Modélisation et extraction des liens complexes entre variables. Application à des données socio-économiques.

Martine Cadot*, Dhouha El Haj Ali**

* Université Henri Poincaré / LORIA, Nancy, France
martine.cadot@loria.fr
http://www.loria.fr/~cadot
** Université Manar I, Faculté des sciences économiques et de gestion de Tunis, Tunisie
elhajali.dhouha@yahoo.fr

Résumé. Nous nous intéressons ici à un type particulier de complexité qui est celle des liaisons entre variables. Il existe des modèles statistiques qui ont été construits pour traiter certains aspects de cette complexité. Ainsi le *modèle linéaire général* (Azaïs et Bardet 2005) permet de rendre compte d'aspects spécifiques de la complexité comme les interactions d'ordre quelconque, les liaisons négatives au même titre que les positives, et les « contrastes ». Mais ces méthodes sont mal adaptées au cas d'un grand nombre de variables et elles exigent une explicitation a priori des liaisons en jeu. Nous présentons notre méthode MIDOVA qui extrait directement des données le même type de liaisons que le modèle linéaire général, sans nécessiter d'hypothèses contraignantes, tout en étant compatible avec un grand nombre de variables, pour l'instant qualitatives. Nous l'illustrons en l'appliquant à des données issues de l'enquête PAPFEM, réalisée en 2001 par l'Office National de la Famille et de la Population en Tunisie, et nous mettons au jour le lien particulièrement complexe entre la pauvreté du ménage et la situation socio-économique des deux conjoints.

1 Introduction

Notre but est d'extraire des données ce que nous appelons les *liaisons complexes* entre variables, par opposition aux liaisons simples, c'est-à-dire entre les variables prises deux par deux . Les données que nous considérons se présentent sous la forme de tableaux individus X variables, c'est-à-dire contenant pour chaque individu sa valeur pour chaque variable. Certains modèles statistiques permettent une représentation des liaisons complexes entre variables. Nous les décrivons dans la section suivante en nous intéressant plus particulièrement au modèle statistique le plus utilisé par les chercheurs en sciences humaines, le modèle linéaire général[1], qui se base sur une décomposition des liaisons complexes en effets simples, interactions, contrastes. Nous décrivons également les conditions d'application de ces modèles statistiques qui les rendent inopérants pour ce que nous souhaitons faire : extraire automati-

[1] Pour Azaïs et Bardet (2005), *le modèle linéaire général*, ou plus *le modèle linéaire*, exprime la variable à expliquer comme combinaison linéaire des *paramètres du modèle*, et non des variables explicatives. Selon leur définition, que nous adoptons, l'équation de régression polynomiale $Y=aX^2+bX+c+\varepsilon$ fait partie du modèle linéaire car elle est linéaire en les paramètres inconnus a, b et c.

quement de grands tableaux individus X variables les liaisons complexes entre variables sans faire aucune hypothèse sur les lois de probabilités suivies par les données. Nous terminons cette deuxième section par un rapide tour d'horizon des méthodes d'EGC (Extraction et Gestion de Connaissances) permettant de brasser davantage de variables.

Dans la section 3, nous détaillons les principes de la dernière méthode présentée dans la section précédente, principes repris et modifiés dans notre méthode MIDOVA afin qu'elle permette de repérer et valider les liaisons complexes entre variables. Dans la section suivante, nous appliquons notre méthode MIDOVA à des données réelles issues d'une enquête portant sur les conditions de vie des ménages tunisiens. La dernière section est consacrée au bilan et aux perspectives.

2 État de l'art des liaisons complexes en traitement des données

2.1 Les liaisons complexes en statistique « classique »

La complexité des relations entre variables n'est pas un fait nouveau en statistique. Dans la figure 1, nous avons représenté quatre modèles statistiques permettant d'exprimer divers aspects de la complexité des liens entre un petit nombre de variables. Les variables y sont représentées par des lettres de A à J, les liens entre elles par des lignes les joignant. Un lien peut comporter selon les cas une flèche, un signe et/ou un libellé.

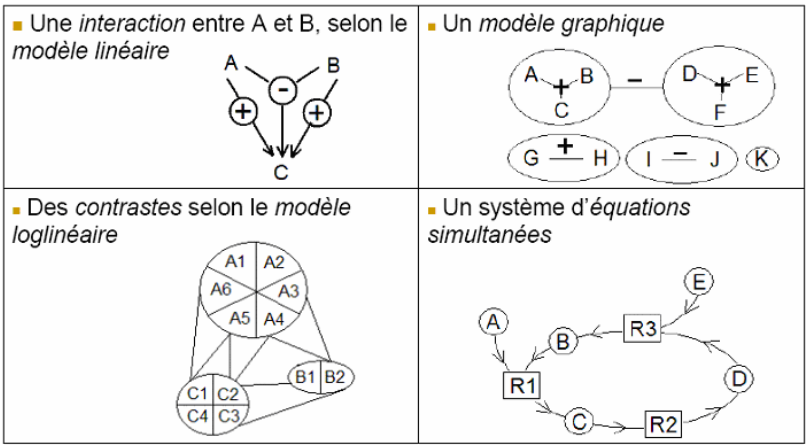


FIG. 1 – Quatre modèles statistiques de liaisons complexes entre variables

Dans la suite nous détaillons pour chacun de ces quatre modèles l'aspect de la complexité sur lequel ils se focalisent et les principes sur lesquels ils s'appuient, avant de faire un bilan.

2.1.1 Le modèle linéaire

Le modèle statistique le plus utilisé pour décrire cette complexité est le modèle linéaire (schéma en haut à gauche de la figure 1) (Fisher 1936) : il permet de décomposer l'influence de quelques variables, ici A et B, sur une autre variable, ici C, en plusieurs constituants indé-

pendants. Dans ce modèle, dont les liaisons sont orientées[2] (l'orientation est indiquée par des flèches, qui vont de A et B vers C), les variables ont deux statuts différents : 1) la variable C est appelée variable *dépendante*, ou *à expliquer*, et doit être *quantitative*, c'est-à-dire mesurée sur une échelle numérique, comme la température par exemple, 2) les variables A et B sont appelées variables *indépendantes* ou *explicatives* et peuvent être quantitatives ou *qualitatives*, c'est-à-dire pouvant prendre plusieurs valeurs différentes qu'on ne peut pas unanimement ordonner, appelées *modalités*, comme les couleurs par exemple. Dans le schéma de la figure 1, on a représenté un cas où A et B n'ont pas le même effet sur C selon qu'ils agissent séparément ou conjointement. Par exemple dans le cas de variables A et B quantitatives de moyennes nulles, l'équation, C=20+3A+5B-10AB+e signifie qu'une augmentation d'une unité de A, indépendamment de B, produit une augmentation moyenne de 3 unités de C, indiquée par un signe + sur le lien allant de A vers C, qu'une augmentation d'une unité de B, indépendamment de A, produit une augmentation moyenne de C de 5 unités, mais qu'une augmentation conjointe d'une unité de A et de B produit en moyenne une augmentation de 3+5 unités de C ainsi qu'une diminution de 10 unités de C, ce qui donne une diminution de 2 unités de C, indiquée par le signe – sur le lien de (A,B) vers C. Cette décomposition linéaire de la valeur de C en composants, ici la moyenne de C, l'effet de A, de B, de l'interaction AB et l'effet individuel e appelé également *résidu*, ne peut être utilisable que si les nombres présents dans l'équation sont des estimations fiables des coefficients réels. L'ANOVA (ANalysis Of VAriance) et la régression linéaire permettent de les estimer au mieux quand diverses conditions concernant la distribution de probabilité des résidus sont vérifiées (Winer 1991). Un modèle linéaire est généralement d'autant plus performant qu'il contient peu d'interactions, et que celles-ci mettent en jeu peu de variables.

Pour rendre plus concrète cette notion d'interaction, citons Snedecor et al. (« Méthodes statistiques », 1966, p. 390), qui relatent le résultat d'une expérience de nutrition de rats visant à estimer de l'effet de deux variables, l'*origine des protéines* et le *niveau de dose*, sur une troisième, le *gain de poids* : « En conséquence de l'interaction, les protéines Animales entraînent un gain de poids beaucoup plus grand que les protéines Céréalières ceci pour le niveau Élevé, mais ne montrent aucune supériorité devant les protéines d'origine Céréale au Faible niveau. ».

2.1.2 Le modèle log-linéaire

Le modèle log-linéaire (schéma en bas à gauche de la figure 1) est un modèle statistique de liaisons complexes entre variables très proche du modèle linéaire qui est apparu après celui-ci (années 1970, Goodman). Il est utilisé pour un petit nombre de variables toutes qualitatives et de même statut[3] et dans sa version la plus élémentaire, il permet de contrôler l'indépendance de deux variables. Pour pouvoir se raccrocher au formalisme du modèle

[2] Les liaisons orientées sont également appelées liaisons de *type causal*, ou de *type cause à effet*, mais l'interprétation des liens en termes de causalité doit s'appuyer sur la connaissance du domaine des données et non sur le modèle linéaire : l'utilisation du modèle linéaire permet d'accepter ou de rejeter l'existence du lien sur des critères numériques et non sémantiques.

[3] Les liaisons entre variables ne sont pas supposées être de type « cause à effet » dans le modèle log-linéaire, ce qui est indiqué par une absence de flèche dans le schéma (figure 1, en bas à gauche) : il n'y a pas de variable à expliquer (i.e. dépendante) et de variables explicatives. Toutefois, la signification des variables peut orienter l'interprétation des liaisons : les difficultés de l'accouchement peuvent avoir une incidence sur le développement ultérieur de l'enfant et non l'inverse.

linéaire, le logarithme de l'effectif correspondant à chaque croisement de modalités est utilisé comme variable quantitative dépendante. Dans le schéma, sont représentées des liaisons entre 3 variables : la variable A, qui dispose de six modalités (A1 à A6), B en ayant deux et C quatre. Les traits du schéma joignent deux à deux des modalités, indiquant par là que chaque lien entre variables peut se décomposer en *contrastes*, qui sont des oppositions entre groupes de modalités[4]. Les contraintes du modèle log-linéaire sont celles du modèle linéaire auxquelles s'ajoutent des exigences d'effectifs suffisants, dès lors que l'on désire établir par des tests appropriés la significativité de certains effets. Nous renvoyons le lecteur intéressé par les détails de ce modèle à l'ouvrage que Morineau et al. (1996) lui ont consacré et nous contentons de montrer dans le paragraphe suivant la difficulté d'utilisation pratique du modèle log-linéaire par un exemple tiré de leur ouvrage.

Les auteurs reprennent une étude de Rabelet (1981) qui donne pour chacune des 326 agressions ayant fait l'objet d'un jugement dans un des 20 états de Floride de 1976 à 1977, la couleur de peau de l'accusé (X : blanche ou noire), celle de la victime (Y : blanche ou noire), et le verdict (Z : condamnation ou non). Les divers modèles log-linéaires possibles liant les 3 variables sont les modèles : m1) pas d'interaction, m2) interaction entre Y et Z, que nous notons Y*Z, m3) interaction X*Z, m4) interaction X*Y, m5) interactions X*Z et Y*Z, m6) interactions X*Y et X*Z, m7) interactions X*Y et Y*Z, m8) interactions X*Y, X*Z et Y*Z, m9) interactions X*Y, X*Z, Y*Z et X*Y*Z. Suite à un test statistique, les modèles m1, m2, m3 et m5 sont rejetés, ce qui invite les auteurs à conclure que l'interaction X*Y entre la couleur de peau de l'accusé et celle de la victime est à considérer. Un autre test de comparaison de modèles emboîtés leur permet de conclure que le modèle m7 semble être tout aussi bon que m8. Un dernier test portant sur le nombre de jugements pour lesquels l'accusé était blanc, la victime était blanche et il y a eu condamnation, a permis de rendre plus probable le modèle m8.

2.1.3 Le modèle graphique

Le schéma situé en haut à droite de la figure 1 est issu d'un *modèle graphique*. Le graphe du schéma a été construit à partir des corrélations significatives entre paires de variables quantitatives en utilisant le coefficient de corrélation linéaire de Bravais-Pearson. La complexité vient ici des regroupements effectués entre variables (groupe G1 formé des variables A, B et C, groupe G2 formé des variables D, E et F, etc.) et des liens entre groupes de variables (lien négatif entre G1 et G2, etc.). Ces regroupements peuvent se faire à partir de notions purement graphiques (par exemple, 3 variables sont regroupées de façon positive si toutes leurs corrélations 2 à 2 sont positives et significatives), ou en calculant des corrélations partielles, conditionnelles ou multiples à partir des corrélations simples. Les modèles graphiques ont de nombreuses variantes ; un état de l'art de divers modèles graphiques fait l'objet d'un chapitre écrit par Jeanne Fine (« Modèles pour l'Analyse des Données Multidimensionnelles », chapitre 9, 1992).

---

[4] Les contrastes ne sont pas une spécificité du modèle log-linéaire : ils interviennent essentiellement dans le modèle linéaire quand les variables explicatives sont qualitatives et se déclinent en contrastes a priori, a posteriori (Howel 1998). Dans l'exemple agricole de Fisher et Snedecor que nous avons cité, la variable appelée *origine des protéines* avait trois modalités : *Bœuf, Porc, Céréale,* et après constat de son absence d'effet significatif sur le *gain de poids*, elle a été remplacée par 2 contrastes a posteriori : *Animale/Végétale* et *Bœuf/Porc*, dont un seul, le premier, a interagi avec la variable *niveau de dose* sur la variable *gain de poids.*

Nous reprenons à titre d'illustration un exemple que Joe Whittaker donne en pages 2 à 6 de son ouvrage consacré aux modèles graphiques (« Graphical models in applied multivariate Statistics », 1990). Les données sont celles de Mardia, Kent et Bibby (1979), et consistent en les notes de 88 élèves à 5 examens de mathématiques : *mécanique, vecteurs, algèbre, analyse, statistique*. Il réalise un modèle graphique à partir de la matrice des corrélations partielles entre ces variables prises deux à deux, sachant les variables restantes. Il joint deux variables par un trait si leur corrélation partielle dépasse un seuil de 0,10 et obtient un graphe qu'il décrit comme ressemblant à un papillon dont le corps est *algèbre* avec une aile composée de *mécanique* et *vecteurs*, l'autre étant composée de *analyse* et *statistique*. Il applique la « propriété globale de Markov » pour résumer la structure des données en « *(mécanique, vecteurs*) est indépendant de (*analyse, statistique*) conditionnellement à *algèbre »*. Il en tire l'interprétation suivante : l'algèbre et l'analyse seront suffisants pour prédire la note de statistique, l'algèbre et les vecteurs seront suffisants pour prédire la note de mécanique, mais il faudra les notes des quatre autres matières pour prédire la note d'algèbre. Il exprime ainsi un lien complexe entre les cinq variables avec un graphe qui << *is nothing but the conditional independence graph for jointly distributed multivariate Normal random variables* >>, ce qui lui permet de « réduire l'objet de dimension 5 à deux objets de dimension 3 » : (*algèbre, mécanique, vecteurs*) et (*algèbre, analyse, statistique*).

2.1.4 Les équations simultanées

Le schéma situé en bas à droite de la figure 1 est un *système d'équations simultanées* formé de trois équations de régression,

R1 : $C=a_1A+a_2B+e_1$, indiquant que C est expliqué en partie par A et B, e_1 étant la partie de C restant non expliquée, appelée *résidu*, ou *erreur*.

R2 : $D=a_3C+e_2$, indiquant que D est expliqué en partie par C, e_2 étant le résidu,

R3 : $B=a_4D+a_5E+e_3$, indiquant que B est expliqué en partie par D et E, e_3 étant le résidu.

Les trois équations s'enchaînent en un système dans lequel certaines variables (B, C et D) prennent un statut intermédiaire entre variables expliquées et explicatives. L'estimation de ces modèles ne peut se faire que sous certaines conditions excluant notamment la possibilité de « cercles vicieux ». La conception de logiciels simples d'emploi, comme LISREL (Joreskog 1984) par exemple, a permis aux chercheurs des sciences humaines d'accéder à la méthodologie des systèmes d'équations simultanées qui était, une quinzaine d'années auparavant, la spécialité des économistes (*l'équilibre des marchés*, comme le marché de l'emploi, s'appuie sur de telles équations).

Ici la complexité des liaisons entre variables vient essentiellement de la variété des statuts des variables. Par exemple, il est bien connu qu'un élève qui travaille dans une discipline qu'il aime a généralement de bons résultats, mais aussi que les bons résultats dans cette discipline et le fait de l'aimer lui donnent envie de travailler, et que les bons résultats suffisent peut-être à lui faire aimer, ou à lui donner envie de travailler. La relation de type causal entre ces trois variables (le travail, les résultats, l'appétence) est complexe, et une voie de résolution consiste à la décomposer selon des intervalles de temps réguliers, en posant par exemple que l'augmentation des notes au temps t entraîne une augmentation du travail et de l'attrait au temps t+1, que l'augmentation du travail au temps t entraîne une augmentation des notes au temps t+1, etc. Des analyses de ce type ont été faites par des chercheurs en sciences humaines, comme Dickes (1996).

2.1.5 Bilan des 4 modèles statistiques

Les quatre modèles que nous venons de décrire présentent diverses facettes de la complexité des liens entre variables vue à travers le formalisme statistique classique. Ils produisent des estimations fiables des composants des liens quand certaines conditions sont réalisées : quelques variables bien choisies, les autres n'étant pas censées intervenir, une grande partie des liens et de leurs composants jugés négligeables pour se concentrer sur les liens et composants restants, et une spécification plus ou moins précise des lois de probabilités suivies par les données.

Pour réaliser notre but, qui est d'extraire et de valider les liaisons complexes entre les variables à partir de données quelconques, nous devons sortir de ce cadre statistique. D'abord, nous ne souhaitons pas privilégier certaines variables, certains liens ou composants, car notre but est de les trouver tous, même ceux qui auraient pu échapper aux experts des sciences humaines qui sont à l'origine des données (Suzuki et al. 1998). Quant aux lois de probabilités suivies par les données, elles s'éloignent souvent des lois préconisées dans les modèles statistiques ; notamment les lois zipfiennes (Breslau et al. 1999) suivies généralement par les données textuelles ne sont pas prises en compte dans les tests statistiques classiques. Si nous choisissions de rester dans ce cadre statistique, il faudrait alors corriger, recoder les données pour se rapprocher des distributions théoriques, et cela aurait pour conséquence d'entraîner la perte d'une partie de l'information, ce que nous souhaitons également éviter. Nous examinons maintenant d'autres modèles contenant plus de variables, et moins de contraintes.

2.2 Les liaisons complexes en EGC

L'analyse et la fouille des données traitent différemment la complexité entre variables. La variabilité intrinsèque des individus figurant dans le tableau individus X variables est d'abord éliminée en le remplaçant par un tableau variables X variables contenant la valeur des liens entre toutes les variables prises 2 à 2. Les analyses factorielles et les classifications font partie des méthodes les plus anciennes et encore les plus utilisées actuellement. Dans l'analyse factorielle, si les données sont quantitatives, c'est le plus souvent le tableau des corrélations qui est factorisé pour créer p nouvelles variables de niveau supérieur (Analyse en Composantes Principales), et si les données sont qualitatives, on factorise le tableau des fréquences de chaque couple de modalités des variables (Analyse Factorielle des Correspondances, Analyse des Correspondances Multiples). Dans la classification hiérarchique ascendante (Lerman, 1995), on part d'un tableau de dissimilarités ou de distances entre variables, et on agrège en q étapes les variables de départ en de nouvelles variables. L'utilisation de ces méthodes requiert moins de conditions sur les données que les méthodes statistiques décrites précédemment et elles sont adaptées à des données dont le nombre de variables est important. Notons toutefois qu'elles ne sont pas très stables, car elles dépendent d'un paramètre (nombre de facteurs, niveau d'agrégation ou nombre de classes), et les résultats diffèrent de façon parfois importante selon les valeurs de ce paramètre, sans parler de l'instabilité due parfois aux choix d'initialisation de l'algorithme. De plus, la seule complexité qui est mise au jour est celle figurant à l'intérieur du tableau de liens 2 à 2. S'il y avait des interactions entre variables de niveau supérieur, elles ne sont plus décelables, et elles peuvent même créer des perturbations dans l'analyse et la rendre moins fiable (Benzecri 1970).

Depuis quelques décennies, l'accroissement constant des capacités de calcul et de stockage des ordinateurs, ainsi que le libre accès sur le Web à des bases de données de plus en

plus volumineuses et nombreuses ont eu pour conséquence l'apparition de nouvelles méthodes de traitement des données. Parmi les méthodes actuelles, les trois plus aptes, à notre connaissance, à prendre en compte la complexité des liaisons entre variables sont 1) les méthodes à noyaux, comme les « Support Vector Machines » (Vapnik 1995), qui permettent de prédire la valeur d'une variable à expliquer par des fonctions bien choisies (les noyaux) des variables explicatives, 2) les réseaux bayésiens (Leray 2004) qui permettent de tenir un raisonnement de type cause à effet en s'appuyant sur une estimation des lois de probabilité de petits groupes de variables supposés sans liens avec les autres et 3) les méthodes d'extraction des motifs qui extraient toutes les associations de k (k>=1) variables dépassant un seuil d'indice de qualité donné. Les SVM font partie des méthodes d'apprentissage supervisé et fonctionnent comme une boîte noire, fournissant un taux global de reconnaissance de la variable à expliquer très fiable, mais sans pouvoir préciser les fonctions de quelles variables ont le plus contribué à ce taux. La construction d'un réseau bayésien à partir des données nécessite de poser certaines hypothèses sur ce réseau (par exemple pas plus de 4 liens qui se suivent), et malgré cela le réseau peut comporter des contresens du point de vue de la causalité, et l'utilisateur doit les corriger avant de pouvoir utiliser le réseau (Cadot 2009). Les méthodes à base d'extraction de motifs ont l'avantage d'être entièrement automatiques, une fois certains paramètres fixés, de donner explicitement la liste des liaisons complexes retenues et de ne pas exiger que les données suivent des lois de probabilités spécifiques. Elles ont été construites initialement pour traiter des données binaires comme les tickets de caisse des supermarchés (Han 2001). C'est la méthode qui nous a paru la plus facilement adaptable à notre recherche de liaisons complexes entre variables.

3 Liaisons complexes entre variables booléennes : des motifs de la fouille de données à notre méthode MIDOVA

Nous faisons dans une première partie de la présente section un rapide tour des principes et fondements de l'extraction des motifs, selon les principaux courants de la fouille de données. Nous donnons dans la partie suivante une idée du type de liaison complexe entre variables que doit représenter un motif intéressant selon nos critères, puis dans la troisième partie nous développons le formalisme de notre méthode MIDOVA qui permet d'extraire les motifs porteurs d'une telle signification. Dans la quatrième partie nous appliquons ce formalisme à un exemple jouet, et les deux dernières parties sont consacré à l'algorithme d'extraction MIDOVA, et à ses différences par rapport à l'algorithme Apriori, selon divers points de vue, notamment numérique et sémantique.

3.1 Les motifs de la fouille de données et les liaisons entre variables

Supposons que l'on dispose d'un ensemble S de N individus, d'un ensemble V de p variables booléennes (c.à.d. à deux valeurs, 1 : Vrai, 0 : Faux), et d'une relation R entre S et V mise sous la forme d'un tableau booléen individus X variables. Un *k-motif* est une association de k variables de V. Il se nomme généralement par la liste de ses variables. Il y a autant de motifs que de sous-ensembles de variables dans V, soit 2^p si l'on compte le motif vide. L'explosion combinatoire est contenue en ne gardant que les motifs les plus « intéressants » selon divers critères. Par exemple dans l'algorithme Apriori (Agrawal et al. 1996), c'est le choix d'un seuil de *support* du motif, nombre d'individus pour lesquels les variables du motif

sont simultanément vraies, qui va enrayer cette explosion. De plus, comme le support d'un k-motif est une fonction non croissante de k, le choix par ces auteurs d'un algorithme d'extraction de motifs fonctionnant *par niveau*, c'est-à-dire construisant tous les k-motifs avant de construire les (k+1)-motifs, permet de rendre cette recherche d'autant plus rapide que le seuil de support choisi est élevé.

A partir des couples de motifs, ont été définies des *règles d'association* (notées RA dans la suite de cette article) permettant de déduire le second motif du premier. Leur qualité a été mesurée par deux indices, le *support* de la règle, nombre d'exemples vérifiant simultanément les deux motifs, et sa *confiance*, proportion d'exemples vérifiant le second parmi ceux vérifiant le premier (Agrawal et al. 1996). Par exemple, la règle AB→C, où A, B et C représentent respectivement l'achat d'écharpe, de bonnet et de paire de gants, se traduit par « les personnes qui ont acheté écharpe et bonnet ont aussi acheté des gants » ; si on suppose que parmi les 50 personnes ayant acheté écharpe et bonnet, 40 ont acheté des gants, le support de la règle est égal à 40 (c'est le support du motif ABC, c.à.d. le nombre de personnes ayant acheté simultanément les trois articles), et sa confiance, rapport entre le support de ABC et celui de AB, est de 40/50 soit 80%. Ces premiers travaux en informatique appliqués dans le domaine de la grande distribution ont produit des règles très proches de celles issues de travaux précédents en statistique dans le domaine de la didactique (règles *d'implication statistique*, Gras 1979) et en algèbre dans le domaine des sciences humaines (règles *d'implication informative*, Guigues et al 1986). Cet héritage multiple a eu des retombées depuis une dizaine d'années non seulement sur les RA, mais également sur les motifs. Par exemple, la qualité des motifs et des RA peut se mesurer actuellement par plus de cinquante indices (Guillet 2004) pris séparément ou combinés (Lenca et al. 2003), qui représentent autant de nuances dans l'interprétation de l'association, selon les statistiques, probabilités, ou la sémantique du domaine d'application. Et il existe actuellement de nombreux algorithmes d'extraction de motifs qui diffèrent par les notions algébriques qu'ils utilisent ou par leur façon de parcourir les données (Cadot 2006) et de stocker l'information au fur et à mesure de sa lecture.

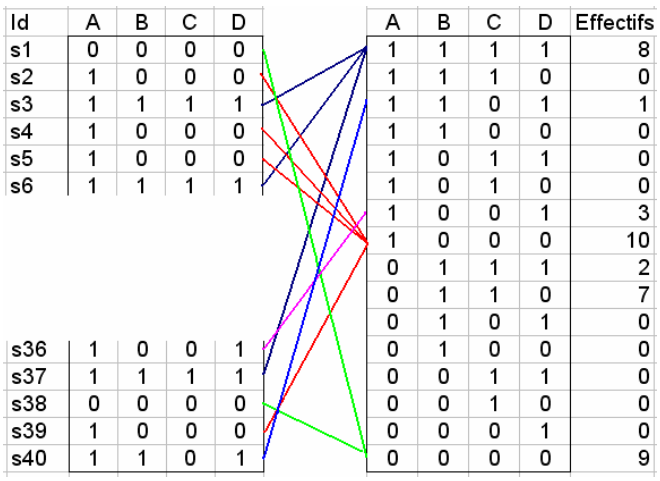
Dans ces travaux, l'extraction de motifs a pour but essentiel de trouver la totalité des agrégats de variables intéressants selon une, voire plusieurs mesures de qualité. À notre connaissance, ces mesures de qualité ne peuvent pas, de par leur conception, rendre compte des liaisons complexes entre plus de deux variables, comme l'interaction par exemple. En effet, dans la plupart des cas, le mode de calcul de la mesure de qualité est défini pour 2 variables et, pour l'appliquer à 3 variables ou plus, on ne le modifie pas mais on l'applique à deux nouvelles variables obtenues par agrégation des précédentes : par exemple, pour l'appliquer au motif ABCD formé de 4 variables binaires, on l'applique au motif XD où X est la variable binaire (A et B et C) obtenue en faisant le produit terme à terme des trois variables binaires A, B et C. L'aspect multidimensionnel des liaisons entre variables se trouve ainsi gommé par de telles mesures de qualité. La méthode MIDOVA a été conçue pour faire ressortir au contraire l'aspect multidimensionnel de type interaction des motifs lors du processus d'extraction.

3.2 Idée intuitive de MIDOVA sur un exemple

Un des premiers buts de notre méthode MIDOVA (Multidimensional Interaction Differential Of VAriance, Cadot 2006) est d'extraire tous les motifs correspondant à des interactions positives ou négatives entre les variables d'un tableau T de type individusXvariables.

Pour ne manquer aucun de ces motifs, nous extrayons par MIDOVA tous les motifs représentant *l'information nécessaire et suffisante* pour pouvoir reconstituer le tableau T, aux permutations des individus près. C'est un point de vue statistique dans la mesure où on ne s'intéresse pas aux individus en eux-mêmes mais aux liens qu'ils créent entre les variables. Selon ce point de vue, le tableau T d'origine peut être remplacé par un tableau T' dans lequel les individus sont seulement décomptés pour chaque croisement de valeurs de variables. C'est à partir du tableau T' que sont exposés les principes de MIDOVA.

Dans la figure 2, on a représenté des données concernant 40 individus, s1 à s40, et 4 variables A, B, C et D. Le tableau à gauche de la figure est un aperçu du tableau T d'origine, qui contient les valeurs des 40 individus aux 4 variables, et le tableau T' à droite contient tous les croisements possibles de valeurs de ces variables, qui sont au nombre de 2^p pour p variables (donc ici 2^4 =16) avec l'effectif correspondant en dernière colonne. Les traits d'une même couleur indiquent que les lignes du tableau de gauche sont fusionnées en une même ligne du tableau de droite. Par exemple les deux traits verts joignent les individus s1 et s38 du tableau T, qui ont toutes leurs variables à 0, à la dernière ligne du tableau T' d'effectif 9 car ils font partie des 9 individus ayant toutes leurs variables à 0.



Id	A	B	C	D
s1	0	0	0	0
s2	1	0	0	0
s3	1	1	1	1
s4	1	0	0	0
s5	1	0	0	0
s6	1	1	1	1
s36	1	0	0	1
s37	1	1	1	1
s38	0	0	0	0
s39	1	0	0	0
s40	1	1	0	1

A	B	C	D	Effectifs
1	1	1	1	8
1	1	1	0	0
1	1	0	1	1
1	1	0	0	0
1	0	1	1	0
1	0	1	0	0
1	0	0	1	3
1	0	0	0	10
0	1	1	1	2
0	1	1	0	7
0	1	0	1	0
0	1	0	0	0
0	0	1	1	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	9

FIG. 2 – À gauche le tableau T pour 40 sujets et 4 variables, à droite le tableau T'

Un premier coup d'œil sur la dernière colonne du tableau T' de la figure 2 fait apparaître un certain déséquilibre des effectifs : plus de la moitié des effectifs sont nuls, et au moins le quart des effectifs dépasse 5 (ils sont de 7, 8, 9, 10). Et la question qu'on se pose de façon naturelle est « est-ce surprenant ? ». Cette question peut s'exprimer de façon plus opérationnelle : si les valeurs de A, B, C et D de chaque sujet étaient attribuées au hasard, pourrait-on arriver à une telle répartition des effectifs ? En simulant ce hasard, on arrive à des réponses d'autant plus vagues que le nombre de simulations faites est petit ; une telle répartition se produit par hasard :

a) *très rarement* si on simule les valeurs des 4 variables de façon indépendante, selon une loi uniforme (même probabilité d'avoir 0 ou 1),
b) *très rarement* si on simule les valeurs des 4 variables de façon indépendante, mais selon des lois qui respectent la répartition de valeurs de chaque variable (22/40 valeurs 1 pour A, 18/40 pour B, 17/40 pour C et 14/40 pour D),

c) *presque toujours* si on simule les valeurs des 4 variables en exigeant que les lois jointes des variables prises par deux suivent les répartitions d'effectifs des données.
d) *presque toujours* si on simule les valeurs des 4 variables en exigeant que les lois jointes des variables prises par trois suivent les répartitions d'effectifs des données.

C'est cette vision statistique qui est à l'origine de la méthode MIDOVA, dans laquelle les simulations sont remplacées par des calculs que nous exposons dans la section suivante. La méthode fournit des résultats cohérents avec les 4 réponses : aucun motif de longueur 4 n'est extrait (réponse d), tous les motifs de longueur 1 et 2 sont extraits (réponses a et b), la moitié des motifs de longueur 3 sont extraits, mais aucun parmi ceux-ci n'est finalement retenu (réponse c). Toute l'information pertinente des données de T' figure dans les k-motifs de longueur $k \leq 2$, les liaisons de plus de 2 variables ne sont pas complexes car elles se ramènent à des liaisons entre 2 variables.

3.3 Formalisme de MIDOVA

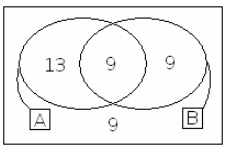
Dans la partie 3.1 qui faisait un rapide état de l'art des motifs en fouille de données, nous avons défini un motif comme une association de variables booléennes, puis nous avons décrit son mode d'utilisation, au travers des règles d'association. La règle AB→C liant les achats de 3 articles, que nous avons donnée en exemple, peut s'interpréter en terme d'interaction telle que définie dans le modèle linéaire décrit précédemment : les achats de gants qu'on peut prédire à partir des achats cumulés d'écharpe et de bonnet diffèrent de ceux qu'on peut prédire en combinant les deux prédictions séparées (achats de gants à partir de d'achats d'écharpes, et achats de gants à partir d'achats de bonnets). Selon le signe de cette différence, l'interaction est positive ou négative. En transposant de façon similaire l'interaction du modèle log-linéaire aux motifs, nous obtenons que la valeur de l'interaction correspondant au 3-motif ABC est relative à une valeur attendue qui dépend des 2-motifs AB, AC et BC et nous proposons de la mesurer par l'écart entre le support observé de ABC et la valeur attendue de ce support sachant AB, AC et BC. Avec MIDOVA la valeur attendue du support se calcule à partir d'un indice de *reste*, noté Mr, qui représente les possibilités de variations du k-motif étant donnés ses sous-motifs de longueur k-1. Pour le définir, nous allons d'abord exposer sur un exemple ce que nous appelons les variations d'un 2-motif AB, étant donnés ses sous motifs A et B de longueur 1.

3.3.1 Représentation d'un 2-motif selon MIDOVA

Si nous prenons un 2-motif de la figure 2, par exemple AB, il peut se représenter de plusieurs façons, comme indiqué en figure 3, en croisant les variables A et B. À gauche de la figure 3, on a utilisé le formalisme ensembliste du diagramme de Venn pour représenter les variables A et B. Les points habituels des individus ont été remplacés par leur l'effectif. L'ensemble S des 40 individus se trouve ainsi réparti dans 4 zones :

- Au milieu du diagramme, l'intersection de A et de B contient les 9 individus qui ont la valeur 1 pour A et pour B (s3, s6, s37, s40 en font partie comme on peut le voir dans le tableau de gauche de la figure 2)
- À gauche du diagramme, la différence A-B contient les 13 individus pour lesquels A=1 et B=0 (comme s2, s4, s5, s36, s39 de la figure 2)

- À droite du diagramme, la différence B-A contient les 9 individus pour lesquels A=0 et B=1
- En bas du diagramme, la différence S-AUB contient les 9 individus pour lesquels A=0 et B=0 (comme s1, s38 de la figure 2)



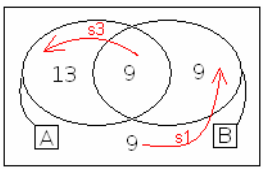
	B		
A	0	1	Total
0	9	9	18
1	13	9	22
Total	22	18	40

FIG. 3 – Deux représentations du motif AB : à gauche le diagramme de Venn, à droite son tableau de contingence

Ces effectifs sont repris dans le *tableau de contingence* à droite de la figure 3, qui contient en intitulé de la première colonne les valeurs de la variable A, en intitulé de la deuxième ligne les valeurs de la variable B, et dans les cellules situées aux croisements des valeurs de A et de B les effectifs correspondants. Dans les *marges*, c'est-à-dire dans la dernière colonne à droite et la dernière ligne en bas, figurent les totaux des 4 cellules : dans la dernière colonne les effectifs de la variable A sont obtenus en additionnant les valeurs des 2 cellules de la ligne correspondante (9+9=18 sujets ont la valeur 0, 13+9=22 sujets ont la valeur 1), et dans la dernière ligne les effectifs de la variable B sont obtenus en additionnant les valeurs des 2 cellules de la colonne correspondante (9+13=22 sujets ont la valeur 0, 9+9=18 sujets ont la valeur 1). Le total général des 4 cellules figure en bas à droite et correspond au nombre total d'individus.

3.3.2 Les variations d'un 2-motif à sous-motifs fixés

La première étape de la recherche des variations du motif AB, quand ses sous-motifs sont fixés, consiste à chercher les valeurs différentes que peut prendre une des quatre cellules du tableau, par exemple la cellule[5] correspondant à A=1 et B=1, appelée support du motif, sans que les effectifs des variables A et B ne soient modifiés. Dans la figure 4 on a représenté un exemple d'une modification possible qui produirait un support de AB de 8.



	B		
A	0	1	Total
0	9 -1	9+1	18
1	13+1	9 -1	22
Total	22	18	40

FIG. 4 – Transformation du motif AB gardant inchangés ses sous-motifs A et B.

À gauche de la figure 4, on voit que le sujet s3 (ses valeurs figurent dans le tableau à gauche de la figure 2), qui faisait partie des 9 individus ayant leurs valeurs de A et de B à 1, a été modifié : sa valeur de B est devenu 0, et il est allé rejoindre les 13 sujets dont la valeur de A est à 1 et la valeur de B à 0. Cela correspond aux deux modifications de la ligne A=1 du

[5] Nous verrons plus loin que n'importe laquelle des cellules peut être choisie à la place de celle relative au support.

tableau de contingence à droite de la figure 4 : 13 devient 13+1=14 et 9 devient 9-1=8. De ce fait, le motif A n'a pas changé, il y a toujours 18 sujets ne vérifiant pas A et 22 le vérifiant. Par contre, B a changé (17 individus vérifient B contre 18 avant), et il faut changer une deuxième valeur du tableau d'origine pour le rétablir à ses valeurs d'origine : le sujet s1 qui avait ses valeurs de A et de B à 0 est modifié : sa valeur de B passe à 1, et il quitte les 9 sujets dont les valeurs de A et de B sont à 0 pour rejoindre les 9 sujets pour lesquels B=1 et A=0. Cela correspond aux modifications de la ligne A=0 dans le tableau de contingence de la figure 4 : 9 devient 9-1=8 et 9 devient 9+1=10. Cette fois les effectifs de B reprennent leurs valeurs initiales. Toutes les valeurs du tableau de contingence ont changé quand le support est passé de 9 à 8, mais les marges sont restées inchangées. Le 2-motif AB a ainsi changé, mais pas ses sous-motifs de longueur 1.

De cet essai de variation du motif AB, on peut tirer quelques remarques :

a) Quand on augmente de 1 une des quatre zones du diagramme, pour garder les mêmes effectifs des 1-motifs A et B, il convient de modifier en conséquence les trois autres zones : les quatre zones sont divisées en deux groupes de deux zones tels que dans l'un, les effectifs ont augmenté de 1 et dans l'autre ils ont diminué de 1. Deux zones d'un même groupe n'ont pas de « bord » commun.
b) Les quatre cellules du tableau de contingence du 2-motif AB sont également divisées en deux groupes de deux cellules, les effectifs d'un groupe ont augmentés de 1, et ceux de l'autre ont diminué de 1. Deux cellules d'un même groupe ne sont pas contiguës.
c) Pour obtenir une modification d'un des effectifs de 1, il faut changer au moins une valeur d'une variable (A ou B) pour au moins deux sujets. Ce qui correspond dans le diagramme de Venn au déplacement de deux sujets entre deux zones n'appartenant pas au même groupe.

3.3.3 Amplitude de variation, indices Mg, Mr d'un 2-motif et de ses sous-motifs

Pour connaître *l'intervalle de variation* du support de AB, avec A et B fixés, il suffit d'essayer de le diminuer le plus possible, et de l'augmenter le plus possible, selon les règles issues des remarques précédentes. Comme il diminue en même temps qu'un autre effectif, situé dans le même groupe que lui, il peut diminuer jusqu'à ce qu'un des deux effectifs atteigne 0. Les deux effectifs étant à 9, ils atteignent en même temps 0 qui est la valeur la plus petite possible du support. Puis on essaie de l'augmenter. Quand il augmente, les effectifs de l'autre groupe diminuent d'autant, ce qui fait qu'on peut l'augmenter jusqu'au moment ou l'un des deux est nul. Ces effectifs étant de 13 et 9, ils peuvent diminuer de 9, et le support est augmenté de 9 et atteint la valeur de 18. L'intervalle de variation du support de AB est donc [0 ; 18], et *l'amplitude de variation* du support est 18-0=18.

Comme le support est 9, il est situé juste au centre de son intervalle de variation. Si sa valeur avait été 0, on aurait pu dire qu'il y avait une forte interaction négative entre A et B, et si sa valeur avait été 18, on en aurait déduit une forte interaction positive. La valeur de 9 étant confondue avec le centre de l'intervalle, nous concluons qu'il n'y a pas d'interaction remarquable entre A et B. Nous calculons un indice de *gain* Mg dont la valeur est proportionnelle à l'écart entre le support et le centre. Le coefficient de proportion est 2^{k-1}, où k est la longueur du motif, afin qu'il exprime le nombre de sujets qui ont été déplacés par rapport au centre pour arriver à cet écart. Nous avons vu que pour obtenir un écart de -1 au centre (support de 8 au lieu de 9), il avait fallu déplacer deux sujets, ce qui donne un gain de -2 dans ce

cas. Ainsi le gain maximal (en valeur absolue) qu'un motif peut atteindre est égal au nombre total de sujets. Mais il ne l'atteint que dans des cas très particuliers (pour des exemples détaillés, voir Cadot 2006).

Le reste Mr indique la variabilité que le motif laisse à ses sur-motifs. Sa formule est $Mr = 2^{k-1}|b-s|$ où b est la borne de l'intervalle de variation du support s la plus proche de s. Plus le gain est élevé en valeur absolue, plus le reste est faible, et inversement. Retournons aux 1-motifs A et B qui ont engendré le 2-motif AB. Le 1-motif A a un support pouvant varier dans l'intervalle [0 ; 40], de centre 20, et son support étant de 22, il a un gain de $2^0(22-20)=2$, et un reste de $2^0(40-22)=18$. Par un calcul identique, on trouve que le 1-motif B a un gain Mg=-2 et un reste Mr=18. De cela, avant même de calculer le 2-motif AB, on pouvait déduire que son amplitude de variation ne pourrait pas dépasser 18, qui est le minimum des restes de ses sous-motifs. Cela est dû à une propriété du reste, qui ne peut que diminuer au fur et à mesure que la longueur du motif augmente. En ce qui concerne le motif AB, nous avons vu que son gain est nul, son reste est donc de 18. Il n'a pas entamé le potentiel de variabilité transmis par ses sous-motifs, il le retransmet intact à ses sur-motifs.

3.4 Extraction par MIDOVA des k-motifs de l'exemple jouet

Motif	k	Support	Max Amplitude	Intervalle	Ampl. variation	Gain Mg	Reste Mr
A	1	22	40	[0 ; 40]	40	2	18
B	1	18	40	[0 ; 40]	40	-2	18
C	1	17	40	[0 ; 40]	40	-3	17
D	1	14	40	[0 ; 40]	40	-6	14
AB	2	9	18	[0 ; 18]	18	0	18
AC	2	8	17	[0 ; 17]	17	-1	16
AD	2	12	14	[0 ; 14]	14	10	4
BC	2	17	17	[0 ; 17]	17	17	0
BD	2	11	14	[0 ; 14]	14	8	6
CD	2	10	14	[0 ; 14]	14	6	8
ABC	3	8	0	[8 ; 8]	0	0	0
ABD	3	9	2	[9 ; 9]	0	0	0
ACD	3	8	2	[8 ; 8]	0	0	0
BCD	3	10	0	[10 ; 10]	0	0	0
ABCD	4	8	0	[8 ; 8]	0	0	0

TAB. 1 – Les k-motifs de l'exemple de la figure 2

Dans le tableau 1, nous avons repris tous les k-motifs de l'exemple de la figure 2 afin d'exposer leur traitement selon MIDOVA. Les contenus des colonnes du tableau 1 sont, de la gauche vers la droite, l'intitulé du motif, sa longueur k, son support, le maximum d'amplitude de variation qu'il peut avoir (qui est le minimum des résidus de ses sous-motifs), son intervalle de variation, l'amplitude de variation, le gain et le reste.

Le motif de longueur 0 est le motif vide, ne contenant aucune variable, mais tous les individus. Il ne figure pas dans le tableau dans la mesure où il n'est soumis à aucune variabilité. Le calcul des motifs de longueur 1 a été détaillé dans la section précédente pour les motifs A et B, il se fait de la même façon pour C et D. Pour les motifs de longueur 2, on a détaillé tous les calculs correspondant au motif AB. Dans le tableau figurent les résultats des autres 2-motifs. Parmi ceux-ci, le 2-motif le plus intéressant est BC, car son gain est le plus élevé,

ce qui indique une forte interaction entre B et C. Notons que son reste est nul, ce qui fait que ce motif ne génèrera que des sur-motifs sans possibilité de variation donc sans intérêt. Les motifs concernés sont ABC, BCD et ABCD. Ils ont été mis dans le tableau malgré tout afin de permettre des comparaisons avec d'autres méthodes d'extraction de motifs. Parmi les deux 3-motifs restants qui pourraient être extraits, détaillons les calculs de ABD. Il est engendré par 3 sous-motifs AB, AD et BD, de restes respectifs 18, 4 et 6. Son amplitude de variation ne peut donc pas dépasser 2=4/2, et on constate même qu'elle est nulle. Examinons les raisons de sa nullité sur le diagramme de Venn représenté en figure 5.

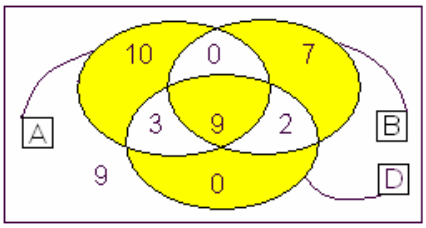


FIG. 5 – Diagramme de Venn du motif ABD.

Le digramme de Venn comprend 8 zones correspondant aux divers croisements des valeurs 0 et 1 de A, B et D. Ces zones sont réparties en deux groupes selon qu'elles varient dans le même sens que l'intersection des 3 ensembles (4 cellules coloriées en jaune) ou dans le sens contraire (4 cellules non coloriées). Nous avons vu que pour construire l'intervalle de variation du support de ABC, on fait varier cet effectif (il est de 9, comme indiqué dans la cellule jaune au centre de la figure 5). Pour pouvoir le diminuer, il faut pouvoir diminuer d'autant tous les effectifs des cellules jaunes, ce qui est impossible, une valeur 0 figurant dans une zone jaune (en bas, correspondant à D=1, A=0 et B=0). Pour pouvoir l'augmenter, il faut pouvoir diminuer d'autant tous les effectifs des cellules non coloriées, ce qui est impossible car l'une d'elle a un effectif de 0 (en haut, correspondant à A=1, B=1 et D=0). Ainsi le support de ABD ne peut pas varier, son amplitude de variation, son gain et son reste sont nuls. Il est sans intérêt car son absence de variabilité indique que la liaison entre les variables A, B et D qu'il exprime peut se déduire de façon logique (par un système d'équations produisant une seule solution) des liaisons entre les variables deux à deux. Autrement dit, l'interaction de niveau 3 entre A, B et D est nulle, et la liaison entre les trois variables n'est pas tri-dimendionnelle, mais de dimension inférieure. Et il en est de même pour le motif ACD.

Pour résumer, seuls les 1-motifs et 2-motifs ont une variabilité non nulle, connaissant leurs sous-motifs ce qui permet de conclure que le tableau de données de la figure 2 ne contient aucune liaison complexe entre plus de deux variables.

3.5 Algorithme de MIDOVA

Une version détaillée de l'algorithme de MIDOVA se trouvant dans Cadot et al. (2010), nous nous contentons de donner ici quelques éléments de son fonctionnement. C'est un algorithme par niveau : les k-motifs sont créés par « assemblage » des (k-1) motifs, la longueur k du motif représente le niveau, et l'algorithme démarre au niveau 1. Un k-motif est *candidat* si tous ses sous-motifs ont été jugés *fertiles* lors de l'étape précédente. Puis le candidat est examiné et jugé *informatif* s'il n'est pas entièrement déterminé par ses sous-motifs. Les k-

motifs informatifs qui laissent quelques possibilités de variation à leur sur-motifs sont déclarés *fertiles*. Par exemple, pour extraire les motifs du tableau T' de la figure 2, on a choisi un seuil de reste de 1, et fait les étapes suivantes :

- Étape 1 : tous les 1-motifs ont une amplitude de variation de N=40, il sont donc informatifs. Comme aucune variable n'est vraie pour tous les individus, ni fausse pour tous les individus, tous les restes sont différents de 0 (voir tableau 1 de la section 3.4), et tous les motifs sont jugés fertiles et conservés pour l'étape suivante.
- Étape 2 : tous les 2-motifs possibles sont obtenus en assemblant les 1-motifs. On obtient ainsi 6 candidats, pour lesquels on calcule les amplitudes de variation et les restes. Ils ont tous une amplitude de variation non nulle, et sont donc informatifs. Un seul reste est nul, il s'agit du motif BC, qui est déclaré *stérile*, les 5 autres 2-motifs étant déclarés fertiles.
- Étape 3 : les cinq 2-motifs fertiles sont assemblés et forment deux 3-motifs candidats. Les amplitudes de variation sont calculées. Comme elles sont toutes deux nulles, les candidats ne sont pas informatifs. Ils ne sont pas conservés, et l'algorithme s'arrête là.

Avec cet algorithme, on a extrait toute l'information du tableau T' sous la forme de 11 motifs, le motif vide, qui donne l'effectif total, les quatre 1-motifs, et les six 2-motifs. Nous avons montré (voir Cadot et al. 2010) qu'avec les motifs extraits par MIDOVA, tout tableau du type de T' peut être reconstitué avec exactitude. Si on choisit un seuil de reste supérieur à 1, on extrait moins de motifs mais avec un potentiel plus élevé de variabilité, et on ne peut reconstituer le tableau T' qu'approximativement.

3.6 Synthèse sur MIDOVA

Il y a de nombreuses similarités entre l'algorithme de MIDOVA et l'algorithme de recherche des motifs fréquents Apriori (Agrawal 1996). Mais c'est le schéma de fonctionnement qui est similaire, pas le résultat recherché. Apriori a pour but de trouver tous les motifs représentant des associations élevées entre variables, c'est-à-dire dont le support dépasse un seuil donné, alors que MIDOVA extrait tous les motifs ayant une variabilité potentielle élevée, c'est-à-dire dont les sous-motifs ont un reste dépassant un seuil donné. Par exemple, pour Apriori, le motif ABCD indiqué dans le tableau 1 est de meilleure qualité que les motifs AC, ABC et ACD de même support car il contient plus de variables, et il permet notamment de générer quatre règles d'association, ABC→D, ABD→C, ACD→B, BCD→A, de support 8 et de confiance ≥ 80%. Pour MIDOVA, sa qualité est nulle, ainsi que celle des quatre règles car ni ce motif ni ces règles n'expriment de liaison complexe (i.e. non triviale) entre les quatre variables.

Inversement, donnons un exemple de motif jugé de très bonne qualité par MIDOVA et non extrait par Apriori. Il n'y en a pas dans le tableau 1, car pour ces données, les seuls motifs jugés de bonne qualité par MIDOVA sont de longueur 2, et de gain positif, correspondant à une interaction positive élevée, et ces deux notions, interaction positive élevée selon MIDOVA et association forte selon Apriori sont très proches quand il n'y a que deux variables. Nous proposons en figure 6 un cas d'école : une interaction fortement négative de niveau 4 qui produit un motif très intéressant pour MIDOVA, le motif ABCD. Ce motif, de support nul, n'est pas extrait par Apriori, alors qu'il extrait tous ses sous-motifs si son seuil de support est inférieur ou égal à 5 par exemple. De l'extraction obtenue par MIDOVA, on déduit qu'il n'y a aucune liaison au sein des groupes de deux variables, ni de trois, et que la

seule liaison non nulle est dans le groupe des quatre variables, correspondant à une forte interaction négative. Elle peut s'interpréter par la règle suivante : si un sujet vérifie trois quelconques de ces variables, il ne vérifie pas la quatrième.

A	B	C	D	Effectifs
1	1	1	1	0
1	1	1	0	5
1	1	0	1	5
1	1	0	0	0
1	0	1	1	5
1	0	1	0	0
1	0	0	1	0
1	0	0	0	5
0	1	1	1	5
0	1	1	0	0
0	1	0	1	0
0	1	0	0	5
0	0	1	1	0
0	0	1	0	5
0	0	0	1	5
0	0	0	0	0

Motif	k	Supp	Amp.Var.	Mgain	Mreste
A	1	20	40	0	20
B	1	20	40	0	20
C	1	20	40	0	20
D	1	20	40	0	20
AB	2	10	20	0	20
AC	2	10	20	0	20
AD	2	10	20	0	20
BC	2	10	20	0	20
BD	2	10	20	0	20
CD	2	10	20	0	20
ABC	3	5	10	0	20
ABD	3	5	10	0	20
ACD	3	5	10	0	20
BCD	3	5	10	0	20
ABCD	4	0	5	-20	0

FIG. 6 – À gauche une interaction négative de niveau 4 et à droite les motifs associés

En pratique MIDOVA produit moins de niveaux qu'Apriori : la valeur de Mr ne peut pas croître, et devient nulle pour les motifs de longueur k>=L avec L tel que $2^{L-1}<=N<2^{L}$. Par exemple si $N<4096=2^{12}$, comme dans l'application relatée dans la section suivante, MIDOVA n'extrait que des k-motifs de longueur k<=12, limite qu'Apriori peut dépasser. Et la longueur maximale des motifs diminue d'autant plus que les liaisons dans les données sont importantes. Si nous nous plaçons dans le cas extrême où toutes les variables sont vraies pour tous les individus, Apriori générera les 2^{p} motifs possibles, alors que MIDOVA ne générera que des 1-motifs ayant tous leur reste à 0. Et en cas de variables toutes identiques sans être toujours vraies, Apriori générera à nouveau tous les motifs possibles, alors que MIDOVA ne construira que les 1 et 2-motifs, ces derniers étant de reste nul. Si un seuil de reste est choisi, plus il est grand, plus le nombre de niveaux a tendance à diminuer. Nous avons fixé un seuil de reste à 10 dans l'application de la section suivante, et la longueur maximale des motifs que nous avons obtenus est de 8.

4 Application à des données socio-économiques

4.1 Description des données

Les données proviennent de l'Enquête Nationale sur la santé de la femme Tunisienne réalisée en 2001 par l'office national de la famille et de la population de la Tunisie (ONFP) et financée par le programme arabe de la santé de la famille. Elle couvre 6691 ménages qui représentent 4346 couples où les femmes sont âgées de 15 à 54 ans. Après une étude préli-

minaire des données, correction des variables manquantes et suppression des variables aberrantes nous avons retenu 4087 couples.

L'enquête est riche en données portant sur les deux membres du couple, avec des informations approfondies et détaillées sur la femme car l'enquête a été réalisée principalement pour étudier des phénomènes liés à la femme. Le questionnaire comporte un nombre important de rubriques dont celle relative au ménage qui apporte des informations sur sa composition, les conditions de vie (type de logement, nombre des chambres, toilettes, source d'eau potable, etc.), les caractéristiques socio-économiques de chaque membre du ménage, la profession, la stabilité du travail, l'âge des conjoints, le niveau d'instruction de chacun d'eux, l'état matrimonial, les résidences et des informations sur les parents et les enfants , etc.

Le mariage est considéré depuis peu comme un des moyens de lutte contre la pauvreté et un moyen de redistribution de richesse (Sigle Rushton et al. 2002). Mais peu d'études économiques et statistiques ont analysé et modélisé la liaison complexe entre la décision du mariage et la pauvreté. Nous utilisons la méthode MIDOVA pour déterminer ce lien dans la société tunisienne.

4.2 Préparation des données pour le traitement

Nous avons repris les 104 variables booléennes telles qu'elles ont été construites (El Haj Ali 2007). Elles sont essentiellement de deux types : 1) la réponse de la personne interrogée à une question de type oui/non a donné une variable codée en 1/0, et 2) la réponse à une question par choix dans une liste a fourni autant de variables que de réponses possibles, la seule variable mise à 1 correspondant au choix de la personne interrogée (i.e. recodage par dichotomisation). Quant à la variable « pauvreté », c'est une variable également booléenne construite à partir du résultat d'une analyse factorielle en composantes principales (ACP) de 23 variables décrivant les conditions de vie des ménages telles que : type du sol, possession de voiture, vélo, climatiseur, source d'eau potable…. Nous avons gardé certaines variables « en double », parfois recodées en sens inverse, ce qui nous a permis de contrôler que les relations significatives entre elles apparaissaient bien comme attendu, avec leur signe et avec le niveau de significativité le plus élevé. Après nettoyage des données (correction par recodage et élimination des valeurs manquantes), nous avons obtenu les valeurs de 4072 individus sur 72 variables binaires. Nous en avons tiré un échantillon de 3300 individus (les 8/10es environ), qui a fourni un tableau booléen complet de 3300 individus et 72 variables et nous avons extrait avec MIDOVA les liaisons complexes entre variables sous forme de k-motifs. Les résultats de cette extraction sont détaillés dans la section suivante, ainsi que l'interprétation des k-motifs les plus significatifs. Puis nous les avons confrontés dans la section suivante aux motifs extraits par MIDOVA sur les données restantes (N=772) afin d'évaluer le pouvoir de généralisation de notre méthode.

4.3 Résultats obtenus et interprétations

Nous avons recopié dans le tableau 2 le nombre de k-motifs extraits de ces données par MIDOVA. Les motifs obtenus ont une longueur k allant de 1 à 8. Le seuil de notre paramètre de reste, Mr a été fixé à 10, mais le support n'a pas été fixé, ce qui fait que les supports de ces motifs peuvent prendre toutes les valeurs comprises entre 0 et 3300. Toutefois dès qu'un k-motif atteint de telles valeurs, son reste est nul et il ne permet pas de construire les (k+1)-motifs (c'est le cas de 4 variables, ou 1-motifs, parmi les 72 de la ligne 1 du tableau, 3 sont

de support inférieur à 10, et une de support 3300). C'est ainsi que les motifs construits par MIDOVA restent raisonnables en longueur et en nombre. Nous décrivons dans la suite de cette section les motifs extraits contenant la variable « pauvreté », et la procédure que nous avons suivie pour choisir les plus significatifs afin de les interpréter. Nous donnons une interprétation détaillée d'un motif de longueur 2, en renvoyant le lecteur intéressé par plus de détails à l'annexe 1, puis d'un motif très significatif de longueur 3, les autres étant en annexe. Nous terminons cette section par l'interprétation d'un 4-motif très significatif (il n'y en plus pour k>4). Nous avons choisi de détailler ce dernier exemple car il montre sur des données réelles le type de relation complexe entre 4 variables qu'extrait la méthode MIDOVA.

Longueur k des motifs	Nombre de Variables		Nombre de k-motifs avec		
	Avant	Après	Mr>=0 (tous)	« pauvreté »	Mr>10
1	72	68	72	1	68
2	68	68	2278	67	1627
3	68	67	22946	830	14945
4	65	65	85801	4305	51357
5	65	64	128160	4615	98263
6	64	59	96075	2189	22155
7	53	42	5715	8	841
8	30	16	39	0	0

TAB. 2 – Les k-motifs extraits des données par Midova

Pour commencer, 67 motifs de longueur 2 contenant la variable pauvreté ont été trouvés. On a procédé pour chaque 2-motif à un test du Chi2 d'indépendance[6] entre la variable pauvreté et l'autre variable du 2-motif. On a indiqué par les notations 2* et -2* le fait que les variables sont liées de façon très significative ($p<0.01$) selon ce test, le signe indiquant le sens du lien, par 1* et -1* quand elles sont liées de façon significative ($p<0.05$, 1* et -1*), et « ns » quand elles ne sont pas significativement liées, ou quand un des effectifs théoriques est inférieur à 5 (comme son bon usage le prescrit). Les 67 variables sont en annexe 1 avec leurs intitulés et leurs numéros d'ordre selon leurs niveaux de significativité (36 significatives, dont 30 très significatives). Nous donnons ci-dessous un exemple d'interprétation du premier lien de significativité -2* de l'annexe 1, qui est un lien négatif très significatif entre la pauvreté et le « milieu urbain », d'autres interprétations se trouvant en annexe à la suite de la liste des 67 variables.

- On observe que la pauvreté est moins importante dans le milieu urbain que dans le « milieu rural ». Ceci s'explique par la multiplicité des possibilités de trouver un travail dans tous les secteurs où les femmes sont mieux rémunérées, alors que celles qui vivent dans les zones rurales travaillent principalement dans l'agriculture où la rémunération des femmes est faible.

[6] Ce test ne figure pas dans la méthode MIDOVA, qui a pour but d'extraire la totalité des liaisons complexes entre les variables des données, sans conditions restrictives sur leurs lois de répartition. Toutefois, pour pouvoir rendre la méthode utile aux chercheurs en sciences humaines, il est nécessaire de se limiter aux liaisons significatives. En attendant la création d'un test adapté à MIDOVA (à l'étude), nous avons choisi le test le moins inadapté aux données de l'application, et il a joué correctement son rôle comme nous le verrons dans la section 4.4.

Parmi les 830 motifs de longueur 3 contenant la variable « pauvreté », seuls 43 indiquent des interactions de niveau 3 significatives (dont 20 très significatives, voir annexe 2) selon un test du Chi2 d'adéquation. Pour réaliser ce test, les 8 effectifs théoriques du tableau de contingence de dimension 3 ont été calculés de telle façon que les effectifs de leurs marges soient conservés et que *l'odds-ratio* (Morineau 1996) du tableau de dimension 3 soit égal à 1. Et la valeur du Chi2 n'a été calculée que pour les motifs dont aucun des effectifs théoriques n'était inférieur à 5. L'interprétation est un peu plus délicate que pour les 2-motifs, car elle porte sur le différentiel entre les deux effets conditionnels et l'effet conjoint, mais elle a pu se faire sans problème pour tous les 3-motifs significatifs. Voici un exemple d'interprétation d'interaction positive très significative (2*, ligne n°5 de l'annexe 2) trouvée entre la pauvreté et deux autres variables : « homme ouvrier spécialisé » et « père ouvrier spécialisé », qui sont chacune liée positivement de façon très significative à la pauvreté, leurs numéros respectifs dans l'annexe 1 étant 10 et 11 ;

- le fait que l'homme soit ouvrier spécialisé augmente la pauvreté de son ménage, et le fait que le père de la femme soit ouvrier spécialisé augmente celle du ménage de la femme. Si ces deux personnes sont mariées ensemble, l'augmentation de la pauvreté de leur ménage est encore plus grande que les deux augmentations réunies. Ce résultat confirme que le mariage des pauvres avec des pauvres aggrave la pauvreté chez cette catégorie.

Nous avons procédé de la même façon pour sélectionner les k-motifs de longueur k>3 contenant la variable pauvreté. Les conditions d'effectifs théoriques supérieurs à 5 (test du Chi2) sont vérifiées par une partie restreinte de ces motifs ; en effet le tableau de contingence croisant les k variables contient 2^k cellules, avec 2^{k-1} cellules correspondant à la valeur positive de la pauvreté, vérifiée par seulement 271 personnes, ce qui ferait en moyenne 33,8 individus par cellule correspondant à la valeur positive de la pauvreté pour k=4 (resp. 16,9 pour k=5 ; 8,4 pour k=6 ; 4,2 pour k=7) s'il y avait équi-répartition pour les autres variables. Il n'est donc pas étonnant que seuls quelques 4-motifs soient significatifs (24 significatifs, dont 10 très significatifs).

	ABCD	ABCnD	ABnCD	AnBCD	nABCD	ABnCnD	AnBCnD	nABCnD	AnBnCD
Eff.ob.	11	12	2	30	24	15	74	57	40
Eff.th.	7,26	15,74	5,74	33,74	27,74	11,26	70,26	53,26	36,26

nABnCD	nAnBCD	AnBnCnD	nABnCnD	nAnBCnD	nAnBnCD	nAnBnCnD	Odds-ratio
52	205	87	104	305	674	1608	14,8
48,26	201,26	90,74	107,74	308,74	677,74	1604,26	1

TAB. 3 – Les 16 cellules d'un 4-motif ABCD très significatif : effectifs observés etthéoriques en cas d'absence de liaison d'ordre 4, et Odds-ratio

Dans le tableau 3, nous avons donné un exemple de 4-motif très significatif : dans la première colonne figurent les intitulés des lignes, dans les 16 suivantes, en première ligne le libellé de chaque cellule, en deuxième ligne les effectifs observés de cette cellule, et les effectifs théoriques en troisième ligne. Les variables sont indiquées par les lettres A, B, C et D, et la lettre n devant la variable indique sa négation. Voici un exemple de lecture de la troisième colonne du tableau : il s'agit de ABCnD, cellule qui correspond aux individus

vérifiant simultanément A, B, C mais pas D. On en a observé 12 parmi les 3300 individus, au lieu des « 15,74 » qu'on aurait pu avoir dans le cas où l'association de ces 4 variables n'aurait pas apporté d'effet supplémentaire à celui obtenu en ajoutant leurs effets individuels ou en association par 2 ou par 3, selon le modèle théorique de MIDOVA. Cette différence de 3,74 individus se retrouve dans chacune des 16 cellules de ce tableau qui n'a qu'un degré de liberté. En dernière colonne figure l'odds-ratio. Il est de 14,8 pour les effectifs observés, ce qui indique un fort effet différentiel du motif ABCD par rapport à ses sous-motifs, et il est de 1 pour les effectifs théoriques, ce qui correspond bien à un effet différentiel nul. Dans ce motif ABCD, les variables sont A : *Pauvreté,* B : *Autre a dépensé pour mariage*, C : *Homme non instruit*, D : *Père artisan/petit commerçant*. Les seuls sous motifs significatifs de ce motif contenant la pauvreté sont AB et AC. L'interprétation est la suivante : une femme qui a épousé un homme non instruit a une grande chance d'être pauvre (AB significatif, effet positif de B sur A), ainsi que celle dont le mariage n'a été payé ni par sa famille ni par celle de son mari (AC significatif, effet positif de C sur A), mais pas celle dont le père est artisan/petit commerçant (AD non significatif, effet nul de D sur A) ; une femme qui vérifie deux de ces trois conditions voit les deux effets correspondants s'ajouter sans effet supplémentaire positif ou négatif (ABC, ABD, ACD non significatifs), mais une femme qui vérifie ces trois conditions simultanément risque d'être encore plus pauvre que ce qu'elle pouvait attendre de l'addition des trois effets (en fait des deux premiers car le troisième est nul).

4.4 Validation des résultats obtenus

Nous avons essayé d'évaluer dans la section précédente la qualité sémantique des meilleurs motifs extraits par MIDOVA, selon une sélection faite à l'aide de deux outils issus des statistiques classiques, l'odds-ratio et le test du Chi2, dont on sait par ailleurs qu'ils ne sont pas très adaptés à la fouille de données (voir note de bas de page n°6 en justification de l'utilisation du test du Chi2). Dans cette partie, nous contrôlons la qualité de ces motifs par une méthode issue de l'apprentissage automatique : les motifs de la section 4.3, qui ont été extraits par MIDOVA d'une partie des données (ensemble d'entraînement « *train* », N=3300), sont confrontés à la partie des données laissée de côté (ensemble « *test* », N=772). Pour cela, nous avons extrait les motifs de l'ensemble *test* selon la procédure MIDOVA avec le même seuil de reste, Mr=10, et nous avons comparé les odds-ratio des motifs obtenus sur l'ensemble *test* aux odds-ratio des mêmes motifs extraits de l'ensemble d'apprentissage « *train* »selon leurs niveaux de significativité (2*, -2*, 1*, -1* et ns).

Les résultats sont tout à fait encourageants comme l'indique la figure 7 représentant les 58 motifs (V1, V2) extraits de l'ensemble *test* dans lesquels V1 est la pauvreté. Les 58 variables V2 font toutes partie des 67 présentes dans les 2-motifs contenant la pauvreté extraits de l'ensemble *train* à l'étape 2 (voir table 2). Pour chacun de ces motifs, nous avons calculé l'odds-ratio dans l'ensemble *test*, et recherché dans les résultats de la section précédente leur odds-ratio dans l'ensemble *train*, ainsi que leur niveau de significativité. Les 58 motifs sont rangés sur l'axe des x par niveau de significativité, puis par ordre lexicographique ; les odds-ratio sont indiqués sur l'axe des y, et à chaque motif correspondent deux points : un losange sur une courbe bleue pour l'odds-ratio sur l'ensemble *train*, un carré sur une courbe rouge pour l'odds-ratio sur l'ensemble *test*. La partie la plus à gauche du graphique (s=2*) contient les 11 motifs indiquant une liaison positive très significative entre V2 et la pauvreté, ayant donc un odds-ratio supérieur à 1 pour l'ensemble *train*. On constate que l'odds-ratio pour l'ensemble *test* est proche de celui de l'ensemble *train*, et toujours supérieur à 1, ce qui indi-

que une liaison positive entre V2 et la pauvreté dans l'ensemble *test*. Dans la partie correspondant à s=-2*, les 19 motifs indiquant une liaison significativement négative entre V2 et la pauvreté sur l'ensemble *train* ont bien tous des odds-ratio inférieurs à 1 pour cet ensemble, et leurs odds-ratio pour l'ensemble *test* sont proches des valeurs pour l'ensemble *train*. Parmi ceux-ci, 17 ont un odds-ratio inférieur à 1 sur l'ensemble *test*, ce qui indique une liaison également négative sur cet ensemble entre V2 et la pauvreté, et 2 ont un odds-ratio supérieur à 1, ce qui indique une relation positive au lieu de la relation négative attendue. Ainsi 28 des 30 motifs jugés très significatifs (2* et -2*) sur l'ensemble *train* ont été retrouvés sur l'ensemble *test*. Pour les motifs de niveau de significativité inférieur, 1* et -1*, le résultat est moins bon (4 sur 6). Quant aux motifs non significatifs (à droite du graphique, odds-ratio sur l'ensemble *train* voisin de 1), ils n'influent pas, de par leur absence de significativité, sur la qualité de généralisation de la méthode. On obtient pour les 2-motifs un total de 32 motifs correctement prédits sur les 36 significatifs, soit 83% au lieu des 95% attendus en moyenne.

Pour les 3-motifs, parmi les 315 extraits de l'ensemble *test* par MIDOVA, on en a retrouvé 32 jugés significatifs sur l'ensemble *train*, dont 17 très significatifs. Parmi ces derniers, 13 sont valides, et en tout 23 sur 32, soit 72%. Pour les 4-motifs, 12 jugés significatifs ont été retrouvés parmi les 472 extraits de l'ensemble test et 9 sont valides, soit 75%.

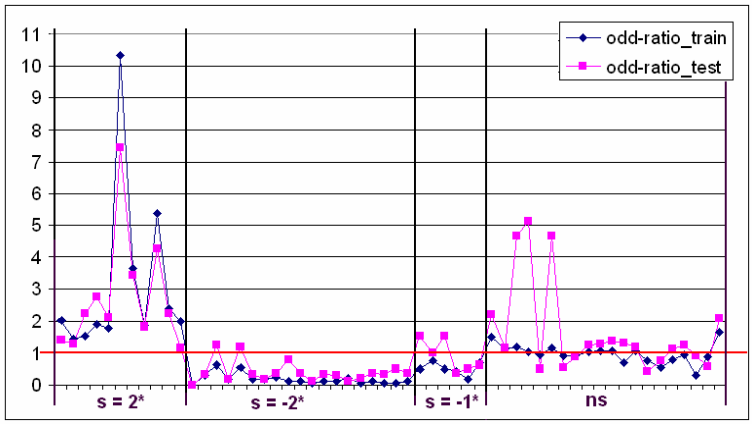


FIG. 7 – Les 2-motifs (V1,V2) de l'ensemble test extraits par MIDOVA, V1 étant la pauvreté ; en abscisse leur rang (ordre : significativité selon l'ensemble d'entraînement puis code de V2), en ordonnée leurs odds-ratio dans les ensembles test et d'entraînement.

Pour conclure, cette confrontation des motifs sur les deux ensembles « *train-test* » montre que notre utilisation de l'odds-ratio et du test du Chi2 sur les données socio-économiques mises à notre disposition nous a permis de sélectionner de façon très convenable les 2-motifs indiquant une liaison complexe significative entre variables, et de façon assez convenable les 3-motifs et 4-motifs significatifs, en attendant un test applicable à tout type de données.

5 Bilan et perspectives

Nous avons présenté dans cet article plusieurs modèles statistiques connus permettant d'exprimer des relations complexes entre variables. Nous avons vu que ces modèles fonc-

tionnent bien quand toutes leurs conditions d'application sont réunies, mais que ce n'est pas souvent le cas pour les données que nous voulons traiter, qui contiennent souvent un nombre important de variables, suivant des lois de probabilité variées. Nous avons ensuite présenté des modèles plus adaptés à ce type de données, qui existent en EGC (Extraction et Gestion de Connaissances), en montrant qu'ils ne permettent pas d'extraire de façon explicite les liaisons complexes comme les interactions de niveau supérieur à 2. Puis nous avons exposé une méthode, appelée MIDOVA, que nous avons construite dans ce but. C'est une variante de la méthode d'extraction de motifs utilisant l'algorithme Apriori. Les motifs que nous avons définis pour MIDOVA sont inspirés des modèles linéaire et log-linéaire des statistiques, et en particulier de l'interaction. Ils diffèrent de par leur nature de ceux d'Apriori : alors que ces derniers sont construits pour trouver les fortes concentrations de valeurs 1 des variables, indiquées par un support élevé, les motifs de MIDOVA sont conçus pour trouver les fortes différences de concentrations de valeurs des variables, indiquées par un gain Mg élevé. La méthode MIDOVA permet, dans tous les cas rencontrés jusqu'ici, d'extraire des motifs moins longs, moins nombreux et plus pertinents. Elle dispose d'un paramètre essentiel, qui est le reste Mr, dont on peut fixer le seuil à 0 pour obtenir l'intégralité des liaisons entre variables, lesquelles permettent de reconstruire le tableau d'origine sans perte d'information (aux permutations des individus près).

Nous avons appliqué notre méthode sur des données de sciences humaines, avec un seuil de reste Mr positif, qui ne permet qu'une reconstruction approximative des données, mais qui élimine le « bruit » des données, c'est-à-dire les liaisons qui s'appuient sur un trop petit nombre d'individus. Puis nous avons sélectionné parmi le grand nombre de motifs obtenus les plus significatifs pour les interpréter. Ils ont tous pu s'interpréter sans problème, mettant au jour les liaisons complexes entre la pauvreté des ménages des femmes mariées de notre étude et les diverses caractéristiques socio-économiques des conjoints. Cela nous permet de conclure que notre méthode MIDOVA a bien extrait, comme attendu, les liaisons complexes, interactions positives comme négatives, entre variables sous une forme explicite. Puis nous avons contrôlé sur le reste des données, mises de côté avant l'application de notre méthode, que sa généralisabilité était convenable, malgré l'utilisation d'outils statistiques imparfaitement adaptés à notre méthode pour décider de la significativité des motifs.

Nous avons montré que notre méthode MIDOVA s'avère utile aux chercheurs en sciences humaines. Il reste à faire un travail complémentaire pour permettre à ces utilisateurs de se l'approprier : l'interprétation des motifs est assez délicate quand leur longueur est supérieure à trois, elle doit être guidée (documentation, tutoriel, etc.). Et il faudrait appuyer le choix des motifs significatifs sur des techniques d'apprentissage automatique (validation croisée par exemple) plus adaptées aux méthodes de fouille de données que les techniques de la statistique classique.

Références

Agrawal R., Mannila H., Srikant R., Toivonen H., Verkamo A.I. Press (1996). Fast discovery of association rules. In Fayyad, U.M. et al., eds., *Advances in Knowledge Discovery and Data Mining*. Menlo Park, California : AAAI Press , MIT Press. pp. 307-328.

Azaïs, J.-M., Bardet, J.-M. (2005). *Le modèle linéaire par l'exemple. Régression, analyse de la variance et plans d'expériences illustrés avec R, SAS et Splus*. Paris. Dunod

Benzécri J.-P. et coll., (1970) *Pratique de l'analyse des données*, Tome 5, Paris, Dunod,

Breslau L., Cao P., Fan L., Phillips G., Shenker S., Web Caching and Zipf-like Distributions : Evidence and Implications, *Proceedings of IEEE Infocom*, NewYork, 1999, p. 126-134.

Cadot M., Lelu A (2010). A Novel Decomposition Algorithm for Binary Datatables: Encouraging Results on Discrimination Tasks. *Fourth IEEE International Conference on Research Challenges in Information Science (RCIS 2010*). p. 57-68.

Cadot, M. (2009). Graphe de règles d'implication statistique pour le raisonnement courant. Comparaison avec les réseaux bayésiens et les treillis de Galois, dans *Analyse Statistique Implicative. Une méthode d'analyse de données pour la recherche de causalités*. Revue RNTI, n° E-16, p. 223-250.

Cadot M. (2006). *Extraire et valider les relations complexes en sciences humaines: statistiques, motifs et règles d'association*. Thèse, université de Franche-Comté.

Dickes P., Kop, J-L. et Tournois J. (1996). Modèles d'équations structurales et sens de la causalité dans les études longitudinales: une application au bien être subjectif. *Bulletin de Méthodologie Sociologique*, n° 50, p. 20-54.

El Haj Ali Dhouha et Zaiem M Hedi (2007). Un test de la théorie du mariage de Becker sur des données tunisiennes. *75ème congrés de l'ACFAS*, Mai 2007, Montreal, Canada.

Fine, J. (1992). Les Modèles Graphiques d'Associations, dans *Modèles pour l'Analyse des Données Multidimensionnelles*, J.J. Droesbeke, B. Fichet et Ph. Tassi éditeurs, ECONOMICA,

Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7: 179-188

Goodman, L.A. (1970), The Multivariate Analysis of Quantitative Data: Interactions Among Multiple Classifications, *Journal of the American Statistical Association*, 65, 226-256.

Gras R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*, Thèse, univ. Rennes I.

Guigues J.L. et Duquenne V. (1986) Familles minimales d'implications informatives résultat d'un tableau de données binaires, *Math. Sci. Hum.* n°95, pp. 5-18

Guillet, F. (2004). Mesure de qualité des connaissances en ECD. *EGC 2004*, France.

Han J. and Kamber M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco.

Joreskog K.G et Sorbom D. (1984) *Lisrel VI, user's guide*, 3ème édition, Mooresville, IN : Scientific Software. 1984

Lenca P., Meyer P., Picouet P., Vaillant B., Lallich S. (2003). Critères d'évaluation des mesures de qualité en ECD, *JS 2003, Proceedings* , Lyon, pp. 647-650.

Leray P., Francois O., (2004) Réseaux bayésiens pour la classification, Méthodologie et illustration dans le cadre du diagnostic médical, *Revue d'intelligence artificielle, RSTI série RIA*, Vol 18, no 2/2004, Lavoisier, Paris, , p. 168-193

Lerman I.C.,(1995) Rôle de l'inférence statistique dans une approche de l'analyse classificatoire des données, *Méthodes d'analyses statistiques multidimensionnelles en didactique des mathématiques*. IUFM de Caen,

Morineau, A., Nakache, J.-pp, Krzyzanowski, C. (1996), *Le modèle log-linéaire et ses applications*, Cisia-Ceresta, Paris.

Sigle Rushton. W., McLanahan, S. (2002), Richer or poorer ? Marriage as an antipoverty strategy in the United States. *Population*, 57(3); p. 519-538;

Snedecor G.W. et Cochran W.G. (1966), *Méthodes Statistiques*, Association de coordination technique agricole, Paris.

Suzuki E., Kodratoff Y., (1998) Discovery of Surprising Exception Rules Based on Intensity of Implication. *Second European Symposium on Principles of Data Mining and Knowledge Discovery*. .Springer-Verlag, London, LN In Computer Science. , p. 10-18.

Vapnik, V. (1995) *The Nature of Statistical Learning Theory*. Springer-Verlag

Winer, B.J., Brown D.R., and Michels, K.M. (1991*). Statistical principles in experimental design* 3rd ed. New York: McGraw-Hill.

Whittaker J., (1990), *Graphical models in applied multivariate Statistics*, John Wiley & Sons

Annexe 1 : 2-motifs (pauvreté et une autre variable V2)

Liste des variables V2 numérotées de 1 à 67 ordonnées par niveau de significativité

2* : 1) Dépenses du mariage Dm : autre a dépensé pour mariage, 2) Dm de la famille du mari, 3) Enfants décédés, 4) Derni. naiss pd 5DA, 5) A connu son mari dans famille, 6) Femme n'a aucun niveau d'instruction, 7) Homme non instruit, 8) Femme ne travaille pas, 9) Homme ne travaille pas, 10) Homme ouvrier spécialisé, 11) Père ouvrier spécialisé

-2* : 12) Milieu urbain, 13) A travaillé avant, 14) Dm : Pere/mere/titulaire/frere, 15) Dm : elle-même, 16) Dm : mari sans crédit, 17) Dm : mari avec crédit, 18) Connaît Sida, 19) Niveau d'instruction de la femme (Nif) : primaire, 20) Nif 2eme cycle primaire, 21) Nif technique 22) Nif secondaire, 23) Nif universitaire, 24) Niveau d'instruction de l'homme (Nih) : technique, 25) Nih secondaire, 26) Nih universitaire, 27) Femme cadre supérieur, 28) Homme cadre moyen, 29) Homme cadre moyen, 30) Père cadre moyen

-1* : 31) Nombre de mariages : une seule fois, 32) Grossesses perdues, 33) nombre de mariage de la femme est 1, 34) Homme de niveau 2eme cycle de base, 35) Femme ouvrière spécialisée, 36) Homme ouvrier,

ns : 37) Dm d'une autre source 38) Homme de profession libérale/ grand agriculteur 39) A participe au programme d'illettré, 40) Dm de sa famille à elle, 41) Ne sait pas, 42) Époux a été marie auparavant, 43) Naissances vivantes, 44) Enfants vivant a la maison, 45) Enfants vivant ailleurs, 46) Au moins une naissance, 47) Entendu parler d'autre MST, 48) Nombre de mariages de l'époux est égal à 1, 49) Ses parents habitent chez elle, 50) Ses beaux-parents habitent avec elle ?: Elle habite seule, 51) Elle habite seule, 52) Homme de niveau d'instruction primaire, 53) Femme ouvrière, 54) Femme artisan/petit commerçant, 55) Femme cadre moyen, 56) Femme profession libérale/grand agriculteur, 57) Homme

artisan /petit commerçant, 58) Père ne travaille pas, 59) Père ouvrier, 60) Père artisan/petit commerçant, 61) Père cadre supérieur, 62) Mère ne travaille pas, 63) Mère ouvrière, 64) Mère ouvrière spécialisée, 65) Mère artisan/petit commerçant, 66) Mère cadre moyen, 67) Mère cadre supérieur

Voici les interprétations qui peuvent en être tirées en termes d'économie de la famille :

5 : Plus le lien de parenté est important plus le taux de pauvreté est élevé, ceci s'explique par l'appartenance des deux membres du couple à la même catégorie sociale. En effet, dans la société tunisienne, les petites familles et « les familles mères » découlant de même racine ont pratiquement le même niveau de vie, et comme cette variable est reliée positivement et fortement à la pauvreté, cela confirme que le mariage des pauvres avec les pauvres aggrave encore la pauvreté des ménages.

6 à 10 : On constate que les ménages composés par des couples non instruits ou chômeurs ou de basse qualification professionnelle (ouvriers) sont pauvres. Un tel résultat est attendu puisque l'éducation est un critère primordial pour occuper une profession et la profession représente la source principale du revenu ou de la richesse de l'individu ou du ménage.

11 : Nous remarquons que le capital initial de la femme (capital de sa famille) a un impact sur ses conditions de vie après le mariage. Plus le capital initial est important, plus la vie de la femme est confortable et inversement. Ce résultat s'explique par deux points : 1- la famille intervient souvent pour aider leur fille (après le mariage) financièrement et socialement, ce résultat est confirmé par la variation de la variable w416A. 2- la fille de capital initial important est plus demandée sur le marché du mariage tunisien (El Hadj Ali 2007), cette compétition lui permet de choisir l'homme avec qui elle maximise son utilité et qui lui garantit une vie confortable.

19 à 29 : Plus la femme et son mari sont éduqués, plus ces couples vivent au dessus de la pauvreté, ceci s'explique par la forte relation positive entre le niveau d'éducation et la profession occupée. Les couples composés des femmes participant au marché de l'emploi et des maris non instruits sont pauvres. Ceci s'explique par la composition du marché du mariage tunisien (El Hadj Ali 2007) où les riches s'apparient avec les riches et les pauvres s'apparient avec les pauvres (la richesse est mesurée en terme de niveau d'éducation). En effet, si le mari est non instruit, sa femme est très probablement non instruite ou de niveau primaire et suite à son niveau d'études elle occupe une profession de faible revenu. Par conséquent le ménage composé par de tels membres est forcément pauvre.

62 à 67 : la situation professionnelle de la mère de la femme n'a pas d'impact sur la pauvreté ou non de la femme.

Annexe 2 : 3-motifs (pauvreté et deux autres variables V2, V3)

Les 3-motifs (V1, V2, V3) ci-dessous sont très significatifs ($p<.01$, test du chi2), ce qui est indiqué en deuxième colonne par 2* ou -2* selon que l'interaction entre les 3 variables, conditionnellement aux liaisons entre les variables prises 2 à 2, est positive ou négative. Ils sont formés de 3 variables V1, la pauvreté, et V2 et V3 dont les libellés respectifs sont indiqués en troisième et quatrième colonne. Dans la cinquième colonne, on a indiqué le niveau de significativité de la liaison entre V1 et V2 correspondant au 2-motif (V1, V2), et dans la dernière colonne celui du 2-motif (V1, V3).

Liaisons complexes entre variables

n°	Sgn3	V2	V3	Sgn2	Sgn2
1	2*	Femme a travaillé avant	Homme non instruit	-2*	2*
2	-2*	Homme ouvrier spécialisé	Père artisan/petit commerçant	2*	ns
3	-2*	Femme a travaillé avant	Mère ne travaille pas	-2*	ns
4	-2*	Enfants vivant ailleurs	Homme de niveau primaire	ns	ns
5	2*	Homme ouvrier spécialisé	Père ouvrier spécialisé	2*	2*
6	-2*	Lieu où a connu son mari : famille	Homme de niveau primaire	2*	ns
7	-2*	Homme non instruit	Femme ne travaille pas	2*	2*
8	-2*	Femme n'a aucun niveau d'instruction	Femme ne travaille pas	2*	2*
9	-2*	Père/mère/titulaire/frère a dépensé pour mariage	Autre a dépensé pour mariage	-2*	2*
10	2*	Femme ne travaille pas	Mère ne travaille pas	2*	ns
11	-2*	Homme artisan /petit commerçant	Père ouvrier spécialisé	ns	2*
12	2*	Enfants décédés	Connaît SIDA	2*	-2*
13	-2*	Homme non instruit	Homme ouvrier spécialisé	2*	2*
14	-2*	Grossesses perdues	Femme ne travaille pas	-1*	2*
15	2*	Homme ouvrier	Père artisan/petit commerçant	-1*	ns
16	-2*	Homme ouvrier	Père ouvrier spécialisé	-1*	2*
17	2*	Femme a travaillé avant	Femme n'a aucun niveau d'instruction	-2*	2*
18	2*	Enfants décédés	Homme artisan /petit commerçant	2*	ns
19	2*	Homme artisan /petit commerçant	Père artisan/petit commerçant	ns	ns
20	-2*	Enfants décédés	Homme de niveau primaire	2*	ns

Summary

We deal in this paper with a particular type of complexity, i.e. the one underlying the links between qualitative variables. Some statistical models have been designed for handling a few aspects of this complexity: the general linear model (GLM) gets to account for interesting types of complexity such as interactions of various orders, negative links, and contrasts. But these methods badly fit to the case of a large number of variables, and they constrain the user to explicit beforehand the links at work. We present MIDOVA, a method for mining the same type of relations as GLM directly from the data, without any constraining hypotheses, while being compatible with a large number of variables. We illustrate this method applying it to data extracted from the PAPFEM survey achieved in 2001 by the Tunisian Office National de la Famille et de la Population; we have brought to light the specially complex link between the household poverty and the socio-economic position of the husband and wife.