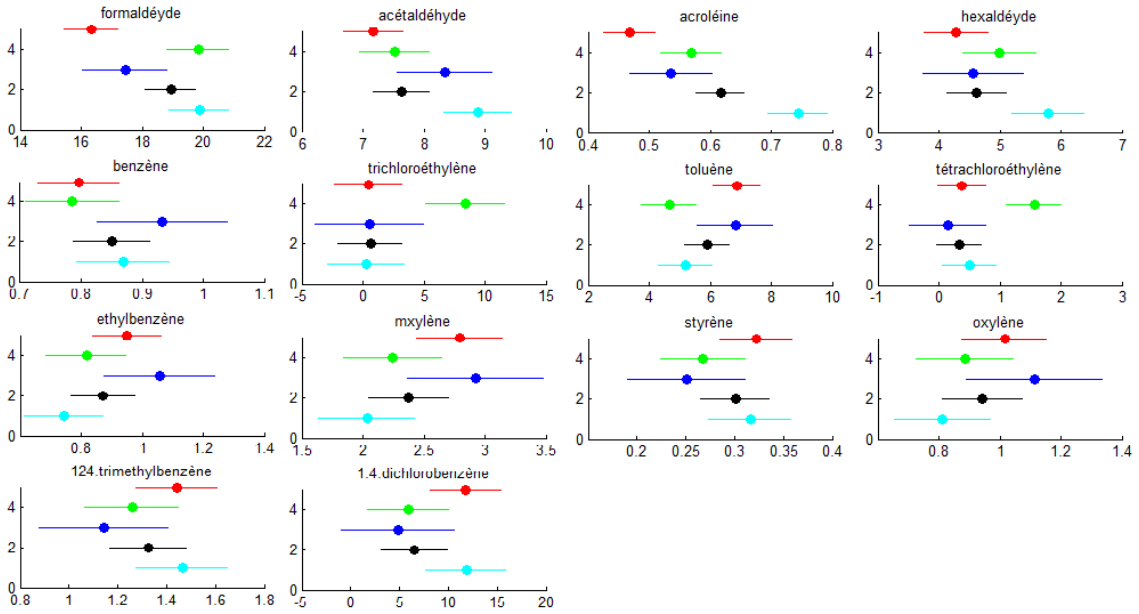


FIGURE 6.10 – Description des classes par rapport aux variables du bloc Polluants

chaque bloc à fait l'objet d'une étude à part entière réaliser par le laboratoire LOCEAN [Kirchner *et al.* 2011]. Les typologies obtenues sur les blocs logement, ménage et habitude ont permis de regrouper les logements en respectivement 6, 7 et 9 classes.

Nous utilisons l"information mutuelle normalisée (NMI) pour mesurer la liaison entre les partitions obtenues sur les blocs avec la typologie fournie par la méthode 2S-SOM. Le tableau 6.2 présente les indices de NMI entre les différentes typologies.

La comparaison des typologies obtenues séparément montre que les blocs logement, ménage et habitude ne partagent pas la même quantité d'information concordante. On remarque de même que la concordance entre la typologie des polluants et celle des blocs logement, ménage et habitude est en générale plus faible, ce qui peut s'expliquer par le caractère multi-factoriel de la pollution de l'air intérieur.

La partition globale obtenue par application de SOM sur l'ensemble des données (concaténation horizontale des différents blocs de variables) est plus concordante avec le bloc logement alors que la concordance est faible avec les blocs ménages et habitudes. À contrario, l'approche de type subspace clustering (2S-SOM) permet d'avoir des concordances

meilleures avec tous les blocs. Ce qui conforte l'idée de recherche des classes dans des sous-espaces de l'espace initial.

6.4 Conclusion

Afin de définir la typologie multi-blocs des données de la CNL, les approches 2S-SOM et CSOM ont été utilisées. La première approche fondée sur un système de poids recherche en ensemble de cartes topologiques qui sont ensuite fusionnées par la méthode CSOM. Dans le cas de cette application, le système défini privilégie le bloc des Polluants. Nous avons ainsi défini une typologie de la base CNL en 5 groupes qui tient compte des 4 blocs de variables et qui met en lien les caractéristiques des observations dans chaque bloc de variables.

Conclusion et perspectives

Dans cette thèse, différents aspects du processus de fouille de données ont été abordés dans le but de lever des verrous concernant l'apprentissage non-supervisé traitement de grands volumes de données complexes.
Dans une première partie, nous avons étudié les méthodes classiques de classification rencontrées dans la littérature et présenté un ensemble de critères d'évaluation de la pertinence d'une partition et d'évaluation du nombre exact de classes ainsi que les méthodes de traitement des données structurées en blocs de variables. Nous nous sommes particulièrement intéressé aux méthodes de type subspace clustering et aux méthodes de recherche de consensus de partitions.

Dans la deuxième partie de ce mémoire, pour surmonter un certain nombre de problèmes liés à la structure multi-blocs des données, à la présence de donnant manquantes et d'"outliers", à la présence de variables non-informative pour la classification (distribution uniforme), nous avons proposé trois méthodes destinées à améliorer les performances globales de classification. Ces méthodes permettent de réaliser une première étude descriptive assez précise des données relativement à la pertinence des variables et des blocs.

La première méthode destinée à traiter les données multi-blocs et basée sur les cartes auto-organisées consiste à définir un système de poids prenant en compte l'importance relative des variables des blocs dans la classification. Cette approche basée sur le critère métrique modifié de SOM, est initialement connue pour la méthode des K-moyennes, nous l'avons utilisée ici en exploitant les avantages des cartes auto-organisatrices. Les résultats obtenus nous semblent satisfaisants, et confirment l'intérêt d'une telle méthode sur des jeux de données complexes comme ceux de l'OQAI.

Les méthodes de cluster ensemble apparaissent comme des outils très prometteurs dans

une optique d'étude de la robustesse des algorithmes de classification. Nous avons présenté, dans le cas des cartes SOM, un premier algorithme de consensus de cartes topologiques pour des données multi-blocs. Son principe consiste à définir une matrice synthétique des données tenant compte de la qualité relative de chaque carte topologique pour améliorer le consensus final des cartes.

La troisième méthode proposée est une méthode de recherche de consensus qui contrairement à la deuxième méthode recherche le consensus des cartes topologiques en tenant compte de la proximité entre les cartes topologiques qui est évaluée par le coefficient de corrélation vectorielles R_V.

Ces approches, ont été appliquées aux données de la CNL et les résultats obtenus ont permis de dégager 5 groupes homogènes de logements caractérisés par les 4 blocs de variables pris en compte dans la CNL. Ainsi, en fonction des paramètres de la méthode 2S-SOM nous avons dégagé les blocs et les groupes pertinents de la classification pour la base CNL.

Plusieurs perspectives de travail nous semblent intéressantes. Puisque les blocs sont préalablement définis par l'utilisateur, on peut s'interroger sur l'influence de ces derniers sur la classification. Il peut alors être intéressant de proposer une alternative à l'application directe de la méthode 2S-SOM. Une première idée serait de redéfinir les blocs à l'aide d'une méthode de classification des variables puis d'appliquer la méthode de partitionnement multi-blocs 2S-SOM sur ces blocs. La seconde idée plus ambitieuse serait d'envisager, comme en classification croisée, la définition des blocs en même temps que la recherche des classes.

Au niveau de la méthode de recherche de consensus R_V-CSOM, les performances sont fortement liées à la corrélation vectorielle entre les cartes de l'ensemble de diversification. Une alternative à cette approche peut être de rendre moins sensible le résultat du consensus lorsque les liaisons sont faibles. Cela peut être fait en liant la définition des poids σ_b à la partition consensus recherchée.

On peut aussi envisager l'extension de nos travaux aux cartes SOM probabilistes dans le contexte plus général de la classification basée sur les modèles de mélanges.

Au niveau application cette méthodologie proposée pourra être directement appliquée aux données des campagnes de L'OQAI plus particulièrement sur les bases CNB.

Bibliographie

Aggarwal C. C, Hinneburg A et Keim D. A : *On the surprising behavior of distance metrics in high dimensional space*. Springer, 2001. 73

Aggarwal C. C, Wolf J. L, Yu P. S, Procopiuc C et Park J. S : Fast algorithms for projected clustering. *In ACM SIGMOD Record*, volume 28, pages 61–72. ACM, 1999. 81

Agrawal R, Gehrke J, Gunopulos D et Raghavan P : Automatic subspace clustering of high dimensional data for data mining applications. *In Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 94–105, New York, NY, USA, 1998. ACM. 36, 37, 73, 84

Alelyani S, Tang J et Liu H : Feature selection for clustering : A review. *Data Clustering : Algorithms and Applications*, page 29, 2013. 76

Anouar F, Badran F et Thiria S : Probabilistic self-organizing map and radial basis function networks. *Neurocomputing*, 20(1):83–96, 1998. 60, 61

Apte M. G et Daisey J. M : Vocs and "sick building syndrome" : Application of a new statistical approach for sbs research to us epa base study data. *Proceedings of Indoor Air*, 99:8–13, 1999. 28

Apte M. G, Fisk W. J et Daisey J. M : Associations between indoor co_2 concentrations and sick building syndrome symptoms in u. s. office buildings : An analysis of the 1994-1996 base study data. *Indoor Air*, 10:246–257, 2000. 28

Aurenhammer F et Klein R : Voronoi diagrams. *Handbook of computational geometry*, 5:201–290, 2000. 103

Ayad H. G et Kamel M. S : Cumulative voting consensus method for partitions with variable number of clusters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30:160–173, 2008. 94

Bache K et Lichman M : UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`. 116

Banfield J. D et Raftery A. E : Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993. 39

Baruque B, Corchado E et Yin H : Visom ensembles for visualization and classification. *In Computational and Ambient Intelligence*, pages 235–243. Springer, 2007. 94, 101

Ben Mena S : Introduction aux méthodes multicritères d'aide à la décision. *Biotechnol. Agron. Soc. Environ*, 4(2):83–93, 2000. 37

Benhadda H et Marcotorchino F : L'analyse relationnelle pour la fouille de grandes bases de données. *Revue des Nouvelles Technologies de l'Information*, pages 149–167, 2007. 100, 101

Berchtold S, Böhm C, Keim D. A et Kriegel H.-P : A cost model for nearest neighbor search in high-dimensional data space. *In Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 78–86, New York, NY, USA, 1997. ACM. 36

Berkhin P : Survey of clustering data mining techniques (2002). *Accrue Software : San Jose, CA*, 2004. 39, 40, 74

Billionnet C : *Pollution de l'air intérieur et santé respiratoire : prise en compte de la multi-pollution*. Thèse de doctorat, Université Pierre et Marie Curie-Paris VI, 2012. 28

Blackmore J. M et Miikkulainen R : Visualizing high-dimensional structure with the incremental grid growing neural network. Mémoire de D.E.A., University of Texas at Austin, 1995. 101

Bluyssen P. M, Cox C, Boschi N, Maroni M, Raw G, Roulet C. A et Foradini F : European project hope (health optimisation protocol for energy-efficient buildings). *Healthy Buildings*, pages 76–81, 2003. 30

Breiman L : Bagging predictors. *Machine learning*, 24(2):123–140, 1996. 90

Bremond-Gignac D, Tixier J, Missotten T, Laroche L et Beresniak A : Évaluation de la qualité de vie en ophtalmologie. *La Presse médicale*, 31(34):1607–1612, 2002. 29

Brightman H, Milton D, Wypij D, Burge H et Spengler J : Evaluating building-related symptoms using the us epa base study results. *Indoor Air*, 18(4):335–345, 2008. 28

Cha S.-H : Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1, 2007. 45

Charrad M et Ben Ahmed M : Simultaneous clustering : A survey. *In* Kuznetsov S, Mandal D, Kundu M et Pal S, éditeurs : *Pattern Recognition and Machine Intelligence*, volume 6744, pages 370–375. Springer Berlin Heidelberg, 2011. 85

Chavent M, Lacomblez C et Patouille B : Critère de rand asymétrique. *Proceedings SFC*, 8:82–88, 2001. 69

Chen N et Marques N. C : An extension of self-organizing maps to categorical data. *In Proceedings of the 12th Portuguese conference on Progress in Artificial Intelligence*, EPIA'05, pages 304–313, 2005. 112

Chen X, Ye Y, Xu X et Huang J. Z : A feature group weighting method for subspace clustering of high-dimensional data. *Pattern Recognition*, 45(1):434–446, 2012. 22, 84, 107, 108, 109, 114

Choi S.-S, Cha S.-H et Tappert C : A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010. 45

Cottrell M, Ibbou S et Letrémy P : Traitement des donnees manquantes au moyen de l'algorithme de kohonen. *arXiv preprint arXiv :0704.1709*, 2007. 59

CSTB : Référentiel technique de certification bâtiments tertiaires - démarche hqe bureau et enseignement, 2005. 29

Dash M et Liu H : Feature selection for clustering. *In* Terano T, Liu H et Chen A, éditeurs : *Knowledge Discovery and Data Mining. Current Issues and New Applications*, volume 1805 de *Lecture Notes in Computer Science*, pages 110–121. Springer Berlin Heidelberg, 2000. 77, 78

Davies D. L et Bouldin D. W : A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):224–227, 1979. 64, 159

De Baudouin C : Qualité de l'air intérieur dans les bâtiments de bureaux : spécificité de la problématique et proposition d'études à mener, 2006. 25, 26, 30

De Soete G et Carroll J. D : K-means clustering in a low-dimensional euclidean space. *In New approaches in classification and data analysis*, pages 212–219. Springer, 1994. 74

Diday E : Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques. *Revue de statistique appliquée*, 19(2):19–33, 1971. 50, 87

Ding C, Li T, Peng W et Park H : Orthogonal nonnegative matrix t-factorizations for clustering. *In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135. ACM, 2006. 145

Domeniconi C, Papadopoulos D, Gunopulos D et Ma S : Subspace clustering of high dimensional data. *In SDM*, 2004. 73, 93

Dubes R et Jain A. K : Clustering techniques : The user's dilemma. *Pattern Recognition*, 8(4):247 – 260, 1976. 62

Dubois C : *Confort et diversité des ambiances lumineuses en architecture : l'influence de l'éclairage naturel sur les occupants.* Thèse de doctorat, Université Laval, 2006. 29

Dudoit S et Fridlyand J : Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9):1090–1099, 2003. 93, 94, 95

EPA U : The inside story : A guide to indoor air quality, 1995. 28

Escofier B : Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. *Cahiers de l'Analyse des Données*, 4(2):137–146, 1979. 37

Fern X. Z et Brodley C. E : Cluster ensembles for high dimensional clustering : An empirical study. *Tufts University, Electrical Engineering and Computer Science,161 College Avenue, Medford, MA 02155, USA.*, 2004. 94

Fern X. Z et Brodley C. E : Random projection for high dimensional data clustering : A cluster ensemble approach. *In ICML*, volume 3, pages 186–193, 2003. 93, 94

Fisher D. H : Knowledge acquisition via incremental conceptual clustering. *Machine learning*, 2(2):139–172, 1987. 99

Fisk W. J, Black D et Brunner G : Benefits and costs of improved ieq in us offices. *Indoor Air*, 21(5):357–367, 2011. 28

Forgy E. W : Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965. 50

Fred A : Finding consistent clusters in data partitions. *In In Proc. 3d Int. Workshop on Multiple Classifier*, pages 309–318. Springer, 2001. 95

Fred A. L et Jain A. K : Robust data clustering. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2:128, 2003. ISSN 1063-6919. 92

Fred A. L et Jain A. K : Combining multiple clusterings using evidence accumulation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(6):835–850, 2005. 96

Friedman J. H et Meulman J. J : Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 66 (4):815–849, 2004. 83

Gentle J. E : *Matrix algebra : theory, computations, and applications in statistics*. Springer, 2007. 142

Georgakis A, Li H et Gordan M : An ensemble of som networks for document organization and retrieval. *In Int. Conf. on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, page 6, 2005. 94, 102, 131

Gordon A. D : A review of hierarchical classification. *Journal of the Royal Statistical Society Series A (General)*, 150(2):119–137, 1987. 48

Gordon A : A survey of constrained classification. *Computational Statistics & Data Analysis*, 21(1):17 – 29, 1996. 112

Gordon A et Vichi M : Partitions of partitions. *Journal of Classification*, 15:265–285, 1998. 90, 94

Govaert G : Classification croisée. *These d'état, Université Paris*, 6, 1983. 15, 85, 86

Govaert G : Classification simultanée de tableaux binaires. *In E. Diday, M. Jambu, L. Lebart, J. Pages, and R. Tomassone, editors, Data analysis and informatics III,North Holland*, 67(337)(1):233–236, 1984. 85, 86, 89

Govaert G : *Classification et modèle de mélange*. Lavoisier, 2003. 60

Guénoche A : Consensus of partitions : a constructive approach. *Advances in data analysis and classification*, 5(3):215–229, 2011. 91

Halkidi M, Batistakis Y et Vazirgiannis M : On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145, 2001. 92

Halkidi M, Batistakis Y et Vazirgiannis M : Cluster validity methods : part i. *SIGMOD Rec.*, 31:40–45, 2002. ISSN 0163-5808. 62

Hartigan J. A : *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 99th édition, 1975. ISBN 047135645X. 86, 89

He X, Cai D et Niyogi P : Laplacian score for feature selection. *In Advances in neural information processing systems*, pages 507–514, 2005. 76

Huang J. Z, Ng M. K, Rong H et Li Z : Automated variable weighting in k-means type clustering. *IEEE Transaction on pattern analysis and machine intelligence*, 27(5):657–668, 2005. 78, 84

Huang Z : Clustering large data sets with mixed numeric and categorical values. *In In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 21–34, 1997. 113

Huang Z : Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2:283–304, 1998. 47, 51

Huang Z et Ng M. K : A fuzzy k-modes algorithm for clustering categorical data. *Fuzzy Systems, IEEE Transactions on*, 7(4):446–452, 1999. 50, 107

Hubert L et Arabie P : Comparing partitions. *Journal of classification*, 2(1):193–218, 1985. 69, 74

I. Guyon A. E : An introduction to variable and feature selection. *J. Mach. Learn. Res*, 3:1157–1182., 2003. 76

Iam-On N, Boongoen T et Garrett S : Refining pairwise similarity matrix for cluster ensemble problem with cluster relations. *In Discovery Science*, pages 222–233. Springer, 2008. 92, 93

Jain A. K, Murty M. N et Flynn P. J : Data clustering : a review. *ACM Comput. Surv.*, 31(3):264–323, septembre 1999a. ISSN 0360-0300. 40

Jain A. K : Data clustering : 50 years beyond k-means. *Pattern Recognition Letters*, 31 (8):651–666, 2010. 15, 51

Jain A. K, Murty M. N et Flynn P. J : Data clustering : a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999b. 73, 74, 91

Jain A : Data clustering : 50 years beyond k-means. *In Machine Learning and Knowledge Discovery in Databases*, volume 5211, pages 3–4. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-87478-2. 62

Jiang Y et Zhou Z.-H : Som ensemble-based image segmentation. *Neural Processing Letters*, 20(3):171–178, 2004. 102

Jing L, Ng M et Huang J : An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *Knowledge and Data Engineering, IEEE Transactions on*, 19(8):1026 –1041, 2007. 22, 84, 107

Jollois F.-X : Contribution de la classification automatique à la fouille de données. *These de Doctorat, Université de Metz*, 12, 2003. 89

Karypis G, Aggarwal R, Kumar V et Shekhar S : Multilevel hypergraph partitioning : Application in vlsi domain. *In Proceedings of the 34th annual Design Automation Conference*, pages 526–529. ACM, 1997. 98, 99

Kaski S : Data exploration using self-organizing maps. *In Acta polytechnica scandinavi : mathematics, computing and management in engineering series No. 82*, 1997. 59

Kaufman L et Rousseeuw P. J : Partitioning around medoids (program pam). *Finding groups in data : an introduction to cluster analysis*, pages 68–125, 1990. 52

Kirchner S, Buchmann A, Cochet C, Dassonville C, Derbez M, Leers Y, Lucas J, Mandin C, Ramalho O et Ouattara J. R. M : *Qualité d'air intérieur, qualité de vie. 10 ans de recherche pour mieux respirer.* CSTB éditions, paris, 2011. 17, 19, 26, 154, 166

Kohonen T : The self-organizing map. *Neurocomputing*, 21(1-3), 1998. 53, 107

Kriegel H.-P, Kröger P et Zimek A : Clustering high-dimensional data : A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data*, 3(1):1 :1–1 :58, mars 2009. ISSN 1556-4681. 37, 73, 74, 80

Krieger A. M et Green P. E : A generalized rand-index method for consensus clustering of separate partitions of the same data base. *Journal of Classification*, 16(1):63–89, 1999. 94, 99

Lavit C, Escoufier Y, Sabatier R et Traissac P : The act (statis method). *Computational Statistics & Data Analysis*, 18(1):97–119, 1994. 143

Law M. H, Topchy A. P et Jain A. K : Multiobjective data clustering. *In Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–424. IEEE, 2004. 93

Lebbah M, Chazotte A, Badran F et Thiria S : Mixed topological map. *ESANN*, 17, 2005. 47, 112

Li T et Ding C : Weighted consensus clustering. *Mij*, 1(2), 2008. 97, 145

Li T, Ding C et Jordan M. I : Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. *In Data Mining*, pages 577–582. IEEE, 2007. 94, 96, 136, 143

Luttrell S. P : Hierarchical self-organising networks. *In Artificial Neural Networks, 1989., First IEE International Conference on (Conf. Publ. No. 313)*, pages 2–6. IET, 1989. 60

Luttrell S. P : A bayesian analysis of self-organizing maps. *Neural Comput.*, 6(5):767–794, septembre 1994. ISSN 0899-7667. 60

MacQueen J : Some methods for classification and analysis of multivariate observation. *Proc. Fifth Berkeley Symp. Math. Statistics and Probability*, pages 281–297, 1967. 50

Maystre L. Y, Pictet J, Simos J et Roy B : *Méthodes multicritères ELECTRE : description, conseils pratiques et cas d'aplication à la gestion environnementale.* Presses polytechniques et universitaires romandes, 1994. 37

Meilă M : The uniqueness of a good optimum for k-means. *In Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 625–632, New York, NY, USA, 2006. ACM. 52

Merlo M : Qualité de l'air intérieur et habitat. analyse des plaintes et de leur suivi. Rapport technique, Centre Scientifique et Technique du Bâtiment, 2002. 28

Mitra P, Murthy C et Pal S. K : Unsupervised feature selection using feature similarity. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):301–312, 2002. 76

Mosqueron L et Nedellec V : Inventaire des données françaises sur la qualité de l'air à l'intérieur des bâtiments, observatoire de la qualité de l'air intérieur. Rapport technique, Observatoire de Qualité de l'Air Intérieur, 2001. 25, 28

Mosqueron L et Nedellec V : Inventaire des données françaises sur la qualité de l'air à l'intérieur des bâtiments : actualisation sur la période 2001-2004, observatoire de la qualité de l'air intérieur. Rapport technique, Observatoire de Qualité de l'Air Intérieur, 2004. 28

Nadif M et Govaert G : Binary clustering with missing data. *Applied stochastic models and data analysis*, 9(1):59–71, 1993. 89

Nakache J.-P et Confais J : *Approche pragmatique de la classification : arbres hiérarchiques, partitionnements.* Editions Technip, 2004. 79

Nguyen N et Caruana R : Consensus clusterings. *In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 607–612. IEEE, 2007. 94, 95

Parat S, Ramalho O, Lahrech R, Mandin C, Gregoire A, Riberon J, Cocher V, Corrales P et Kirchner S : Qualité de l'air intérieur, santé, confort et performance énergétique dans les immeubles de bureaux en france : étude de cadrage de la campagne nationale. Rapport technique, Centre Scientifique et Technique du Bâtiment, 2009. 29, 30

Parsons L, Haque E et Liu H : Subspace clustering for high dimensional data : a review. *SIGKDD Explor. Newsl.*, pages 90–105, 2004. ISSN 1931-0145. 73, 74, 80

Plasse M : *Utilisation conjointe des méthodes de recherche de règles d'association et de classification : contribution à l'amélioration de la qualité des véhicules en production grâce à l'exploitation des systèmes d'information.* Thèse de doctorat, 2006. Thèse de doctorat dirigée par Saporta, Gilbert Informatique Paris, CNAM 2006. 80

Régnier S : Sur quelques aspects mathématiques des problèmes de classification automatique. *Mathématiques et Sciences humaines*, 82:13–29, 1983. 90, 94

Richard D et Jain A. K : Validity studies in clustering methodologies. *Pattern Recognition*, 11(4):235 – 254, 1979. 62

Robert P et Escoufier Y : A unifying tool for linear multivariate statistical methods : the rv-coefficient. *Applied statistics*, pages 257–265, 1976. 141, 143

Roulet C.-A : Santé et qualité de l'environnement intérieur dans les bâtiments. *PPUR, Lausanne*, page 362 p, 2004. 28

Roulet F. FC. A and, Foradini F, Bluyssen P, Cox C et Aizlewood C : Multi-criteria analysis of health, comfort and energy-efficiency of buildings. *Building Research Information*, pages 475–482, 2006. 30

Rousseeuw P. J : Silhouettes a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 63

Saavedra C, Salas R, Moreno S et Allende H : Fusion of self organizing maps. *In Computational and Ambient Intelligence*, pages 227–234. Springer, 2007. 102, 103, 131

Sabine H : Pollution de l'air intérieur : état des connaissances concernant les effets sanitaires et faisabilité d'une étude épidémiologique en ile-de-france : Rapport de stage [en ligne]. *Mémoire d'ingénieur filière du Génie Sanitaire. Rennes : EHESP*, 68, 2005. 26

Saporta G : *Probabilités, analyses des données et statistiques.* Editions Technip, 2006. 37, 52, 99

Smilde A. K, Kiers H. A, Bijlsma S, Rubingh C et Van Erk M : Matrix correlations for high-dimensional data : the modified rv-coefficient. *Bioinformatics*, 25(3):401–405, 2009. 143

Strehl A et Ghosh J : Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617., 2002. 90, 92, 93, 98, 100, 136, 143, 145

Topchy A, Jain A. K et Punch W : A mixture model of clustering ensembles. *In Proc. SIAM Intl. Conf. on Data Mining*, 2004a. 91, 100

Topchy A, Jain A. K et Punch W : Clustering ensembles : Models of consensus and weak partitions. *IEEE transactions on pattern analysis and machine intelligence*, pages 1866–1881, 2005. 93

Topchy A. P, Law M. H, Jain A. K et Fred A. L : Analysis of consensus partition in cluster ensemble. *In Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*, pages 225–232. IEEE, 2004b. 90

Tuomainen M, Smolander J, Kurnitski J, Palonen J et Seppanen O : Modelling the cost effects of the indoor environment. *Proceedings of Indoor Air*, pages 814–819, 2002. 28

Wal Van der J. F, Hoogeveen A. W et Wouda P : The influence of temperature on the emission of volatile organic compounds from pvc flooring, carpet, and paint. *Indoor Air*, 7(3):215–221, 1997. 26

Vega-Pons S et Ruiz-Shucloper J : A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03):337–372, 2011. 37, 74

Vesanto J, Himberg J, Alhoniemi E et Parhankangas J : *SOM toolbox for Matlab 5*. Citeseer, 2000. 112

Vesanto J, Sulkava M et Hollmén J : On the decomposition of the self-organizing map distortion measure. *In Proceedings of the workshop on self-organizing maps (WSOM'03)*, pages 11–16, 2003. 66

Vichi M et Kiers H. A : Factorial k-means analysis for two-way data. *Computational Statistics and Data Analysis*, 37(1):49 – 64, 2001a. 74

Vichi M et Kiers H. A : Factorial k-means analysis for two-way data. *Computational Statistics and Data Analysis*, 37(1):49 – 64, 2001b. 74

Vigneau E et Qannari E : Clustering of variables around latent components. *Communications in Statistics-Simulation and Computation*, 32(4):1131–1150, 2003. 79

Witten D. M et Tibshirani R : A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490), 2010. 78, 79

Yacoub M, Badran F et Thiria S : *A Topological Hierarchical Clustering : Application to Ocean Color Classification*. 2001. 57

Zhao Z et Liu H : Spectral feature selection for supervised and unsupervised learning. *In Proceedings of the 24th international conference on Machine learning*, pages 1151–1157. ACM, 2007. 76

Annexes

Les blocs de variables

Variables du bloc Logement		
Code question	Libellé question	Réponses possibles
NPVe1	Nombre de pièces de vie	
HPRF	Quelle est la hauteur moyenne des pièces de votre logement, en dehors de celle des pièces professionnelles et des pièces annexes ? (cm)	
HSRF	Quelle est la surface totale de votre logement, en dehors de celle des pièces professionnelles et des pièces annexes ?	
MATER2	Les murs principaux de l'immeuble sont-ils constitués d'un seul matériau ou de plusieurs ?	1=Un seul matériau ;2=Deux Matériaux ; 3=Trois matériaux
NIACe1	Ancienneté de l'immeuble ou du logement date numérique	
NSQ41	Nombre d'appareils àcombustion indépendants non raccordés àun conduit de fumé dans le logement	
NSQ31	Nombre d'appareils àcombustion indépendants raccordés à un conduit de fumé dans le logement	
TMM1	Taux de menuiseries en bois dans le logement (pièces de vie)	
TMM2	Taux de menuiseries en PVC dans le logement (pièces de vie)	
TMM3	Taux de menuiseries autre que bois et PVC dans le logement (pièces de vie)	
TMMB1	Taux de présence de meubles en bois massif dans le logement (pièces de vie)	

Code question	Libellé question	Réponses possibles
TMMB2	Taux de présence de meubles en bois aggloméré dans le logement (pièces de vie)	
TMMB3	Taux de présence de meubles en bois mixte dans le logement (pièces de vie)	
TRM0	Taux de murs en crépi dans le logement (pièces de vie)	
TRM1	Taux de murs en papier peint dans le logement (pièces de vie)	
TRM2	Taux de murs en tissu dans le logement (pièces de vie)	
TRM3	Taux de murs en carrelage dans le logement (pièces de vie)	
TRM4	Taux de murs en lambris dans le logement (pièces de vie)	
TRM5	Taux de murs en peinture dans le logement (pièces de vie)	
TRM7	Taux de murs sans revêtement dans le logement (pièces de vie)	
TRP0	Taux de plafonds en bois dans le logement (pièces de vie)	
TRP1	Taux de plafonds en papier peint dans le logement (pièces de vie)	
TRP2	Taux de plafonds en tissu dans le logement (pièces de vie)	
TRP3	Taux de plafonds en peinture dans le logement (pièces de vie)	
TRP5	Taux de plafonds sans revêtement dans le logement (pièces de vie)	
TRS1	Taux de sols en moquette dans le logement (pièces de vie)	
TRS2	Taux de sols en plastique dans le logement (pièces de vie)	
TRS3	Taux de sols en carrelage dans le logement (pièces de vie)	
TRS4	Taux de sols en parquet bois massif dans le logement (pièces de vie)	
TRS5	Taux de sols en parquet stratifié dans le logement (pièces de vie)	
TRS6	Taux de sols en peinture dans le logement (pièces de vie)	

Code question	Libellé question	Réponses possibles
TRS8	Taux de sols sans revêtement dans le logement (pièces de vie)	
ACTP1	Existe-t-il une activité professionnelle dans l'immeuble où est situé le logement ?	1=Oui ;2=Non
CHEM1	Votre logement est-il équipé d'une cheminée ?	1=Oui ;2=non
DBRI	Disposez-vous d'un endroit où vous pouvez bricoler (atelier, hangar, garage) ?	1=Oui ATTENANTau logement ; 2=Oui NON ATTENANT au logement ;3=Non
DCA3	Existence d'une cave	1=oui communiquante ; 2=oui non-communiquante ; 3=non
DGG2b	Existence d'un garage ?	1=oui attenant et communiquant avec le logement ; 2=oui non attenant ; 3=non
FC3	Votre logement est :	1=maison individuelle ; 2=appart dans immeuble collectif ;3=studio dans immeuble collectif ; 4= pièce indépendante ;5 logement-foyer personnes âgées ; 6=ferme,bât d'exploitation agricole ;7 chambre hôtel ;8=construction provisoire ; 9=logement dans un immeuble àusage professionnel
FC8	Occupez-vous ce logement comme :	1=Fermier ou métayer ;2=Propriétaire ou accédant àla propriété ; 3=Logé gratuitement ;4=Locataire ou sous-locataire
FC9b	Quel est votre propriétaire ?	1=organisme HLM (office, société ou OPAC) ; 2=une autre société du secteur public ou privé ; 3=une administration ; 4=une association ; 5=un membre de votre famille ;6=un autre particulier ;7=Autre cas
HCU1	Avez-vous une cuisine ?	1=Oui cuisine FERMEE ;2=Oui cuisine OUVERTE (américaine) ;3=Non pas de cuisine mais une installation pour faire la cuisine (avec évacuation des eaux usées) ; 4=Non pas dinstallation pour faire la cuisine

Code question	Libellé question	Réponses possibles
HCU3	Votre cuisine est-elle équipée d'une hotte aspirante ?	1=Oui avec filtre pour lair recyclé ;2=Oui avec rejet de lair àlextérieur du logement (orifice en faÃ§ade) ;3=Oui avec rejet de lair raccordé àun conduit de ventilation ;4=Non pas de hotte
HPLBO	Le plancher le plus bas des pièces d'habitation est en :	1= Plancher en bois ;2=Plancher en béton ou similaire
KCC1	Quel est le combustible principal utilisé pour votre installation de chauffage (y compris chauffage électrique type convecteurs) ?	1=Fioul domestique ; 2=Butane ou propane (GPL) en citerne ; 3=Butane ou propane (GPL) en bouteille ; 4=Gaz de réseau (gaz de ville) ; 5=Charbon ;6=Bois ; 7=Charbon+bois ; 8=Electricité ; 9=Autre cas
KCC	Votre logement est-il équipé d'un système de chauffage ?	1=Aucun ;2=Urbain ;3=Electrique mixte (câbles+convecteurs fixes) ;4=Electrique individuel (convecteurs fixes) ;5=Chauf. central individuel (+chaudières électriques) ;6=Chauf. central collectif dimmeuble ; 7=Chauf. central collectif de groupe d'immeubles
KEIU1	Le logement est équipé d'appareils indépendants de production d'eau chaude de type Chauffe-eau SANS BALLON (production instantanée) ?	1=Oui ;2=Non
KEIU2	Le logement est équipé d'appareils indépendants de production d'eau chaude de type "Ballon, cumulus ou autre appareil àaccumulation" ?	1=Oui ;2=Non
KVNT2	Votre logement est-il aéré au moyen de :	11=Une ventilation mécanique générale simple flux ; 12=Une ventilation mécanique générale double flux ;2=Des moteurs de ventilateurs placés dans quelques pièces ;3=Des conduits ou des grilles daération (ventilation naturelle) ;4=Aucun dispositif particulier
MATER0	Les murs principaux de l'immeuble sont constitués de :	1=1- Bois ;2=2- Brique ; 3=3- Béton ;4=4- Parpaing ; 5=5- Granit ; 6=6- Pierre ;7=7- Autre

Code question	Libellé question	Réponses possibles
REAB	L'immeuble a-t-il fait l'objet de travaux de réhabilitation ?	1=Oui, il y a moins de 5 ans ;2=Oui, il y a plus de 5 ans et moins de 10 ans ;3=Oui, il y a plus de 10 ans ;4=Non, jamais

Variables du bloc Ménage		
Code question	Libellé question	Réponses possibles
DIPLOM2	Diplôme le plus élevé obtenu simplifié	1=Sans diplôme ;2=fin du premier cycle enseignement général ; 3=fin du second cycle enseignement général ; 4=enseignement technique court ; 5=enseignement technique long ; 6= fin études primaire ;7=enseignement supérieur bac+2 ; 8=enseignement supérieur bac+3/4 ; 9=enseignement supérieur bac+5 &+
PROFES1	Profession (ou dernière profession exercée) niveau 1	1=agriculteur ; 2=artisans ; 3=cadre supérieur ; 4=prof intermédiaire ; 5=employé ; 6=ouvrier ; 7=autre
NOCCUA	Occupation actuelle	1=Exerce une profession ;2=Chômeur ; 3=Etudiant ; 4=Retraite ou pré-retraite/Retiré des affaires ; 5=Au foyer ; 6= Autre inactif
REV1	Quelle est la ressource principale de votre ménage ?	1=Salaires ou traitements ; 2=Revenus dune activité indépendante ; 3=Préretraite, retraite, pensions et rentes diverses ; 4=Indemnités de chômage / RMI et allocations sociales ; 5=Revenus des actifs fonciers ou financiers
STRMEN	structure du ménage	1=Salaires ou traitements (0.17) ; 2=Revenus dune activité indépendante (0.05) ; 3=Préretraite, retraite, pensions et rentes diverses (0.24) ; 4=Indemnités de chômage / RMI et allocations sociales (0.04) ; 5= Revenus des actifs fonciers ou financiers (0.04)
AGE	1 Age de la personne (en années)	
REV3	Revenu du ménage ? (variable quantitative)	

Code question	Libellé question	Réponses possibles
FC10	Combien d'enfants de moins de 10 ans ?	
FC11	Combien d'enfants de moins de 10 ans ?	
FC11b	Combien d'enfants de plus de 10 ans ?	

Variables du bloc Habitude		
Code question	Libellé question	Réponses possibles
DGG3n	Rentrez-vous une ou plusieurs voitures dans ce garage attenant ?	1 =tous les jours (0.18) ;2=Parfois (0.03) ; 3=Rarement ou jamais (0.04) ; 4=question pas posée (0.75)
FUMEUR	Des personnes du ménage fument-elles àl'intérieur de votre logement	1=Non, pas de fumeur (0.57) ; 2=Il y a des fumeurs mais personne ne fume à l'intérieur du logement (0.11) ; 3=Un fumeur (0.22) ; 4=Deux fumeurs ou Trois fumeurs et plus (0.11) ;
LVSbn	A cette époque de l'année faites vous habituellement sécher votre linge à l'intérieur de votre logement ;	1=non (0.32) ; 2=oui (0.68)
LVS4n	Au cours des 4 dernières semaines, avez-vous utilisé un sèche linge	1= Tous les jours (0.06) ; 2= Plus de 2 fois par jour (0.9) ; 3 =1 à 2 fois par semaine (0.08) ; 4= Moins d'une fois par semaine (0.07) ; 5= Jamais au cours des 4 dernières semaines (0.69) ; 6=question pas posée (0.02)
QNS1	Au cours des 4 DERNIERES SEMAINES, avez-vous fait nettoyer A SEC (pressing) puis récupéré des vêtements ou des textiles ?	1=Oui 1 fois par semaine ou plus (0.03) ; 2=Oui moins d'une fois par semaine (0.15) ; 3=Non pas au cours des 4 dernières semaines (0.82)
AMN1n	Faites-vous appel régulièrement à une personne extérieure au ménage pour vous aider dans vos tâches ménagères (ménage, repassage, etc) ?	1 =Non (0.82) ; 2 = Oui (0.18)
QTA1n	Au cours des 12 DERNIERS MOIS, avez-vous introduit un tapis neuf ou une descente de lit neuve dans votre logement ? (hors carpettes et tapis de bains de petites dimensions)	1 = Non (0.88) ; 2 = Oui (0.12)

Code question	Libellé question	Réponses possibles
QTARIn	Au cours des 12 DERNIERS MOIS, avez-vous nettoyé ou fait nettoyer A SEC des tissus d'ameublement ou des tapis ?	1=Non (0.92) ; 2=Oui (0.08)
QPV	Combien de plantes avez-vous à l'intérieur de votre logement ?	1=aucune plante (0.19) ; 2= de 1 à 4 plantes (0.45) ; 3= de 5 à 14 plantes (0.29) ; 4=15 plantes et plus (0.07)
ANI21n	Avez-vous des chiens ?	1= Non ou question pas posée (0,71) ; 2 =Oui (0.29)
ANI22n	Avez-vous des chats ?	1=Non ou question pas posée (0.72) ; 2=Oui (0.28)
ANTCPn	Traitement contre les parasites pour animaux domestiques chien ou chat ?	1= Oui (0.16) ; 2 =Non (0.31) ; 3 = question pas posée (0.53)
ANI25n	Avez-vous des hamster, cochon d'Inde, lapin, chinchilla ?	1=Non ou question pas posée (0.92) ; 2=Oui (0.08)
ULn	Utilisation d'un insecticide contre blattes/cafards, ou fourmis ou puces, ou termites au cours des 12 derniers mois	1=Non (0.83) ; 2= Oui ou Oui en cours (0.17)
QOM1	A cette époque de l'année, vous sortez vos ordures ménagères à l'extérieur du logement :	1=tous les jours (0.48) ; 2=2 à 3 fois par semaines (0.46) ; 3=tous les autres cas (0.07)
QOM2n	Avant d'être sorties, vos ordures ménagères sont-elles stockées dans une pièce à l'intérieur du logement	1 =Non (0.16) ; 2= Oui (0.84)
QOM4	Les ordures ménagères sont-elles stockées dans :	1=Un meuble sous évier) dans une poubelle (0.29) ; 2=Une poubelle sans couvercle (0.15) ; 3 =Une poubelle avec couvercle (0.48) ; 4=Autres cas (0.07)
TMG6n	Au cours des 4 dernières semaines, vous avez bricolé :	1=Rarement ou jamais (0.49) ; 2= Moins d'une fois/semaine (0.12) ; 3=Une fois par semaine (0.11) ; 4=plusieurs fois/semaine (0.18) ; 5=Tous les jours (0.09)
TMG7n	Au cours des 4 dernières semaines, vous avez jardiné :	1=Rarement ou jamais (0,60) ; 2=Moins d'1 fois/semaine (0.11) ; 3=Une fois par semaine (0.12) ; 4=plusieurs fois/semaine (0.10) ; 5=Tous les jours (0.06)
LIMHO	Ce matelas est-il enveloppé dans une housse ou un protège matelas ?	1=Pas de housse simple (0.94) ; 2=Housse simple (0.06)

Code question	Libellé question	Réponses possibles
LIMHO13	Ce matelas est-il enveloppé dans une housse ou un protège matelas ?	1=Pas de housse imperméable (0.92) ; 2=Housse imperméable (0.08)
LVLb	Au cours des 4 DERNIERES SEMAINES, avez-vous fait des lessives ? (à la main ou en machine)	
ICOS1	Indice quantitatif d'utilisation de déodorants dans le ménage	
ICOS2	Indice quantitatif d'utilisation d'eau de toilette dans le ménage	
ICOS3	Indice quantitatif d'utilisation de produits de soin cheveux dans le ménage	
ICOS4	Indice quantitatif d'utilisation de produits de soin visage et corps dans le ménage	
ICOS5	Indice quantitatif d'utilisation de vernis à ongle, disssolvants	
NQ142	Nombre de pièces de vie avec stockage de produits chimiques	
QPE1b	Au cours de la semaine, avez-vous utilisé dans votre logement un NETTOYANT DE SURFACE destinés aux sols, murs, vitres, ameublements (carrelage, céramique, faience, plastique, vitres, cire, dépoussiérant) ?	
QPE2b	Au cours de la semaine, vous avez utilisé dans votre logement un AUTRE TYPE DE NETTOYANT (fours-plaques de cuisson, anti-calcaire, anti-moisissure, débouche évier, détachant, shampoing moquette) ?	
QPD1b	Au cours de la semaine, vous avez utilisé dans votre logement un DESODORISANT ou des PARFUMS D'AMBIANCE sous forme d'AEROSOL ou VAPORISATEUR ou PISTOLET ?	

Code question	Libellé question	Réponses possibles
QPD2b	Au cours de la semaine, vous avez utilisé dans votre logement un AUTRE TYPE DE DESODORISANT(diffuseur mèche, bougie, lampe, encens, pot pourri, désodorisant pour aspirateur, bloc WC, solide, gel, etc.) ?	
QPP1b	Au cours de la semaine, vous avez utilisé dans votre logement des INSECTICIDES ou PESTICIDES sous forme d'AEROSOL (pour végétaux, animaux, domestiques) ?	
QPP2b	Au cours de la semaine, vous avez utilisé dans votre logement un INSECTICIDE ou PESTICIDE sous forme de piège, contaminateur, poudre, bloc, gel, plaquette, diffuseur, etc.) ?	
QME1b	14 Au cours de la semaine, combien de fois avez-vous NETTOYE LES SOLS PAR ASPIRATION mécanique ?	
QME2b	Au cours de la semaine, combien de fois avez-vous NETTOYE LES SOLS à l'aide d'un BALAI ou d'une SERPILLERE ?	
QME3b	Au cours de la semaine, combien de fois avez-vous NETTOYE d'AUTRES SURFACES que le sol (mobilier, vitrines) ?	
CUI1b	Au cours de la semaine, combien de fois avez-vous cuit des aliments AU FOUR TRADITIONNEL ?	
CUI2b	Au cours de la semaine, combien de fois avez-vous cuit des aliments A L'EAU ?	
CUI3b	Au cours de la semaine, combien de fois avez-vous cuit des aliments A LA VAPEUR ?	
CUI4b	Au cours de la semaine, combien de fois avez-vous cuit des aliments PAR FRITURE ?	

Code question	Libellé question	Réponses possibles
CUI5b	Au cours de la semaine, combien de fois avez-vous cuit des aliments PAR GRILLADE ?	

Glossaire

- ADEME : Agence de l'Environnement et de la Maîtrise de l'Energie
- CEDRIC : Centre d'Etude et de Recherche en Informatique et Communication
- CNAM : Conservatoire National des Arts et Métiers
- BASE : Building Assessment Survey and Evaluation
- BRS : Building Related Symptom
- BSI : Building Symptom Index
- COV : Composé Organique Volatil
- CSTB : Centre Scientifique et Technique du Bâtiment
- HOPE : Health Optimisation Protocol for Energy-efficient buildings
- ppb : est une manière d'exprimer les concentrations et les proportions en général et correspond à un rapport de 10^{-9}
- OQAI : Observatoire de la Qualité de l'Air Intérieur
- OMS : Organisation Mondiale de la Santé
- $PM_{2.5}$: Matière Particulaire dont le diamètre est inférieur à 2,5 μm
- PM_{10} : Matière Particulaire dont le diamètre est inférieur à 10 μm
- QAI : Qualité de l'Air Intérieur
- SBS : Sick Building Syndrome
- US-EPA : United States Environmental Protection Agency

Mory Ouattara

Développement et mise en place d'une méthode de classification multi-blocs

Application aux données de l'OQAI

Résumé :
La multiplication des sources d'information et le développement de nouvelles technologies ont engendré des bases données complexes, souvent caractérisées par un nombre de variables relativement élevé par rapport aux individus. L'objectif de ce travail a été de développer des méthodes de classification adaptées à ces jeux de données de grande dimension et structurées en blocs de variables.
La première partie de ce travail présente un état de l'art des méthodes de classification en général et dans le cas de la grande dimension. Dans la deuxième partie, trois nouvelles approches de classification d'individus décrits par des variables structurées en blocs ont été proposées. La méthode 2S-SOM (Soft Subspace-Self Organizing Map), une approche de type subspace clustering basée sur une modification de la fonction de coût de l'algorithme des cartes topologiques à travers un double système de poids adaptatifs défini sur les blocs et sur les variables. Nous proposons ensuite des approches CSOM (Consensus SOM) et Rv-CSOM de recherche de consensus de cartes auto-organisées basées sur un système de poids déterminés à partir des partitions initiales. Enfin, la troisième partie présente une application de ces méthodes sur le jeu de données réelles de la campagne nationale logement (CNL) menée par l'OQAI afin de définir une typologie des logements au regard des thématiques : qualité de l'air intérieur, structure du bâtiment, composition des ménages et habitudes des occupants.

Mots clés :
classification, multi-blocs, subspace clustering, consensus, SOM

Abstract :
The multiplication of information source and the development of news technologies generates complex databases, often characterized by relatively high number of variables compared to individuals. However, in case of high dimensional data, classical clustering algorithms are not efficient to find clusters which may exist in subspaces of the original space. The goal of this work is to develop clustering algorithms adapted to high dimensional data sets with multi-block structure. The first part of the work shows the state of art on clustering methods. In the second part, three new methods of clustering : the subspace clustering method 2S-SOM (Soft Subspace-Self Organizing Map) is based on a modified cost function of the Self Organizing Maps method across a double system of weights on the blocks and the variables. Then we propose two approaches to find the consensus of self-organized maps CSOM (Consensus SOM) and Rv-CSOM based on weights determined from initial partitions. The last part presents an application of these methods on the OQAI data to determine a typology of dwellings relatively to the following topics : indoor air quality, dwellings structure, household characteristics and habits of the inhabitants.

Keywords :
clustering, multi-block, subspace clustering, cluster ensemble, SOM