Chapitre 3

Réorganisation des cubes de données par une approche factorielle

" Ne réorganisez jamais sauf pour une bonne raison. Mais si cela fait un moment que vous ne l'avez pas fait, c'est une bonne raison. "

John Akers

3.1 Introduction

Dans un système décisionnel, les données sont organisées selon un modèle, en *étoile* ou en *flocon de neige*, dédié à l'analyse et traduisant un contexte d'étude ciblé [Inm96, Kim96]. Autour d'une table de faits centrale contenant une ou plusieurs mesures à observer, existent plusieurs tables de dimensions comprenant des descripteurs. Une dimension peut comporter plusieurs hiérarchies exprimant différents niveaux de granularités dans la description de chaque fait. Cette organisation est particulièrement adaptée pour créer des cubes de données destinées à l'analyse OLAP. Dans un cube, un fait est ainsi identifié par un ensemble de modalités des différentes dimensions. Le fait est observé par une ou plusieurs mesures ayant des propriétés d'additivité.

L'analyse en ligne est un outil basé sur la visualisation permettant la navigation et l'exploration dans ces cubes de données. L'objectif est d'observer des faits, à travers une ou plusieurs mesures, en fonction de différentes dimensions. Il s'agit, par exemple, d'observer les niveaux de ventes en fonction des produits, des périmètres commerciaux (localisations géographiques) et de la période d'achat.

De cette visualisation dépend la qualité de l'exploration des données. Or, différents facteurs peuvent dégrader cette visualisation. D'une part, la représentation

multidimensionnelle engendre une éparsité, puisqu'à l'intersection de différentes modalités de dimensions, il n'existe pas forcément de faits correspondants. Cette éparsité peut être accentuée par la présence d'un grand nombre de dimensions (forte dimensionnalité) et/ou d'un grand nombre de modalités dans chacune des dimensions.

D'autre part, les modalités des dimensions sont généralement représentées selon un ordonnancement lexical pré-établi qui correspond souvent à un ordre naturel (ordre chronologique pour les dates et alphabétique pour les libellés par exemple.) Par conséquent, dans la plupart des cas, les points associés aux faits observés (les cellules pleines) sont éparpillés dans l'espace des dimensions d'un cube de données.

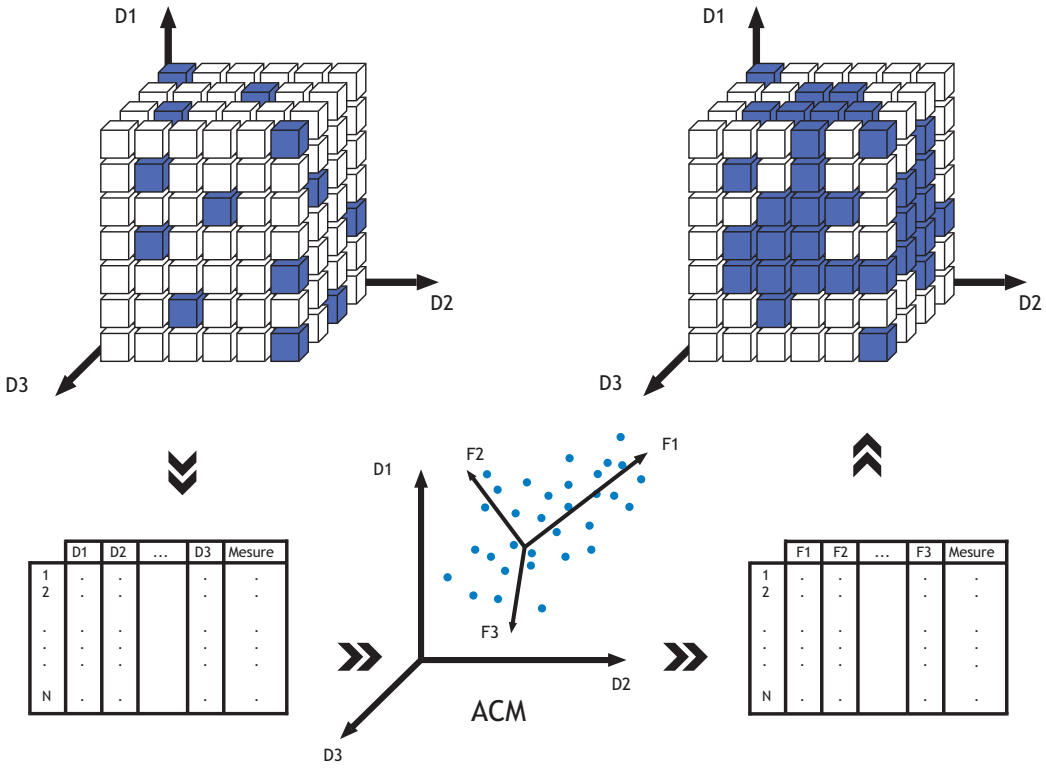


FIG. 3.1 – Étapes de la réorganisation d'un cube de données par approche factorielle

Dans ce chapitre, nous proposons d'améliorer la visualisation des données dans les cubes. Notre proposition consiste à coupler l'analyse en ligne avec l'analyse des correspondances multiples (ACM) [Ben73]. Nous adoptons la première approche du couplage, présentée dans le chapitre 2, qui se base sur la transformation des données multidimensionnelles en données tabulaires afin de les exploiter par des algorithmes de fouille. Nous résumons les étapes de notre approche selon l'aperçu général de la figure 3.1. La première étape dans ce processus consiste à transformer les données

initiales d'un cube en tableau *individus-variables* selon un codage binaire spécifique à l'ACM. Ensuite, nous appliquons l'ACM aux données transformées. Nous obtenons ainsi des axes factoriels qui représentent aux mieux les faits OLAP et qui traduisent des relations avec les modalités des dimensions du cube. Chaque axe factoriel (ou facteur) est caractérisé par une valeur propre indiquant l'inertie (dispersion) des individus dans la direction définie par cet axe [LMP00].

Dans notre approche, nous essayons d'établir via ce couplage une méthode efficace de *réorganisation* des données multidimensionnelles afin de réduire l'effet de leur éparsité. D'une manière plus générale, nous exploitons l'ACM comme un outil d'aide à la construction de cubes de données ayant de meilleures caractéristiques pour la visualisation. Il est à noter que l'objectif de notre présente méthode n'est pas de diminuer l'éparsité des cubes, comme par exemple dans [NNT03], mais d'atténuer plutôt son effet négatif sur la visualisation. Notre idée consiste à *regrouper les cellules pleines* et à les séparer le mieux possible des cellules vides dans l'espace de représentation d'un cube de données. Pour ce faire, nous proposons d'arranger l'ordre des modalités dans chaque dimension du cube étudié étant donné que leurs ordres initiaux n'engendrent pas forcément une bonne visualisation.

Dans [MRB05], nous avons déjà amorcé une réflexion sur l'usage de l'analyse factorielle dans un contexte OLAP où nous avons montré que l'ACM construit des axes factoriels qui offrent de meilleurs points de vue du nuage de points des faits d'un cube. Dans ce chapitre, nous exploitons les associations fournies par l'ACM entre les modalités du cube étudié et les axes factoriels construits. Ces associations se traduisent par les contributions des modalités dans la construction des axes factoriels. Nous exploitons ces contributions afin d'arranger les modalités dans chaque dimension du cube selon deux façons. Alors que la première arrange les modalités selon leurs *projections* sur les axes factoriels [MAF05], la seconde façon les arrange selon leurs *valeurs-test* [MBR05, MBR06d, MBR06b].

Ce chapitre est organisé de la manière suivante. Nous exposons notre contexte de travail, nos objectifs et nos motivations dans la section 3.2. Un état de l'art sur les travaux ayant particulièrement traité du problème des représentations multidimensionnelles des données est fourni dans la section 3.3. Dans les deux sections suivantes, nous présentons les notations générales, les définitions et les formalismes que nous adoptons. Les deux façons de réorganiser des faits OLAP dans les cubes de données sont présentées dans la section 3.6. Afin évaluer la qualité des nouvelles représentations des données générées par notre approche, nous proposons un indice d'homogénéité dans la section 3.7. Deux études de cas sur des jeux de données réelles et des études de performance font l'objet des deux sections suivantes. Enfin, dans la dernière section, nous concluons et présentons de nouvelles directions de recherche pour notre approche.

3.2 Objectifs et motivations

La vocation de l'OLAP est de fournir à l'utilisateur un outil visuel pour explorer et naviguer dans les données d'un cube afin d'y découvrir des informations pertinentes. Toutefois, dans le cas de données volumineuses, telles que les données bancaires ou les données démographiques considérées dans notre étude, l'analyse en ligne n'est pas une tâche facile pour l'utilisateur. En effet, un cube à forte dimensionnalité comportant un grand nombre de modalités, présente souvent une structure éparse difficile à exploiter visuellement. De plus, l'éparsité, souvent répartie de façon aléatoire dans le cube, altère davantage la qualité de la visualisation et de la navigation dans les données.

Prenons l'exemple de la figure 3.2 qui présente un cube de données à deux dimensions : les localités géographiques d'agences bancaires $(L_1, \ldots, L_8)$ et les produits de la banque $(P_1, \ldots, P_{10})$. Les cellules grisées sur la figure sont pleines et représentent la mesure des faits existants (chiffres d'affaires, par exemple) alors que les cellules blanches sont vides et correspondent à des faits inexistants. D'après la figure 3.2, la répartition des cellules pleines dans la représentation (a) ne se prête pas facilement à l'interprétation. En effet, visuellement, l'information est éparpillée dans l'espace de représentation des données. En revanche, dans la représentation (b), les cellules pleines sont concentrées dans une zone centrale du cube. Cette représentation offre des possibilités de comparaison et d'analyse des valeurs des cellules pleines (les mesures des faits) plus aisées et plus rapides pour l'utilisateur.

Notons que les deux représentations de la figure 3.2 correspondent au même cube de données. La représentation (b) est obtenue par simples permutations de lignes et de colonnes de la représentation (a). Dans la plupart des serveurs OLAP, les modalités d'une dimension sont présentées selon un ordre arbitraire. En général, cet ordre est alphabétique pour les libellés des modalités et chronologique pour les dimensions temporelles. Malheureusement, dans le cas des cubes éparses et volumineux, ce choix entraîne des représentations de données inadaptées à l'analyse, voire même difficilement exploitables, comme c'est le cas de la représentation (a) de la figure 3.2.

La composante visuelle de l'OLAP est primordiale dans un processus décisionnel. En effet, de la qualité et de la clarté de celle-ci dépendent les orientations de l'utilisateur dans son exploration du cube. Ceci détermine l'intérêt de l'analyse en ligne. En se basant sur notre idée de l'arrangement des modalités des dimensions illustrée dans cet exemple, nous proposons une méthode permettant à l'utilisateur d'améliorer automatiquement la qualité de la représentation des données. Nous souhaitons produire une meilleure visualisation homogénéisant au mieux le nuage des faits (cellules pleines) et mettant en avant des points de vue intéressants pour l'analyse. Notre idée de réorganisation consiste à rassembler géométriquement les cellules pleines dans l'espace de représentation des données. Dans ce travail, nous ne cherchons pas à diminuer l'éparsité du cube, mais à le réorganiser de manière à atténuer l'impact négatif sur la visualisation qu'elle engendre.

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
L_1	32	18		24	81		16	52		18
L_2								43		
L_3		16		20			28	15		
L_4		74						43		
L_5		61		22				14		53
L_6				31		13				
L_7		44	65	49			67	21	43	
L_8				12						

(a)

	P_1	P_3	P_5	P_7	P_8	P_4	P_2	P_{10}	P_9	P_6
L_2					43					
L_6						31				13
L_3				28	15	20	16			
L_1	32		81	16	52	24	18	18		
L_7		65		67	21	44	44		43	
L_5					14	22	61	53		
L_4					43		74			
L_8						12				

(b)

FIG. 3.2 – Exemple de deux représentations d'un espace de données

Pour des raisons de complexité de traitements, nous avons exclu la recherche d'un optimum global, voire même local, de l'indice de qualité selon une exploration exhaustive des configurations possibles du cube ; c'est à dire, toutes les combinaisons des arrangements possibles des modalités des dimensions du cube. En effet, considérons le cas d'un cube à trois dimensions où chaque dimension comporte seulement 10 modalités. Le nombre de configurations possibles pour ce cube est égal à $A_{10}^{10} \times A_{10}^{10} \times A_{10}^{10} = 10! \times 10! \times 10! \simeq 4,7 \cdot 10^{19}$.

Afin de parvenir à un arrangement convenable des modalités du cube, sans passer par une recherche exhaustive d'un optimum, nous avons choisi d'utiliser les résultats d'une analyse des correspondances multiples (ACM) [Ben73]. L'ACM est considérée comme une heuristique appliquée à la volée aux données du cube

que l'utilisateur cherche à visualiser. Les individus et les variables de l'ACM correspondent, respectivement, aux faits et aux dimensions du cube. En construisant des axes factoriels, l'ACM fournit une représentation d'associations entre individus et d'associations entre variables dans un espace réduit. Ces axes factoriels permettent d'ajuster au mieux le nuage de points des individus et des variables.

Cependant, l'ACM s'applique classiquement sur un *tableau disjonctif complet* obtenu en remplaçant chaque variable qualitative par l'ensemble des variables indicatrices des différentes modalités de cette variable. Ainsi, une adaptation des données multidimensionnelles du cube en un tableau disjonctif complet s'impose pour notre méthode. Cette adaptation des données multidimensionnelles en données tabulaires est un élément clé dans le cadre de notre couplage général entre l'analyse en ligne et la fouille de données. Nous décrivons en détail dans la section 3.5 cette étape d'adaptation des données ainsi que les différentes étapes de l'ACM sur les données transformées. À présent, dans la section suivante de ce chapitre, nous exposons un état de l'art des différents travaux relatif au problème de la représentation des données multidimensionnelles.

3.3 Représentation des données multidimensionnelles

Les travaux de recherche qui se sont intéressés à l'étude de l'espace de représentation ont été menés suite à des motivations différentes. Tandis que certains se sont penchés sur des aspects d'optimisation technique (stockage, temps de réponse, etc.), d'autres s'intéressent plutôt à l'aspect de l'analyse en ligne, et particulièrement à la visualisation. Notre travail se rapproche davantage des seconds travaux. Tout d'abord, nous présentons les travaux qui ont abordé l'approximation des cubes de données, leur compression et l'optimisation des calculs d'agrégats.

3.3.1 Compression des cubes de données

En se basant sur le principe d'approximation par ondelettes (*wavelets*), Vitter *et al.* [VW99] proposent un algorithme pour construire un cube de données compact. L'algorithme proposé fournit des résultats meilleurs que ceux de l'approximation par histogrammes ou par échantillonnage aléatoire [VWI98]. Dans le même ordre d'idées, Barbara et Sullivan [BS97] ont proposé l'approche Quasi-Cube qui matérialise une partie, au lieu de la totalité, du cube en se basant sur une description incomplète mais suffisante des données. Les données non matérialisées sont ensuite approximées par une régression linéaire.

Une technique de compression basée sur la modélisation statistique de la structure des données d'un cube a été proposée dans [SFB99]. Après estimation de la densité de probabilité des données, les auteurs construisent une représentation compacte des

données capable de supporter des requêtes d'agrégation. Cette technique n'a de sens que dans le cas de cubes présentant des dimensions continues.

La méthode de compression Dwarf proposée dans [SDRK02], réduit l'espace de stockage d'un cube de données. Cette méthode consiste à identifier les n-uplets redondants dans la table de faits. Les redondances de données sont ensuite remplacées par un seul enregistrement. Wang *et al.* [WLFY02] proposent de factoriser ces redondances par un seul n-uplet de base appelé BST (*Base Single Tuple*). À partir du BST, les auteurs construisent un cube de données de moindre taille MinCube (*Minimal condensed BST Cube*). Cette approche requiert des temps de traitement relativement longs. En vue de remédier à cette limite, Feng *et al.* [FFD04] ont repris l'approche en introduisant une nouvelle structure de données PrefixCube. Ils suggèrent de ne plus utiliser tous les BST dans la construction du cube mais plutôt de se contenter d'un seul BST par dimension. En contre partie, ils proposent l'algorithme BU-BST pour la construction d'un cube compressé (*Bottom Up BST algorithm*). Cet algorithme est une version améliorée de l'algorithme BUC (*Bottom Up Computation algorithm*) proposé à l'origine dans [BR99].

Lakshmanan *et al.* [LPH02] proposent la méthode Quotient Cube pour la compression d'un cube de données en résumant son contenu sémantique et en le structurant sous forme de partitions de classes. La meilleure partition n'est pas seulement celle qui permet de réduire la taille du cube mais aussi celle qui permet de conserver une structure de treillis valide donnant la possibilité de naviguer avec les opérations de forage vers le haut (*Roll-Up*) et de forage vers le bas (*Drill-Down*) dans le cube réduit. Malheureusement, la technique des Quotient Cube fournit des structures peu compactes. De plus, ces structures ne sont pas adaptées aux mises à jours des données. Dans [LPZ03], Lakshmanan et *al.* proposent une nouvelle version améliorée QC-Tree (*Quotient Cube Tree*) qui pallie les limites de la technique des Quotient Cube. QC-Tree permet de rechercher les structures compactes de données dans un cube, d'extraire et de construire les cubes intéressants à partir des données mises à jour.

Feng *et al.* [FAAM04] proposent la méthode Range CUBE pour la compression des cubes en se basant sur les corrélations entre les cellules du cube. Cette approche consiste à créer un arrangement des cellules d'un cube selon un certain formalisme d'appartenance introduit dans les nœuds du treillis du cube original. Cet arrangement permet de produire une nouvelle structure du cube plus compacte et moins coûteuse en stockage et en temps de réponse.

Ross et Srivastava [RS97] traitent le problème de l'optimisation du calcul d'agrégats dans les cubes de données éparses. Les auteurs proposent l'algorithme Partitioned-Cube qui partitionnent les relations entre les données d'un cube en plusieurs fragments de façon à ce qu'ils tiennent en mémoire centrale. Cette mesure permet de réduire le coût des entrées/sorties. Les fragments de données sont ensuite traités indépendamment, un par un, afin de calculer les agrégats possibles et de générer des sous-cubes de données. Cette notion de fragment est reprise dans les

travaux de Li *et al.* [LHG04]. Leur méthode, appelée Shell Fragment, partitionne un ensemble de données de forte dimensionnalité en sous-ensembles disjoints de données de dimensionnalités moins importantes appelés *"fragments"*. Pour chaque fragment, un cube de données local est calculé. Les identifiants des n-uplets participant à la construction de cellules non vides dans un fragment sont enregistrés. Ces identifiants sont utilisés pour lier différents fragments et reconstruire de petits cubes (*cuboïdes*) nécessaires à l'évaluation d'une requête. Le cube de données de départ est assemblé via ces fragments.

3.3.2 Organisation des cubes de données

À notre connaissance, peu de travaux se sont intéressés à l'étude de l'espace de représentation en vue d'améliorer la visualisation des cubes de données. Néanmoins, citons les travaux de Choong *et al.* [CLLM04, CLM03] qui ont une motivation similaire à la nôtre. Les auteurs utilisent les règles floues (combinaison d'un algorithme de règles d'association et de la théorie des sous-ensembles flous) afin de faciliter la visualisation et la navigation dans l'espace de représentation des cubes de données. Leur approche, consiste à identifier et à construire des blocs de données similaires au sens de la mesure du cube. Cependant, cette approche ne prend pas en compte le problème d'éparsité du cube. De plus, elle se base sur le comptage du nombre d'occurrences des mesures où ces dernières sont considérées comme des nombres de type entier.

3.4 Notations générales

Nous adoptons dans ce chapitre, ainsi que dans les deux chapitres suivants, les mêmes notations générales relatives à la structure d'un cube de données. Pour faciliter la compréhension des formalismes des nos différentes propositions, nous utilisons également le même exemple du cube de données des *ventes* de la figure 3.3.

Soit donc $\mathcal{C}$ un cube de données ayant les propriétés suivantes :

- $\mathcal{C}$ est constitué d'un ensemble non vide de d dimensions $\mathcal{D} = \{D_i\}_{(1 \leq i \leq d)}$;
- $\mathcal{C}$ contient un ensemble non vide de m mesures $\mathcal{M} = \{M_q\}_{(1 \leq q \leq m)}$;
- chaque dimension $D_i \in \mathcal{D}$ contient un ensemble non vide de n_i niveaux hiérarchiques. Nous considérons que H_j^i est le $j^{ième}$ niveau hiérarchique de la dimension D_i. Par exemple, dans la figure 3.3, la dimension *Lieu* (D_1) contient deux niveaux ($n_1 = 2$) : *Continent* et *Pays*. Le niveau *Continent* est noté H_1^1 et le niveau *Pays* est noté H_2^1 ;
- le niveau d'agrégation totale *All* dans une dimension correspond au niveau hiérarchique zéro. Par exemple, dans la dimension D_1 ce niveau est noté H_0^1 ;

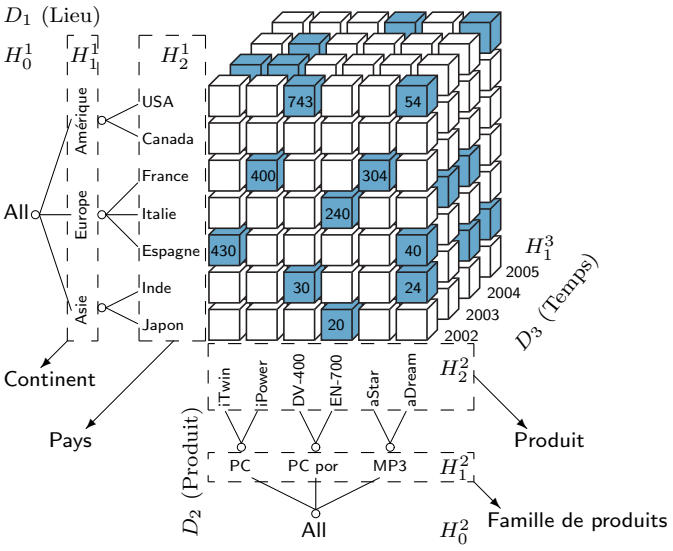


FIG. 3.3 – Exemple d'un cube de données de ventes

- $\mathcal{H}_i = \{H_j^i\}_{(0 \leq j \leq n_i)}$ représente l'ensemble des niveaux hiérarchiques de la dimension D_i, Par exemple, dans la figure 3.3, l'ensemble des niveaux hiérarchiques de D_2 est $\mathcal{H}_2 = \{H_0^2, H_1^2, H_2^2\} = \{All, Famille\ de\ produits, Produit\}$;
- chaque niveau hiérarchique $H_j^i \in \mathcal{H}_i$ consiste en un ensemble non vide de l_{ij} modalités. Nous considérons que a_t^{ij} est la $t^{ième}$ modalité du niveau H_j^i. Par exemple, dans le cube de la figure 3.3, le niveau *Famille de produits* (H_1^2) de la dimension *Produit* (D_2) contient trois modalités ($l_{21} = 3$) : *PC*, notée a_1^{21}, *PC por*, notée a_2^{21} et *MP3*, notée a_3^{21} ;
- $\mathcal{A}_{ij} = \{a_t^{ij}\}_{(1 \leq t \leq l_{ij})}$ représente l'ensemble des modalités du niveau hiérarchique H_j^i de la dimension D_i. Par exemple, dans la figure 3.3, l'ensemble des modalités du niveau *Produit* de D_2 est $\mathcal{A}_{22}$ =$\{$*iTwin*, *iPower*, *DV-400*, *EN-700*, *aStar*, *aDream*$\}$;
- pour le niveau d'agrégation total d'une dimension, nous considérons que *All* est la seule modalité de ce niveau. Ainsi, pour une dimension D_i, on note que $a_1^{i0} = All$ et $\mathcal{A}_{i0} = \{All\}$.

3.5 Définitions et formalisation

Notre approche peut s'appliquer directement au cube de données $\mathcal{C}$ ou à une vue partielle de $\mathcal{C}$ (un sous-cube). L'utilisateur est libre de choisir les dimensions qui l'intéresse, fixer un niveau hiérarchique dans chacune de ces dimensions afin d'observer les données multidimensionnelles selon une mesure M_q qu'il aurait sélectionnée. Ainsi, dans le but d'améliorer la qualité de représentation de cette configuration du cube de données, l'utilisateur peut appliquer notre approche factorielle.

Dans la suite, supposons que le point de départ de notre méthode correspond à la configuration du cube $\mathcal{C}$ à d dimensions $(D_1, \ldots, D_i, \ldots, D_d)$ et n faits OLAP observés selon la mesure quantitative M_q. Dans le but d'alléger les notations, nous assimilons volontairement une dimension D_i à son niveau hiérarchique H_j^i $(0 < j \leq n_i)$ sélectionné par l'utilisateur. Ainsi, on notera que chaque dimension D_i contient l_i modalités catégorielles au lieux de l_{ij}. Soit donc $\{a_1^i, \ldots, a_t^i, \ldots, a_{l_i}^i\}$ l'ensemble des modalités de la dimension D_t. On note aussi que $l = \sum_{i=1}^{d} l_i$ est le nombre total de toutes les modalités de $\mathcal{C}$.

Nous considérons également qu'une cellule A dans un cube $\mathcal{C}$ est pleine (respectivement, vide) si elle contient une mesure d'un fait existant (respectivement, ne contient pas de faits).

3.5.1 Tableau disjonctif complet

Classiquement, une analyse de correspondance multiple (ACM) ne peut opérer que sur des données catégorielles codées en binaire selon un tableau disjonctif complet. Ainsi, afin d'appliquer l'ACM sur $\mathcal{C}$, nous sommes amenés à transformer ce dernier et à le représenter sous forme d'un tableau disjonctif complet.

Pour chaque dimension D_i $(i \in \{1, \ldots, d\})$, nous générons une matrice Z_i à n lignes et l_i colonnes. Z_i est telle que sa $k^{ième}$ ligne contenant $(l_i - 1)$ fois la valeur 0 et une fois la valeur 1 dans la colonne correspondant à la modalité que prend le fait f_k $(k \in \{1, \ldots, n\})$. Z_i est un sous-tableau disjonctif qui décrit la partition des n faits induite par les modalités de la dimension D_i. Le terme général de la matrice Z_i s'écrit :

$$z_{kt}^i = \begin{cases} 1 & \text{si le fait } f_k \text{ prend la modalité } a_t^i \text{ de la dimension } D_i \\ 0 & \text{sinon} \end{cases} \tag{3.5.1}$$

En juxtaposant les d matrices Z_i, nous construisons la matrice Z à n lignes et l colonnes. $Z = [Z_1, Z_2, \ldots, Z_i, \ldots, Z_d]$ est un tableau disjonctif complet qui décrit les d positions des n faits du cube $\mathcal{C}$ par un codage binaire. Dans l'algorithme 1, nous résumons cette transformation du cube de données en tableau disjonctif complet

(ligne 1 à 11). La figure 3.4 (a) représente un exemple simple d'un cube de données à 3 dimensions $D_1 : \{L_1, L_2\}$, $D_2 : \{T_1, T_2\}$ et $D_3 : \{P_1, P_2, P_3\}$. Ce cube est transformé en un tableau disjonctif complet selon la matrice Z de la figure 3.4 (b).

Id	D_1	D_2	D_3	M_1
1	$L1$	$T2$	$P1$	9
2	$L2$	$T2$	$P3$	5
3	$L2$	$T1$	$P2$	6
4	$L1$	$T1$	$P3$	7

(a)

	Z						
	Z_1		Z_2		Z_3		
Id	$L1$	$L2$	$T1$	$T2$	$P1$	$P2$	$P3$
1	1	0	0	1	1	0	0
2	0	1	0	1	0	0	1
3	0	1	1	0	0	1	0
4	1	0	1	0	0	0	1

(b)

FIG. 3.4 – Exemple de transformation d'un cube de données en tableau disjonctif complet

3.5.2 Tableau de contingence

À partir du tableau disjonctif complet Z, nous construisons le tableau symétrique $B = Z'Z$, où Z' désigne la transposée de Z. B est une matrice d'ordre (l, l) qui rassemble les croisements deux à deux de toutes les dimensions du cube $\mathcal{C}$. B est appelé tableau de contingence de *Burt* associé à Z (ligne 13 dans l'algorithme 1).

La matrice B contient en diagonale d sous-matrices diagonales correspondant chacune à une dimension. Chacune de ces sous-matrices contient en diagonale les effectifs marginaux de chaque modalité de la dimension en question. En dehors de ces sous-matrices, la matrice B contient tous les croisements possibles des effectifs des modalités des d dimensions du cube de données $\mathcal{C}$. Par exemple, la figure 3.5 (b) représente le tableau de contingence de *Burt* obtenu à partir du tableau disjonctif complet Z de la figure 3.5 (a).

3.5.3 Équation factorielle

Soit X la matrice diagonale, d'ordre (l, l), ayant les mêmes éléments diagonaux que le tableau de contingence de *Burt* B et des zéros ailleurs (ligne 14 à 22 dans l'algorithme 1). Afin de trouver les axes factoriels, l'ACM consiste à diagonaliser une matrice $S = \frac{1}{d} Z'ZX^{-1}$ dont le terme général s'écrit selon l'équation suivante :

$$s_{tt'} = \frac{1}{dz_{.t'}} \sum_{k=1}^{n} z_{kt} z_{kt'} \tag{3.5.2}$$

$$Z = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad B = Z'Z = \begin{pmatrix} 2 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 2 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 2 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 2 \end{pmatrix}$$

(a) (b)

FIG. 3.5 – Exemple de transformation d'un tableau disjonctif complet en tableau de contingence de *Burt*

Après diagonalisation de la matrice S (ligne 24 à 26 dans l'algorithme 1), l'ACM fournit $(l-d)$ valeurs propres notées λ_α, où $\alpha \in \{1, \ldots, (l-d)\}$. Chaque valeur propre λ_α correspond à un axe factoriel F_α, de vecteur directeur u_α et vérifiant dans $\mathbb{R}^l$ *l'équation factorielle* suivante :

$$Su_\alpha = \lambda_\alpha u_\alpha \tag{3.5.3}$$

3.5.4 Contribution d'une modalité

Les modalités d'une dimension D_i sont projetées sur les $(l-d)$ axes factoriels. Soit φ_α^i le vecteur des projections des l_i modalités de D_i sur F_α.

$$\varphi_\alpha^i = \begin{pmatrix} \varphi_{\alpha 1}^i \\ \vdots \\ \varphi_{\alpha t}^i \\ \vdots \\ \varphi_{\alpha l_i}^i \end{pmatrix}$$

Nous désignons par φ_α le vecteur des d projections des modalités de toutes les dimensions sur l'axe factoriel α.

$$\varphi_\alpha = \begin{pmatrix} \varphi_\alpha^1 \\ \vdots \\ \varphi_\alpha^i \\ \vdots \\ \varphi_\alpha^d \end{pmatrix}$$

Entrée cube de données $\mathcal{C}$
Sortie : valeurs propres λ_α

pour $i = 1$ à d **faire**
  $Z_i \leftarrow 0$
  pour tout modalité a_t^i dans D_i **faire**
    pour tout fait f_k dans $\mathcal{C}$ **faire**
      si le fait f_k prend la modalité a_t^i **alors**
        $z_{kt}^i \leftarrow 1$
        arrêter pour
      fin si
    fin pour
  fin pour
  $Z \leftarrow$ **joindre**(Z,Z_i)
fin pour
$B \leftarrow ZZ'$
pour $t = 1$ à l **faire**
  pour $t' = 1$ à l **faire**
    si $t \neq t'$ **alors**
      $x_{tt'} \leftarrow 0$
    sinon
      $x_{tt'} \leftarrow b_{tt'}$
    fin si
  fin pour
fin pour
$S \leftarrow \frac{1}{d} Z'ZX^{-1}$
$S \leftarrow$ **diagonaliser**(S)
pour $\alpha = 1$ à $l - d$ **faire**
  $\lambda_\alpha \leftarrow s_{\alpha\alpha}$
fin pour

Algorithme 1: Construction des axes factoriels à partir d'un cube de données

φ_α vérifie l'équation suivante :

$$\frac{1}{d} X^{-1} Z' Z \varphi_\alpha = \lambda_\alpha \varphi_\alpha \tag{3.5.4}$$

La contribution d'une modalité a_j^t dans la construction de l'axe α est évaluée par :

$$Cr_\alpha(a_t^i) = \frac{z_{.t}^i \varphi_{\alpha t}^{i\ 2}}{nd\lambda_\alpha} \tag{3.5.5}$$

où $z_{.t}^i = \sum_{k=1}^{n} z_{kt}^i$ correspond au nombre de faits dans le cube $\mathcal{C}$ ayant la modalité a_t^i. En d'autre termes, $z_{.t}^i$ correspond au poids de la modalité a_t^i dans le cube $\mathcal{C}$. $Cr_\alpha(a_t^i)$ représente la part d'inertie due à la modalité a_t^i dans la construction de l'axe factoriel F_α.

Dans notre approche d'arrangement des modalités selon leurs projections, nous prenons en compte les contributions des dimensions dans l'inertie de chaque axe factoriel. La contribution d'une dimension D_i dans la construction du facteur α est la somme des contributions de ses modalités :

$$Cr_\alpha(D_i) = \sum_{t=1}^{l_i} Cr_\alpha(a_t^i) = \frac{1}{nd\lambda_\alpha} \sum_{t=1}^{l_i} {z_{.t}^i \varphi_{\alpha t}^i}^2 \tag{3.5.6}$$

On repère ainsi les dimensions du cube initial qui ont participé à la définition de chaque axe factoriel de l'ACM. La contribution d'une dimension à un axe factoriel est un indicateur de liaison entre la dimension et le facteur.

3.5.5 Valeur-test d'une modalité

Soit $I(a_t^i)$ l'ensemble des faits ayant pris la modalité a_t^i pour la dimension D_i. Notons aussi n_t^i le nombre de ces faits dans $I(a_t^i)$. En d'autre termes, n_t^i correspond au nombre de faits dans le cube de données $\mathcal{C}$ ayant a_t^i comme modalité, ce qui correspond aussi au poids de la modalité a_t^i dans le cube.

$$n_t^i = \text{Card}(I(a_t^i)) = \sum_{k=1}^{n} z_{kt}^i \tag{3.5.7}$$

Nous notons $\psi_{\alpha k}$ la coordonnée du fait f_k selon l'axe factoriel F_α. Par conséquent, la coordonnée de la modalité a_t^i selon l'axe F_α s'exprime selon l'expression suivante :

$$\varphi_{\alpha t}^i = \frac{1}{n_t^i \sqrt{\lambda_\alpha}} \sum_{f_k \in I(a_t^i)} \psi_{\alpha k} \tag{3.5.8}$$

Supposons, sous une hypothèse nulle H_0, que si les n_t^i faits sont choisit aléatoirement dans l'ensemble des n faits du cube, alors la moyenne des coordonnées sur l'axe factoriel F_α de ces n_t^i faits peut être représentée selon une variable aléatoire centrée $Y_{\alpha t}^i$:

$$Y_{\alpha t}^i \;=\; \frac{1}{n_t^i} \sum_{f_k \in I(a_t^i)} \psi_{\alpha k} \tag{3.5.9}$$

où la variance de cette variable aléatoire est exprimée selon :

$$\text{VAR}_{H_0}(Y_{\alpha t}^i) \;=\; \frac{n - n_t^i}{n-1} \; \frac{\lambda_\alpha}{n_t^i} \tag{3.5.10}$$

Sachant que $\varphi_{\alpha t}^i \;=\; \frac{1}{\sqrt{\lambda_\alpha}} \; Y_{\alpha t}^i$ et que la moyenne de $Y_{\alpha t}^i$ est nulle, alors la moyenne de la variable $\varphi_{\alpha t}^i$ est nulle aussi. Par conséquent, la variance de $\varphi_{\alpha t}^i$ est exprimé selon :

$$\text{VAR}_{H_0}(\varphi_{\alpha t}^i) \;=\; \frac{n - n_t^i}{n-1} \; \frac{1}{n_t^i} \tag{3.5.11}$$

Enfin, la valeur-test de la modalité a_t^i selon l'axe factoriel F_α s'exprime selon :

$$V_{\alpha t}^i = \sqrt{n_t^i \, \frac{n-1}{n-n_t^i}} \, \varphi_{\alpha t}^i \tag{3.5.12}$$

Dans notre deuxième type d'arrangement des modalités, nous prenons en compte l'ordre d'importance de leurs valeurs-test sur les axes factoriels. Une valeur-test $V_{\alpha t}^i$ mesure en nombre d'écart-types la distance entre la modalité a_t^i et l'origine de l'axe factoriel F_α. Ainsi, la position d'une modalité est intéressante dans une direction α données si le sous-nuage qu'elle constitue occupe une zone étroite dans cette direction et si cette zone est éloignée du centre de gravité du nuage. La valeur-test est un critère qui permet d'apprécier rapidement si une modalité a une position *significative* sur un axe [LMP00].

3.6 Arrangement des modalités

Rappelons que, pour arranger les modalités dans les dimensions du cube, nous exploitons leurs contributions dans la construction des axes factoriels de l'ACM. Néanmoins, nous adoptons deux façons pour réarranger les modalités. La première exploite les projections des modalités sur les axes factoriels et la deuxième façon utilise les valeurs-test des modalités sur ces axes. Nous exposons dans la suite le principe de chaque type d'arrangement.

3.6.1 Arrangement des modalités selon leurs projections

Cet arrangement de modalités consiste à associer à chaque dimension initiale D_i le meilleur axe factoriel F_α possible. Pour cela, nous exploitons les contributions relatives des dimensions dans la construction des axes factoriels.

Pour une dimension D_i donnée, nous cherchons, parmi les axes factoriels F_α, celui qui a été le mieux expliqué par les modalités de cette dimension. Nous cherchons à maximiser la valeur de $\lambda_\alpha Cr_\alpha(D_i)$. Il s'agit donc de chercher l'axe F_{α^*} pour lequel la somme des carrés des projections pondérées des modalités de la dimension D_i est maximale. En d'autres termes, nous cherchons l'indice α^* vérifiant l'équation suivante :

$$\lambda_{\alpha^*} Cr_{\alpha^*}(D_i) = \max_{\alpha \in \{1,\ldots,l-d\}} (\lambda_\alpha Cr_\alpha(D_i)) \tag{3.6.1}$$

Nous récupérons ensuite les coordonnées des modalités a_t^i de la dimension D_i sur l'axe factoriel F_{α^*} le mieux expliqué par ces dernières. Ces coordonnées correspondent aux l_i projections $\varphi_{\alpha^* t}^i$ sur F_{α^*}. Selon un tri croissant de ces coordonnées, nous

obtenons un nouvel ordre des indices t avec lequel nous arrangeons les modalités a_t^i dans la dimension D_i.

L'intérêt de cet arrangement est de converger vers une répartition des modalités de la dimension suivant l'axe factoriel. Cet arrangement a pour effet de concentrer les cases pleines au centre du cube et d'éloigner les cases vides vers les extrémités. Sans diminuer l'éparsité, cette méthode nous permet néanmoins d'améliorer la répartition des données dans le cube.

3.6.2 Arrangement des modalités selon leurs valeurs-test

Une valeur-test $V_{\alpha t}^i$ mesure le nombre d'écarts types entre la modalité a_t^i (le centre de gravité des n_t^i faits) et le centre de gravité de l'axe factoriel F_α. Ainsi, la position d'une modalité est intéressante dans la direction d'un axe factoriel F_α si le sous-nuage qu'elle constitue occupe une zone étroite dans cette direction et si cette zone est éloignée du centre de l'axe F_α. La valeur-test est un critère qui permet d'apprécier si une modalité a une position *significative* sur un axe factoriel.

Dans une représentation classique d'un cube de données, les modalités des dimensions sont généralement organisées selon un ordre lexical tels que l'ordre alphabétique pour des dimensions géographiques ou encore l'ordre chronologique pour des dimensions temporelles. D'autres représentations n'adoptent pas d'ordre particulier dans l'organisation de ses modalités. Dans ce cas, les modalités de chaque dimension sont organisées d'une manière aléatoire.

Dans le cadre de notre approche, nous proposons d'exploiter les valeurs-tests des modalités afin d'organiser différemment les faits d'un cube de données. La nouvelle organisation permet de mettre en valeur dans un cube de données une représentation intéressante qui se prête mieux à l'analyse en ligne OLAP. Cette organisation a d'autant plus d'intérêt lorsque les cubes de données sont éparses et ont un grand volume.

Pour chaque dimension, nous trions ses modalités selon l'ordre croissant de leurs valeurs-test. Or, pour une modalité données, on associe s valeurs-tests dont chacune correspond à un des s premiers axes factoriels choisis par l'utilisateur. Une valeur-test d'une modalité est plus importante lorsqu'elle indique la position de cette dernière sur un axe factoriel important (ayant une grande valeur propre).

Pour cela, nous proposons de trier les modalités d'une dimension selon l'ordre croissant de leurs valeurs-test sur le premier axe factoriel F_1, puis sur le deuxième axe factoriel F_2, jusqu'au tri des valeurs-test sur le $s^{ième}$ axe factoriel F_s. Par exemple, pour une dimension D_i, nous trions les l_i modalités de cette dimension selon les valeurs-test V_{1t}^i, puis selon les valeurs-test V_{2t}^i, jusqu'aux valeurs-test V_{st}^i.

3.7 Évaluation de la qualité de représentation multidimensionnelle

Nous proposons un indice permettant de mesurer l'homogénéité de la répartition géométrique des cellules dans un cube de données [MBR05]. Grâce à cet indice, nous pouvons évaluer le gain induit par l'arrangement des modalités des dimensions. Nous considérons que plus les cellules pleines (ou bien vides) sont concentrées, plus le cube est dit *"homogène"*.

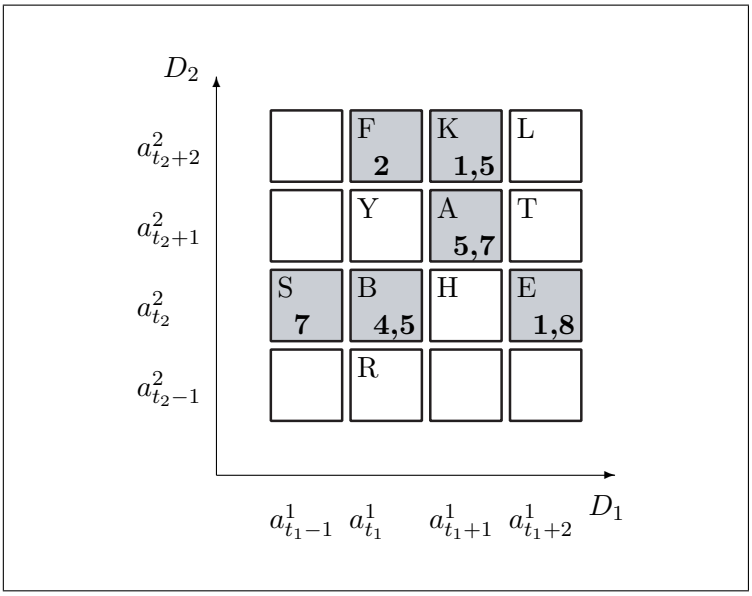


FIG. 3.6 – Exemple en 2 dimensions de la notion de voisinage des cellules d'un cube de données

3.7.1 Voisinage d'une cellule dans une représentation multidimensionnelle

Les modalités des dimensions constituent les coordonnées des cellules dans le cube. Soit $A = (a^1_{t_1}, \ldots, a^i_{t_i}, \ldots, a^d_{t_d})$ une cellule dans le cube $\mathcal{C}$, avec $i \in \{1, \ldots, d\}$ et $t_i \in \{1, \ldots, l_i\}$. t_i est l'indice de la modalité que prend la cellule A pour la dimension D_i.

Nous considérons que toutes les modalités des dimensions D_i sont géométriquement ordonnées dans l'espace de représentation des données selon l'ordre des indices t_i. C'est-à-dire, la modalité $a^i_{t_i-1}$ précède $a^i_{t_i}$, qui, à son tour, précède $a^i_{t_i+1}$ (voir l'exemple

de la figure 3.6). L'ordre des indices t_i correspond à l'ordre dans lequel sont arrangées dans l'espace les modalités de la dimension D_i. Nous définissons à présent la notion de voisinage pour les cellules d'un cube.

Définition 3.7.1 (Cellules voisines) *Soit* $A = (a^1_{t_1}, \ldots, a^i_{t_i}, \ldots, a^d_{t_d})$ *une cellule dans un cube* $\mathcal{C}$*. La cellule* $B = (b^1_{t_1}, \ldots, b^i_{t_i}, \ldots, b^d_{t_d})$ *est dite voisine de* A*, notée* $B \dashv A$*, si* $\forall i \in \{1, \ldots, d\}$*, les coordonnées de* B *vérifient :* $b^i_{t_i} = a^i_{t_i-1}$ *ou* $b^i_{t_i} = a^i_{t_i}$ *ou* $b^i_{t_i} = a^i_{t_i+1}$*. Dans le cas où* $\forall i \in \{1, \ldots, d\}$ $b^t_{t_i} = a^i_{t_i}$*,* B *n'est pas considérée comme une cellule voisine de* A *car* $B = A$*.*

Dans l'exemple de la figure 3.6, la cellule B est voisine de A ($B \dashv A$). Y est aussi voisine de A ($Y \dashv A$). En revanche, les cellules S et R ne sont pas voisines de A. Ceci nous ramène à définir le voisinage d'une cellule.

Définition 3.7.2 (Voisinage d'une cellule) *Soit* A *une cellule du cube* $\mathcal{C}$*, nous définissons le voisinage de* A*, noté* $\mathcal{V}(A)$*, par l'ensemble de toutes les cellules* B *de* $\mathcal{C}$ *qui sont voisines de* A*.*

$$\mathcal{V}(A) = \{B \in \mathcal{C} \text{ tel que } B \dashv A\}$$

Par exemple, dans la figure 3.6, le voisinage de la cellule A correspond à l'ensemble $\mathcal{V}(A) = \{F, K, L, Y, T, B, H, E\}$.

3.7.2 Similarité entre cellules dans une représentation multidimensionnelle

Définition 3.7.3 (Similarité de deux cellules) *Soient deux cellules* A *et* B *d'un cube de données* $\mathcal{C}$*. La similarité de* A *et* B*, notée* $\delta(A, B)$*, est un scalaire dans* $\mathbb{R}$ *défini comme suit :*

$$\delta(A, B) = \begin{cases} 1 - (\frac{||A| - |B||}{\max(\mathcal{C}) - \min(\mathcal{C})}) & \text{si } A \text{ et } B \text{ sont pleines ;} \\ 0 & \text{sinon.} \end{cases}$$

où $||A| - |B||$ *est la valeur absolue de la différence des mesures contenues dans* A *et* B*.* $\max(\mathcal{C})$ *(respectivement,* $\min(\mathcal{C})$*) est la valeur maximale (respectivement, la valeur minimale) de la mesure dans* $\mathcal{C}$*, avec* $\min(\mathcal{C}) \neq \max(\mathcal{C})$*.*

Dans le cube de la figure 3.6, où les cellules grises sont pleines et les cellules blanches sont vides, la mesure maximale du cube correspond à la cellule S ($\max(\mathcal{C}) = 7$) et la mesure minimale correspond à la cellule K ($\min(\mathcal{C}) = 1,5$). Par conséquent, la similarité des cellules A et B de la figure 3.6 est $\delta(A, B) = 1 - (\frac{|5,7-4,5|}{7-1,5}) \simeq 0,78$.

En revanche, la similarité des cellules A et Y est nulle vue que la cellule Y est vide. Il est à noter que notre définition de la similarité de deux cellules n'est pas applicable dans le cas où les cellules du cube $\mathcal{C}$ comportent la même valeur de la mesure. Ceci explique la condition $\min(\mathcal{C}) \neq \max(\mathcal{C})$.

Nous introduisons maintenant la notion de la similarité au voisinage Δ.

Définition 3.7.4 (Similarité au voisinage) *Soit une cellule A d'un cube de données $\mathcal{C}$. La similarité de A à son voisinage, notée $\Delta(A)$, est un scalaire dans $\mathbb{R}$ défini comme suit :*

$$\Delta(A) = \sum_{B \in \mathcal{V}(A)} \delta(A, B)$$

$\Delta(A)$ correspond à la somme des similarités de la cellule A avec toutes ses cellules voisines dans le cube de données. Par exemple, la similarité au voisinage de la cellule A de la figure 3.6 se calcule selon :

$$\begin{array}{rcl}
\delta(A,F) & = & 1 - (\frac{|5,7-2|}{7-1,5}) \simeq 0.33 \\
\delta(A,K) & = & 1 - (\frac{|5,7-1,5|}{7-1,5}) \simeq 0.24 \\
\delta(A,L) & = & 0 \\
\delta(A,T) & = & 0 \\
\delta(A,E) & = & 1 - (\frac{|5,7-1,8|}{7-1,5}) \simeq 0,29 \\
\delta(A,H) & = & 0 \\
\delta(A,B) & = & 1 - (\frac{|5,7-4,5|}{7-1,5}) \simeq 0,78 \\
\delta(A,Y) & = & 0 \\
\hline
\Delta(A) & \simeq & 1,64
\end{array}$$

3.7.3 Indice d'homogénéité d'une représentation multidimensionnelle

Définition 3.7.5 (Indice d'homogénéité brut) *Soit un cube de données $\mathcal{C}$. L'indice d'homogénéité brut du cube $\mathcal{C}$, noté $IHB(\mathcal{C})$, est défini comme suit :*

$$IHB(\mathcal{C}) = \sum_{\substack{A \in \mathcal{C} \\ |A| \neq NULL}} \sum_{B \in \mathcal{V}(A)} \delta(A, B) = \sum_{\substack{A \in \mathcal{C} \\ |A| \neq NULL}} \Delta(A)$$

L'indice d'homogénéité brut d'un cube est la somme des similarités de tous les couples de ses cellules qui sont à la fois pleines et voisines. Par exemple, l'indice d'homogénéité brut du cube de la figure 3.6 se calcule selon $IHB(\mathcal{C}) = \Delta(F) + \Delta(K) + \Delta(A) + \Delta(S) + \Delta(B) + \Delta(E) \simeq 6,67$.

Il est à noter que, par construction, cet indice est croissant en fonction de la qualité de la représentation d'un cube de données. En effet, plus les cellules d'un cube sont homogènes en terme de voisinage et de similarité, plus la valeur de l'indice d'homogénéité brut est grande. La représentation la plus homogène d'un cube de données $\mathcal{C}$ correspond au cas où ce dernier ne contient pas de cellules vides et que toutes les cellules ont des mesures égales. Dans ce cas, les similarités aux voisinages sont toutes égales à 1. Par conséquent, l'indice d'homogénéité brut atteint sa valeur maximale $IHB_{max}(\mathcal{C})$.

$$IHB_{max}(\mathcal{C}) = \sum_{A \in \mathcal{C}} \sum_{B \in \mathcal{V}(A)} 1$$

Dans l'algorithme 2, nous résumons les traitements nécessaires pour le calcul de l'indice d'homogénéité brut et l'indice d'homogénéité brut maximal d'un cube de données (lignes 1 à 16).

Définition 3.7.6 (Indice d'homogénéité) *Soit un cube de données $\mathcal{C}$. L'indice d'homogénéité du cube $\mathcal{C}$, noté $IH(\mathcal{C})$, est défini comme suit :*

$$IH(\mathcal{C}) = \frac{IHB(\mathcal{C})}{IHB_{max}(\mathcal{C})} = \frac{\sum\limits_{\substack{A \in \mathcal{C} \\ |A| \neq \mathsf{NULL}}} \Delta(A)}{\sum\limits_{A \in \mathcal{C}} \sum\limits_{B \in \mathcal{V}(A)} 1}$$

L'indice d'homogénéité d'un cube $\mathcal{C}$ représente le rapport de l'indice d'homogénéité brut de ce dernier par son indice d'homogénéité maximale (ligne 17 dans l'algorithme 2). Il mesure la qualité de représentation d'un cube de données. Cette qualité est d'autant meilleure quand les cellules pleines et ayant des mesures proches sont géométriquement voisines et rassemblées dans des régions particulières de l'espace de représentation du cube. Par exemple, sachant que l'indice d'homogénéité brut maximum de cube $\mathcal{C}$ de la figure 3.6 est $IHB_{max}(\mathcal{C}) = 84$, l'indice d'homogénéité est dans ce cas égal à : $IH(\mathcal{C}) = \frac{6,67}{84} \simeq 0,08$.

Cependant, cet indice est intrinsèquement lié à la configuration d'un cube de données. En d'autres termes, la valeur de l'indice est relativement liée à un cube de donnée et ne peut pas renseigner sur une qualité de représentation universelle pour toutes les structures multidimensionnelles des données. En revanche, avec cet indice, on peut mesurer l'apport d'une réorganisation de la même représentation d'un cube de données en évaluant le gain de la qualité induit par cette réorganisation. Nous introduisons dans la suite la notion du gain d'homogénéité.

3.7.4 Gain d'homogénéité

Pour mesurer l'apport de l'arrangement des modalités sur la représentation d'un cube de données $\mathcal{C}$, nous calculons le gain d'homogénéité, noté g, selon la formule :

$$g = \frac{IH(\mathcal{C}_{arr}) - IH(\mathcal{C}_{ini})}{IH(\mathcal{C}_{ini})} \tag{3.7.1}$$

où $IH(\mathcal{C}_{ini})$ est l'indice d'homogénéité de la représentation du cube initial et $IH(\mathcal{C}_{arr})$ est celui de la représentation réorganisée selon notre méthode. Notons que, pour le même type d'arrangement des modalités (selon les projections ou selon les valeurs-test), quelle que soit la représentation initiale du cube, nous obtenons toujours la même réorganisation par notre méthode. En effet, l'ACM est une méthode déterministe qui n'est pas sensible à l'ordre des variables en entrée.

Entrée cube de données $\mathcal{C}$
Sortie : indice d'homogénéité IH

$IHB \leftarrow 0$
$IHB_{max} \leftarrow 0$
pour tout cellule A dans $\mathcal{C}$ **faire**
  si $|A| \neq$ NULL **alors**
    pour tout cellule B dans $\mathcal{V}(A)$ **faire**
      si $|B| \neq$ NULL **alors**
        $IHB \leftarrow IHB + (1 - (\frac{||A|-|B||}{\max(\mathcal{C})-\min(\mathcal{C})}))$
      fin si
      $IHB_{max} \leftarrow IHB_{max} + 1$
    fin pour
  sinon
    pour tout cellule B dans $\mathcal{V}(A)$ **faire**
      $IHB_{max} \leftarrow IHB_{max} + 1$
    fin pour
  fin si
fin pour
$IH \leftarrow \frac{IHB}{IHB_{max}}$

Algorithme 2: Calcul de l'indice d'homogénéité d'un cube de données

3.8 Études de cas

Pour tester et valider nos deux propositions de réorganisation des cubes de données, nous exposons dans cette section deux études de cas. La première concerne la réorganisation d'un cube de données bancaires en arrangeant ses modalités selon leurs projections sur les axes factoriels [MAF05]. La deuxième étude est dédiée à un cube de données démographiques. Ce dernier fait l'objet d'une réorganisation selon les valeurs-test de ses modalités [MBR06d, MBR06b].

3.8.1 Cas de l'arrangement des modalités selon leurs projections

Nous utilisons un jeu de données bancaires extrait du système d'information du *"Le Crédit Lyonnais" (LCL)*. À partir de ces données, nous avons construit un cube de données correspondant à un contexte d'analyse. Un fait du cube correspond au comportement d'achat d'un client. Nous disposons dans ce cube de $n = 311\,959$ faits mesurés par le *produit net bancaire* (M_1) et le *montant des avoirs* (M_2). Le tableau 3.1 détaille la description des dimensions considérées pour observer ces mesures.

Dimension	l_i	Description
D_1 : catégorie socio-professionnelle	$l_1 = 58$	profil professionnel du client
D_2 : produit	$l_2 = 25$	détention de formule(s) qui sont des offres combinées de produits bancaires
D_3 : unité commerciale	$l_3 = 65$	localisations géographiques de vente
D_4 : segment	$l_4 = 15$	potentiel commercial du client
D_5 : âge	$l_5 = 12$	variable discrétisée selon des tranches d'âge de dix ans ([0-10], [11-20], [21-30], etc.)
D_6 : situation familiale	$l_6 = 6$	exemple : marié, divorcé, etc.
D_7 : type client	$l_7 = 4$	origine du client (par exemple, client membre du personnel du Crédit Lyonnais)
D_8 : marché	$l_8 = 4$	une vente réalisée auprès d'un client est faite sur le marché "particulier des professionnels" si le client est artisan ou exerce une profession libérale, etc., ou sur le marché "particulier" sinon

TAB. 3.1 – Description des dimensions du cube des données bancaires

Pour rendre plus claire la suite de notre exposé, notre étude de cas porte sur un cube à deux dimensions ($d = 2$) : la dimension "*catégorie socio-professionnelle*" (D_1) et la dimension "*produit*" (D_2). La mesure observée est "*le montant des avoirs*". Nous générons les matrices Z_1 et Z_2 selon un codage binaire disjonctif des modalités des deux dimensions. Le tableau disjonctif complet $Z = [Z_1, Z_2]$ est à $n = 311\,959$ lignes et $l = l_1 + l_2 = 83$ colonnes.

En appliquant l'ACM sur le tableau Z, on obtient $l - d = 81$ axes factoriels F_α. La figure 3.7 montre le premier plan factoriel obtenu à partir des faits du cube des données bancaires. Chaque axe est caractérisé par sa valeur propre λ_α et les contributions apportées par les dimensions : $Cr_\alpha(D_1)$ et $Cr_\alpha(D_2)$. Nous cherchons, pour chaque dimension, l'axe qui est le mieux contribué par cette dernière. Nous obtenons les résultats suivants :

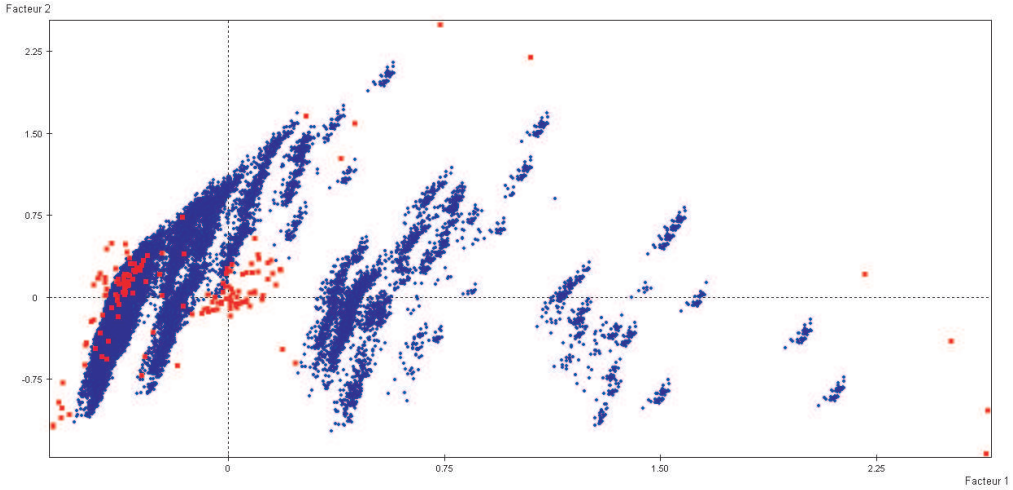


FIG. 3.7 – Premier plan factoriel construit par l'ACM à partir du cube des données bancaires

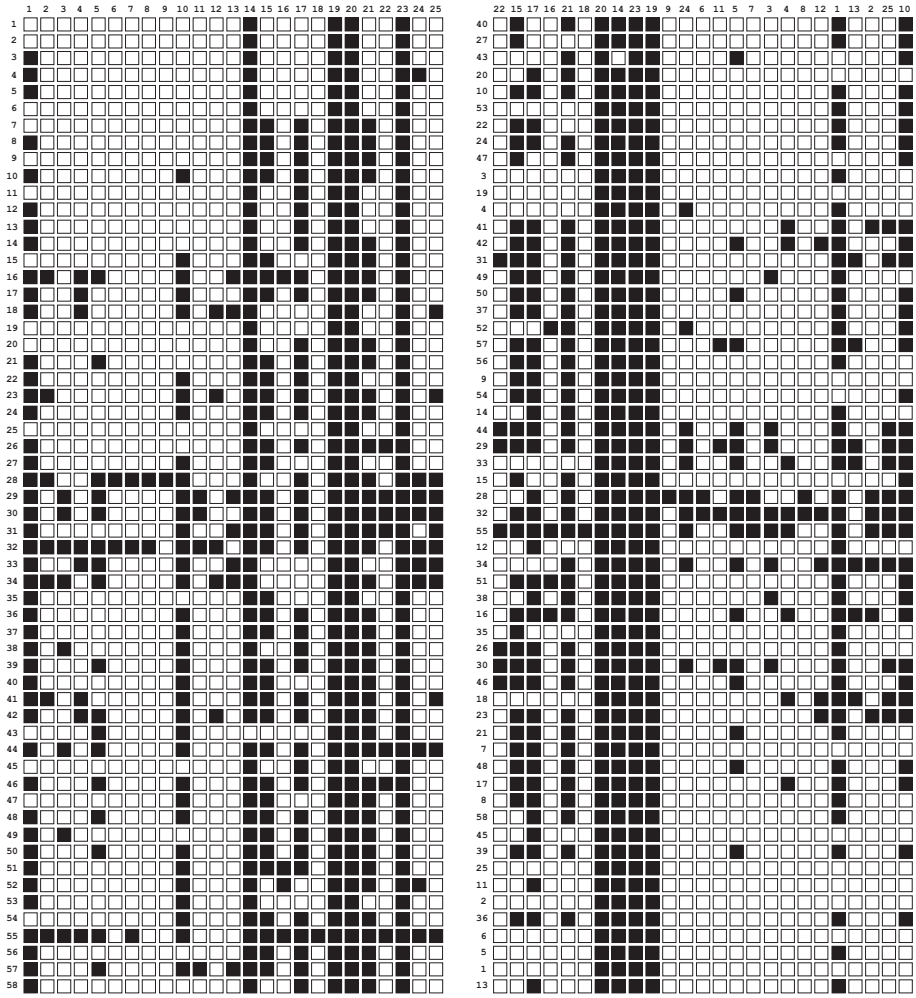
Pour la dimension D_1 : $\lambda_{45}Cr_{45}(D_1) = \max_{\alpha\in\{1,\ldots,81\}}(\lambda_\alpha Cr_\alpha(D_1))$ avec $\lambda_{45} = 0.5$ et $Cr_{45}(D_1) = 99.9\%$.

Pour la dimension D_2 : $\lambda_1 Cr_1(D_2) = \max_{\alpha\in\{1,\ldots,81\}}(\lambda_\alpha Cr_\alpha(D_2))$ avec $\lambda_1 = 0.83$ et $Cr_1(D_2) = 50\%$.

Ainsi, la dimension D_1 est associée à l'axe F_{45} et D_2 à l'axe F_1. Les modalités de D_1 (respectivement, D_2) sont arrangées suivant l'ordre croissant de leurs projections sur F_{45} (respectivement, F_1). Dans la figure 3.8, nous présentons le résultat de cet arrangement. La représentation (a) correspond à l'arrangement initial du cube selon l'ordre alphabétique des libellés des modalités. La représentation (b) correspond à l'arrangement obtenu par l'ordre croissant des projections des modalités sur les axes factoriels suscités. Pour des raisons de confidentialité des données du *LCL*, nous masquons les libellés des modalités de chaque dimension ainsi que les valeurs des mesures. Nous remplaçons les libellés par des codes chiffrés et les mesures existantes par des cases noires. Les cases blanches du cube représentent les creux correspondant à des cellules vides. Le taux d'éparsité étant égal au rapport entre le nombre de cases vides et le nombre total des cases du cube, sur cet exemple, le taux d'éparsité du cube est égal à 64%. La valeur de l'indice d'homogénéité est de $17,75\%$ pour la représentation (a) et de $20,60\%$ pour la représentation (b). Nous obtenons donc un gain d'homogénéité de $16,38\%$ par rapport à la représentation initiale du cube.

Nous avons également appliqué notre méthode sur un cube à trois dimensions : "*catégorie socio-professionnelle*" (D_1), "*produit*" (D_2) et "*âge*" (D_5). Ce cube, dont

le taux d'éparsité est égal à $87,94\%$, contient plus de cellules vides comparé au cube précédent. L'arrangement des modalités correspond à l'ordre alphabétique pour D_1 et D_2, et à l'ordre croissant des tranches d'âge pour D_5. Le cube initial a un indice d'homogénéité de $IH(\mathcal{C}_{ini}) = 5,12\%$. Le cube arrangé a un indice d'homogénéité de $IH(\mathcal{C}_{arr}) = 6,11\%$. Nous obtenons ainsi un gain de $g = 19,33\%$.



(a) cube initial **(b)** cube réorganisé

FIG. 3.8 – Représentations du cube des données bancaires (a) avant et (b) après arrangement des modalités

3.8.2 Cas de l'arrangement des modalités selon leurs valeurs-test

Pour notre deuxième méthode d'arrangement selon les valeurs-test des modalités, nous étudions le cas de données démographiques. Nous avons construit un cube à 5 dimensions ($d = 5$) dont les données sont extraites à partir de la base *Census-Income*

Database[1] concernant un *recensement sur les revenus* de la population des États-Unis d'Amérique entre 1994 et 1995. Le cube étudié contient $n = 199\,523$ faits OLAP où chaque fait représente un profil d'une sous-population d'employés mesuré par le *salaire par heure* (M_1). Le tableau 3.2 détaille la description des cinq dimensions prises en compte pour observer ces faits.

Dimension	l_i
D_1 : niveau d'éducation	$l_1 = 17$
D_2 : catégorie socio-professionnelle	$l_2 = 22$
D_3 : état de résidence	$l_3 = 51$
D_4 : situation du ménage	$l_4 = 38$
D_5 : pays de naissance	$l_5 = 42$

TAB. 3.2 – Description des dimensions du cube des données démographiques

Selon un codage binaire disjonctif des modalités de chaque dimension du cube, nous générons le tableau disjonctif complet $Z = [Z_1, Z_2, Z_3, Z_4, Z_5]$. Z contient 199523 lignes et $l = \sum_{i=1}^{5} l_i = 170$ colonnes. En appliquant l'ACM sur Z, on obtient $l - d = 165$ axes factoriels F_α. Chaque axe est associé à une valeur propre λ_α. Supposons que, selon l'histogramme des valeurs propres, l'utilisateur retient les trois premiers axes factoriels ($s = 3$). Ces trois premiers axes, expliquent 15.35% de l'inertie totale du nuage des faits du cube étudié. Cette contribution à l'inertie totale peut sembler insignifiante dans le cas absolu. Cependant, en prenant en compte le nombre d'axes construits par l'ACM, cette contribution devient relativement importante. En effet, dans le cas d'une distribution uniforme des variables à l'inertie totale sur tous les axes factoriels, chaque axe devrait avoir une contribution seulement égale à $\frac{1}{l-d} = 0.6\%$. En d'autres termes, dans notre cas d'application, les trois premiers axes factoriels sont 25 fois plus importants que le cas d'une distribution uniforme des variables. La figure 3.9 montre le premier plan factoriel obtenu à partir des faits du cube des données démographiques.

Le cube réorganisé est obtenu en triant les modalités de chacune de ses dimensions. Pour chaque dimension D_i, ses modalités sont triées selon l'ordre croissant de leurs valeurs-test V^i_{1t}, puis selon les valeurs-test V^i_{2t} et enfin selon V^i_{3t}. Par exemple, la table 3.3 montre le nouvel ordre des modalités de la dimension "*catégorie socio-professionnelle*" (D_2). Notons que, d'après ce tableau, t est l'indice de l'ordre alphabétique des noms des modalités initialement établi.

Les figures 3.10 et 3.11 montre l'effet visuel que produit l'arrangement des modalités sur la représentation d'une vue partielle du cube des données démographiques. Cette vue résulte du croisement de la dimension "*catégorie socio-professionnelle*" (D_2) en colonnes avec la dimension "*pays de naissance*" (D_5) en lignes. La figure 3.10

[1] `http://kdd.ics.uci.edu/databases/census-income/census-income.html`

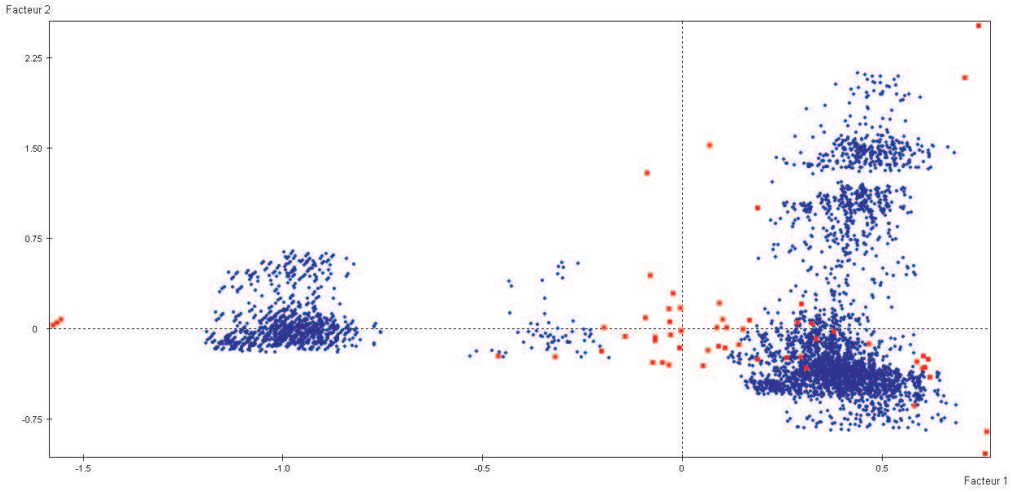


FIG. 3.9 – Premier plan factoriel construit par l'ACM à partir du cube des données démographiques

montre la représentation initiale de ce croisement qui respecte l'ordre alphabétique des noms des modalités dans chaque dimension. La figure 3.11 montre la représentation de ce croisement suite à l'arrangement des modalités de chaque dimension selon l'ordre de leurs valeurs-test.

Rappelons que le but de notre approche factorielle n'est pas de compresser ou de réduire la dimensionnalité des cubes de données. Nous ne cherchons pas non plus à réduire l'éparsité des données dans un cube. Cependant, nous réduisons plutôt l'effet négatif de cette éparsité sur la qualité de la représentation visuelle des cubes de données. Pour cela, nous arrangeons différemment les modalités dans chaque dimension du cube de sorte à mieux organiser la répartition des faits dans l'espace de représentation. L'objectif de notre approche se résume par le fait de regrouper les cellules pleines et de les séparer le mieux possible des cellules vides dans l'espace de représentation d'un cube de données.

D'une manière intuitive, la représentation de la figure 3.11 est visuellement mieux adaptée à l'analyse que celle de la figure 3.10. En effet, dans la figure 3.11, nous pouvons clairement remarquer l'existence de régions denses caractérisées par une forte concentration de cellules pleines au dépend d'autres régions dans cet espace de représentation qui sont quasiment vides. Ainsi, l'utilisateur peut focaliser son exploration du cube sur ces régions denses qui sont en réalité plus intéressantes vu qu'elles regroupent plus de faits OLAP. Ces régions sont, par conséquent, caractérisées par des taux d'homogénéité supérieurs aux autres régions du cube.

		Valeurs-test		
t	Modalité	$V_{1t}i$	$V_{2t}i$	$V_{3t}i$
9	*Hospital services*	-99.90	-99.90	-99.90
14	*Other professional services*	-99.90	-99.90	99.90
17	*Public administration*	-99.90	-99.90	99.90
12	*Medical except hospital*	-99.90	99.90	-99.90
5	*Education*	-99.90	99.90	99.90
7	*Finance insurance*	-99.90	99.90	99.90
19	*Social services*	-99.90	99.90	99.90
8	*Forestry and fisheries*	-35.43	-8.11	83.57
3	*Communications*	-34.05	-99.90	99.90
15	*Personal services except private*	-21.92	-5.50	10.28
13	*Mining*	-6.59	-99.64	-5.25
16	*Private household services*	7.77	51.45	11.68
6	*Entertainment*	40.04	99.90	96.23
1	*Agriculture*	68.66	3.39	-27.38
4	*Construction*	99.90	-99.90	-99.90
10	*Manufact. durable goods*	99.90	-99.90	-99.90
11	*Manufact. nondurable goods*	99.90	-99.90	-99.90
21	*Utilities and sanitary services*	99.90	-99.90	-99.90
22	*Wholesale trade*	99.90	-99.90	-24.37
20	*Transportation*	99.90	-99.90	99.90
18	*Retail trade*	99.90	99.90	-99.90
2	*Business and repair*	99.90	99.90	99.90

TAB. 3.3 – Nouvel ordre des modalités de la dimension D_2 du cubes des données démographiques

Ce constat est confirmé par la mesure de la qualité de représentation par l'indice d'homogénéité. En effet, pour un taux d'éparsité de 63,42% dans le cube des données démographiques, l'indice d'homogénéité est de $IH(\mathcal{C}_{ini}) = 14,22\%$ pour la représentation initiale de la figure 3.10. Pour la représentation organisée de la figure 3.11, la valeur de l'indice d'homogénéité est de $IH(\mathcal{C}_{arr}) = 17,67\%$. Notre méthode d'arrangement selon les valeurs-test des modalités permet donc de réaliser un gain d'homogénéité égal à $g = 24,26\%$.

3.9 Expérimentations et performances

Nous avons réalisé une série d'expérimentations de notre approche sur le cube des données démographiques. Afin de mesurer l'impact de l'éparsité des données, nous avons tiré plusieurs échantillons aléatoires à partir de la population des $n = 199\,523$ faits du cube initial. Ainsi, en variant le taux d'échantillonnage, nous parvenons à faire varier l'éparsité du cube. Nous avons mené des expériences sur les deux types de réorganisation que nous avons proposés.

La figure 3.12 (a) montre l'évolution de l'indice d'homogénéité du cube initial et du cube arrangé selon les projections des modalités. La figure 3.13 (a) montre l'indice

	Agriculture	Business and repair services	Communications	Construction	Education	Entertainment	Finance insurance	Forestry and fisheries	Hospital services	Manufact. durable goods	Manufact. nondurable goods	Medical except hospital	Mining	Other professional services	Personal services except private	Private household services	Public administration	Retail trade	Social services	Transportation	Utilities and sanitary services	Wholesale trade
Cambodia										125.0										750.0		
Canada		35.0		93.1	54.1			112.5	253.1	182.3		373.4		22.2			169.2	94.0		267.6	11.1	350.0
China	622.0			40.7	50.1		105.0	566.7	336.8	46.7	64.2	60.7					833.8	21.6			329.0	206.3
Columbia							79.0			46.6		80.3					175.0					
Cuba			501.5						31.8			19.0						28.9				
Dominican-Republic			375.0		116.7				146.0	92.7	38.1							35.1	75.0			
Ecuador	107.2	109.1	250.0	205.6	515.0		206.7	68.8			128.1	265.6	100.0				300.0	41.9	175.0		333.3	212.5
El-Salvador	55.6	46.1		36.1	81.0	950.8	344.0			184.7	19.4	120.0			79.5			20.7	400.0	36.9		365.6
England		77.9	222.7	418.1	90.2	50.0	46.9		383.0	257.1	365.0			194.7			136.4	26.3	198.9			
France	450.0										394.8							229.0				
Germany		115.0	200.0	157.1			97.9		417.2	152.3	31.7	128.6		22.2		218.9	108.7	77.9		253.1	428.2	
Greece					257.1					300.0	150.0			241.7			400.0	52.4	400.0			63.6
Guatemala				121.8						47.5	39.8				136.2	25.8						
Haiti									90.0		80.6				178.7							###
Holand-Netherlands											21.4											
Honduras															151.7		945.0					
Hong Kong	125.4			190.5			590.4	183.3			100.0			225.0	###			150.0		566.7	55.1	484.3
Hungary																400.0						
India		94.2			101.2		17.9	228.1	157.2	145.9								100.0			167.1	81.3
Iran		95.8		225.0			66.7			160.7				311.1	100.0		316.7	90.0		159.0		
Ireland			500.0	100.0							533.3											
Italy					80.3										27.8			32.9				
Jamaica	250.0	158.8	###	100.0	147.0		79.2		343.1		571.4	106.0		55.6	91.7	100.0	803.8		604.7	533.3	19.4	
Japan		107.1			63.5	425.0		192.1	678.9	50.9	164.6							26.4	150.0	273.3	107.5	
Laos						500.0				116.6				350.0				71.4				
Mexico	34.5	89.6	75.0	95.0	155.2	46.5	67.6		122.2	61.9	59.8	89.7		159.1	59.9	17.1		52.9	40.3	140.3	121.7	82.1
Nicaragua	159.5		83.3		140.0			47.6		340.0	76.5	65.6		74.1		160.0	178.3	81.0			85.7	
Outlying-U S								###										93.8				200.0
Panama														452.5								
Peru	225.0	699.6	69.7		106.3		47.0	450.0	166.7	215.4	76.2			134.5				127.3	124.2	86.4	20.0	32.0
Philippines	200.0	122.7	265.0	270.0	317.8	62.5	165.0		331.1	66.7	166.1	95.6					77.8	134.7		197.3		322.7
Poland		252.9	175.6				105.0		325.0	185.5	92.6	175.2					180.0	196.2		187.5		212.5
Portugal				166.7	155.6		107.1		141.1													236.7
Puerto-Rico		87.8	250.0	54.2		66.7	80.7	250.0	37.5	122.3	48.3	420.7		40.0			110.1	23.9	43.5	163.8	142.9	33.6
Scotland				87.5		725.0	300.0		785.0	95.2	14.0		23.9					131.3	350.0	173.6	700.0	36.5
South Korea																					870.0	
Taiwan																				46.2		
Thailand												150.0						43.8				
Trinadad&Tobago	66.3	243.8		63.8		920.0	333.3	89.3		466.7		175.0						453.0	200.0			250.0
United-States	37.8	92.6	153.4	130.6	75.4	117.9	71.1	84.3	214.4	165.4	146.9	141.7		76.0			142.1	99.3	96.0	157.0	199.9	84.4
Vietnam			###				75.0		327.5	173.8				250.0	32.1							
Yugoslavia		42.1																				

FIG. 3.10 – Représentation du cube des données démographiques avant l'arrangement des modalités

d'homogénéité du cube initial et du cube arrangé selon les valeurs-test des modalités. D'une manière générale, nous remarquons que les valeurs de l'indice sont décroissantes en fonction de l'éparsité du cube. Ceci est naturellement dû à la construction de cet indice dont la valeur dépend fortement, d'une manière croissante, du nombre de cellules pleines dans le cube.

Cependant, notons que pour des taux d'éparsité élevés, supérieurs à 60%, le cube obtenu selon nos deux types d'arrangement est toujours de meilleure qualité que le cube initial au sens de notre indice d'homogénéité. Dans les cas d'une forte éparsité, nous réalisons toujours un gain d'homogénéité lors de l'arrangement du cube.

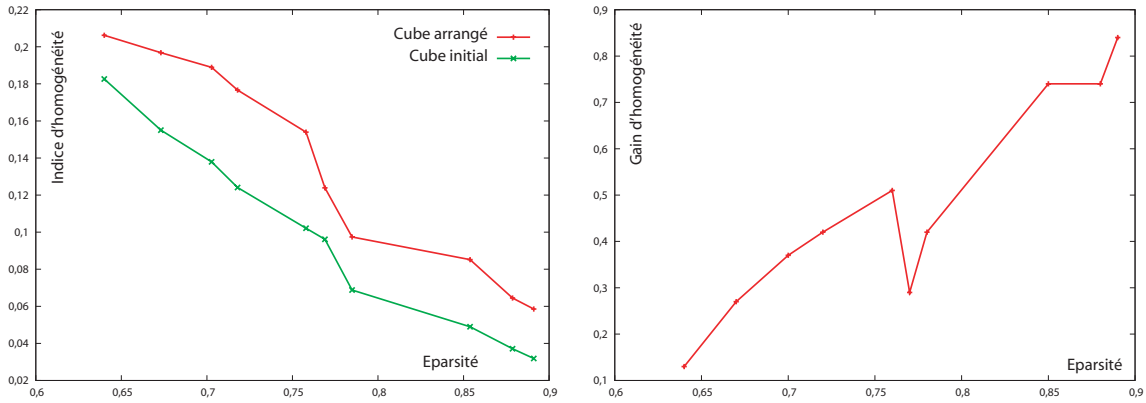
D'après la figure 3.12 (b) et 3.13 (b), pour des éparsités supérieures à 60%, le gain en homogénéité a une tendance générale croissante en fonction de l'éparsité du cube.

	Hospital services	Other professional services	Public administration	Medical except hospital	Education	Finance insurance	Social services	Forestry and fisheries	Communications	Personal services except private	Mining	Private household services	Entertainment	Agriculture	Construction	Manufact. durable goods	Manufact. nondurable goods	Utilities and sanitary services	Wholesale trade	Transportation	Retail trade	Business and repair services
Philippines	331.1		77.8	95.6	317.8	165.0			265.0				62.5	200.0	270.0	66.7	166.1		322.7	197.3	134.7	122.7
India	157.2				101.2	17.9		228.1								145.9		167.1	81.3		100.0	94.2
Canada	253.1	22.2	169.2	373.4	54.1			112.5							93.1	182.3		11.1	350.0	267.6	94.0	35.0
Jamaica	343.1	55.6	803.8	106.0	147.0	79.2	604.7		###	91.7		100.0		250.0	100.0		571.4	19.4		533.3		158.8
Iran		311.1	316.7			66.7				100.0					225.0	160.7				159.0	90.0	95.8
Japan	678.9				63.5		150.0	192.1					425.0			50.9	164.6	107.5		273.3	26.4	107.1
China	336.8		833.8	60.7	50.1	105.0		566.7						622.0	40.7	46.7	64.2	329.0	206.3		21.6	
Hong Kong		225.0				590.4		183.3		###				125.4	190.5		100.0	55.1	484.3	566.7	150.0	
Greece		241.7	400.0		257.1		400.0									300.0	150.0		63.6		52.4	
Germany	417.2	22.2	108.7	128.6		97.9			200.0			218.9			157.1	152.3	31.7	428.2		253.1	77.9	115.0
Scotland	785.0					300.0	350.0				23.9		725.0		87.5	95.2	14.0	700.0	36.5	173.6	131.3	
Poland	325.0		180.0	175.2		105.0			175.6							185.5	92.6		212.5	187.5	196.2	252.9
England	383.0	194.7	136.4		90.2	46.9	198.9		222.7				50.0		418.1	257.1	365.0				26.3	77.9
Haiti	90.0									178.7							80.6		###			
Taiwan																				46.2		
Panama		452.5																				
Outlying-U S								###											200.0		93.8	
Thailand				150.0																	43.8	
Italy					80.3					27.8											32.9	
Hungary												400.0										
Vietnam	327.5	250.0				75.0			###	32.1						173.8						
Holand-Netherlands																	21.4					
Portugal	141.1				155.6	107.1									166.7				236.7			
Yugoslavia																						42.1
South Korea																		870.0				
Honduras			945.0							151.7												
Cuba	31.8			19.0					501.5												28.9	
France														450.0			394.8				229.0	
Cambodia																125.0				750.0		
Dominican-Republic	146.0				116.7		75.0		375.0							92.7	38.1				35.1	
Laos		350.0											500.0			116.6					71.4	
Guatemala										136.2		25.8			121.8	47.5	39.8					
Columbia			175.0	80.3		79.0										46.6						
Ireland									500.0						100.0		533.3					
Trinadad&Tobago				175.0		333.3	200.0	89.3					920.0	66.3	63.8	466.7			250.0		453.0	243.8
Puerto-Rico	37.5	40.0	110.1	420.7		80.7	43.5	250.0	250.0				66.7		54.2	122.3	48.3	142.9	33.6	163.8	23.9	87.8
Ecuador			300.0	265.6	515.0	206.7	175.0	68.8	250.0		100.0			107.2	205.6		128.1	333.3	212.5		41.9	109.1
Peru	166.7	134.5			106.3	47.0	124.2	450.0	69.7					225.0		215.4	76.2	20.0	32.0	86.4	127.3	699.6
Nicaragua		74.1	178.3	65.6	140.0			47.6	83.3			160.0		159.5		340.0	76.5	85.7			81.0	
Mexico	122.2	159.1		89.7	155.2	67.6	40.3		75.0	59.9		17.1	46.5	34.5	95.0	61.9	59.8	121.7	82.1	140.3	52.9	89.6
El-Salvador				120.0	81.0	344.0	400.0			79.5			950.8	55.6	36.1	184.7	19.4		365.6	36.9	20.7	46.1
United-States	214.4	76.0	142.1	141.7	75.4	71.1	96.0	84.3	153.4				117.9	37.8	130.6	165.4	146.9	199.9	84.4	157.0	99.3	92.6

FIG. 3.11 – Représentation du cube des données démographiques après l'arrangement des modalités

En effet, plus le cube est éparse, plus nous avons une meilleure marge de manœuvre pour concentrer les données et les regrouper ensemble autour des axes factoriels de l'ACM. Notons aussi que le gain en homogénéité, pour les fortes éparsités, peut fléchir localement. Ceci est inhérent à la structure des données. C'est-à-dire, si les données du cube initial sont déjà dans une représentation homogène, l'application de notre méthode n'apportera pas de gain considérable. En effet, dans ce cas, la méthode n'aura qu'un effet de translation du nuage des faits vers les zones centrales des axes factoriels.

D'après la figure 3.13 (b), nous remarquons que, pour de faibles valeurs de l'éparsité, le gain d'homogénéité est oscillant autour de la valeur nulle. En effet, quand l'éparsité est inférieure à 60%, le gain n'a pas de sens de variation constant.



(a) **(b)**

FIG. 3.12 – Évolution de (a) l'indice d'homogénéité et (b) du gain d'homogénéité en fonction de l'éparsité (arrangement selon les projections des modalités)

Le gain en homogénéité enregistre même des valeurs négatives. Ceci s'explique par le fait général que notre approche n'apporte pas de valeur ajoutée à la qualité de représentation des cubes de données denses. En effet, l'utilisation de notre méthode sur des données peu éparses n'a pas toujours une efficacité significative. Ceci s'explique par la nature de l'indice d'homogénéité qui privilégie le nombre de cellules pleines et qui est, par conséquent, fortement dépendant de l'éparsité des données étudiées. Ceci s'explique aussi par la structure des données denses qui, par définition, correspondent à de bonnes propriétés de représentation au sens de l'homogénéité géométrique et visuelle que nous proposons via notre indice. En d'autres termes, les cubes denses ont déjà une bonne répartition de leurs données dans leurs espaces de représentation multidimensionnelle. En résumé, notre approche de réorganisation des faits dans l'espace de représentation des cubes de données est plutôt efficace lorsque ces derniers sont volumineux et éparses.

3.10 Conclusion et perspectives

Dans ce chapitre, nous avons proposé une approche factorielle apportant une solution au problème de la visualisation des données dans un cube éparse. Sans réduire l'éparsité, nous cherchons à réorganiser l'espace multidimensionnel des données en regroupant géométriquement les cellules pleines dans un cube. La recherche d'un arrangement optimal du cube est un problème complexe et coûteux en temps de calcul. Nous avons choisi d'utiliser les résultats de l'ACM comme heuristique pour réduire cette complexité. Notre approche consiste à arranger les modalités des dimensions

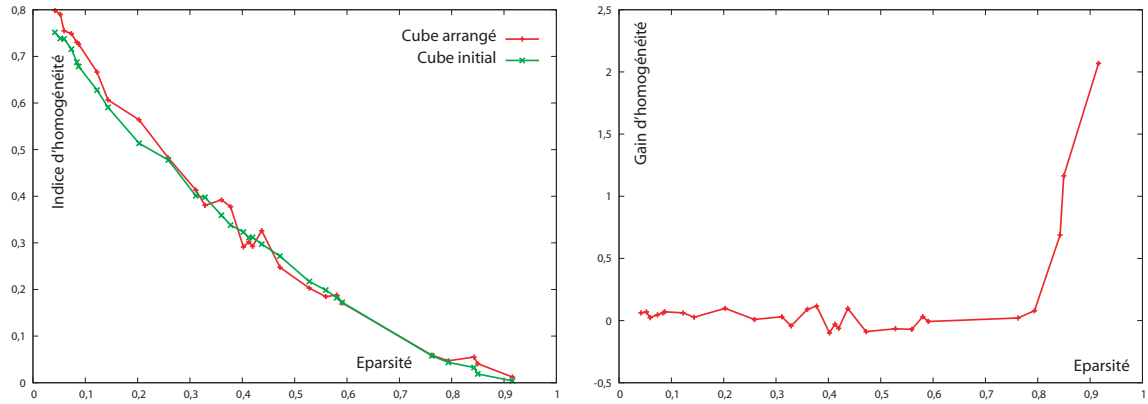


FIG. 3.13 – Évolution de (a) l'indice d'homogénéité et (b) du gain d'homogénéité en fonction de l'éparsité (arrangement selon les valeurs-test des modalités)

d'un cube en tenant compte des contributions des modalités dans la construction des axes factoriels de l'ACM. Nous proposons deux types d'arrangement des modalités du cube étudié : un arrangement selon leur projections sur les axes factoriels et un arrangement selon leurs valeurs-test. Ces deux types arrangements établissent une réorganisation des faits dans l'espace de représentation du cube de données.

Pour évaluer l'apport de cette nouvelle représentation, nous avons proposé un indice d'homogénéité basé sur le voisinage. La comparaison des valeurs de l'indice entre les représentations initiale et arrangée du cube nous permet d'évaluer l'efficacité de notre approche. Les différents tests sur un jeu de données démographiques nous ont montré que notre approche est pertinente pour les cubes de données éparses. En effet, pour des données d'une éparsité supérieure à 60%, le gain d'homogénéité est globalement croissant en fonction de l'éparsité et son amplitude est également inhérente à la structure des données. Cependant, pour des données plutôt moins éparses, notre approche ne fournit pas de résultats intéressants vu que les données denses possèdent déjà une bonne propriété de représentation au sens de notre indice d'homogénéité.

Suite à ce travail, plusieurs perspectives sont à prévoir. Tout d'abord, nous devons étudier la complexité de notre méthode. Cette étude doit prendre en compte aussi bien les propriétés du cube (taille, éparsité, cardinalités, etc.) que l'impact de l'évolution des données (rafraîchissement de l'entrepôt de données).

Ensuite, à ce stade de nos travaux, pour appliquer l'ACM, nous tenons seulement compte de la présence ou de l'absence des faits du cube dans la construction des axes factoriels. Nous envisageons alors d'introduire la valeur de la mesure comme

pondération des faits (poids des individus de l'ACM). Ceci permettra de construire des axes factoriels qui traduisent mieux la représentation des faits du cube selon leur ordre de grandeur. Dans ce cas, il serait également intéressant d'introduire la notion de distance entre cellules voisines en fonction des valeurs de la mesures qu'elles contiennent.

Dans le même ordre d'idées, nous souhaitons utiliser les résultats de l'ACM afin de faire émerger des régions intéressantes pour l'analyse à partir d'un cube de données initial. En effet, l'ACM permet de concentrer dans les zones centrales des axes factoriels les individus ayant un comportement moyen, et d'éloigner ceux ayant des comportements atypiques vers les zones extrêmes. Nous pouvons déjà exploiter les résultats de l'arrangement des modalités du cube dans le cadre de la distinction de régions correspondant à ces comportements caractéristiques.

Nous voulons aussi comparer la visualisation obtenue par notre approche avec celle proposée dans [CR98]. Cette dernière représente les résultats d'une analyse factorielle sous forme d'un *diagramme de Bertin* [Ber81, Ber99] adapté à la visualisation et à l'interprétation. L'objectif de cette méthode est de proposer une visualisation optimisée d'un tableau de contingence. Cependant, elle se limite à des tableaux à deux dimensions sans données manquantes et ne peut pas s'appliquer à des cubes de forte dimensionnalité. Notre approche peut être considérée comme une extension de cette méthode concernant la dimensionnalité du cube et l'éparsité de ses données.

Par ailleurs, la matérialisation des cubes de données permet le pré-calcul et le stockage des agrégats multidimensionnels de manière à rendre l'analyse OLAP plus performante. Cela requiert un temps de calcul important et génère un volume de données élevé lorsque le cube matérialisé est à forte dimensionnalité. Au lieu de calculer la totalité du cube, il serait judicieux de calculer et de matérialiser que les parties intéressantes du cube (fragments contenant l'information utile). Comme l'information réside dans les cellules pleines, le cube arrangé obtenu par l'application de l'ACM serait un point de départ pour déterminer ces fragments. Ainsi, comme dans [BS97], chaque fragment donnera lieu à un cube local. Les liens entre ces cubes permettront de reconstruire le cube initial.

Agrégation par classification dans les cubes de données

Résumé

Nous présentons dans ce chapitre notre deuxième proposition dédiée à la structuration et la classification des données multidimensionnelles. Nous adoptons l'approche du couplage entre l'analyse en ligne et la fouille de données qui exploite les outils OLAP afin d'extraire les données nécessaires à la construction de l'algorithme de fouille.

Notre présente proposition fait l'objet d'une nouvelle agrégation des faits d'un cube en se basant sur la classification ascendante hiérarchique (CAH). Celle-ci permet d'obtenir de nouveaux agrégats sémantiquement plus riches que ceux fournis par les opérateurs OLAP classiques.

Sommaire

Publications

[MBR04] Messaoud R.B., Boussaid O., Rabaséda S., « A New OLAP Aggregation Based on the AHC Technique », in *Proceedings of the 7^{th} ACM International Workshop on Data Warehousing and OLAP (DOLAP'2004)*, pp. 65–72, Washington D.C., VA, USA : ACM Press. November 2004.

[MBR06a] Messaoud R.B., Boussaid O., Rabaséda S.L., « A Data Mining-Based OLAP Aggregation of Complex Data : Application on XML Documents », *International Journal of Data Warehousing and Mining*, 2(4) :1–26. 2006.

[MRBB04] Messaoud R.B., Rabaséda S., Boussaid O., Bentayeb F., « OpAC : Opérateur d'analyse en ligne basé sur une technique de fouille de données », in $4^{èmes}$ *Journées francophones d'Extraction et de Gestion des Connaissances (EGC'2004)*, volume 2 de *Revue des Nouvelles Technologies de l'Information*, pp. 35–46, Clermont-Ferrand, France. Janvier 2004.