COMMENT EXTRAIRE DES CONNAISSANCES A PARTIR DES CONCEPTS DE VOS BASES DE DONNEES ?

LES DEUX ETAPES DE L'ANALYSE DES DONNEES SYMBOLIQUES.

E. Diday

Université Paris-Dauphine

diday@ceremade.dauphine.fr

Résumé

Vos bases de données contiennent des concepts sous-jacents. Ils sont associés aux catégories issues de produits cartésiens de variables qualitatives ou de classifications automatiques. Ces concepts constituent alors des unités d'étude d'un niveau de généralité supérieur aux données initiales. Ce niveau est souvent désiré par les utilisateurs mais freinés par le carcan des données classiques qui ne tiennent pas compte de la variation des instances de ces concepts. L'analyse des données symboliques (ADS) a pour objectif dans une première étape de constituer ces concepts et de les décrire en prenant en compte leur variation interne par des variables dites « symboliques » (à valeur intervalle, histogramme, lois etc.) car non manipulables comme des nombres. La seconde étape d'une ADS consiste à les analyser. Pour cela on est amené à étendre les méthodes de la statistique exploratoire et de la fouille de données aux données symboliques (ces méthodes deviennent alors des cas particuliers d'ADS) et de développer des outils nouveaux spécifiques. On montre que ces données ne peuvent pas être réduites à des données classiques. On décrit les quatre espaces de la modélisation sous-jacente où les concepts sont modélisés par des objets symboliques, puis la modélisation mathématique des données (sous forme de variables à valeur variable aléatoire) et des classes ainsi que de leur structure en généralisant les treillis de Galois, hiérarchies, pyramides classiques aux données symboliques. On introduit leur classification spatiale étendant les cartes de Kohonen à des données et des structures pyramidales plus riches. On termine enfin par une application industrielle et la présentation du logiciel SODAS issu de deux projets européens d'EUROSTAT.

Mots clés : Data Mining, fouille de données, analyse des données, statistique descriptive, analyse des données exploratoire, données symboliques, classification automatique, analyse factorielle, treillis de Galois stochastiques, pyramides, analyse de concepts, classification spatiale.

1) Introduction

Les progrès de la technologie informatique dans le recueil et le transport de données, font que dans tous les grands domaines de l'activité humaine, on recueille maintenant des données en quantité souvent gigantesque et de toutes sortes (numériques, textuelles, graphiques,...). Partout dans le monde, il se constitue ainsi des gisements de connaissances considérables. Résumer ces données, à l'aide de « concepts » sous-jacents, afin de mieux les appréhender et d'en extraire de nouvelles connaissances en réduisant la taille des données et en considérant les unités statistiques au niveau de généralité désiré par l'utilisateur, constitue une question cruciale et abordable grâce à Analyse de Données Symboliques (ADS). Toute ADS est basée sur une modélisation du monde réel supposé constitué d'individus et de concepts. Un concept est défini par une "intension" et une "extension". L'intension est un ensemble de propriétés caractéristiques du concept, l'extension est l'ensemble des individus appelés « instances » du concept qui satisfont ces propriétés. Par exemple, le concept nommé « assuré social n° 153 » a pour extension l'ensemble des feuilles de maladies envoyées à la sécurité sociale sur une période donnée par cet assuré; le concept « crédits d'une région » a pour extension l'ensemble des usagers d'une banque ayant demandé un crédit dans une région. Le concept « trajectoire de patient » a pour extension l'ensemble des patients ayant eu un même parcours dans les hôpitaux d'une région (voir Touati et al. (2006)). Les individus sont modélisés dans un espace de description qui exprime leurs propriétés à l'aide des variables (certains disent aussi "attributs") qui les caractérisent (leur âge, leur taille,…). Une catégorie est une modalité d'une variable qualitative ou d'un produit cartésien de telles variables. Une classe d'individus formée d'instances d'un concept, est modélisée dans l'espace des descriptions à l'aide d'un opérateur de généralisation des descriptions des individus qui la constituent (l'intervalle de variation, l'histogramme ou loi de probabilité de leur âge, de leur taille etc.).

Le premier grand principe de l'ADS consiste à considérer que dans toute base de données et dans tout tableau de données il existe des concepts qu'il suffit de dégager pour en faire des unités statistiques d'étude au même titre que les individus habituelles mais cette fois munis de description tenant compte de la variation des instances de ces concepts. En effet, toute variable quantitative peut être transformée en variable qualitative dont les modalités (ou catégories) peuvent être associés à des concepts. Plus généralement, chaque catégorie d'un produit cartésien de variables qualitatives peut être associée à un concept. Ces concepts sont alors décrits par les autres variables transformées en variables dites « symboliques » car munies de la variation induite par les instances de chaque concept, ces instances étant constituées des individus de la catégorie associée au concept.

La conséquence pratique de ce principe est que la première étape d'une ADS consiste à transformer la table initiale des individus décrits par des données classiques en une table de concepts décrits par des variables symboliques exprimant leur variation interne par des intervalles, des lois, des courbes etc. Plus précisément, la première étape d'une ADS consiste à former puis décrire des concepts soit à partir de catégories (ou combinaison logique de catégories) fournies par un expert du domaine, soit à partir de classes construites par classification automatique ou toute autre méthode fournissant des classes. Par exemple, si 200 personnes habitent une région donnée et font partie d'une ethnie particulière, on peut associer à ces personnes un concept (par exemple, « norvégiens ») puis le caractériser par l'intervalle interquartile ou l'âge minimum et maximum, le diagramme de fréquence du sexe, l'ensemble des types d'emplois, de scolarité etc. Si l'on définit un opérateur de calcul de l'extension de ce concept à partir de ses propriétés, il devra induire au moins ces deux cents personnes.

Le second grand principe de l'ADS consiste à considérer un ensemble de concepts, comme des individus de niveau supérieur, en prenant en compte leur variation interne. Cette variation est exprimée sous forme d'une statistique propre, provenant par exemple de données répétées dans l'espace ou dans le temps. Quand la variation n'est pas prise en compte, on se trouve dans le cas des données classiques. Il en résulte, que selon ce second principe toute méthode classique de statistique, d'Analyse des données ou de Data Mining sur des données classiques doit pouvoir être étendue au cas de données symboliques pour enrichir ainsi le champ des méthodes d'ADS et en devenir un cas particulier.

Comme conséquence pratique de ce second principe, la seconde étape d'une ADS consiste à analyser le tableau de données symboliques issu de la première étape en étendant au moins les méthodes classiques aux données symboliques. On parle alors de « Knowledge Mining » (Bock, Diday 2000, Diday, Noirhomme (2007)) et de « Conceptual Statistics » (Billard, Diday (2006). Ainsi, dans sa seconde étape, l'ADS n'a pas pour but d'analyser une sorte de données complexes mais au contraire d'analyser des concepts décrivant la variation interne de leurs instances décrites par des données de toutes sortes qui peuvent être complexes, imprécises, incertaines, floues,… . Par exemple, la description d'une image peut être très complexe et peut être analysée par une analyse des données classiques une fois que ces données sont transformées en données classiques. De façon complémentaire, une fois que les images ont été transformées en données classiques l'ADS peut alors avoir pour objectif de décrire d'abord dans sa première étape, des types d'images (représentant la mer, la montagne, la forêt etc.) considérés comme concepts, sous forme de données symboliques puis de les analyser dans sa seconde étape.

Le troisième grand principe consiste à considérer que toute sortie d'une ADS s'exprime en termes de données symboliques et peut être utilisée comme entrée d'une ADS de niveau supérieur. Il consiste à considérer que les résultats obtenus doivent eux-mêmes s'interpréter en termes de données symboliques ou "d'objets symboliques" (définis au paragraphe 5) pour modéliser les classes de concepts, autrement dit, dans des termes plus riches que ceux utilisés en AD classique tout en restant intelligibles par l'expert puisque ce sont ceux qu'il a utilisé en entrée.

Depuis les premières publications annonçant les principes de l'ADS (Diday 1987 a, 1987 b, 1989, 1991), beaucoup de travaux ont eu lieu. On peut d'abord mentionner des ouvrages collectifs Bock and Diday (Springer, 2000), Diday, Noirhomme (Wiley, 2008), un livre qui se veut pédagogique avec de nombreux exercices Billard, Diday (Wiley, 2006) ainsi que plusieurs articles synthétiques Diday (2000, 2002 a, 2005), Billard, Diday (2003), Diday, Esposito (2003). L'ADS doit étendre de façon inéluctable les outils de l'Analyse des données classiques et du Data Mining (voir par exemple, Tukey (1958), Benzécri (1973), Diday et al. (1984), Saporta (2006), Lebart et al. (2006), Mirkine (2005), Larose, Vallaud (2005), Han, Kamber (2006), P.B. Cerrito (2007)) sur des unites statistiques de plus haut niveau (ou « concepts ») décrits par des variables dites symboliques car elles prennent en compte la variation interne de ces nouveaux type d'unités. Par exemple les notions classiques de moyenne, variance, corrélation etc. sont étendues à ce nouveau type de variables dans: De Carvalho (1995), Bertrand and Goupil (chapter 6 dans Bock Diday (2000)), Billard, Diday (2003), Billard (2004), Billard and Diday (2006), Gioia and Lauro (2005). De même l'Analyse en composantes principales a été étendue dans Cazes, Chouakria, Diday, Schektman (1997), Lauro, Verde, Palumbo (2000), Irpino, Verde, Lauro (2003), Irpino (2006), Lauro, Gioia (2006). Le multidiensional scaling dans Groenen et al. (2006). L'extension des dissimilarities aux données symboliques peut être trouvée dans Gowda et al (1991), Bock, Diday (2000, chapitre 8 de Esposito et al.), ainsi que dans Esposito, Malerba , Semerano (1991, 1992), Malerba, Esposito, Monopoli (2002), Diday, Esposito (2003), Bock (2005). Concernant l'extension de la classification automatique aux données symboliques on trouvera De Souza, De Carvalho (2004), Bock (2005), Diday,

Murty (2005), De Carvalho, De Souza, Chavent, Lechevallier (2006), De Carvalho, Brito, Bock (2006). Pour le problème de la détermination du nombre de classes on pourra se reporter à Hardy (2005) ou à son chapitre dans Diday, Noirhomme (2008) où l'on trouvera aussi un chapitre de P. Bertrand sur la stabilité des classes (voir aussi, Bertrand P., Bel Mufti (2006)). Pour les arbres de decision on pourra se reporter à Ciampi et al. (2000), Bravo et al. (2000), Bravo (2001), Mballo et al. (2006), Limam et al. (2004). Pour l'extension des treillis de Galois aux données symboliques, on peut citer Diday(1991), Diday, Emilion (1997, 2003), Pollaillon (1998), Brito, Polaillon (2005). En classification hiérarchique et pyramidale voir Brito (2002), Diday (2004), Diday (2008). En discrimination voir les articles de Duarte Silva, Brito (2006), Appice et al (2006). En régression symbolique, voir Rodriguez (2000), Afonso, Billard, Diday (2004), De Carvalho et al. (2004). En décomposition de mélanges de vecteurs de distributions Diday (2001), Diday, Vrac (2005), Cuvelier et al. (2005). En extraction de règle et extension de l'algorithme apriori on peut citer au moins Afonso, Diday (2005), pour la sélection de variables voir par exemple Ziani (1996). En visualisation de données symboliques, Noirhomme-Fraiture (2002), Irpino et al (2003). En séries chronologiques Prudêncio et al. (2004), en météorologie Vrac et al (2004), en Web Mining, on peut citer Caruso et al (2005), Da Silva et al (2006), Meneses, Rodríguez-Rojas (2006).

Cet article est organisé de la façon suivante. On distingue d'abord en section 2, les données décrivant les individus, des connaissances décrivant des concepts. On explique ensuite le processus de réification qui consiste à transformer les individus et concepts en unités appelées individus du premier et second ordre qui peuvent être décrits de façon exhaustive par restriction des descriptions des individus et des concepts du monde. On termine cette section par un exemple de construction d'un tableau de données symbolique à partir d'un tableau de données classiques qui introduit ainsi à la section 3 suivante. Cette section 3 est consacrée à la première étape d'une ADS. On montre d'abord comment de façon générale, on obtient un fichier de données symbolique à partir d'une base de données classique. Ce fichier a le même format que les données symboliques d'entrée et peut être utilisé pour une ADS de niveau supérieure. On discute ensuite de la façon de conserver des informations perdues lors du passage des données initiales portant sur les individus aux données symboliques portant sur les concepts. On montre en particulier comment conserver la corrélation et même l'expliquer à l'aide de données symboliques. On montre ensuite comment on peut par l'approche symbolique concaténer des tables qui n'ont ni mes mêmes unités statistiques, ni les mêmes variables pour les décrire sauf une variable commune induisant les mêmes catégories. On montre enfin comment on peut étendre les tableaux de contingences classiques à des tableaux de contingences symboliques où des histogrammes apparaissent dans les cases plutôt que des fréquences. La section 4 est consacrée à la seconde étape de l'ADS qui consiste à faire l'analyse du tableau des données symboliques obtenu lors de la première étape. Dans cette section, on explique d'abord pourquoi on n'a pas intérêt à considérer uniquement le tableau de données classique induit du tableau de données symbolique en transformant par exemple, les variables intervalles en deux variables (celle des min et celle des max) ou les variables à valeur histogramme en autant de variables numériques qu'il y a de modalités dans l'histogramme. On montre ensuite que la statistique des individus n'est pas la statistique des concepts et on en déduit que les deux modes d'analyse sont complémentaire. Autrement dit que l4analyse des Données ou le Data Mining classique ne sont ni meilleur ni moins intéressants qu'une ADS, les deux approches sont simplement différentes et complémentaires. On termine cette section en donnant la liste actuelle des méthodes classiques qui ont été étendues au cas symbolique puis des méthodes symboliques en perspective.

La section suivante concerne la modélisation des concepts par des objets symboliques. On décrit d'abord le schéma à quatre espaces qui montre comment on modélise des individus des classes d'individus et des concepts du monde respectivement par des descriptions classiques, des descriptions symboliques et des « objets symboliques ». On décrit ensuite un processus d'apprentissage permettant d'améliorer la modélisation des concepts par des objets symboliques. On termine cette section en présentant « le cibleur » qui est à la fois un moyen graphique pour représenter un objet symbolique et un moyen de mesurer sa qualité par l'écart de son volume par rapport à celui d'un cône. Dans la section 6 on donne plusieurs façon de modéliser les données et les classes de concepts en considérant qu'un tableau de données symboliques est une réalisation d'un tableau dont les variables sont à valeur « variables aléatoires » contrairement aux tableaux de données classiques où les variables sont à valeur numérique ou qualitative. On présente deux formes de modélisations stochastiques, celle basée sur les capacités et l'autre basée sur un modèle de copule. On s'intéresse ensuite à différentes structures de classes : celles des treillis de Galois bien adaptée à une modélisation stochastique par des capacités puis celle des partitions obtenues par décomposition de mélange de distributions de distributions. On propose ensuite une nouvelle structure de classe, celle des « pyramides spatiales » basées sur une grille et dont les nœuds peuvent être associés à un concept modélisé par un objet symbolique. Dans la section 7, on s'intéresse au champ d'application de l'ADS qui est très vaste puisqu'il concerne tous les domaines où des données sont recueillies et doivent être étudiées pour mieux les comprendre et améliorer les performances décisionnelles. Un exemple d'application industrielle concernant des anomalies à détecter sur un pont suite au passage de TGV illustrant bien l'intérêt de plusieurs méthodes d'ADS est enfin présenté. La dernière section concerne la présentation succincte du logiciel SODAS issu de deux projets européens sur l'ADS. La conclusion résume les grandes lignes de cet article, donne quelques principes de « moralité » et les perspectives de recherche. On termine enfin par la forme de diffusion actuelle de l'ADS.

2) Des individus aux concepts : des données aux connaissances

2.1 Données décrivant des individus et connaissances décrivant des concepts

Nous distinguons les données (au sens classique), des connaissances. Les « données classiques» sont des grandeurs ou des qualités décrivant des entités du monde appelées « individus ». Par exemple, on peut décrire une personne considérée comme un « individu » décrit par le fait qu'elle pèse « 90 kilos » et qu'elle est « grande ». Par contre, les «connaissances » sont des informations d'ordre intensionnel décrivant des entités du monde appelées « concepts » qui portent sur des individus du monde nommés « instances », en les caractérisant par des propriétés communes appelées « intension » exprimant la variation des instances. L'ensemble de ses instances forment « l'extension » d'un concept. Par exemple, une chaîne de magasins ayant pour instance l'ensemble de ses magasins peut être décrite par le fait que le nombre de pulls de marque Rodier vendus dans une période donnée a varié entre 87 et 363 parmi ses magasins et que le pourcentage de pulls bleus et rouges vendus a été respectivement de 30% et 45%. Ces données qui prennent en compte la variation seront dites « symboliques » car elles ne peuvent pas être traitées comme des nombres et contiennent donc le cas particulier des données purement qualitatives ou quantitatives. Toute Analyse de Données Symboliques (ADS) est basée sur une modélisation du monde supposé constituée d'individus et de concepts. On peut illustrer ces notions par beaucoup d'autres exemples : une assurance et ses assurés, un hôpital et ses patients, une région et ses habitants recensés etc.

2.2 La réification des individus et des concepts en individus du premier et du second ordre

En Analyse ou Fouille des Données, on part d'entités du monde réel appelées sujets (être libres, conscients, responsables) ou objets (choses du monde réel ou virtuel, le contraire des sujets) que l'on appelle « individus » qu'ils soient sujets ou objets. On considère également que les concepts définis par une intension et une extension formées d'individus constituent des entités du monde réel ou virtuel. On remarque alors qu'aucune description exhaustive de tels individus ou concepts n'est possible : toute définition partielle est destructrice (Rimbaud remarque joliment que « je est autre »). La réification permet de sortir de ce dilemme en transformant les individus ou les concepts du monde réel ou virtuel en « objets d'étude » munis d'une description exhaustive.

On débouche ainsi sur des individus réifiés appelés ici « individus du premier ordre » décrits de façon exhaustive par des données classiques définies par des variables numériques ou qualitatives. Bien sûr ces descriptions sont très restrictives de la véritable nature des individus initiaux du monde réel ou virtuel. Cette première réification se fait dans la pratique de façon tellement standard en Analyse ou fouille de données qu'il n'en est pratiquement jamais question.

Par contre, la réification des concepts constitue l'une des originalités de notre approche. En effet, le procédé de réification qui s'applique de façon habituelle aux individus peut s'appliquer aussi aux concepts en leur associant des « individus de second ordre » décrivant restrictivement les concepts au moins par les mêmes variables que les individus de premier ordre mais en tenant compte cette fois-ci de la variation des valeurs prises par les instances de chaque concept.

2.3 La description des individus du second ordre

Ces individus de second ordre, considérés comme nouvelles unités statistiques, sont décrits par des données plus complexes que celles habituellement rencontrées en statistique. Elles sont dites « symboliques », car en exprimant la variation interne inéluctable des concepts (due à leurs instances) et en tenant compte de connaissances supplémentaires au niveau des concepts, elles ne peuvent pas être décrites uniquement par des variables numériques ou qualitatives. En effet, elles nécessitent l'utilisation de variables à valeur intervalle, histogramme, lois, ensemble de valeurs parfois pondérées et munies de règles et de taxonomies. Dans ce contexte, pour analyser un ensemble d'individus de second ordre on est conduit au moins à étendre à de telles données, les méthodes de « l'Analyse des Données Exploratoires » et plus généralement, de la « Statistique Multidimensionnelle » ou de la Fouille de Données. Cependant, avant d'arriver à la seconde étape d'analyse il faut d'abord passer par une première étape qui consiste à construire le tableau de données symboliques qui décrit les individus de second ordre. Ces deux étapes sont décrites dans les paragraphes suivants.

Par la suite, pour simplifier on parlera parfois simplement d'individus et de concepts plutôt que d'individus du premier et du second ordre qui les représentent en les réifiant. Notons cependant que bien qu'il y ait une application bijective entre les individus (resp. les concepts) et les individus du premier ordre (resp. du second ordre), ces bijections n'existent pas avec les descriptions des individus du premier ordre (resp. du second ordre) car dans un cas comme dans l'autre ces descriptions (restrictives de la réalité du monde où sont plongés les individus et les concepts), peuvent être identiques pour deux individus (resp. concepts) différents.

2.4 Un exemple de construction d'un tableau de données symboliques à partir d'un tableau de données classiques

Pour fixer les idées, prenons un exemple tiré d'une étude faite à la MSA (Mutuelle Sociale Agricole) voir (Diday, Pelc, Wagne (2004)). On dispose de feuilles de maladies qui indiquent pour chaque assuré bénéficiaire, les médicaments appelés « occurrences » qu'il a achetés durant une période donnée. On dispose de plus de la date de la première année où le remboursement a été accepté par la CNAM (Caisse Nationale d'Assurance Maladie) ainsi que le code de prise en charge du médicament parmi beaucoup d'autres informations. Dans le tableau de données classiques de la figure 1, les « individus de premier ordre » sont les numéros d'occurrence, dans les colonnes suivantes on a le numéro des bénéficiaires suivie de l'année de remboursement et du code de prise en charge.

Occurrences	Bénéficiaire	Année Rembour (intervalle)	Prise en Charge (diagramme)
111111	236	1996	21
111112	236	1996	31
111113	236	2002	31
111114	362	1995	1
111115	362	1996	21
111116	235	1994	1
111117	235	2000	31

Figure 1 Les individus de premier ordre sont les occurrences, les concepts sont les bénéficiaires.

Bénéficiaire	Année Rembour (intervalle)	Prise en Charge (diagramme)	Age
236	[1996,2002]	21(33.3), 31(66 ,6)	72
362	[1995,1996]	1(50%), 21(50%)	85
235	[1994,2000]	1(50%), 31(50%)	65

Figure 2 Les individus de second ordre sont les bénéficiaires

La MSA voudrait étudier le comportement des bénéficiaires réifiés en « individus de second ordre ». Par application du premier principe, on obtient ainsi le tableau de données symboliques de la figure 2 où les lignes sont associées aux bénéficiaires et les colonnes sont des variables symboliques. L'année de remboursement est devenue une variable à valeur intervalle exprimant le min et le max des années de remboursement pour chaque bénéficiaire, la prise en charge est une variable à valeur histogramme exprimant le pourcentage de chaque code pour chaque bénéficiaire. Dans le paragraphe suivant, nous allons voir comment de façon plus précise, le premier principe est mis en pratique.

3) Première étape d'une ADS : des données classiques au tableau de données symboliques

3.1 Construction d'un fichier de données symboliques à partir d'une base de données

Tout logiciel d'Analyse des Données Symboliques doit avoir un module qui permet de passer des données initiales d'une base de données décrivant des individus de premier ordre aux données symboliques décrivant des individus de second ordre sous forme d'une base de données réutilisable. Dans le logiciel SODAS ce module s'intitule DB2SO (voir Stéphan 1998) et le chapitre 2 de Lechevallier et al dans Diday, Noirhomme (2008)) qui veut dire « from Data Base to Symbolic Objects ». Il produit à partir d'une base de données classique un fichier XML ou SDS (Symbolic Data System) qui a le même format que les données symboliques d'entrée et qui peut être à son tour utilisé dans une ADS de plus haut niveau (appliquant ainsi le troisième grand principe de l'ADS énoncé dans l'introduction). Nous verrons plus bas que les objets symboliques sont un moyen de modéliser les concepts. Les différentes étapes pour parvenir à la construction du tableau de données symboliques sont décrites dans la figure 3.

D'abord, on part d'une base de données relationnelles pour construire une table dont les lignes décrivent les individus de premier ordre par des variables classiques, cette table contient en seconde colonne pour chaque ligne le nom du concept dont l'individu associé à cette ligne est une instance. Les autres colonnes sont des variables descriptives de ces individus. A partir de cette table le module DB2SO construit la table des individus de second ordre qui portent le nom d'un concept et qui sont décrits par des variables symboliques prenant en compte la variation comme dans le tableau de la figure 2.

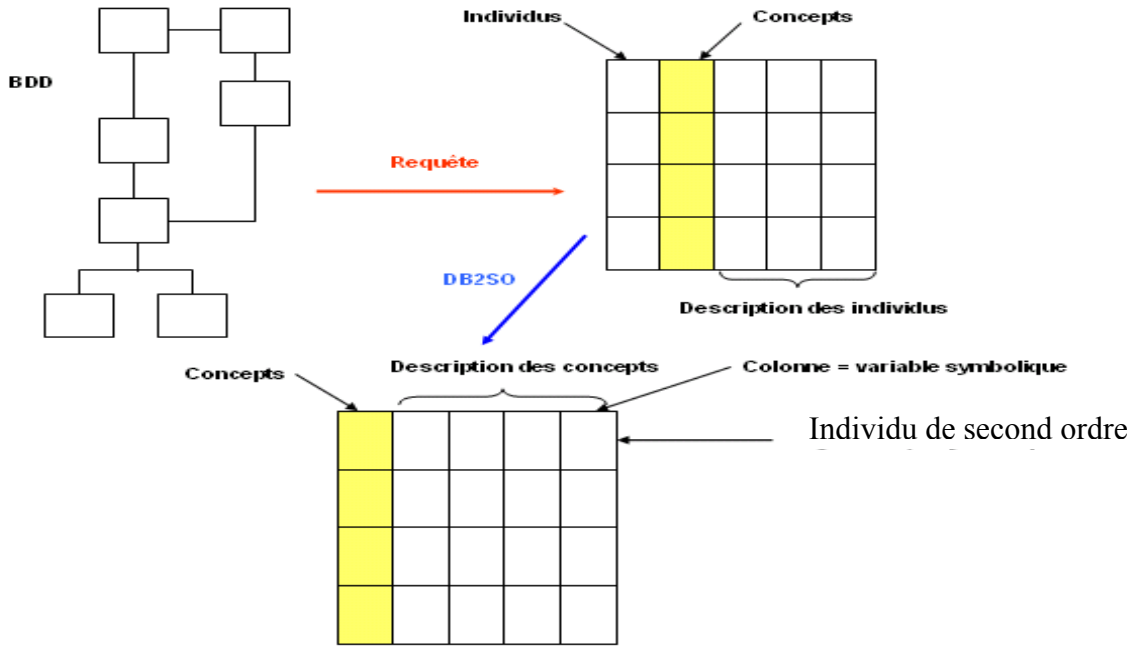


Figure 3 Les différentes étapes de la construction d'un tableau de données symbolique à partir d'une base de données relationnelle.

3.2 Comment conserver des informations perdues par le passage des individus aux concepts

La transformation du tableau de données classiques en données symboliques fait perdre un certain nombre d'informations qui étaient contenues dans le tableau de départ. On peut cependant s'efforcer d'en conserver par exemple, sous forme de règles, de taxonomies ou de corrélations.

Le logiciel SODAS permet d'ajouter des règles perdues qui étaient implicites dans le tableau de départ. Par exemple si tous les individus instances d'un concept suivaient la règle « si la variable 1 prend la valeur 2 alors la variable 6 prend la valeur 4 », cette règle peut être conservée et être éventuellement utilisée dans les différents algorithmes d'Analyse des Données Symboliques. De même la corrélation entre deux variables calculée sur l'ensemble des instances d'un concept peut être conservée comme le montre les tableaux des figures 4 puis 5 suivants où les concepts expriment le fait d'avoir été à la World Cup ou pas. Ainsi non seulement on peut conserver la corrélation mais de plus on peut l'expliquer ultérieurement par une méthode d'ADS explicative.

Joueur	**Equipe**	**Age**	**Poids**	**Taille**	**Nationalité**	**_ _ _ _**	**World Cup**
Manuel	Spain	29	85	1. 84	Spanish		oui
Rodriguez	Spain	23	90	1.92	Brazilian		oui
Mballo	France	25	82	1.90	African		oui
Zedane	France	27	78	1.85	French		oui
_ _ _ _ _	_ _ _ _	_ _ _	_ _ _ _	_ _ _ _	_ _ _ _ _	_ _ _ _	
Elyès	Egypte	23	91	1. 75	Spanish		non
Bernard	France	29	84	1.84	Brazilian		non
Marcelle	France	24	83	1.83	African		non
Younès	Maroc	30	81	1.81	French		non

Figure 4 Description des individus de premier ordre

World Cup	**Age**	**Poids**	**Taille**	**Cor(Poids, taille)**
Oui	[21, 26]	[78, 90]	[1.85, 1.98]	0.85
Non	[23, 30]	[81, 92]	[1.75, 1.85]	0.65

Figure 5 Prise en compte des corrélations.

3.3 Construction d'un tableau de données symboliques par concaténation de tables différentes

L'un des avantages de l'approche symbolique est de permettre de réunir des informations éparses contenues dans des tables n'ayant ni les mêmes unités statistiques ni les mêmes variables descriptives, à conditions que les mêmes concepts apparaissent. Ainsi par exemple, dans la table de la figure 6, on décrit des écoles alors que dans la figure 7, on décrit des hôpitaux. Cependant, le fait que ces deux tables contiennent le concept de ville permet de les concaténer en une seule table permettant d'avoir une description des villes selon les deux types d'informations : les écoles et les hôpitaux. Il suffit de concaténer la table symbolique décrivant les villes obtenue à partir de la table 6 des écoles puis de la table 7 des hôpitaux.

Cette concaténation est donnée figure 8

Ecole	Ville	Nb d'élèves	Type d'école	Niveau
Lamartine	Paris	320	Public	1
Condorcet	Paris	450	Public	3
St Louis	Lyon	200	Public	2
St Hélène	Lyon	380	Privée	3
St Sernin	Toulouse	290	Public	1
St Hilaire	Toulouse	210	Privée	2

Figure 6 Description des Ecoles

Hôpital	Ville	Code Nb de lits	Code Specialité
Lariboisière	Paris	750	5
St Louis	Paris	1200	3
Herriot	Lyon	650	3
Besgenettes	Lyon	720	2
Purpan	Toulouse	520	6
Marchant	Toulouse	450	2

Figure 7 Description des hôpitaux

Ville	Nb délèves	Type décole	Niveau	C Nb lits	C. Spec
Paris	[320, 450]	(100%)Public	{1, 3}	[750, 1200]	{3, 5 }
Lyon	[200, 380]	(50%)Public , (50%)Privée	{2, 3}	[650, 720]	{2, 3}
Toulouse	[210, 290]	(50%)Public , (50%)Privée	{1, 2}	[450, 520]	{2, 6}

Figure 8 Description symbolique des villes par les variables des écoles et des hôpitaux.

3.4 Construction d'un tableau de données symboliques par croisement de plusieurs variables qualitatives ou symboliques

Si l'on dispose dans un tableau de données classiques de trois variables qualitatives, on peut croiser deux d'entre elles et faire apparaître dans chaque case (i, j) de ce tableau le diagramme de fréquence des catégories de la troisième variable dans les catégories i et j

respectivement associés à chacune des deux variables que l'on croise. Autrement dit, si on désire croiser la première et la seconde variable, on obtient ainsi un tableau de données symboliques dont chaque ligne est associé à une catégorie de la première variable, chaque colonne à une catégorie de la seconde variable et chaque case contient le diagramme des fréquences des catégories de la troisième variables pour les catégories de la ligne et de la colonne correspondante. Dans le cas où la troisième variable est numérique, on peut faire apparaître dans chaque case du tableau croisé un intervalle de variation exprimant la valeur minimum et la valeur maximum atteintes dans la catégorie de la première et celle de la seconde variable respectivement associées à cette case. Si la troisième variable est symbolique à valeur histogramme, on peut faire apparaître dans la case, la moyenne et la variation (min et max) des histogrammes associés à ces catégories. Cette idée de croisement a été appliquée dans le cas de données évolutives (chap 1 de Diday, Kodratoff, Brito, Moulet (2000)).

Exemple

On dispose au départ d'un tableau de données dont les lignes décrivent des lieux géographiques à un instant donné par des variables de pollution (taux d'oxyde de carbone dans l'air, d'ozone, ...) et des variables environnementales (orientation du vent, saison, ...) comme l'indique le tableau de la figure 9.

	Y: POLLUTION							**Z:** ENVIRONNEMENT							
	y_1				y_p			z_1			z_p				
TOUR EIFFEL (T1)															
.....															
TOUR EIFFEL (Tn)															
.....															
DAUPHINE (T1)															
.....															
DAUPHINE (Tk)															
.....															

Figure 9: Le tableau des données initiales

On utilise un algorithme de classification automatique sur les variables de pollution pour obtenir des classes de pollution (notées POLi) homogènes et associer comme indiqué dans le tableau de la figure10 à chaque couple (lieu , instant) une classe de pollution. On procède de même avec les variables d'environnement pour obtenir des classes d'environnement homogènes que l'on note ENVi.

	POLLUTION	ENVIRONNEMENT
TOUR EIFFEL(T1)	POL 9	ENV 4
TOUR EIFFEL(T2)	POL 5	ENV 7
.....		
TOUR EIFFEL(Tn)	POL 11	ENV 9
.....		
UNIV.DAUPHINE(T1)	POL 5	ENV 9
.....		
UNIV.DAUPHINE(Tk)	POL 6	ENV 7

Figure 10: Chaque couple (lieu , instant) est associé à une classe POLi caractérisée par les variables de pollution et une classe ENVi caractérisée par des variables d'environnement.

Finalement, chaque individu de second ordre est un lieu décrit pour différents types (i.e. classes) d'environnements (qui jouent le rôle de variables) par la fréquence de chaque type de pollution.

Ainsi en considérant trois lieux dans trois types d'environnements, on obtient, le tableau de la figure 11.

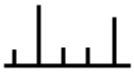
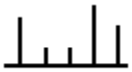
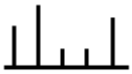
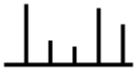
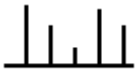
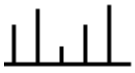
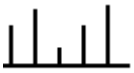
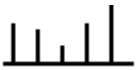
	ENV 1	ENV 2	ENV 3
TOUR EIFFEL			
UNIVERSITE DAUPHINE			
PLACE DE LA CONCORDE			

Figure 11: Chaque individu est un lieu dont la description est un histogramme de classes de pollution pour chaque type d'environnement.

Chaque histogramme exprime pour un lieu et un type d'environnement donnés la fréquence d'occurrence de chaque type de pollution parmi les instants où ce lieu a été observé.

A partir du tableau de données symboliques de la figure 11, on peut obtenir une classification symbolique (voir De Carvalho et al, chap 11 de Diday, Noirhomme (2008)) en zones de Paris subissant le même type de pollution lors d'une même situation d'environnement.

3.5 Construction de données symboliques à partir de données complexes ou de données floues

3.5.1 Des données complexes aux données symboliques

Les données complexes qui se présentent, par exemple, sous formes d'images, de textes ou de signaux issus de capteurs sont transformées et condensées sous formes de données numériques. On peut ensuite désirer décrire puis analyser des catégories d'images (par exemple, la mer, la montagne etc.) ou des catégories de textes (parties d'un texte biblique, par exemple) ou les deux à la fois pour décrire des catégories de patients par exemple, eux-mêmes connus par des images radiographiques et des commentaires médicaux textuelles. C'est aussi le cas de clients caractérisés à la fois par des caractéristiques sociaux économiques et des données textuelles issues par exemple de conversations téléphoniques (voir Touati et al. (2006). Toutes ces catégories considérées comme des concepts peuvent être réifiées en individus de second ordre inéluctablement décrits par des données symboliques pour tenir compte de la variation des instances de ces concepts et ne pas trop déformer la réalité.

3.5.2 Des données floues aux données symboliques

Dans le chapitre 1 de Diday, Noirhomme (2008), on développe un exemple montrant clairement que la sémantique des données floues n'est pas le même que celle des données symboliques. En effet, dans le premier cas il s'agit de transformer des données numériques décrivant des individus sous forme de catégories munies d'une pondération exprimant un degré d'appartenance de chaque individu à chaque catégorie. Ces catégories munies de ces degrés d'appartenance sont appelées « données floues » ou « ensemble floues » au sens de Zadeh (1978), (voir aussi Dubois, Prades (1988), Bandemer et al. (1992)). Par contre, en ADS, il s'agit de décrire des classes d'individus par des données symboliques qui expriment la variation des individus de ces classes, qu'ils soient décrits par des données qualitatives, quantitatives ou floues. Par exemple, si Paul est « grand » avec un degré 0.7 d'appartenance à cette catégorie et Pierre est « grand » avec un degré d'appartenance 0.3 et si la catégorie des « blonds » se réduit à ces deux individus, alors cette catégorie sera associée au concept « être blond » qui sera modélisé par l'individu de second ordre appelé « blond » dont la description sera « grand » avec un degré d'appartenance appartenant à l'intervalle [0.3, 0.7]. Une classe d'individus décrits par des données floues se décrit inéluctablement par des données symboliques (ici, par une variable à valeur intervalle) exprimant la variation des individus de cette classe si on ne veut pas trop déformer la réalité. On voit ainsi que les deux points de vu (symbolique et flou) sont complémentaires et non remplaçable l'un par l'autre.

4) Seconde étape de l'ADS : l'analyse des tableaux de données symboliques

4.1 Pourquoi on ne code pas les données symboliques sous forme de données classiques ?

Actuellement dans la pratique, les utilisateurs n'étudient pas en général les classes ou les catégories en tant que nouvelles unités statistiques, ils se contentent simplement de les décrire. Si toutefois ils désirent les étudier en tant que tels, ils les décrivent par des moyennes ou des fréquences de façon à se ramener à un tableau de données classiques que les outils du marché savent traiter (un peu comme celui qui a perdu sa montre une nuit dans une ville et la cherche sous un lampadaire car c'est le seul endroit éclairé !). Pour aller dans ce sens la première idée serait donc de recoder les données symboliques pour les transformer en données classiques puis d'appliquer les méthodes standards. Cela peut apporter des résultats utiles mais a le désavantage de démultiplier le tableau de données, de perdre les variables symboliques initiales et de perdre la variation.

Pour fixer les idées, prenons l'exemple de la figure 12 où l'on décrit le concept « très bon » joueur, obtenu à partir d'un tableau de données classiques de joueurs de football. Les autres concepts du tableau de données symboliques sont : « bons », « moyens », « très faibles » et « faibles » mais ne sont pas représentés dans cette figure.

Catégorie de joueurs	Poids	Taille	Nationalité
Très Bons	[80, 95]	[1.70, 1.95]	{0.6 Eur, 0.3 Afr, 0.1 Amer}

Figure 12 Description symbolique du concept « Très bon » joueur.

Catégorie de joueurs	Poids Min	Poids Max	Taille Min	Taille Max	Eur	Afr	Amer
Très Bons	80	95	1.70	1.95	0. 6	0.3	0.1

Figure 13 Transformation des données symboliques de la figure 12 en données classiques.

Dans la figure 13 on donne les données classiques issues de la façon la plus naturelle possible des données symboliques de la figure 12.

A l'aide de trois méthodes d'analyse de données symboliques : les arbres de segmentation et l'Analyse en Composantes Principales et les histogrammes symboliques, nous allons montrer l'avantage qu'il y a de conserver les données sous la forme symbolique pour les traiter sous cette forme, plutôt que de les transformer en données classiques pour les traiter par des méthodes classiques.

Exemple des arbres de segmentation classiques et symboliques :

En codage classique la variable « Taille » n'existe plus car seules « Taille Min » et « Taille max » demeurent. De même la variable Nationalité est remplacée par les variables induites des trois modalités Eur, Afr et Amer. Donc c'est seulement chacune de ces 5 variables qui pourra par des questions binaires du type Taille Min < 1.80 ou Prob(Eur) < 0.4 fournir chaque nœud de l'arbre de segmentation classique.

Par contre, l'arbre de segmentation symbolique (voir Chavent chap 11.2 et Périnel, Lechevallier chap. 10.3 dans Bock, Diday (2000)), pourra en posant des questions binaires du type : Taille < 1.80 ou Prob(Eur $\cup$ Amer) < 0.4, utiliser les deux variables initiales elles-mêmes pour la segmentation (voir figure 14).

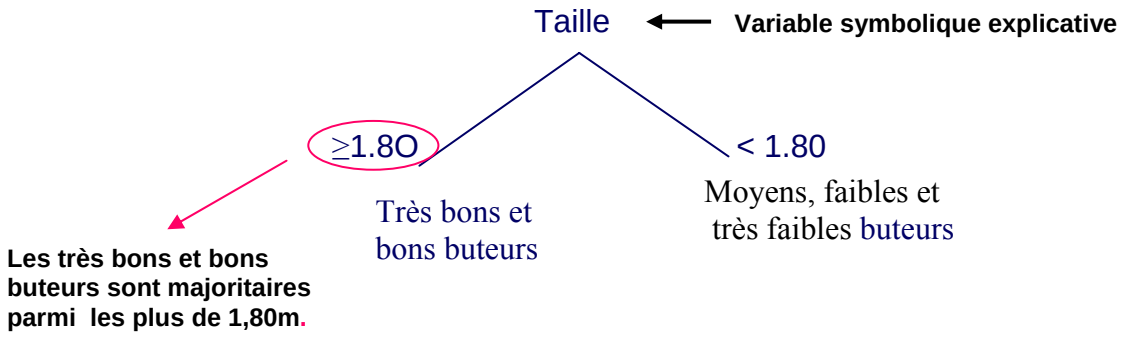


Figure 14 Arbre de segmentation symbolique : la taille n'est ici ni une taille minimum, ni une taille max., ni une moyenne, ni une modalité d'un histogramme etc., mais la valeur de l'âge qui discrimine au mieux les types de joueurs.

Exemple de l'Analyse en composantes principales classique et symbolique :

On considère ici le tableau des mêmes concepts décrits par plusieurs variables à valeur intervalles (la taille, le poids, l'âge, le nombre de buts marqués, le nombre de sélections, etc.).

En codage classique comme indiqué en figure 15, chaque concept est représenté par un point du plan factoriel d'une ACP classique. Ainsi, dans cette figure, on voit apparaître les points « très bons », « bons », « moyens », « très faibles » et « faibles ».

En codage symbolique une ACP symbolique (voir Chouakria et al. Chap. 9 de Diday , Bock (2000) ou C. Lauro et al. chap. 15 de Diday, Noirhomme (2008)): chaque concept est représenté par une surface, ici (voir la figure 15) un rectangle exprimant la variation du concept (de la valeur min. à la valeur max. prise par les individus inclus dans le concept). Notons que chaque concept peut être décrit par une conjonction de propriétés réduite aux axes factoriels retenus en sortie, tenant compte ainsi de la variation, alors que l'ACP classique ne donnerait en sortie qu'un vecteur de nombres associé à chaque axe factoriel.

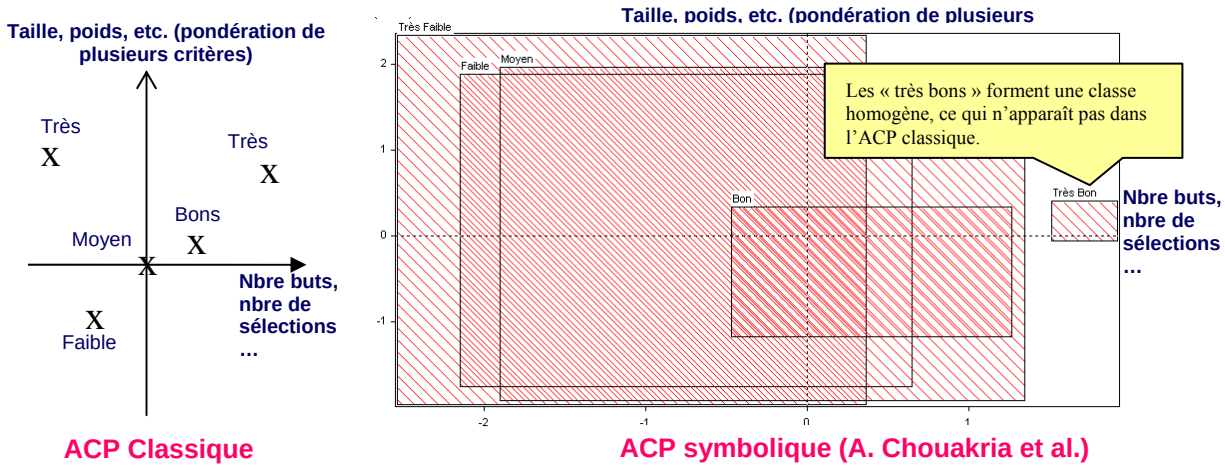


Figure 15. ACP classique sur des données symboliques transformées en données classiques et ACP symbolique sur les données symboliques.

Exemple de l'histogramme classique et symbolique

L'approche classique perd la notion de variable à valeur intervalle et ne permet de construire un histogramme que sur les min ou les max. Dans l'exemple suivant, on voit d'abord le diagramme classique de la variable « nbre de miles » induit du tableau de données classique de la même figure 16. Figure 17 (a) on représente le tableau de données symboliques issu du tableau de données initiales représenté figure 16 où les deux clients constituent les concepts. Figure 17 (b) on voit «le diagramme symbolique» réalisé à partir du tableau de données symboliques donné en 17 (a). Elle se distingue du diagramme classique de la figure 16 réalisé sur les données initiales. Il est construit en cumulant dans chacune des classes, les proportions d'intervalle associées à chaque concept « client ». Les deux diagrammes du min et du max qui seraient construits à partir des données classiques du tableau (c) de la figure 17 issues des données symboliques du tableau (a) de la même figure, n'auraient aucun

intérêt avec seulement 2 valeurs pour chacun. Au cas où l'on aurait plus de concepts on obtiendrait deux diagrammes du min et du max complémentaires et non redondants avec celui qui serait obtenu avec l'approche symbolique.

Achat 1	M. Dupont	150
Achat 2	M. Dupont	180
Achat 3	M. Dupont	250
Achat 1	Mme Fabre	270
Achat 2	Mme Fabre	245
Achat 3	Mme Fabre	310
Achat 4	Mme Fabre	320
Achat 5	Mme Fabre	315

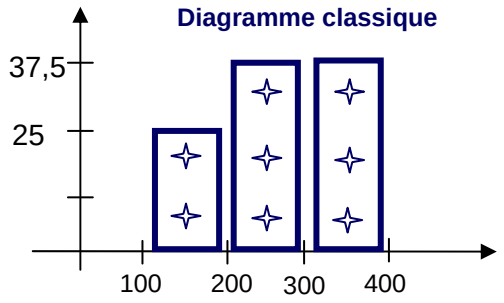


Figure 16 Diagramme classique sur données classiques pour la variable numérique « Nbre de Miles »

Concept Client	**Nombre de miles**
Dupont	[150, 250]
Fabre	[245, 320]

(a)

Concept Client	**Nbre Minimum de miles**	**Nbre Maximum de miles**
Dupont	150	250
Fabre	245	320

(c)

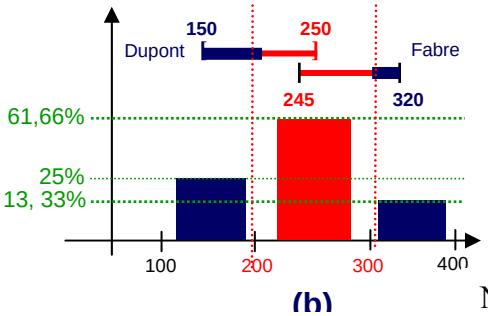


Figure 17 (a) Données symboliques induites du tableau initial de la figure (16).

(b) Diagramme symbolique sur la variable symbolique « Nbre de Miles » : dans la classe [200, 400], le cumul des proportions d'intervalle est le plus grand. En (c) le tableau classique induit du tableau symbolique (a) d'où deux diagrammes complémentaires pourraient être extraits.

4.2 La statistique des individus n'est pas la statistique des concepts : complémentarité des approches classiques et symboliques

Nous allons illustrer le fait que la statistique des individus n'est pas la statistique des concepts à l'aide d'un exemple simple :

Dans une enquête réalisée par une entreprise de crédit se trouvent 400 clients ayant opéré un crédit de type A et habitant un iris 1 (i.e. une zone d'habitation définie par l'INSEE), 100 clients ayant réalisé un prêt de type B habitant un iris 2, 100 habitants un iris 3 et ayant également réalisé un prêt de type B. En plus de ces deux variables, on indique pour chaque client, dans le tableau de la figure 18 (a), le type d'habitat (pavillon ou HLM) ainsi que son revenu. Ainsi, la variable « Catégorie de client » du tableau de la figure 18 (b), croisant le type de crédit et le numéro d'iris, définit 3 concepts : prêt de type A x iris 1, prêt de type B x iris 2, prêt de type B x iris 3. La «catégorie de client » devient la nouvelle unité statistique dans le tableau de données symboliques de la figure 18 (b). On ne s'intéresse plus aux 600 clients en tant que tels dans ce nouveau tableau. Dans ce tableau, les variations dues aux individus (i.e. clients) inclus dans l'extension de chaque concept (i.e. catégorie de client) sont conservées sous forme de diagramme de fréquences ou d'intervalle respectivement pour les variables « Habitat » et « Revenu ». Au niveau de ce second tableau, on peut ajouter des variables « conceptuelles » qui ne s'appliquent bien au niveau des concepts. Dans ce sens, on ajoute ici: le nombre d'habitants de chaque iris.

Dans la figure 19, en (a) on représente le diagramme de fréquence des clients pour la variable type de prêt et en (b) celui des catégories de clients pour la même variable. On voit qu'ils sont inverses. Cet exemple montre donc clairement que la statistique des individus n'est pas celle des concepts.

Remarquons au passage qu'il est bien sûr facile de calculer un diagramme de fréquence sur les concepts mais que par contre, c'est beaucoup moins trivial de le définir et donc de le construire pour des variables symboliques telles que « type d'habitation » ou « revenu » dont les valeurs sont respectivement des diagrammes et des intervalles. C'est l'un des objectifs de l'ADS de résoudre ce genre de questions (voir par exemple, Billard, Diday (2003, 2006))

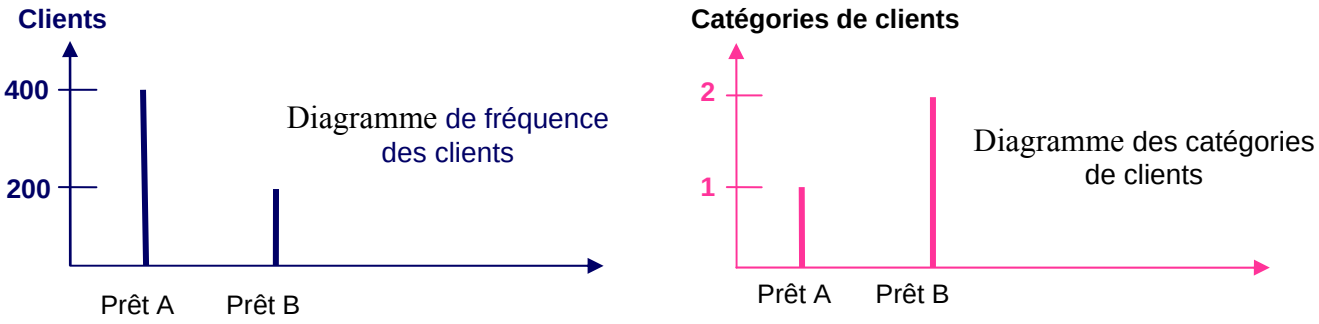
Client	Catégorie de client	Iris	Prêt	Habitat	Revenu
1	Prêt B iris 2	2	B	Pavillon	80
2	Prêt A iris 1	1	A	HLM	70
600	Prêt B iris 3	3	B	HLM	125

(a)

Catégorie de client	Iris	Prêt	Habitat	Revenu	Nombre d'habitants de l'iris
Prêt A iris 1	1	A	0.3Pav,0.7 HLM	[60, 85]	10000
Prêt B iris 2	2	B	0.7Pav,0.3HLM	[70, 120]	4000
Prêt B iris 3	3	B	0.9Pav,0.1HLM	[78, 200]	2000

(b)

Figure 18 (a) Tableau des données classiques initiales. (b) Tableau des descriptions symboliques des concepts.



(a) (b)

Figure 19 Les diagrammes de fréquence sur les individus (clients) et les concepts catégories de clients sont inversés.

4.3 Extension des méthodes classiques aux données symboliques

Comme les données symboliques contiennent comme cas particulier les données classiques (un nombre est un cas particulier d'intervalle, une modalité est un cas particulier d'histogramme qui aurait une fréquence égale à 100%, etc.) il est naturel de penser à étendre au minimum la statistique descriptive ainsi que l'analyse et la fouille des données traditionnelles aux tableaux de données symboliques.

On a commencé ainsi à étendre les méthodes suivantes (on trouvera beaucoup de choses dans Bock, Diday (2000), Diday, Monique (2008)):

- En statistique descriptive (Diagrammes, Histogrammes, Corrélations, biplots)
- Typologie (hiérarchies, pyramides, Nuées dynamiques, Cartes de Kohonen ,…)
- Décomposition de mélange de lois
- Arbres de Décision,
- Calcul et Représentation de dissimilarités
- Inférence de règles ou d'arbres de causalités
- Méthodes de visualisation (étoîles, cibleur,…)
- Analyse factorielle (ACP, AFC, …)
- Régression classique, PLS
- Réseaux neuronaux, VSM (Vector Support Machine), Etc.
- Treillis de Galois (données binaires)
- Tableur de données symboliques (extension d'EXCEL à des données symboliques)

Remarquons que les sorties des méthodes ainsi étendues doivent aussi s'exprimer en terme de concepts modélisés par exemple par des descriptions symboliques munies d'un opérateur de comparaison et d'une fonction de reconnaissance formant un triple appelé « objet symbolique » voir Bock, Diday (2000). Ces objets symboliques permettent de décrire les concepts par leurs propriétés communes mais aussi de calculer leur extension dans

l'ensemble des individus qu'ils représentent. Nous détaillons cette modélisation des concepts dans le paragraphe suivant.

4.4 Autres méthodes à développer

En dehors des méthodes classiques à étendre au cas symbolique, beaucoup d'autres méthodes spécifiques à l'analyse des données symboliques ont été développées mais là aussi beaucoup reste à faire. Par exemple le trie automatique de données symboliques. Autrement dit, l'ordonnancement de variables à valeur intervalle ou à valeur histogramme par exemple (voir Mballo (2005), Mballo, Diday (2006)), la recherche de descriptions symboliques de volume minimum recouvrant sous forme discriminantes et séparantes l'ensemble des descriptions initiales qu'elles soient classiques ou symboliques (voir Limam et al. (2004)).. Beaucoup reste à faire également pour étendre et améliorer les dissimilarités (Hausdorff, …) entre descriptions symboliques, la construction de prototypes pour représenter une classe de concepts, ainsi que les représentations graphiques pour la représentation concepts ou de structures de classes de concepts (treillis, pyramides spatiales,…). Dans la section suivante nous soulevons la question de la modélisation des concepts et de son apprentissage qui reste également très ouvert.

5) La modélisation des concepts par des objets symboliques

5.1 Le schéma à 4 espaces

Pour être cohérent avec les entrées qui sont des concepts il faudrait de plus obtenir également en sortie des concepts. On est donc conduit à modéliser les concepts découverts en utilisant les descriptions symboliques de même type que celles qui ont été utilisées en entrée. De façon plus générale, pour modéliser des concepts nous utilisons quatre espaces schématisés dans la figure 20: à gauche l'espace des individus et l'espace des concepts dits du « monde réel », à droite dans le « monde modélisé »: l'espace des individus du premier ordre et l'espace des individus du second ordre. Chaque individu de premier ordre est muni d'une description classique, chaque classe d'individus du premier ordre est muni d'une description symbolique, chaque individu de second ordre est muni d'un « objet symbolique » qui modélise le concept auquel il est associé. Par définition d'un concept, une telle modélisation doit être capable de définir l'intension d'un concept et son extension. Plus précisément, nous allons demander à l'objet symbolique de fournir la fonction qui permet de calculer l'extension ainsi que la façon de la construire. Cette fonction dite « de reconnaissance » ou d'appartenance (en anglais membership function), notée a_C dans la figure 20, associe à chaque individu une valeur vraie ou faux ou plus généralement, un degré d'appartenance au concept dans un espace noté L. Cette valeur qu'elle soit booléenne ou numérique permet de savoir dans quelle mesure un individu fait partie de l'extension d'un concept : il en fait partie dans le cas booléen si la valeur est vrai et dans le cas numérique il en fait d'autant plus partie que sa valeur est proche de 1.

La fonction d'appartenance dépend de trois choses : i) de la variable notée « y » qui associe à un individu du monde réel, un individu du premier ordre lui-même associé à une description. ii) la description notée d_C du concept obtenue par généralisation des descriptions des individus du premier ordre qui sont images d'instances du concept. L'opérateur de généralisation noté T dans la figure peut être par exemple une T-norme (voir dans le premier chapitre de Diday, Kodratoff et al (2000) comment elle est utilisée). iii) une relation de comparaison notée R_C, dite parfois d'appariement (ou de « matching ») entre la description d'un individu y(w) et celle d'un concept d_C. Une fois ces trois éléments connus, on peut construire la fonction « a_C » de différentes façons (voir Chap 1 de Bock, Diday (2000)). Sous sa forme la plus simple, elle s'écrit:

$a_C(w) = [y(w)\ R_C\ d_C]$ qui mesure l'adéquation entre la description d'un individu et la description d'un concept d_C à l'aide de la relation R_C et permet de modéliser la façon de calculer l'extension

du concept tout en précisant le procédé constructif. Réduire la modélisation d'un concept à la seule donnée de a_C serait très réducteur. En effet, plusieurs choix de y , R_C et d_C peuvent conduire à la même fonction a_C. Ainsi, par exemple, un oiseau qui chante peut aussi bien être reconnu par sa voix que par sa vue. Dans ces deux cas la fonction y qui associe l'oiseau à sa description (selon qu'elle est visuelle ou auditive), la description du concept d_C (selon qu'elle est visuelle ou auditive par exemple, de l'espèce Rossignol), R_C la relation de comparaison entre la description de l'individu (l'oiseau) ou celle du concept (Rossignol) selon qu'elle est visuelle ou auditive, seront complètement différentes mais les résultats de la fonction a_C seront identiques (les deux procédés de reconnaissances conduiront à l'identification pour un oiseau donné de l'espèce rossignol). La réduction de la modélisation d'un concept à sa seule description symbolique (pour tenir compte de la variation de ses instances) serait également très réductrice, car elle ne permettrait pas de calculer l'extension du concept. Ainsi la modélisation d'un concept doit non seulement tenir compte de la fonction: $a_{C,}$ mais aussi de la façon dont elle est construite au moins partiellement. En supposant que la fonction de description « y » est fournie, on a pris l'habitude de définir un objet symbolique par le triplet $S_C = (a_C, R_C, d_C)$. Remarquons que « y » associe un individu dit du monde réel à un individu de premier ordre du monde modélisé, mais ne lui associe pas directement sa description car deux individus du premier ordre pourrait avoir la même description et ne seraient donc pas distingués bien que représentant deux individus différents.

Exemple

L'objet symbolique $S_C = (a_C, R_C, d_C)$ modélise un concept C associé à la catégorie « iris 122 » décrit par l'âge des habitants de cet iris et le type d'emploi (supposé ici partagé entre « employé » et « paysan » pour simplifier) . Avec d_C = [18, 52] x {0.1employés, 0.9 paysans}] obtenu par un opérateur de généralisation T à partir des individus de cette catégorie. L'opérateur d'appariement est défini par la relation binaire $R = (\subseteq , \subseteq_C)$ où $\subseteq_C$ exprime l'appariement entre deux descriptions pondérées : celle d'un individu (il se peut qu'un individu soit à temps partiel sur chacun de ces emplois) et celle du concept. Enfin, la fonction de reconnaissance a_C est définie à l'aide d'un opérateur dit « d'agrégation » $\wedge_C$. Le choix de cet opérateur peut se faire de diverses façons qui sont par exemple développés dans Diday (1995). Elle peut s'écrire sous la forme :

$a_C(w)$ = [age(w) $\subseteq$ [18, 52] $\wedge_C$ [CSP (w) $\subseteq_C$ {0.1 employés, 0.9 paysans}] selon que l'on désire une décision booléenne ou plus fléxible avec un degré d'appartenance, on aura a_C (w) $\in$ {Vrai, Faux} ou a_C (w) $\in$ [0, 1]. On trouvera dans Esposito et Al : Chapitre 8 de Diday, Noirhomme (2008), différentes façon de réaliser ces calculs.

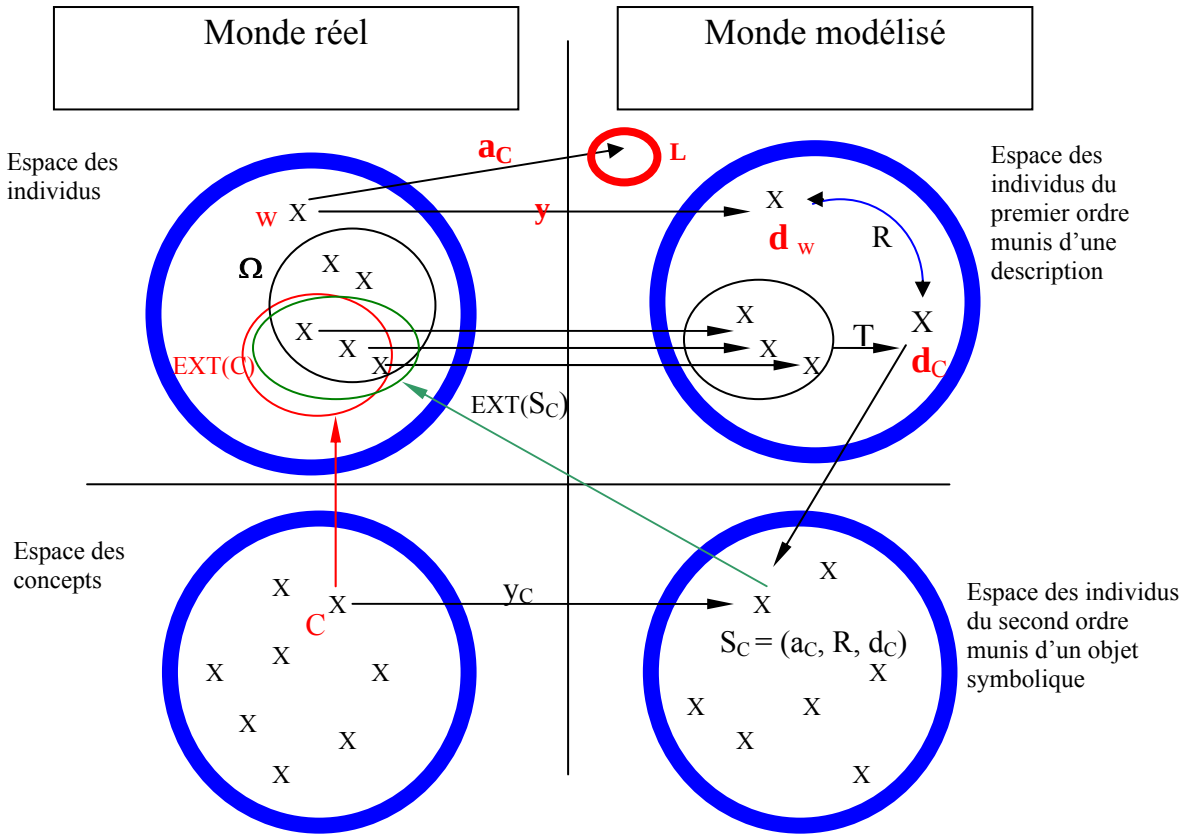


Figure 20 Modélisation par un objet symbolique d'un concept connu par son extension dans Ω.

5.2 Apprentissage de la modélisation d'un concept

A partir du schéma de la figure 20, l'apprentissage de la modélisation d'un concept C par un objet symbolique $S_C = (a_C, R_C, d_C)$ va se baser sur l'adéquation entre l'extension du concept dans le monde réel et son extension obtenue par son modèle S_C par : $EXT(S_C) = \{w \in \Omega / a_C(w) = Vrai\}$ dans le cas booléen et $EXT_\alpha(S_C) = \{w \in \Omega / a_C(w) \geq \alpha\}$ dans le cas dit « modal » ou « fléxible » à un seuil donnée α. On est conduit à obtenir des erreurs de deux espèces : les individus qui sont dans l'extension de S_C mais pas dans l'extension du concept C, les individus qui sont dans l'extension du concept C mais pas dans l'extension de S_C, ou leurs combinaisons (voir Chapitre 1 de Diday , Noirhomme (2008)). En se basant sur ces deux critères, on peut mettre aux point différentes stratégies pour changer les choix des différents opérateurs : y, T_C, R_C, d_C, a_C qui définissent l'objet symbolique y compris par un système coopératif multi-agent combinant des descriptions d_C selon différents environnements.

5.3 Modélisation graphique des objets symboliques et apprentissage: le cibleur

Il est possible de représenter graphiquement un objet symbolique par un "cibleur" qui doit associer à un concept un objet symbolique et fournir une cible graphique qui peut visualiser la performance de chaque individu ou de l'ensemble des individus. Le "cibleur" est l'opérateur qui associe à un concept l'objet symbolique s = (a, R, d) qui le modélise dans le sens où cet objet symbolique constitue une cible que les individus de l'extension du concept doivent atteindre. Cette cible peut être représentée graphiquement par un cercle dont les rayons sont les variables et

dont la position prise notée $a_i(w) = [y(w)\ R_i\ d_i]$ par un individu pour chaque variable y_i est d'autant plus proche du cercle que l'individu satisfait (au sens de R_i) à la description d_i du concept associée à cette variable. On peut de plus construire un axe gradué perpendiculaire à ce cercle et passant par son centre. Sur cet axe la valeur 0 est atteinte au centre du cercle. On représente sur cet axe la valeur a(w) qui est donc confondue avec le centre du cercle quand a(w) = 0, c'est à dire quand w ne satisfait pas le concept et qui est maximale sur l'axe quand w satisfait au concept (ie fat partie de ses instances). En normalisant, la valeur maximum de a est 1 pour chaque variable. Donc quand pour une variable donnée un individu se positionne sur le cercle de rayon 1, cela veut dire qu'il satisfait au concept du point de vu de l'objet symbolique s qui le modélise. On note A la position de a(w) sur l'axe et B_i la position de $a_i(w)$ sur le rayon associé à la variable y_i. La pyramide géométrique de sommet A et de base $B_1,\ldots,B_p$ a un volume d'autant plus grand que l'individu w satisfait au concept du point de vu de l'objet symbolique s qui le modélise. Notons que ce volume est majoré par le cône de sommet 1 et de base le cercle de rayon 1.

La représentation graphique d'un cibleur sur l'ensemble des individus peut se faire à l'aide d'un cercle (de la cible) dont chaque rayon exprime le degré d'adéquation d'un individu au concept pour cette variable suivant qu'il est plus ou moins près de la circonférence du cercle. On peut alors construire la pyramide géométrique de sommet A dont la base relie toutes les valeurs minimales atteintes par l'ensemble des individus. Dans le processus d'apprentissage, il faut que cette pyramide tendent vers le cône de sommet A et de base le cercle de la cible. L'écart entre le volume de la pyramide minimale et le volume du cône peut être utilisé pour mesurer la qualité de l'objet symbolique.

6) Le modèle mathématique associé aux données symboliques

6.1 Les données

Le tableau de données symboliques est modélisé, dans sa case de ligne i et de colonne j, par une variable aléatoire X_{ij} de distribution f_{ij} . On note Ω l'ensemble des individus de second ordre et l'on considère que les variables aléatoires X_{ij} définies sur Ω prennent leur valeur dans O_j. On trouvera plus de détails formels dans la note de E. Diday et R. Emilion présentée à l'Académie des Sciences par G Choquet en 1997 ou dans Diday, Emilion (2003) et Diday, Vrac (2005). En remplaçant chaque variable aléatoire par sa distribution exprimée sous forme d'histogramme (si O_j est numérique) ou de diagramme (si O_j est discret), ou par son intervalle interquartile, etc., on obtient un tableau de données symboliques exprimant la variation des instances de chaque concept associé à chaque élément de Ω.

Pour fixer les idées, considérons les enquêtes sociaux démographiques de l'INSEE elles portent au premier niveau sur différentes sortes d'unités statistiques : des habitants (âge, sexe, type d'emploi,…), des foyers (nombre d'enfants, de voitures, …), des habitations (par type : HLM , Pavillon,…, nombre d'étage,…), des écoles, des hôpitaux etc.. Ainsi, ces enquêtes produisent des tables qui n'ont ni les mêmes unités statistiques ni le mêmes variables. Il est ainsi difficile de faire apparaître des liens entre les variables des différentes tables. Par contre, en passant au niveau des IRIS qui sont des zones géographiques définies par l'INSEE (la France est ainsi découpée en plus de 50 000 IRIS), on peut réunir toutes les tables en une seule, dont chaque ligne décrit un individu de second ordre associé à un IRIS considéré ici comme un concept. Chaque case de ce tableau contient la variable aléatoire X_{ij} de distribution f_{ij} qui associe par exemple à l'IRIS i, l'âge d'un habitant w si l'âge est la j ième variable. Ainsi, si chaque variable aléatoire X_{ij} est représentée par un diagramme de classes d'âge, la variable symbolique « âge » associera à chaque IRIS un diagramme d'âge. La seconde étape de l'ADS, permet ensuite d'analyser les IRIS décrits par les variables

symboliques ainsi définies, (par exemple, une centaine de variables symboliques, plutôt que considérer comme des variables les 1500 modalités qui leur sont associées).

6.2 Les modèles de classes

On s'est intéressé à deux formes de modélisation de classes d'individus de second ordre. Le premier est basé sur une représentation d'une classe par l'enveloppe supérieure et inférieure des descriptions de ses individus (voir Diday, Emilion (1996, 2003, 2005)), la seconde est basée sur une représentation d'une classe par des distributions de distributions des individus de la classe, elles mêmes reliées par un modèle de copules (voir Diday, Vrac (2005).

Dans le premier cas on utilise une approche non paramétrique. On suppose que l'espace des descriptions $d_i = (d_{ij})$, des individus de second ordre est muni d'un ordre et que toute partie de cet espace a une borne supérieure et une borne inférieure. La difficulté d'ordonner des mesures de probabilités nécessite l'introduction de la théorie des capacités de Choquet car l'enveloppe supérieure d'un ensemble de mesure de probabilités n'est pas une mesure de probabilité. Notons que si les d_{ij} sont considérées comme des mesures positives σ-additives sur O_j alors le $\sup_{i\in A} d_{ij}$ (resp $\inf_{i\in A} d_{ij}$) est une capacité σ-sous-additive (resp-sur-addidive). Pour les liens avec les capacités voir Diday (1995), Diday, Emilion (1995, 2003)).

Dans le second modèle, on utilise une approche plus paramétrique, on considère que les descriptions d_i sont des vecteurs de distributions. L'idée de base (Diday (2002 b)) est d'introduire la notion de distribution de distributions. Une distribution de distribution est définie par $G_i(x) = \text{Prob}(\{w \in \Omega / F_w(t_i) \leq x\})$ où F_w est la distribution associée à w pour une variable donnée. Soit H la distribution jointe de k distributions de distributions $G_1, \ldots G_k$. On peut alors démontrer grâce à un théorème de Sklar (voir Schweizer, Sklar (1983, 2005)) qu'il existe une fonction C appelée « copule » de $[0, 1]^k$ dans [0, 1]: telle que pour tout $(x1, \ldots, xk) \in [0, 1]^k$, on ait : $H(x_1, \ldots, x_k) = C(G_1(x1), \ldots, G_k(x_k))$. Il existe une grande famille de copules paramétrées (voir par exemple, Nelsen (1998)) dont par exemple le copule de Frank (1979) :

$F_b(u, v) = -1/b \text{ Log}(1 + (e^{-bu} - 1)(e^{-bv} - 1)/(e^{-b} - 1))$où $b \in R\backslash\{0\}$. Les paramètres qui permettent de modéliser une classe sont par exemple déterminés par maximisation de vraisemblance (voir Diday, Vrac (2005), Genest, Frank (1987)).

6.3 Les structures des classes : : Treillis de Galois, Partitions, Pyramides spatiales

Le premier modèle conduit à une structure de treillis de Galois maximal. Par la suite E est le treillis des parties de Ω noté $< P(\Omega), \subseteq, \cup, \cap >$ et F est le treillis des descriptions noté $<D, \leq, \vee, \wedge >$ supposé σ-complet (i.e. toute partie dénombrable de D a un sup et un inf calculés à l'aide des opérateurs ∨ et ∧), tout élément $w \in \Omega$ admet une description $y(w) \in D$ notée d_W.

On trouvera dans Diday, Emilion (1997) le résultat suivant et sa démonstration:

Les applications $f : P(\Omega) \to D$ et $g : D \to P(\Omega)$, définies par $f(A) = \wedge_{W\in A} d_W$ et $g(d) = \{w \in \Omega / d \leq d_W\}$ donnent le meilleur treillis de Galois (dit « des descriptions ») au sens où f et g sont maximales parmi les applications décroissantes telles que h et k soient extensives. Les ensembles I_h et I_k sont alors des treillis en bijection par f et le treillis de Galois est le treillis $\{(A,f(A)) / A \in I_h\}$ ou $\{(g(d), d)) / d \in I_k\}$ avec $I_h = \{A \in E / gof(A) = A\}$ et $I_k = \{B \in F / fog(B) = B\}$.

Ce treillis sera dit "inférieur" par opposition au treillis de Galois dit "supérieur" qui est obtenu en posant $f(A) = \vee_{w \in A} d_W$ et $g(d) = \{w \in I / d_W \geq d\}$. On obtient aussi un treillis de Galois à la fois supérieur et inférieur en posant:

$f(A) = (\wedge_{w \in A} d_W, \vee_{w \in A} d_W)$ et $g(d, d') = \{i \in \Omega / d \leq d_i \leq d'\}$.

On retrouve le treillis de Galois binaire standard (par exemple, Birkhoff (1967), Barbut, Monjardet (1970), Wille (1983), Duquenne (1996)) ou des valeurs multiples Brito (1991, 1994), Pollaillon (1998)) comme cas particuliers du treillis de Galois inférieur.

Notons que l'on peut associer à chaque élément du treillis à la fois inférieur et supérieur un objet symbolique défini par le triplet $s = (a, R, (d,d'))$ avec $a(w) =$ vrai si $d \leq d_w \leq d'$) donc l'ensemble des w tels que $a(w) =$ vrai est identique à $g(d,d')$. Autrement dit on a $g(d,d') = Ext(s)$. D'autre part, f permet de calculer l'intention de cette extension par $f(g(d, d')) = (d,d')$. Les objets symboliques satisfaisant cette propriété de point fixe sont dits complets (Diday (1989, 1991)).

Dans le cas stochastique, ce treillis converge vers une position qui se stabilise à mesure que la connaissance (i.e. la description) se précise (voir Diday, Emilion (1997)). Autrement dit, l'étude de l'évolution de ce treillis montre sous certaines conditions, que les objets symboliques complets et donc les concepts qu'ils représentent s'organisent et se stabilisent en convergent vers un treillis de Galois maximal à mesure que la connaissance des lois de distributions s'améliore par l'arrivée de nouvelles données.

La seconde approche où l'on représente les classes par des modèles de copules conduit dans Diday (2001) à la construction d'une structure de partitions par Nuées Dynamiques. Dans Diday, Vrac (2005), elle conduit à une structure de partition floue par EM.

Les pyramides qu'elles soient linéaires (Diday (1984), (1986)) en généralisant les hiérarchies (par la possibilité de classes qui peuvent être d'intersection non vide sans être nécessairement incluses l'une dans l'autre) ou spatiales Diday (2008) sont une solution intermédiaire entre les treillis trop lourd dés que le nombre d'individu devient important et les partitions trop simples pour exprimer la structure des classes entre elles. Elles font l'objet du paragraphe suivant.

6.4) Les pyramides classifiantes

Une pyramide spatiale basée sur une grille pour un ensemble fini Ω est un ensemble de parties appelées « paliers » satisfaisant les propriétés suivantes:

1) $\Omega \in P$

2) $\forall w \in \Omega, \{w\} \in P$.

3) $\forall (h, h') \in P \times P$ on a $h \cap h' \in P \cup \varnothing$

4) Il existe une grille dont les nœuds sont des éléments de Ω et les paliers sont des convexes de la grille.

On trouvera dans Diday (2008) une définition plus générale. Les hiérarchies sont le cas particulier où l'on impose que $h \cap h' \neq \varnothing$ implique $h \subset h'$ ou $h' \subset h$. Les pyramides standards sont le cas particulier où la condition 4) est remplacée par la nécessité d'avoir un ordre tel que tous les paliers soient connexes selon cet ordre. Dans la figure 20 on visualise

les hiérarchies, les pyramides et les pyramides spatiales de façon à mieux les positionner. On voit que dans chaque cas, chaque palier sous-tend une partie convexe de son support (ici une partie est dite convexe si les chemins les plus courts qui relient deux sommets de cette partie sont dedans).

On sait qu'il existe une bijection entre l'ensemble des hiérarchies indicées et l'ensemble des ultramétriques, entre l'ensemble des pyramides indicées et l'ensemble des disimilarités robinsoniennes. On a généralisé ces résultats dans Diday (2008), en montrant qu'il existe une bijection entre les pyramides spatiales et un nouveau type de dissimilarité appelé « yadidienne ».

Ces différentes bijections permettent de calculer la qualité de la classification par l'écart avec la dissimilarité initiale. On peut montrer que l'ensemble des ultramétriques est inclus dans l'ensemble des robinsoniennes qui est lui-même inclus dans l'ensemble des yadidiennes. Dans la figure 21, on représente successivement la hiérarchie, la pyramide et la pyramide spatiale obtenues à partir de la dissimilarité entre les sommets d'un carré de côté de longueur 1. On voit sur cet exemple que l'écart entre la dissimilarité initiale se réduit de la hiérarchie à la pyramide puis de la pyramide à la pyramide spatiale (vue de dessus).

Les pyramides peuvent être construites directement à partir d'une dissimilarité par un algorithme ascendant en autorisant au maximum un ascendant unique pour chaque palier d'une hiérarchie, deux ascendants dans le cas d'une pyramide et 4 dans le cas d'une pyramide spatiale. Une fois construites on peut associer un objet symbolique à chaque palier, en considérant qu'il modélise un concept dont l'extension contient au moins ce palier. Il se peut que l'extension de l'objet symbolique ainsi construit ne contienne pas uniquement les éléments du palier ou à l'inverse qu'il ne couvre pas complètement le palier. Pour éviter cela, d'autres algorithmes de constructions pyramidales (voir le chapitre de P. Brito et al. dans Diday, Noirhomme (2008)) ont été conçus afin de ne retenir à chaque étape de la construction que les paliers auxquels on peut associer des objets symboliques complets. On construit de cette façon des paliers qui sont associées aux fermés d'un treillis de Galois

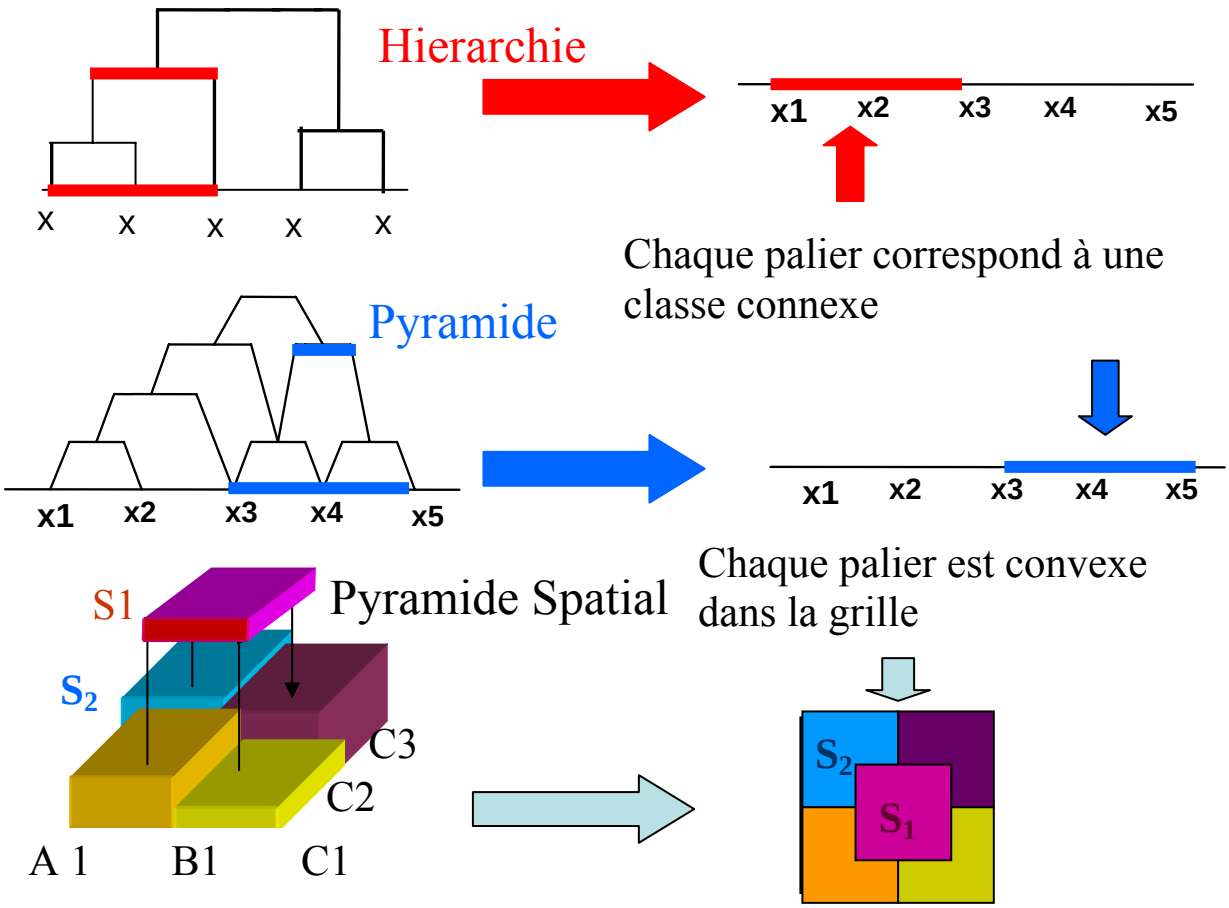
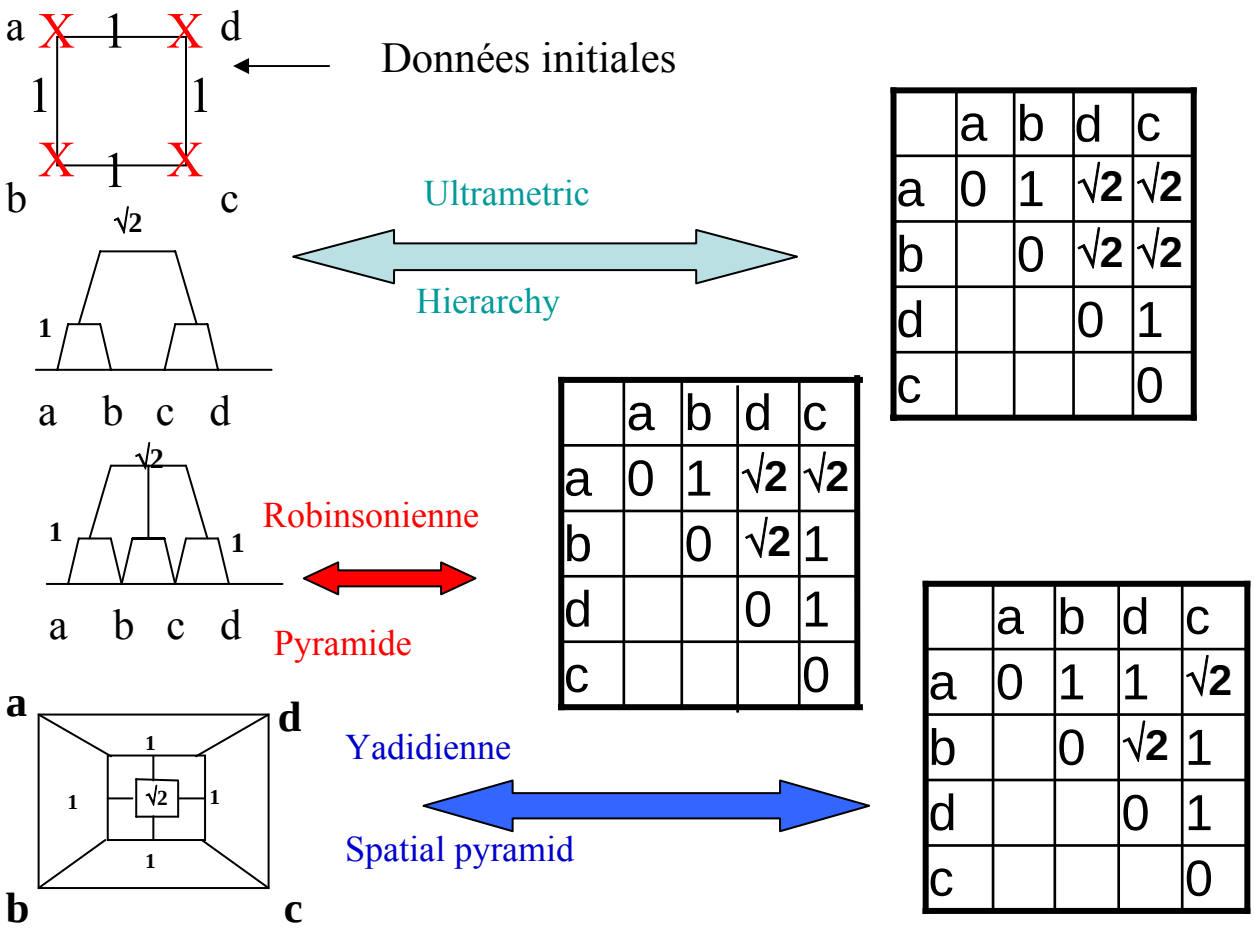


Figure 20 On voit que chaque palier sous-tend une partie convexe de son support.



	a	b	d	c
a	0	1	√2	√2
b		0	√2	√2
d			0	1
c				0

	a	b	d	c
a	0	1	√2	√2
b		0	√2	1
d			0	1
c				0

	a	b	d	c
a	0	1	1	√2
b		0	√2	1
d			0	1
c				0

Figure 21 On voit qu'avec seulement deux niveaux, la pyramide spatiale induit une distance exacte des sommets du carré alors que la hiérarchie fait deux erreurs avec 3 niveaux et la pyramide en fait une avec 4 niveaux.

Exemple d'application :

L'approche pyramidale en marketing permet de détecter les segments de clientèles intermédiaires. Par exemple, entre deux segments de clientèles bien déterminées à ne pas changer de comportement, c'est sur la classe intermédiaire (dont les individus sont caractérisés par un objet symbolique), fournie par la pyramide spatiale que l'on pourra utiliser une campagne de publicité afin de conduire les clients de cette classe à modifier leur comportement dans le bon sens. La baisse des effectifs de cette classe constatée par sondage, dans un sens ou dans l'autre des deux classes dont elle est intermédiaire, permettront de mesurer l'efficacité de la campagne.

7) Applications industrielles

7.1 Le champs d'applications

Il est difficile de donner une liste exhaustive du champ d'application de l'ADS, tant ses domaines d'application potentiels sont nombreux puisqu'ils s'adressent à tout domaine (économique, social, médical, industriel etc.) où des données sont recueillies.

On a ainsi appliqué l'ADS à des domaines aussi divers que l'analyse de performances, la détection de la fraude bancaire, à la recherche et l'identification de comportements atypiques et à la détection et caractérisation d'anomalies par exemple sur des bâtiments industriels. On l'a appliqué à la segmentation de clients ou de produits et à la création d'indicateurs pertinents pour l'analyse d'un marché ou d'un segment et ses perspectives. On l'a appliqué à l'analyse géomarketing (classification de zones suivant leur attractivité) à l'explication de comportements (classes de client, populations…), à la définition de scénarios (scénarios d'accident, d'évolution, …), à la définition de prototypes (zone-type, client-type, scénario-type, etc.), à l'évaluation et la mise en œuvre de nouveaux leviers d'action (sur les clients, les salariés, la concurrence, etc.). On l'a appliquée pour la caractérisation de trajectoires de patients dans les hôpitaux, l'analyse de trajectoires de transport Origine-Destination. De façon générale, les applications potentielles sont considérables pour tous les domaines où il s'agit de réduire la taille des données par des résumés décrivant des entités (ie concepts) intéressantes pour l'utilisateur, à l'aide d'un faible nombre de variables symboliques puis d'analyser et fouiller..

7.2 Un exemple d'application industrielle : détection d'anomalies détectées sur un pont suite au passage de TGV

7.2.1 Les données

Les données fournies par le LCPC (Laboratoire Central des Ponts et Chaussées) sont constituées d'un ensemble de 14 TGV qui en passant à une température donnée sur un pont déclenchent des signaux de 9 capteurs répartis à différents endroits du pont. En entrée (voir la figure 22), on dispose d'un tableau de données symboliques qui contient dans la case (i, j) le signal déclenché par le capteur (i.e. accélérateur) j pour le TGV i à une température précisée: chaque case peut contenir jusqu'à 800.000 valeurs. Il s'agit de détecter des anomalies parmi les TGV et de les caractériser.

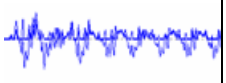
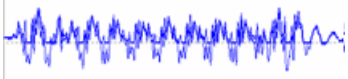
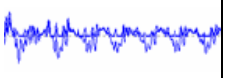
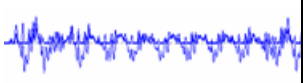
TGV	Accélérateur 1	Accélérateur 2		Accélérateur 9
TGV1 23°23				
TGV 14 24°40				

Figure 22. Chaque ligne représente un TGV passant sur un pont à une certaine température: chaque case du tableau contient jusqu'à 800.000 valeurs de signaux.

7.2.2 Transformation des signaux en variables à valeur histogramme

Dans la figure 23 on représente un tableau d'histogrammes associés à chaque TGV pour chaque capteur (ie « accélérateur »). Ces histogrammes peuvent être obtenus de divers façons ; par exemple, par projection sur l'axe des ordonnées ou par ondelettes. Comme la représentation est très concentrée, on ne voit pas bien les détails des histogrammes. Néanmoins, il est possible de zoomer plus ou moins des cases choisies de ce tableau comme le montre la figure 23.

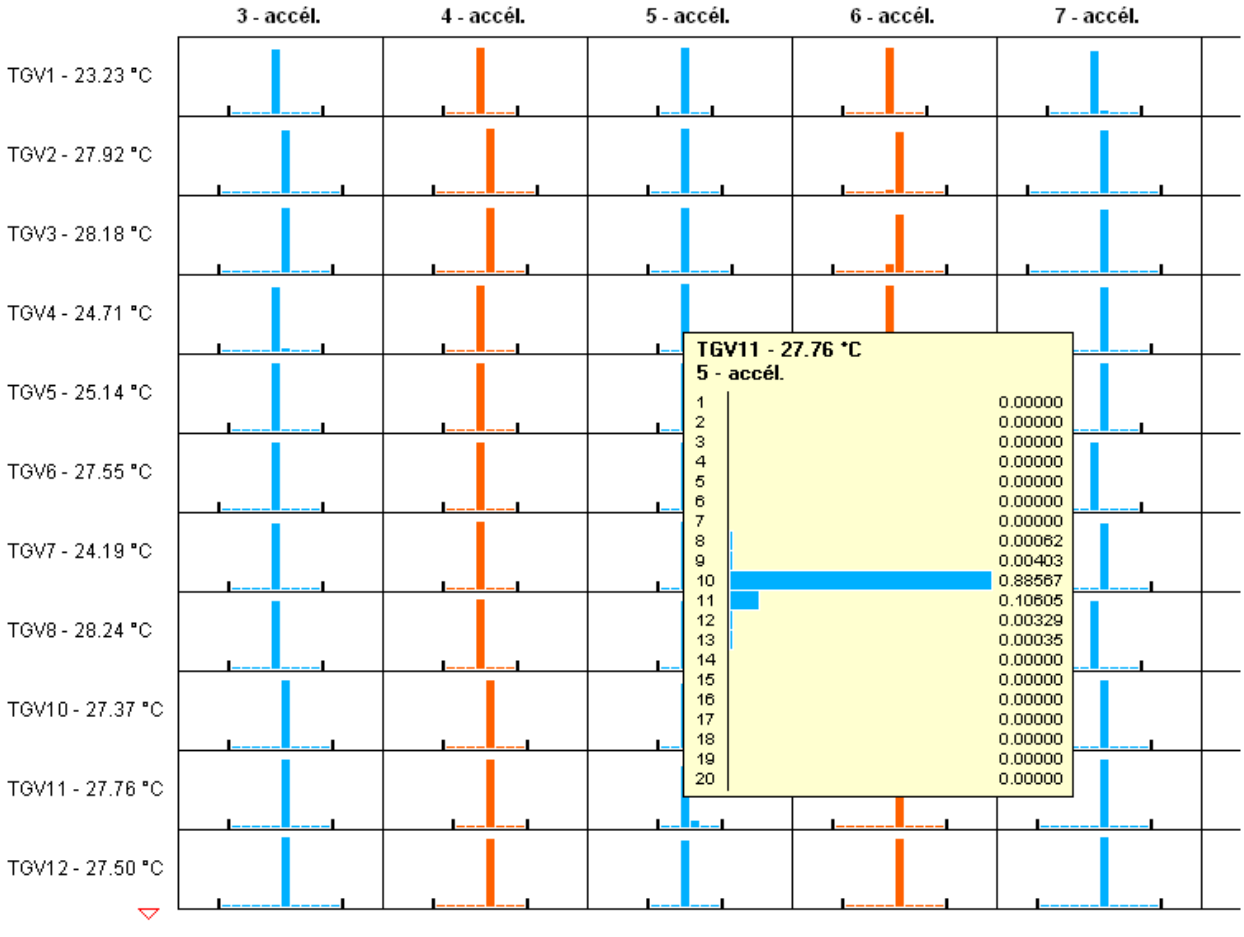


Figure 23 Représentation du tableau de la figure 22 par des histogrammes associés à chaque capteur et pour chaque TGV**.** Le TGV 11 pour le capteur 5 est zoomé.

On peut également, afficher la variation des fréquences pour chaque intervalle de l'histogramme d'une variable à valeur histogramme donnée. Ainsi en utilisant le tableau des histogrammes de la figure 23 et la variable à valeur histogramme associée à l'accélérateur 6, on obtient la figure 24. Dans cette figure est représentée la variation des histogrammes de cet accélérateur pour tous les TGV. Plus précisément, pour chaque intervalle de ces histogrammes le sup, le inf et la moyenne sont représentés. On voit que c'est pour les intervalles 10 et 11 que la variation des fréquences est la plus forte.

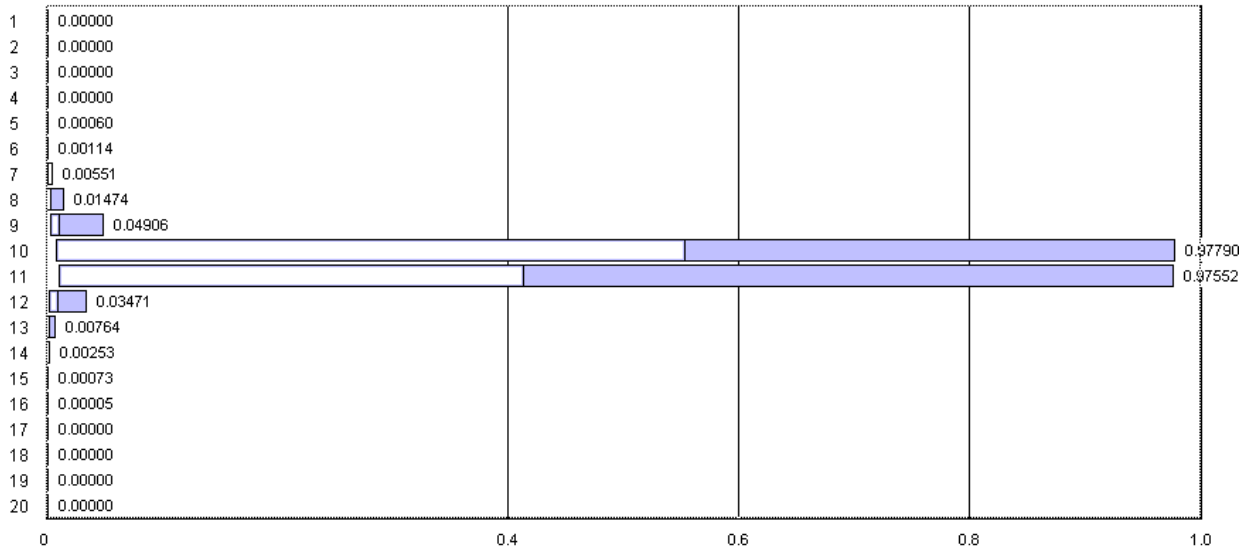


Figure 24 Variation des histogrammes de la variable à valeur histogramme associée à l'accélérateur 6 du tableau de la figure 23.

7.2.3 Analyse en composantes principales de variables à valeur intervalle

L'analyse factorielle en composantes principales de variables à valeur intervalle a pour objectif de positionner les concepts sur un plan en respectant au mieux leur « ressemblance » et en tenant compte de la variation de leurs instances. Le lecteur intéressé par cette méthode pourra se reporter par exemple à Cazes et al (1997) ou à Lauro et al. (chapitre de Diday, Noirhomme (2007)). Ici chaque concept est un TGV décrit par l'intervalle interquartile de chaque histogramme associé à chaque capteur. Le premier plan factoriel appliqué au tableau des intervalles ainsi obtenu exprime plus de 90% de l'inertie. Il est représenté en figure 25. On voit clairement apparaître l'anomalie du TGV1 (noté ici TVG1) qui est en dehors de son groupe de basse température (TGV7, TGV4, TGV5, TGV14) et l'anomalie du TGV 14 qui recouvre la classe des TGV de basse température. Notons qu'une ACP classique où chaque TGV aurait été représenté par un point n'aurait pas permis de détecter aussi facilement l'anomalie du TGV 14.

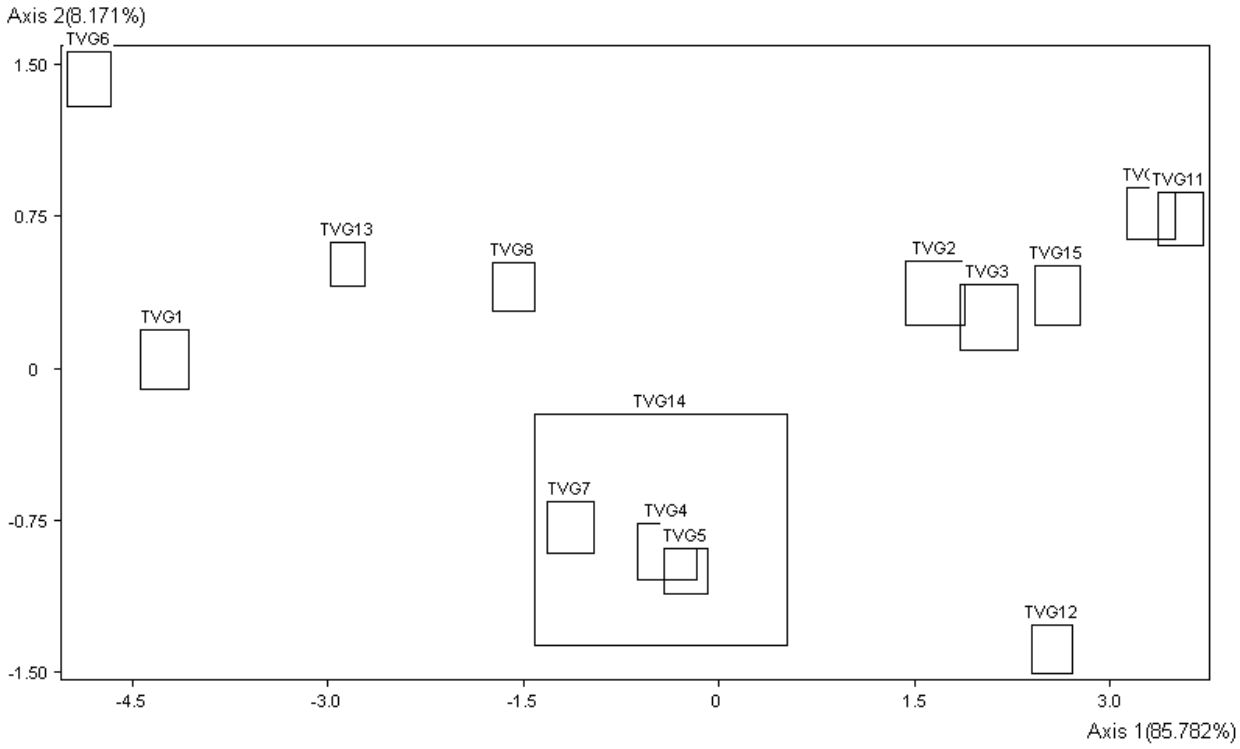


Figure 25. ACP symbolique. Le TGV1 (noté ici TVG1) est en dehors de son groupe de température et l'anomalie du TGV 14 qui recouvre la classe des basses températures n'aurait pas été détecté par une ACP classique où chaque TGV aurait été représenté par un point.

7.2.4 L'approche pyramidale

L'approche pyramidale permet de caractériser les classes en les organisant sous forme de paliers. La pyramide permet de représenter les classes recouvrantes et découvrir des ordres et sous ordres dans une population.

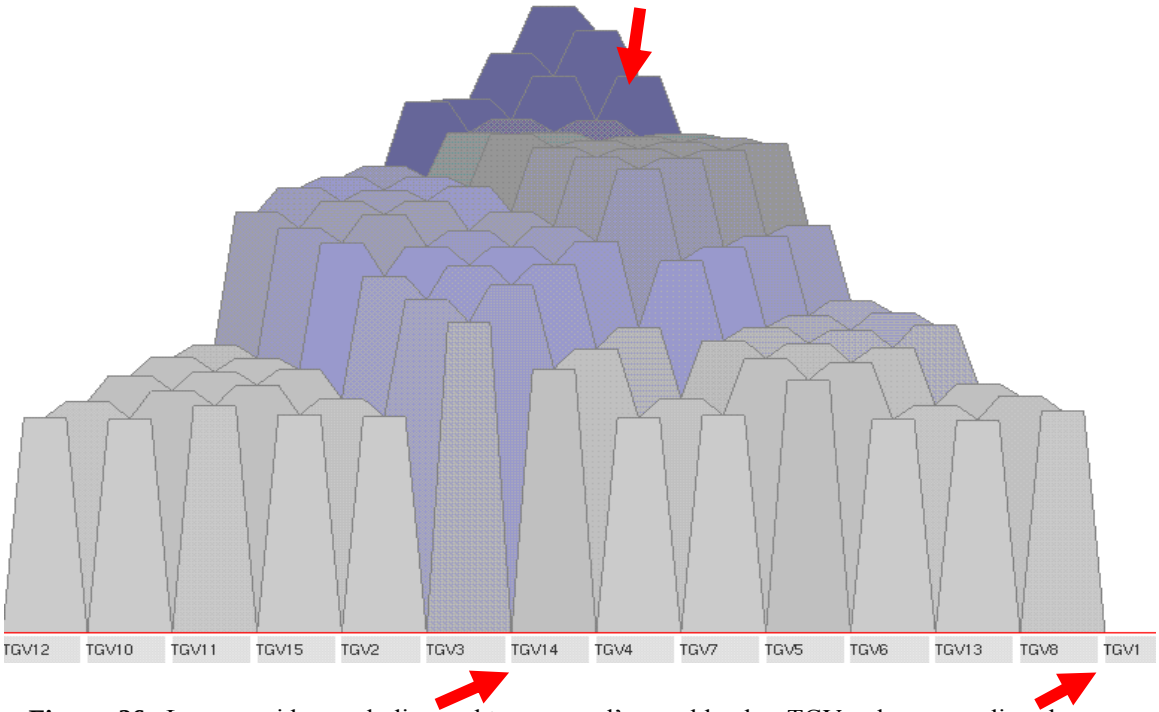


Figure 26. La pyramide symbolique obtenue sur l'ensemble des TGV : des anomalies de regroupement contraires à la température sont constatées.

En entrée de cette méthode, l'utilisateur doit choisir les variables qui seront utilisées pour construire la pyramide. Ces variables peuvent être continues (des valeurs réelles), des intervalles de valeurs réelles ou bien des histogrammes. L'utilisateur est invité à choisir entre les différents types de variables (discrètes, continues, à valeur histogramme, intervalle…) mais il lui est également possible de les mélanger.

La Figure 26 montre le résultat obtenu en appliquant le module HIPYS (voir Pak et al (2005)) sur le tableau de la figure de la figure 23. Les différents paliers de la pyramide peuvent être décrits sous forme de conjonction de propriétés (i.e. descriptions symboliques) définies par les capteurs. La pyramide symbolique fait apparaître le TGV1 en dehors de son groupe de température et on voit aussi que le TGV 14 couvre tous les TGV de son groupe de température. On peut associer à chaque palier un objet symbolique permettant d'identifier la classe de nouveaux TGV.

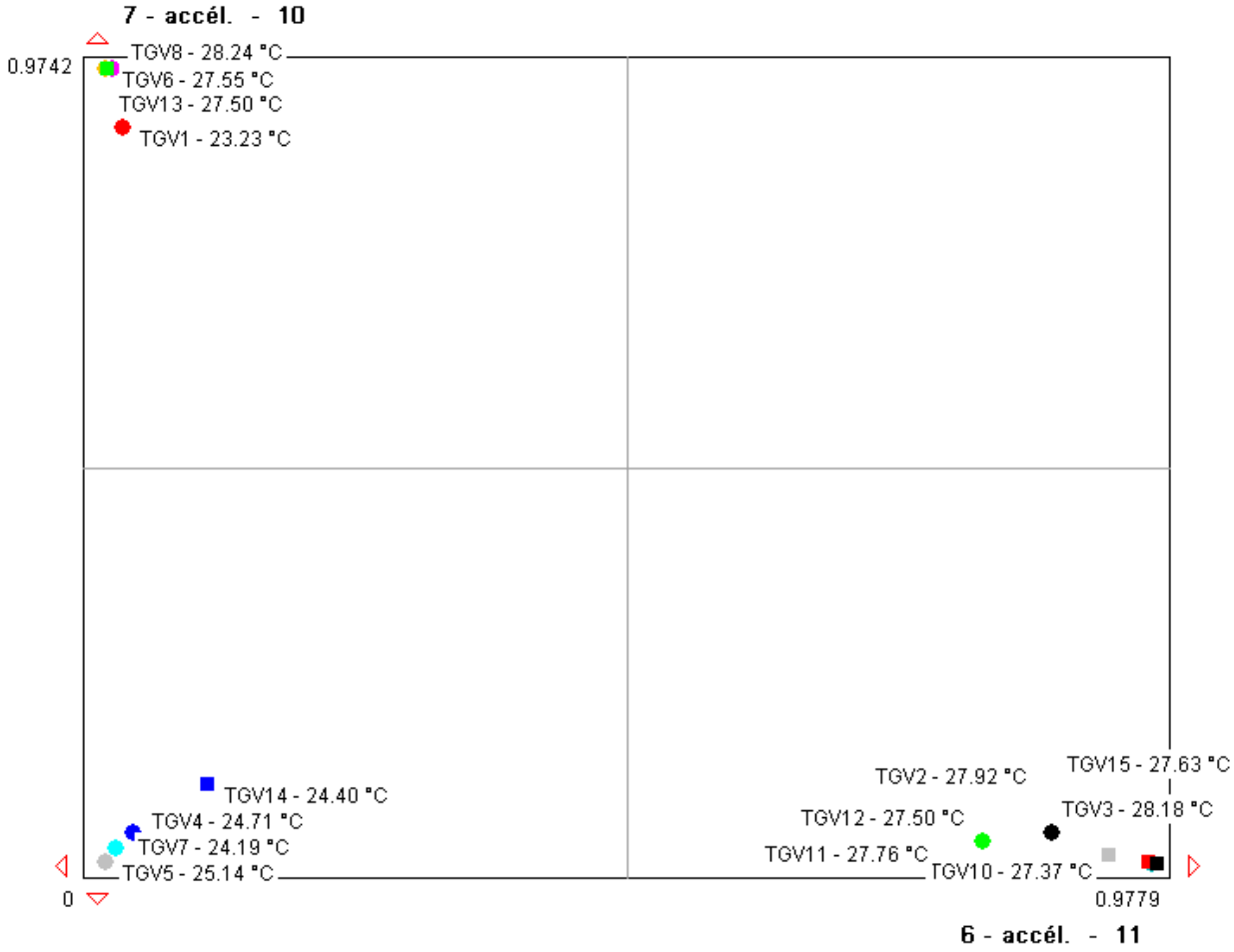


Figure 27 Représentation des TGV munis de leur température selon la modalité 11 du capteur 6 et la modalité 10 du capteur 7.

5.2.5 Réduction de variables par croisements de variables symboliques

En utilisant la variation des histogrammes parmi les TGV pour chaque capteur, on s'est aperçu que ce sont les intervalles 10 et 11 qui varient le plus. En croisant les variables

associées aux capteurs deux à deux, on voit apparaître des croisements intéressants. Ainsi, dans la Figure 27, on peut constater que les capteurs 6 et 7 et les modalités 10 et 11 de leurs histogrammes suffisent à distinguer clairement trois groupes de TGV.

Le groupe des TGV 1, 6, 8, 13 montre une fois de plus le positionnement anormal du TGV 1 de température 23, 23 ° dans un groupe de température supérieure à 27°50. Ce groupe est caractérisé par une forte fréquence de la modalité 10 du capteur 7 et une faible fréquence de la modalité 11 du capteur 6.

Le groupe des TGV 5, 7, 4, 14 de température basse où le 14 est excentré. Ce groupe est caractérisé par une faible fréquence des modalités 10 et 11 des capteurs 7 et 6. Enfin, un dernier groupe de TGV à température supérieur à 27°30 qui est caractérisé par une faible fréquence de la modalité 10 du capteur 7 et une grande fréquence de la modalité 11 du capteur 6.

Ainsi**,** l'ADS a permis de découvrir des groupes caractéristiques de TGV faisant ressortir des anomalies et de montrer que peu de capteurs suffisent à les distinguer. On a utilisé pour cela une représentation des signaux fournis par les capteurs sous forme d'histogramme et sous forme d'intervalle interquartile. Cette étude a été présentée à EGC 2008 (voir Diday, Crémona et al (2008)).

4) La plateforme SODAS d'Analyse de Données Symboliques

La plateforme SODAS est issue de deux projets européens soutenus par EUROSTAT impliquant des laboratoires de 9 pays européens: Namur (FUNDP, Belgique), Naples (DMS, Italie), Paris (Dauphine, France), Aachen (RWTH, Allemagne), Porto(FEP, Portugal), Bari (DIB, Italie), Athens (UOA, Grèce), Madrid (UC, Espagne), Luxembourg (CRP, Luxembourg). quatre Instituts Nationaux de Statistique: INE (Portugal), STATFI (Finlande), EUSTAT (Espagne), ONS (Royaume Uni), un centre de recherche: INRIA (Rocquencourt, France), quatre compagnies: CISIA (Paris)), TES (Luxembourg), EDF (Clamart, France), THOMSON (St Quentin en Yvelines, France)

Une fois que la base de données (ACCESS, ORACLE, ..) est constituée, que les individus (du premier ordre) et concepts sont définis par une requête dans un tableau de données classique, on peut appliquer la première étape de l'ADS et créer le tableau de données symboliques grâce au module DB2SO (voir la thèse de V. Stéphan ou le chapitre 2 dans Diday, Noirhomme (2008) et la section 3.1 dans cet article). Pour la seconde étape de l'ADS, il faut créer une chaîne ou « filière » dans SODAS en précisant la base sur laquelle les analyses vont s'effectuer. Une filière (comme dans le logiciel SPAD de CISIA) est une succession de méthodes d'ADS appliquées au tableau de données symboliques.

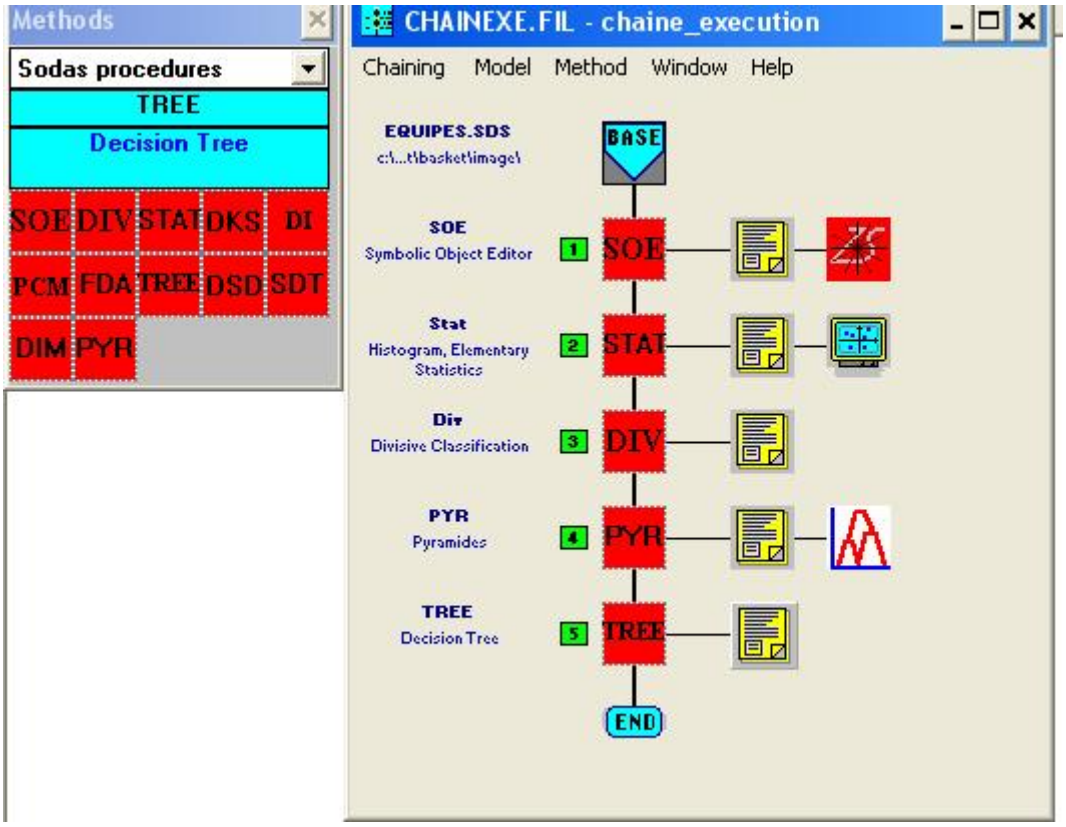


Figure 28 Une filière de la plateforme SODAS.

En haut d'une filière de SODAS (voir figure 28) se trouve une case appelée « base » dans laquelle on peut cliquer pour avoir des détails sur la base des données symboliques constituée d'un fichier XML aux normes de SODAS. Ensuite, on insère des méthodes (ici, les méthodes SOE, STAT, DIV etc.). Les méthodes en rouge ont déjà été exécutées, contrairement à celles en grise. Pour fixer les paramètres d'une méthode, il suffit de faire un clic droit sur son nom, puis de choisir *Parameters*. Lorsque ceux-ci sont définis, il faut lancer la méthode (*Run Method*) pour pouvoir ensuite visualiser les données. Le fichier en jaune correspond aux données résultats (par exemple, pour la méthode SPCA, on aura les valeurs et vecteurs propres ainsi que les différents indices de qualité). Les graphes en rouge ou en bleu permettent quant à eux de visualiser graphiquement les résultats des méthodes exécutées.

Remarquons enfin que dans SODAS, il est possible d'attribuer un pois à chaque individu de second ordre. Ce poids peut exprimer un degré d'appartenance au concept auquel il est associé ou bien l'importance à lui donner de façon générale du fait d'une enquête par sondage par exemple.

Dans le site de SODAS au CEREMADE à Dauphine (cliquer sur LISE) ou faire directement www.ceremade.dauphine.fr/%7Etouati/sodas-pagegarde.htm, on peut charger SODAS et on pourra trouver aussi plusieurs mémoires d'étudiants expliquant pas à pas l'utilisation du logiciel en l'illustrant par des exemples d'application dont les bases de données associées sont mises à disposition..

9) Développement et diffusion de l'Analyse des Données Symboliques

En ce qui concerne la diffusion de l'ADS, deux ouvrages collectifs (Bock, Diday (2000) chez Springer, Diday, Noirhomme (2008) chez Wiley) résume l'ensemble des travaux réalisés

dans le cadre des deux projets européens SODAS et ASSO. On trouvera aussi beaucoup de sections consacré aux dernières recherches en ADS dans le livre édité chez Springer par P. Brito, P. Bertrand, G. Cucumel, F. De Carvalho en 2007. Une revue internationale D'Analyse de Données Symboliques a été créée: Electronical Journal of SDA (JSDA) at www.jsda.unina2.it/newjsda/volumes/index.htm.

Une entreprise SYROKKO (www.syrokko.com) a été créée pour valoriser SODAS et l'ADS former à l'ADS, créer un logiciel complémentaire et faire des études pour l'industrie et la fonction publique.

Conclusion

Par rapport aux approches classiques, l'analyse des données symboliques présente les caractéristiques et ouvertures suivantes. En entrée elle part d'une base de données relationnelle d'où elle constitue des concepts qu'elle décrit par des données symboliques (variables à valeurs multiples, intervalle, histogramme, distribution de probabilité, de possibilité, capacité...) munies de règles et de taxonomies et peut fournir en sortie des connaissances nouvelles par exemple, sous forme d'objets symboliques. Elle utilise des outils adaptés à la manipulation d'objets symboliques de généralisation et spécialisation, d'ordre et de treillis, de calcul d'extension, d'intension et de mesures de ressemblances ou d'adéquation tenant compte des connaissances sous-jacentes basées sur les règles et taxonomies. Elle fournit des représentations graphiques exprimant entre autres la variation interne des concepts. Par exemple, en analyse factorielle, un concept sera représenté par une zone (elle même exprimable sous forme d'objet symbolique) et pas seulement par un point.

Ses principaux avantages sont d'abord de permettre d'étudier les bonnes unités statistiques: les assurés plutôt que les feuilles de maladies ou des régions plutôt que des habitants, etc.. Le second avantage est la réduction de la taille des données en considérant comme unités d'étude, des classes plutôt que les individus. Ainsi, en passant des feuilles de maladies aux assurés on divise la taille (i.e. le nombre de lignes) par le nombre de feuilles de maladie.

Le troisième avantage est la réduction du nombre de variables du fait qu'elles sont à valeur symbolique. Par exemple, à valeur « histogramme » plutôt qu'à valeur «fréquence d'une catégorie» ou à valeur intervalle plutôt qu'à valeur « borne d'intervalle ». Une variable à valeur histogramme à 20 classes ce n'est pas 20 variables à valeur numériques. La transformation en histogramme réduit considérablement: dans l'exemple des TGV : on est ainsi passé de 800. 000 valeurs à un histogramme à 20 modalités.

Les principaux avantages des descriptions symboliques peuvent se résumer comme suit :

- Elles fournissent un résumé de la base plus riche que les données agrégées habituelles car ils tiennent compte de la variation interne et des règles sous-jacentes aux classes décrites, ainsi que des taxonomies fournies.
- Elles sont explicatives, puisqu'ils s'expriment sous forme de propriétés des variables initiales ou de variables significatives obtenues (axes factoriels).
- Obtenues en sorties des méthodes elles sont au même format que les entrées et permettent donc de construire un nouveau tableau de données de plus haut niveau sur lequel une analyse de données symbolique de second niveau peut s'appliquer.

Moralité : dans votre travail vérifiez si vos unités d'étude sont des individus ou des concepts. Si ce sont des individus demandez-vous s'il n'y aurait pas des catégories d'individus (induits par des variables qualitatives intéressantes ou une typologie) à étudier en tant que concepts.

Si ce sont des concepts pensez à prendre en compte leur variation interne (i.e. des individus de leur extension) pour les décrire par des variables symboliques munies de connaissances supplémentaires.

Nous avons montré que la représentation des données et connaissances n'est pas seulement un domaine d'utilisation normal des outils standards de la Statistique, de la Fouille de Données (Data Mining) ou de l'Analyse des Données plus ou moins complexes, mais de plus, le fait de s'intéresser aux connaissances et aux descriptions de concepts qui en forment les atomes en tant qu'unités d'étude remet totalement en cause ces outils et nécessite leur renouvellement complet aussi bien dans leur théorie que dans leur pratique et dans la façon de les penser.

Si on peut dire que l'Analyse des données a rendu les individus à la statistique, alors on peut dire aussi que l'Analyse des Données Symboliques lui rend les concepts.

Le champs de recherche et d'application est immense puisqu'il faut tout reprendre en AD, STAT et Data Mining en pensant autrement, c'est à dire en termes de concepts et de données symboliques plutôt que d'individus décrits par des données classiques ou complexes: on manque de bras!

References

Afonso F., Billard L., Diday E. (2004): Symbolic linear regression with taxonomies, Studies in classification, Data Analysis and Knowledge organization: Classification, Clustering and Data Mining Applications (Proc. IFCS'2004), D. Banks, L. House, F. R. McMorris, P. Arabie, et W. Gaul eds, p. 429-437, Springer.

Afonso F., Diday E. (2005): Extension de l'algorithme Apriori et des règles d'association aux cas des données symboliques diagrammes et intervalles. Revue RNTI, Extraction et Gestion des Connaissances (EGC 2005), Vol. 1, pp 205-210, Cépadues, 2005.

Appice A., D'Amato C., Esposito F., Malerba D. (2006): Classification of Symbolic Objects: A Lazy Learning Approach. Intelligent Data Analysis, 10 (4), 301 – 324.

Barbut M., Monjardet B. (1970) : Ordre et Classification, Hachette, Paris.

Bandemer, H., Nather, W. (1992): Fuzzy Data Analysis. Kluwer, Dordrecht.

Bertrand P., Bel Mufti G. (2006): Loevinger's measures of rule quality for assessing cluster stability, Computational Statistics and Data Analysis, vol. 50/4, pp 992-1015.

Benzécri, J.P. and al. (1973): l'Analyse de Données. Vol. 1, 2. Dunod, Paris.

Billard L., Diday E. (2003): From the statistics of data to the statistic of knowledge: Symbolic Data Analysis. JASA . Journal of the American Statistical Association. Juin, Vol. 98, N° 462.

Billard L. (2004): Dependencies in bivariate interval-valued symbolic data.. In: Classification, Clustering and New Data Problems . Proc. IFCS'2004. Chicago. Ed. D. Banks. Springer Verlag, 319-354.

L. Billard, E. Diday (2006) Symbolic Data Analysis: conceptual statistics and data Mining. 321 pages. Wiley series in computational statistics. Wiley. ISBN 0-470-09016-2.

Birkhoff G., (1967): Lattice Theory, AMS Colloq. Public. Vol. XXV.

Bock H.H. (2005): Optimization in symbolic data analysis: dissimilarities, class centers, and clustering. In: D. Baier, R. Decker, L. Schmidt-Thieme (eds.): Data Analysis and Decision

Support.Studies in Classification, Data Aanlysis, and Knowledge Organization. Springer, Heidelberg, 3-10.

Bock H.H., Diday (2000) E.: Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data. Springer Verlag, Heidelberg, 425 pages, ISBN 3-540-66619-2.

Bravo M.C., García-Santesmases J.M. (2000): Symbolic Object Description of Strata by Segmentation Trees, Computational Statistics, 15, Physica-Verlag, 13-24.

Bravo M.C. (2001): Análisis de Segmentación en el Análisis de Datos Simbólicos. Ed. Universidad Complutense de Madrid. Servicio de Publicaciones. ISBN:8466917918. (http://www.ucm.es/BUCM/tesis/mat/ucm-t25329.pdf)

Brito, P. (1991): Analyse des données symboliques. Pyramides d'héritage. Thèse de doctorat. Université Paris – Dauphine. France.

Brito, P. (1994): Order structure of symbolic assertion objects . IEEE Trans. On Knowledge and Data Engineering 6,5.

Brito, P. (2002): Hierarchical and Pyramidal Clustering for Symbolic Data, Journal of the Japanese Society of Computational Statistics, Vol. 15, Number 2, pp. 231-244.

Brito, P., Polaillon G., (2005) : Structuring Probabilistic Data by Galois Mathématiques et Sciences Humaines, 43ème année, nº 169, (1), pp. 77-104.

P. Brito, P. Bertrand, G. Cucumel, F. De Carvalho éditeurs (2007): Selected Contributions in Data Analysis and Classification. Springer Verlag, Heidelberg, 634 pages, ISBN 978-3-540-73558-8.

Caruso C., Malerba D., Papagni D. (2005): Learning the daily model of network traffic. In M.S. Hacid, N.V. Murray, Z.W. Ras, S. Tsumoto (Eds.) Foundations of Intelligent Systems, 15th International Symposium, ISMIS'2005, Lecture Notes in Artificial Intelligence, 3488, 131-141, Springer, Berlin, Germania.

P.B. Cerrito (2007): Introduction to Data Mining Using SAS Enterprise Miner, SAS Publishing, 486p. Carolyn K. Hamm (2006), Oracle Data Mining: Mining Gold from Your Warehouse, Rampant Techpress, 300p.

Cazes, P., Chouakria, A., Diday, E. Schektman, Y. (1997): Extension de l'analyse en composantes principales à des données de type intervalle, Revue de Statistique Appliquée XIV(3), 5–24.

Ciampi A., Diday E., Lebbe J., Perinel E., R. Vignes (2000): Growing a tree classifier with imprecise data. Pattern Recognition letters 21, pp 787-803.

Cuvelier, E., Noirhomme-Fraiture, M. (2005): Clayton copula and mixture decomposition, In Jacques Janssen and Philippe Lenca, editors, Applied Stochastic Models and Data Analysis ASMDA 2005, pages 699-708. Brest, France 17-20 May

Da Silva A., De Carvalho F., Lechevallier Y., Trousse B. (2006): Mining Web Usage Data for Discovering Navigation Clusters. In: XI IEEE Symposium on Computers and Communications (ISCC 2006), Pula-Cagliari, Italy.

De Carvalho F.A.T., Eufrasio de Lima Neto A., P.Tenerio (2004): A new method to fit a linear regression model for interval-valued data. In: Advances in Artificial Intelligence: Proceedings of the Twenty Seventh German Conference on Articial Intelligence (eds. S. Biundo, T. Fruchrirth, and G. Palm). Springer-Verlag, Berlin, 295-306.

De Carvalho F.A.T., De Souza R., Chavent M., Y. Lechevallier (2006): Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. Pattern Recognition Letters, 27 (3), 167-179

De Carvalho, F. Brito, P. & Bock, H.-H. (2006): Dynamic Clustering for Interval Data Based on L2 Distance. Computational Statistics, 21 2, pp 231-250.

De Carvalho, F. A. T. (1995): Histograms In Symbolic Data Analysis. Annals of Operations Research,Volume 55, Issue 2, 229-322.

De Souza, R. M. C. R. and De Carvalho, F. A. T. (2004): Clustering of Interval Data based on City-Block Distances. Pattern Recognition Letters, Volume 25, Issue 3, 353-365.

E. Diday, C. Crémona, F. Goupil, F. Afonso, M. Rahal (2008): Principes d'Analyse des données symboliques et application à la détection d'anomalies sur des ouvrages publics. Actes EGC '2008. Extraction et Gestion de Connaissances". INRIA Sophia-Antipolis.

Diday E. (1987 a): The symbolic approach in clustering and related methods of Data Analysis. In "Classification and Related Methods of Data Analysis", Proc. IFCS, Aachen, Germany. H. Bock ed.North-Holland.

E. Diday, M. Noirhomme editors (2007): Symbolic Data Analysis and the SODAS software livre à paraître cette année Wiley.

E. Diday et al. (1980): Optimisation en classification automatique, E. Diday et Coll. INRIA publisher (900 pages).

E. Diday (1977): Analyse canonique du point de vu de la classification automatique, Rapport Laboria n°293. INRIA, 78150 Rocquencourt, France.

E. Diday, M. Noirhomme éditeurs et co-auteurs(2008) : Symbolic Data Analysis and the SODAS Software. 457 pages Wiley. ISBN 978—0-470-01883-5. www.wiley.com

E. Diday, C. Crémona, F. Goupil, F. Afonso, M. Rahal (2008): Principes d'Analyse des données symboliques et application à la détection d'anomalies sur des ouvrages publics. Actes EGC '2008. Extraction et Gestion de Connaissances". INRIA Sophia-Antipolis.

E. Diday, Y. Kodratoff, P. Brito, M. Moulet (2000): Induction symbolique numérique à partir de données. Cépadues. 31100 Toulouse. www.editions-cepadues.fr. 442 pages

E. Diday (2005): Categorization in Symbolic Data Analysis. In handbook of categorization in cognitive science. Edited by H. Cohen and C. Lefebvre. Elsevier editor. http://books.elsevier.com/elsevier/?isbn=0080446124

E. Diday (2008): Spatial classification. DAM (Discrete Applied Mathematics) Volume 156, Issue 8, Pages 1271-1294.

E. Diday, N. Murty (2005) Symbolic Data Clustering in Encyclopedia of Data Warehousing and Mining . John Wong editor . Idea Group Reference Publisher.

E. Diday, Mathieu Vrac (2005): Mixture decomposition of distributions by Copulas in the symbolic data analysis framework. Discrete Applied Mathematics (DAM). Volume 147, Issue 1, 1 April, Pages 27-41

E. Diday , R. Emilion (2003): Maximal and stochastic Galois Lattices. Journal of Discrete Applied Mathematics . 127 , 271-284.

Diday E. (1987 b): Introduction à l'approche symbolique en Analyse des Données. Première Journées Symbolique-Numérique. Université Paris IX Dauphine. Décembre 1987.

Diday E. (1989): Introduction à l'Analyse des Données Symboliques. Rapport de Recherche INRIA N° 1074 (August 1989). INRIA Rocquencourt 78150. France.

Diday E. (1991) : Des objets de l'Analyse des Données à ceux de l'Analyse des Connaissances. In « Induction Symbolique et Numérique à partir de données ». Y. Kodratoff,

Diday E. Editors. CEPADUES-EDITION.ISBN 2.85428.282 5.

Diday E. (2000): L'Analyse des Données Symboliques : un cadre théorique et des outils pour le Data Mining. In : E. Diday, Y. Kodratoff, P. Brito, M. Moulet « Induction symbolique numérique à partir de données ». Cépadues. 31100 Toulouse. www.editions-cepadues.fr. 442 pages.

Diday E., Lemaire J., Pouget J., Testu G. (1984): Eléments d'Analyse des Données. Dunod.

Diday E. (2002 a): An introduction to Symbolic Data Analysis and the Sodas software. Journal of Symbolic Data Analysis. Vol. 1, n° 1. International Electronic Journal. www.jsda.unina2.it/JSDA.htm.

Diday E. (2002 b): Mixture Distributions of Distributions by Copulas. Proceedings of IFCS'2002 (Cracovia, Poland). In Krzysztof Jajuga et al (Eds.): Data Analysis, Classification and Clustering Methods. Heidelberg, Springer-Verlag.

Diday E., Esposito F. (2003): An introduction to Symbolic Data Analysis and the Sodas Software IDA. International Journal on Intelligent Data Analysis". Volume 7, issue 6.

Diday E., Emilion R. (2003): Maximal and stochastic Galois Lattices. Journal of Discrete Applied Mathematics, Vol. 127, pp. 271-284.

Diday E., Emilion R. (1996): Lattices and Capacities in Analysis of Probabilist Objects. Proceed. of OSDA (Ordinal and Symbolic Data Analysis Conference). Springer Verlag Editor

E. Diday, R. Emilion (1997): Treillis de Galois Maximaux et Capacités de Choquet. C.R. Acad. Sc. t.325, Série 1, p 261-266. Présenté par G. Choquet en Analyse Mathématiques.

Diday E., Pelc A., Wagne I., (2004): La dispensation des médicaments chez les personnes âgées de 70 ans et plus : approche par la méthode des données symboliques. Journées d'études sur l'Assurance Maladie (26 Mars 2004, Paris).

Diday E. (2004): Spatial Pyramidal Clustering Based on a Tessellation. Proceedings IFCS'2004, In Banks andal. (Eds.): Data Analysis, Classification and Clustering Methods Heidelberg, Springer-Verlag.

Diday E., Vrac M. (2005): Mixture decomposition of distributions by Copulas in the symbolic data analysis framework. Discrete Applied Mathematics (DAM). Volume 147, Issue 1, 1 April, Pages 27-41.

E. Diday (2005): Categorization in Symbolic Data Analysis. In handbook of categorization in cognitive science. Edited by H. Cohen and C. Lefebvre. Elsevier editor. http://books.elsevier.com/elsevier/?isbn=0080446124

Diday E.(1995): Probabilist, possibilist and belief objects for knowledge analysis. Annals of Operations Research. 55, pp. 227-276.

Diday E., Emilion R. (1995): Capacities and credibilities in Analysis of Probabilist Objects. Proceed. of OSDA (Ordinal and Symbolic Data Analysis Conference). Studies in Classification.. Springer

Diday E. (2001): A generalisation of the mixture decomposition problem in the symbolic data analysis framework, CEREMADE Rapport, vol. 112, May 2001, pp. 1–14.

Diday E., Murty N. (2005): Symbolic Data Clustering. In Encyclopedia of Data Warehousing and Mining . John Wong editor . Idea Group Reference Publisher.

Duarte Silva, A. P., Brito, P. (2006): Linear Discriminant Analysis for Interval Data,

Computational Statistics, 21(2):289-308, June 2006.

Dubois D., Prades H. (1988): Possibility theory. Plenum. New-York.

Duquenne V. (1996): On lattice approximations: syntactic aspects. Social Networks 18, pp 217-230.

Esposito, F., Malerba, D., Semeraro (1991): Classification of incomplete structural descriptions using a probabilist distance measure. In: Diday and Lechevallier (eds.): Symbolic-Numeric Data Analysis and Learning. Nova Science Publisher, Nerw York, 469-482.

Esposito, F., Malerba, D., Semeraro (1992): Classification in noisy environments using a distance measure between structural symbolic descriptions. IEEE Transactions on Pattern Analysis and Machine intelligence PAMI-14 (3) , 390-402.

Frank M.J. (1979): On the simultaneous associativity of $f(x, y)$ and $x + y - f(x, y)$, Aequationes Math. 19 53–77.

Genest C (1987): Frank's family of bivariate distributions, Biometrika 74 549–555.

Gioia F.,, Lauro C. (2005): Basic Statistical Methods for Interval Data, Statistica applicata, 1.

Gowda K.C., Diday E. (1991): Symbolic Clustering Using a New Similarity Measure, IEEE Tr. on Systems, Man and Cybernetics. Vol. 22, N° 2, 1992.

Groenen, PJF, Winsberg, S, Rodriguez, O, and Diday, E (2006): "I-Scal Multidimensinal scaling of interval dissimilarities" . Computational Statistics and Data Analysis, 51, 360-378.

Han J. , Kamber M. (2006), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 800p. IBM Redbooks (199), Intelligent Miner for Data Applications Guide, IBM, 172p.

Hardy, A. (2005): Validation in unsupervised symbolic classification, In Jacques Janssen and Philippe Lenca, editors, Applied Stochastic Models and Data Analysis ASMDA 2005, 379-386.Brest, France17-20 May.

Irpino, A. (2006): Spaghetti PCA analysis: An extension of principal components analysis to time dependent interval data. Pattern Recognition Letters, Volume 27, Issue 5, 504-513.

Irpino A., Verde R., Lauro N. C. (2003): Visualizing symbolic data by closed shapes, Between Data Science and Applied Data Analysis, Shader-Gaul-Vichi eds., Springer, Berlin, pp. 244-251.

Larose D.T., Vallaud T. (2005): Des données à la connaissance : Une introduction au data-mining, Vuibert, 223p.

Lebart L., Piron M., Morineau A. (2006): Statistiques exploratoire multidimensionnelle : Visualisations et inférences en fouille de données, Dunod, Collection Sciences sup, 464p.

Mballo C., Diday E. (2006): The criterion of Smirnov-Kolmogorov for binary decision tree : application to interval valued variables. Intelligent Data Analysis. Volume 10, Number 4 . pp 325 – 341.

Mballo C. (2005) : Ordre, codage et extension du critére de Kolmogorov-Smirnov pour la segmentation de données symboliques. Thèse. Université Paris Dauphine.

Mirkin B. (2005): Clustering For Data Mining: A Data Recovery Approach,Chapman & Hall, 296p.

Nelsen R.B. (1998): An Introduction to Copulas, in: Lecture Notes in Statistics, Springer, NewYork, 1998.

Pak K., Rahal M.C., Diday E.. (2005): Élagage et aide à l'interprétation symbolique et graphique d'une pyramide. Congrés d'extraction et géstion des connaissances, EGC 18-21 Janvier 2005 Paris, Editions Cepadues.

Saporta G. (2006): Probabilités, analyses des données et statistiques,Editions Technip

Lauro N.C., Verde R., Palumbo F. (2000): Factorial Data Analysis on Symbolic Objects under cohesion constrains In: Data Analysis, Classification and related methods, Springer-Verlag, Heidelberg.

Lauro N.C., Gioia F.(2006): Dependence and interdependence analysis for interval-valued variables. IFCS'2006: V. Batagelj, H.H. Bock, A. Ferligoj, A. Ziberna (eds) Data Science and Classification (2006). Berlin: Springer-Verlag. Pages 171-183.

Limam M., Diday E., Winsberg S. (2004): Symbolic Class Description with Interval Data. Journal of Symbolic Data Analysis, 2004, Vol 1

Malerba D., Esposito F., Monopoli M. (2002): Comparing dissimilarity measures for probabilistic symbolic objects.In A. Zanasi, C. A. Brebbia, N.F.F. Ebecken, P. Melli (Eds.)Data Mining III, Series Management Information Systems, Vol 6, 31-40, WIT Press, Southampton, UK.

Meneses E., Rodríguez-Rojas O. (2006): Using symbolic objects to cluster web documents. WWW 2006: 967-968.

Noirhomme-Fraiture, M. (2002): Visualization of Large Data Sets : the Zoom Star Solution, Journal of Symbolic Data Analysis, vol. 1, July.

<http://www.jsda.unina2.it/>http://www.jsda.unina2.it

Pollaillon G. (1998): Organisation et interprétation par les treillis de Galois de données de type multivalué, intervalle ou histogramme. Thèse de doctorat. Université Paris – Dauphine. France.

Prudêncio R. B. C., Ludermir T., De Carvalho F. de A. T. (2004): A Modal Symbolic Classifier for selecting time series models. Pattern Recognition Letters, 25 (8), 911-921.

Rodriguez O. (2000): Classification et modèles linéaires en Analyse des Données Symboliques. Thèse de doctorat, University Paris 9 Dauphine.

Saporta G. (2006): Probabilités, analyse des données et statistique. Editions Technip, Paris. France.

Schweizer B., Sklar A. (1983): Probabilistic Metric Spaces, Elsevier, North-Holland, NewYork,

Schweizer B. , Sklar A. (2005): Probabilist metric spaces . Dover Publications INC. Mineola, New-York.

Stéphan V. (1998): "Construction d'objets symboliques par synthèse des résultats de requêtes". (1998). Thesis. Paris IX Dauphine University.

Touati M., Rahal M., Quantin C., Le Teuff G., Andreu N., Diday E., Afonso F., Battaglia G., M. Limam (2006) : Analyse de Trajectoires Hospitalières de Patients atteints d'un Infarctus Aigu du Myocarde, Revue des Nouvelles Technologies de l'Information (RNTI), Extraction et Gestion des Connaissances (EGC 2006), Cépadues, 2006.

Tukey, J.W. (1958): Exploratory Data Analysis. Addisson Wesley,READING, Mass. USA.

Vrac M, Diday E., Chédin A. (2004): Décomposition de mélange de distributions et application à des données climatiques. Revue de Statistique Appliquée, 2004, LII (1), 67-96.

Wille R. (1983): Subdirect decomposition of concept lattices. Algebra Universalis 17.

Zadeh L.A. (1978): Fuzzy sets a basis for a theory of possibility. Fuzzy sets and systems, n°1, 3-28.

Ziani D. (1996): Sélection de variables sur un ensemble d'objets symboliques. Thèse, Paris 9 Dauphine.