Analyse de données objectivo-subjectives : approche par la théorie des sous-ensembles flous

Thierry-Marie Guerra

L.A.I.H.

N° d'ordre 9110

THESE

présentée à

L'UNIVERSITE DE VALENCIENNES
ET DU HAINAUT CAMBRESIS

pour l'obtention du

DOCTORAT

spécialité

AUTOMATIQUE INDUSTRIELLE ET HUMAINE

par

Thierry Marie GUERRA
Ingénieur ENSIMEV

ANALYSE DE DONNEES OBJECTIVO-SUBJECTIVES : APPROCHE PAR LA THEORIE DES SOUS-ENSEMBLES FLOUS

Soutenue le 05 juillet 1991 devant la commission d'examen :

Mme	Bernadette	**BOUCHON-MEUNIER**	(Rapporteur)
Mrs	Bernard	**DUBUISSON**	
	Arnold	**KAUFMANN**	
	Noël	**MALVACHE**	
	Dominique	**ROGER**	
	Marc	**ROUBENS**	(Rapporteur)
	Didier	**WILLAEYS**	(Rapporteur)

L.A.I.H.

N° d'ordre 9110

THESE

présentée à

L'UNIVERSITE DE VALENCIENNES
ET DU HAINAUT CAMBRESIS

pour l'obtention du

DOCTORAT

spécialité

AUTOMATIQUE INDUSTRIELLE ET HUMAINE

par

Thierry Marie GUERRA
Ingénieur ENSIMEV

ANALYSE DE DONNEES OBJECTIVO-SUBJECTIVES : APPROCHE PAR LA THEORIE DES SOUS-ENSEMBLES FLOUS

Soutenue le 05 juillet 1991 devant la commission d'examen :

Mme	Bernadette	**BOUCHON-MEUNIER**	(Rapporteur)
Mrs	Bernard	**DUBUISSON**	
	Arnold	**KAUFMANN**	
	Noël	**MALVACHE**	
	Dominique	**ROGER**	
	Marc	**ROUBENS**	(Rapporteur)
	Didier	**WILLAEYS**	(Rapporteur)

à Jeremie, à Murielle et à leur patience
devant cet "accouchement difficile" ...

"La liste n°5, six maillots, six caleçons et six mouchoirs a toujours intrigué les chercheurs et fondamentalement pour la totale absence de chaussettes"

Woody ALLEN, Getting Even, New York
Random House, 1966, «The Metterling List», p.8

Merci ...

Que tous ceux qui veulent bien se reconnaître
ici s'y reconnaissent ...

PLAN

S'il est intéressant de traiter séparément des ensembles de données issus d'une même expérience, objectives et subjectives par exemple, il est également nécessaire de vérifier l'adéquation entre ces différents groupes de données. Le formalisme flou utilisé dans le deuxième chapitre permet dans le troisième chapitre le développement d'une méthode multidimensionnelle basée sur l'inférence déductive et le Modus Ponens généralisé visant à mettre en relation des données de natures différentes.

Le chapitre quatre présente une application des méthodes développées dans les deux chapitres précédents à une analyse ergonomique d'un poste de travail bureautique où des données de nature objectives et subjectives ont été recueillies.

Enfin, le dernier chapitre a trait aux diverses améliorations possibles des méthodes développées et tente de les situer par rapport au problème général de l'analyse et de la modélisation des systèmes.

CHAPITRE I

LES DIFFERENTS ASPECTS DE L'EVALUATION ET DE LA MODELISATION DES SYSTEMES

L'approche pour l'analyse du système à composante humaine étant multicritère - aspects objectif et subjectif - il faut construire des ensembles de variables pertinentes pour exprimer localement chacun des critères et susceptibles de décrire globalement le système. L'évaluation des actions sur le système et l'analyse de son comportement nécessitent alors, d'une part, de bâtir le système d'observations et plus largement, de représenter les données par une symbolique appropriée et, d'autre part, traiter les résultats, voire même simuler à partir d'un modèle théorique, construire des consignes ou des normes... Toutes ces actions sont des fonctions remplies par l'approche mathématique qui peuvent être donc réparties en deux catégories de rôles : des "fonctions de représentation" et des "fonctions d'opérateur".

Dans la pratique, leurs rôles respectifs sont bien souvent confondus, ce qui peut entraîner certaines ambiguités.

I.1 - LA FONCTION DE REPRESENTATION

La première fonction est celle qui sert au minimum de support symbolique pour représenter une information, qu'il s'agisse de codification, de mesure, ou plus largement de tout moyen de description.

Cette fonction est relative à la notion de "mesure" au sens large. Celle-ci s'établit par une mise en correspondance de deux séries, l'une empirique, les aspects objectifs mais aussi subjectifs traduisant l'état du système, et la seconde, formelle, l'échelle de mesure. Pour chacune d'elles, il existe une diversité de relations possibles entre les éléments qui la définissent. La mise en correspondance revient à "*l'affectation de nombres à des objets ou à des événements en fonction de certaines règles*" /STEVENS 74/.

Les relations entre les différents éléments de la série empirique sont représentées par un système de relations formelles, relation d'équivalence, d'ordre, de pré-ordre..., qui répondent à un certain nombre de propriétés telles que symétrie, réflexivité, transitivité,... /DIDAY 82/. La correspondance entre les deux séries repose sur des relations identiques entre les éléments formant chacune d'elles. Il peut s'agir d'une correspondance terme à terme, voire même d'isomorphisme. L'échelle de mesure peut avoir plusieurs formes, plusieurs niveaux de représentation, depuis l'échelle nominale - les rubriques non structurées susceptibles de caractériser des états - jusqu'à l'échelle d'intervalles, en passant par tous les degré d'ordre - ordre partiel, ordre total. Ceci nécessite la mise en oeuvre d'un système d'analyse susceptible de rendre les relations empiriques - les données - compatible avec tel ou tel système formel représentatif - l'échelle de mesure.

Afin de faire apparaître cette compatibilité il faut, au cours de l'examen de l'organisation des données, passer au crible les propriétés et les relations de la série formelle susceptible de lui correspondre. Cette recherche ne prend toute sa signification que lorsque ces propriétés sont utilisées, sans en oublier et sans en rajouter.

En fonction de ces propriétés, l'ensemble des échelles de mesure peut se diviser en quatre grandes familles /THOLE 79/ /CHANDON et PINSON 81/ qui vont être décrites.

I.1-1 L'échelle nominale

Cette échelle, discrète ou qualitative, permet de diviser une population donnée en classes pour lesquelles la variable a une même valeur. Par exemple, à une question relative au lieu d'habitation :

Quel est votre lieu d'habitation :
- ☐ zone rurale
- ☐ ville de moyenne importance (5000 à 20000 hab.)
- ☐ grande ville (>20000 hab.)

les réponses pourront être codées par 1, 2 et 3 dans l'ordre. Bien entendu, il n'est pas possible d'introduire une relation d'ordre sur les classes, 3 > 2 ne voulant rien dire. La seule relation mathématique existante sur une telle échelle est l'équivalence des individus à l'intérieur d'une même classe et les seules transformations mathématiques qui laissent invariantes les données devront conserver cette équivalence, elles correspondent aux fonctions discrètes bijectives.

Enfin, les opérations mathématiques autorisées sur un tel type d'échelle ne peuvent être basées que sur l'équivalence, à savoir la recherche du mode, le comptage d'éléments par classe, le tableau de fréquence,... , les relations métriques, telle que la moyenne n'ayant elles aucun sens.

I.1-2 L'échelle ordinale

L'échelle ordinale comme l'échelle nominale permet de diviser une population en différentes classes, en introduisant en plus une notion d'ordre entre les modalités des variables.

L'utilisation de cette échelle est fréquente dans beaucoup de domaines, tant en psychologie qu'en ergonomie. En effet, une évaluation subjective, peut être effectuée à l'aide de questions formulées de la manière suivante :

comment trouvez-vous la hauteur d'assise | trop haute | bonne | trop basse |

ou, pour l'analyse des tâches l'utilisation de l'échelle de COOPER-HARPER /KAMOUN 89/ est fréquente, figure I.1.

Il est également possible d'obtenir des informations qui permettent de minimiser le rôle de l'expérimentateur par le biais de notions de préférence /DIDAY 82/. Celles-ci se ramènent à la comparaison d'une paire {x,y} en fonction d'un certain nombre d'éventualités.

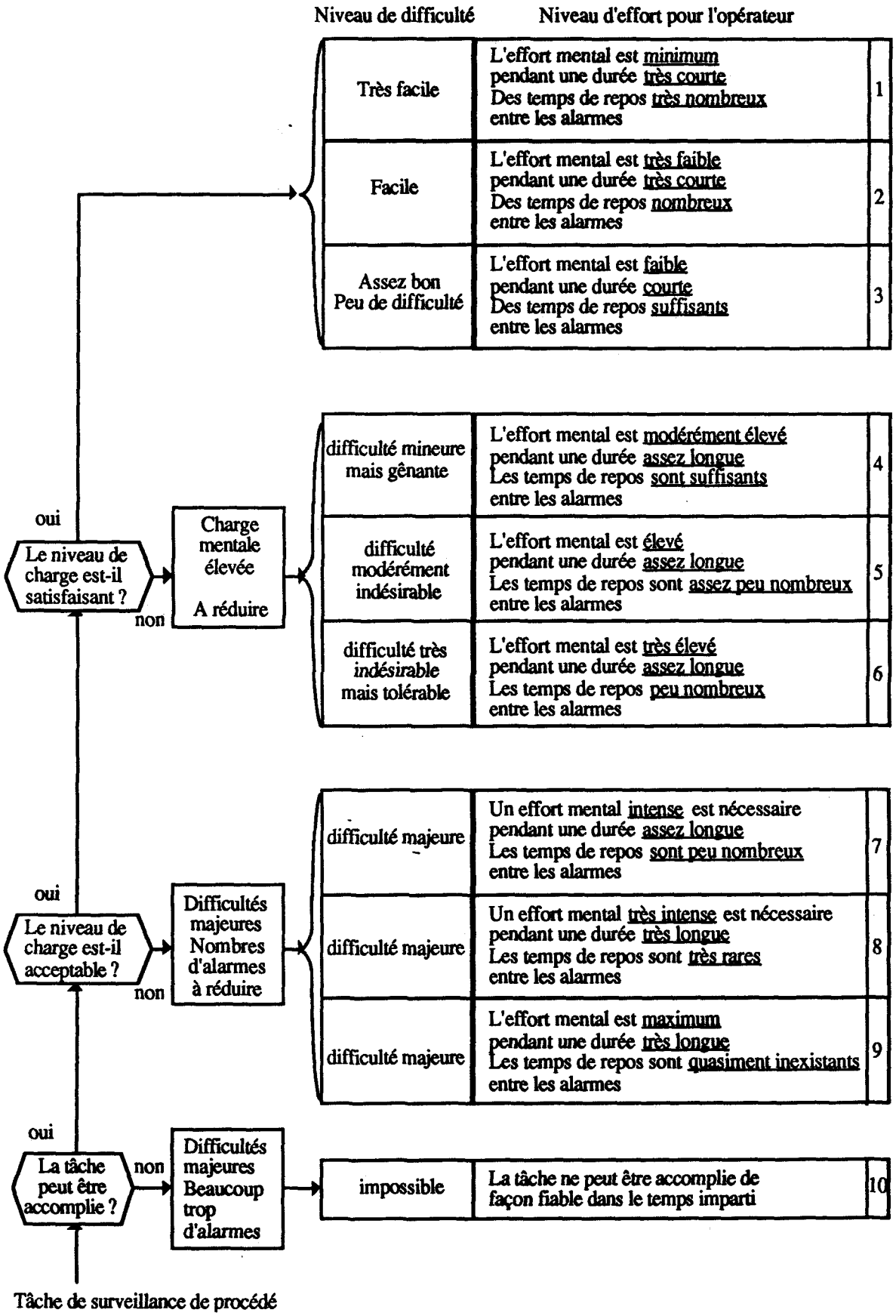


Figure I.1 : Exemple d'échelle de COOPER-HARPER d'après /KAMOUN 89/

Avec 3 éventualités par exemple :

- x préféré à y
- y préféré à x
- x et y indifférents

Comme l'indique sa dénomination l'échelle ordinale induit une notion d'ordre qui peut être de plusieurs types :

- préordre - égalité des préférences autorisée
- ordre partiel ou de pré-ordre partiel - seule une partie des objets est classée
- ordre ou de pré-ordre total.

Les transformations mathématiques qui laissent invariantes les données devant conserver une de ces notions se ramènent aux fonctions monotones croissantes.

Toutes les opérations mathématiques autorisées sur les variables nominales sont autorisées sur les variables ordinales, il convient de rajouter la médiane, les quartiles, les déciles et quelques coefficients de rang.

Enfin il est à noter que la différence entre deux variables n'a aucun sens et que, par conséquent tout emploi d'une distance ou d'un indice de proximité sur des variables ordinales est à déconseiller ou, tout du moins, sans justification préalable. Comme le notent CHANDON et PINSON /CHANDON et PINSON 81/ "*Il est regrettable de remarquer que certaines études en sciences sociales traitent les données ordinales comme des données métriques sans aucune justification*".

I.1-3 L'échelle d'intervalle

Elle permet de donner en plus un sens à la différence entre deux variables et au rapport des différences de variables. Les exemples les plus classiques sont les échelles de température (C°, F°) ou les dates du calendrier.

Ce type d'échelle est caractérisé par un choix arbitraire du zéro qui interdit de donner un sens au rapport des variables et qui entraîne que seules les transformations du type $f(x) = a\,x + b$ $(a > 0)$ permettent de laisser les données invariantes. L'exemple classique des échelles de température permet de résumer ces propriétés :

le changement d'échelle entre degrés Celsius et degrés Farenheit correspond à :

$T_F = 1.8\ T_C + 32$ et il vient alors : $T_F - T_{F'} = 1.8\ (T_C - T_{C'})$
et bien entendu le rapport des différences entre variables reste constant :

$$\frac{T_F - T_{F'}}{T_F - T_{F''}} = \frac{T_C - T_{C'}}{T_C - T_{C''}} .$$

D'autre part, en choisissant $T_{C1} = 50°C$ et $T_{C2} = 100°C$ il vient : $\frac{T_{C1}}{T_{C2}} = \frac{1}{2}$ alors que par changement d'échelle $T_{F1} = 122°F$ $T_{F2} = 212°F$ et $\frac{T_{F1}}{T_{F2}} = \frac{122}{212} \neq \frac{1}{2}$

Tous les calculs statistiques sont alors autorisés (moyenne, écart-type, variance...) sur cette échelle en sus des autres opérations déjà citées sur les autres échelles. L'emploi d'un tel type d'échelle est intéressant en analyse des données car elle permet un éventail d'opérations assez large.

Dans de nombreux domaines où il est nécessaire de recueillir des impressions subjectives, si des questionnaires sont utilisés, ils le sont en général, sous les formes précédemment citées et il est alors nécessaire de faire très attention lors de leur traitement. Pour pallier ce problème il est possible d'imaginer des questionnaires faisant intervenir des échelles d'intervalle.

D'une part, concernant un jugement relatif, une notion de certitude ou de confiance peut être associée au choix de la réponse /FABRE 80/ /DEWITTE 86/ /GUERRA 88/.

Quel réglage préférez-vous ? x ☐ y ☐

Avec quel degré de préférence ? nul |——————————| absolu

D'autre part pour évaluer des attitudes ou des opinions, il est possible d'avoir recours à des questions utilisant des différenciateurs sémantiques dont l'axe est continu. Ce type de question permet d'éviter la discontinuité de l'échelle et les problèmes que celle-ci pose, à savoir, le caractère forcé de la réponse ou le choix du nombre de modalités. Par exemple :

Avez-vous mal au dos beaucoup |——————————| pas du tout

Le zéro est fixé pour ces deux types d'échelle, ce qui semble les faire correspondre à des échelles de rapport. Néanmoins, le zéro ne peut être perçu par tous les individus de la

même façon, ce qui permet de penser que ces échelles s'apparentent plutôt à des échelles d'intervalle.

I.1-4 L'échelle de rapport

Cette échelle est caractérisée par la possession d'un zéro naturel qui indique une absence de phénomène, vitesse nulle ou masse nulle par exemple. Pour ce type d'échelle le rapport a un sens et en conséquence, toutes les opérations mathématiques sont applicables. De ce fait, c'est l'échelle la plus aisée à utiliser en analyse des données. Enfin, les transformations qui laissent les données invariantes sont toutes celles du type f(x) = a.x (1 m = 100 x 1 cm par exemple).

Les différentes propriétés de chaque type d'échelle ont été résumées figure I.2.

	propriété	fonction mathématique	opérations
échelle nominale	équivalence des individus	discrète bijective	- mode - comptage
échelle ordinale	ordre / pré-ordre } partiel / total	monotone croissante	- médiane - quartile,...
échelle d'intervalle	équivalence des individus	f(x)=ax+b (a>0)	- moyenne - écart-type,...
échelle de rapport	équivalence des individus	f(x)=ax	- moments d'ordre h,...

Figure I.2 : propriétés des différentes échelles de mesure /CHANDON et PINSON 81/

Notons enfin que la prise en compte simultanée de variables de nature parfois très différentes nécessite, en amont de la phase de traitement des données, de passer par une étape d'homogénéisation.

Effectivement, l'utilisateur qui choisit l'échelle comme la plus appropriée pour mesurer les différentes caractéristiques des objets doit le faire en sachant qu'il peut-être parfois

difficile de travailler sur des variables trop hétérogènes. Pour les homogénéiser il y a deux choix, soit les appauvrir, soit les enrichir.

Dans le premier cas, l'opération est relativement aisée mais, il est essentiel de garder à l'esprit qu'elle fait obligatoirement perdre de l'information, ce qui peut poser un certain nombre de problèmes. Dans le deuxième cas s'il existe certaines méthodes permettant d'enrichir les variables, il s'agit de faire très attention aux hypothèses supplémentaires que l'on est obligé de formuler.

En conclusion que ce soit pour le choix des échelles ou l'homogénéisation des variables il convient d'être très rigoureux car toutes les méthodes de traitement, ainsi que les résultats en découlent.

I.2 - LA FONCTION D'OPERATEUR

D'une manière générale, il s'agit de réunir, de comparer, de construire une fonction statistique ou d'effectuer des calculs plus complexes comme la mise en relation de variables ou de groupes de variables.

Les mathématiques proposent outre les différents moyens de représentation des données, des systèmes d'opérateurs généraux pouvant se réécrire pour répondre à des besoins précis de manipulation de données et de construction de modèles de référence. Ces modèles se présentent sous des formes très variées, dépendant des actions mathématiques affectant les données. On distingue les quatre grandes classes d'opérateurs suivantes :

I.2-1 Opérateurs portant sur les ensembles

Les principaux sont l'intersection et l'union, ils traduisent les notions de "et" et de "ou" logiques. Dans le cas de sous-ensembles vulgaires l'intersection correspond au produit booléen "*" et l'union à la somme booléenne "+".

En définissant une notion d'appartenance, à savoir :

Soit E un référentiel quelconque, A un sous-ensemble de E et x un élément de E, l'appartenance de x à A se traduit par la fonction μ_A suivante :

$$\forall x \in E \quad \mu_A(x) = 1 \text{ si } x \in A$$
$$= 0 \text{ si } x \notin A$$

et pour deux sous-ensembles A et B :

$$\forall x \in E \quad \mu_{A \cap B}(x) = \mu_A(x) * \mu_B(x) \text{ et } \mu_{A \cup B}(x) = \mu_A(x) + \mu_B(x)$$

Il existe une autre manière d'appréhender ces opérateurs. En effet, ZADEH /ZADEH 65/ propose de créer une fonction d'appartenance valuée, c'est à dire telle que la fonction caractéristique μ_A puisse prendre une quelconque valeur dans [0,1]. C'est le concept de sous-ensemble flou qui se définit par :

Définition : Soit E un référentiel quelconque, un sous-ensemble flou A de E va être déterminé par sa fonction d'appartenance μ_A :

$$\mu_A : E \rightarrow [0,1]$$
$$x \rightarrow \mu_A(x)$$

$\mu_A(x)$ représente le degré d'appartenance de x au sous-ensemble A.

Soit, par exemple, le sous-ensemble flou A des réels proches de 0, pour
$0 \leq x_1 \leq x_2$ on a : $\mu_A(x_1) \geq \mu_A(x_2)$

C'est cette possibilité de graduer l'appartenance d'un élément à un sous-ensemble flou qui permet la représentation de notions imprécises. Pour l'exemple cité, la figure I.3 représente deux courbes possibles permettant de représenter le prédicat "proche de 0".

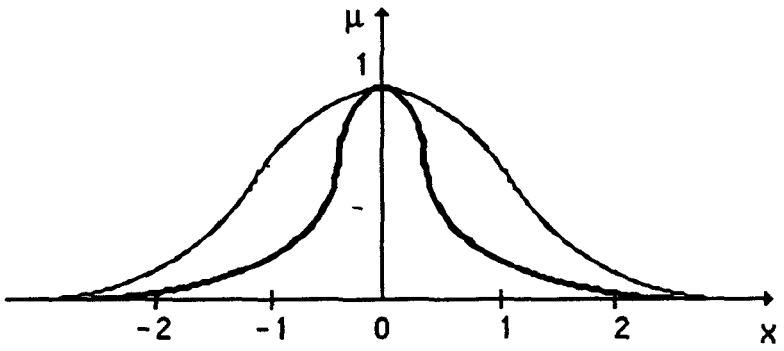


Figure I.3 : 2 représentations possibles du prédicat "proche de 0"

Les opérateurs "et" et "ou" sur ces ensembles peuvent alors être définis de plusieurs manières. L'opérateur d'intersection de deux sous-ensembles flous A et B est défini par le plus grand sous-ensemble flou contenu à la fois dans A et B c'est à dire le minimum. L'opérateur d'union est quant à lui défini par le plus petit sous-ensemble flou qui contient A et qui contient B c'est à dire le maximum. Il vient :

$$\forall x \in E \qquad \mu_{A \cap B}(x) = \text{MIN}(\mu_A(x), \mu_B(x))$$
$$\text{et} \qquad \mu_{A \cup B}(x) = \text{MAX}(\mu_A(x), \mu_B(x))$$

Cette façon d'appréhender les ensembles permet de représenter l'imperfection des informations d'une manière différente. On dispose traditionnellement de la théorie des probabilités qui se base sur un certain nombre d'axiomes :

Soit E un référentiel fini, $P(E)$ l'ensemble de ses parties :

$\forall A \in P(E)$	$P(A) \geq 0$
$\forall (A,B) \in P(E)^2$	$A \cap B = \phi \Rightarrow P(A \cup B) = P(A) + P(B)$
	$P(E) = 1$
$\forall A \in P(E)$	$P(A) + P(\neg A) = 1$ ($\neg A$ représente l'événement contraire de A)

A partir du formalisme flou on fait appel aux mesures de possibilité /ZADEH 78/. Une mesure de possibilité Π est une application de $P(E)$ dans [0,1] qui vérifie :

	$\Pi(\phi) = 0$
	$\Pi(E) = 1$
$\forall (A,B) \in P(E)^2$	$\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$

On vérifie ensuite que :

$$\max(\Pi(A), \Pi(\neg A)) = 1$$

ce qui signifie que de deux événements contraires, l'un au moins est toujours complètement possible. Il vient alors :

$$\Pi(A) + \Pi(\neg A) \geq 1$$

On voit ici une différence fondamentale avec la théorie des probabilités. Effectivement, la possibilité d'un événement ne détermine pas obligatoirement celle de son contraire, alors que la probabilité de l'un détermine celle de l'autre. Comme le notent DUBOIS et PRADE /DUBOIS et PRADE 87/ "*les mesures de probabilité synthétisent naturellement un corps de connaissances précises et différenciées, tandis que les mesures de possibilité sont le reflet de connaissances imprécises mais cohérentes.*"

Les opérateurs sur les ensembles apportent donc des différences fondamentales quant à la prise en compte des données, et toutes manipulations sur les données s'en trouvent modifiées.

Pour les autres types d'opérateurs le problème reste le même que ce soit pour coder, décrire ou comparer les individus ou les variables.

I.2-2 Opérateurs de codage de l'information

Ils permettent principalement d'effectuer des changements d'échelle, mais aussi de rendre comparable des variables de nature mathématique différente. Ce dernier problème a été évoqué rapidement au I.1 pour l'homogénéisation des variables. On rencontre deux formes principales de codage de variable :

* le codage par transformation linéaire

Il s'agit d'homogénéiser les variables à l'aide d'une translation et d'une dilatation, ou d'une réduction, d'échelle. Le modèle de changement de variable est alors le suivant :

$f(v) = \frac{v - a}{b}$ où a et b sont des valeurs choisies par l'utilisateur, le plus souvent la moyenne pour a et l'écart-type de v pour b.

* le découpage en modalités

Le problème du découpage en modalités est complexe car il entraîne obligatoirement une perte d'information. Celle-ci est d'autant plus grande que le nombre de classes est faible. Il s'agit de trouver un bon compromis entre, d'une part, le nombre de classes à choisir et d'autre part, garder un effectif par classe significatif. Effectivement, si le nombre de classes augmente il est nécessaire de ne pas obtenir des effectifs trop peu significatifs. Si la variable ne comporte pas de modes à priori /VOLLE 81/, c'est lorsque les effectifs par classe sont égaux que l'information apportée est maximale.

Il reste alors à déterminer quel type de fonctions d'appartenance choisir pour découper la variable. Celles-ci peuvent être binaires ou "floues" /GALLEGO 82/ figure I.4. Dans ce dernier cas, si le choix est encore plus complexe, il permet de réduire sensiblement la perte d'informations. Il est également possible d'imaginer des codages non symétriques. En reprenant l'exemple des différenciateurs continus, on peut considérer que la symétrie des termes utilisés sur le segment n'est qu'illusoire /FABRE 80/ et que l'effet de l'orientation verbale est propre à chaque individu. A partir de ces constatations un codage barycentrique peut être utilisé, figure I.5, en utilisant par exemple le minimum, le maximum et la moyenne des réponses d'un individu ou d'un groupe d'individus /BEHRAKIS et NICOLAPOULOS 87/ /LOSLEVER 88/.

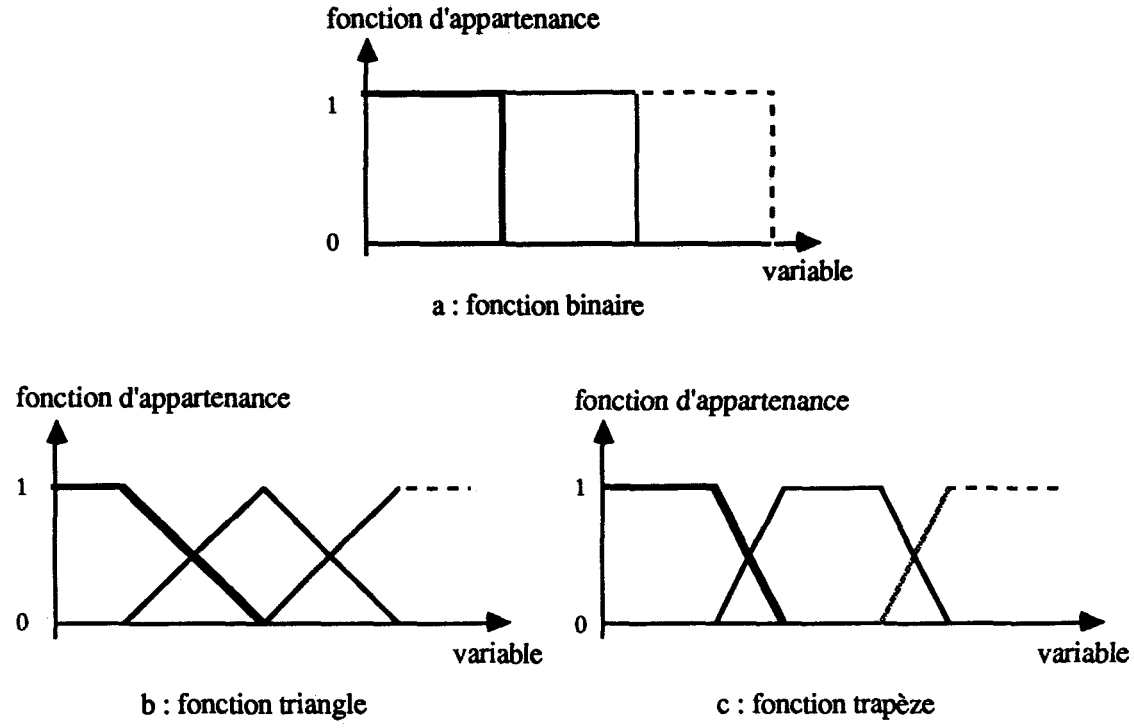


figure I.4 : Codage par découpage; fonctions d'appartenance

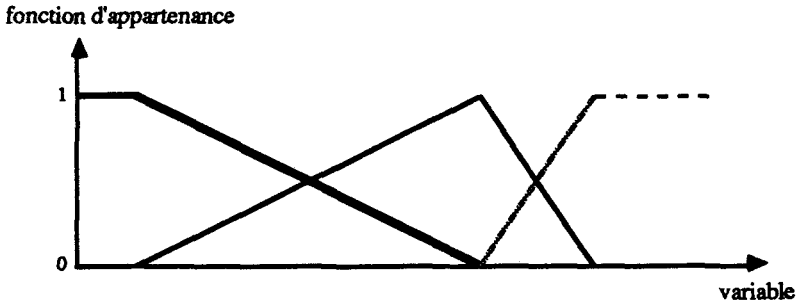


Figure I.5 : Fonctions d'appartenance pour le codage en 3 nuances d'un segment sémantique

I.2-3 Opérateurs de synthèse de l'information

Ce sont tous les opérateurs qui permettent de décrire une ou plusieurs variables, ou un ou plusieurs individus. En statistique on rencontre, bien sur, les opérateurs classiques qui peuvent s'appliquer sur les variables et sur les individus tels que :

- la moyenne $\text{moy}(x) = \sum_{i=1}^{n} p_i x_i$ avec x vecteur de dimension n $(x_1,\ldots,x_n)$
- la variance $\text{var}(x) = \sum_{i=1}^{n} p_i (x_i - \text{moy}(x))^2 = \sum_{i=1}^{n} p_i x_i^2 - \text{moy}(x)^2$
- l'écart-type $s(x) = (\text{var}(x))^{1/2}$

- la covariance $cov(x,x') = \sum_{i=1}^{n} p_i (x_i - moy(x))(x'_i - moy(x'))$

$$= \sum_{i=1}^{n} p_i x_i x'_i - moy(x)\, moy(x') \quad \text{avec } x' : (x'_1,\ldots,x'_n)$$

- la corrélation $cor(x,x') = \frac{cov(x,x')}{s(x).s(x')}$

Pour décrire les variables ou les individus l'utilisation de graphiques, faciles à assimiler et à mémoriser, est un des moyens les plus répandus. Quelques méthodes visuelles élémentaires permettant de synthétiser l'information sont présentées ci-après.

* Cas d'une variable

Les histogrammes ou les fonctions de répartition, figure I.6, permettant de représenter l'ensemble de n valeurs $x_1,\ldots,x_n$ prises par une variable donnée, sont les plus souvent rencontrés en pratique.

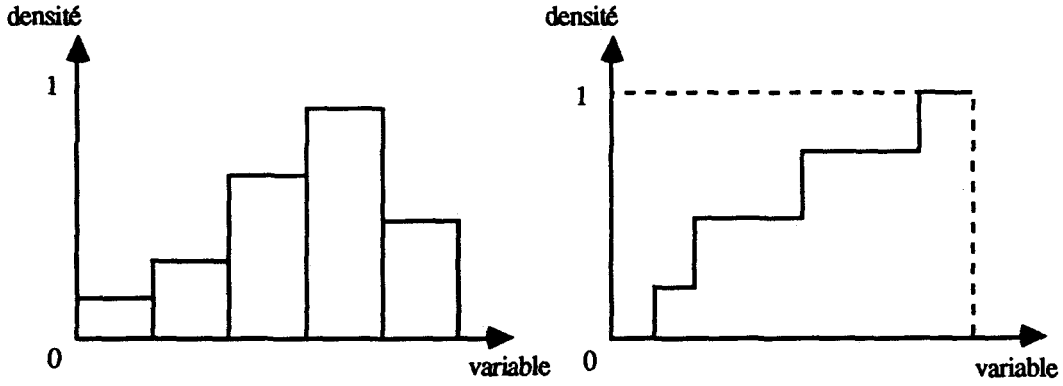


Figure I.6 : Exemple d'histogramme et de fonction de répartition

* Cas de plusieurs variables

Lorsque le nombre de variables est peu important (<10) la représentation directe de leur évolution est possible sans étude statistique préalable. Par exemple, certains auteurs /DIDAY 82/ proposent d'adopter une représentation par polygones ou la distance de chaque sommet au centre est proportionnelle à l'amplitude de chaque variable. Certains seuils caractéristiques (moyenne, minimum, optimal,...) peuvent alors etre superposés à la représentation, figure I.7.

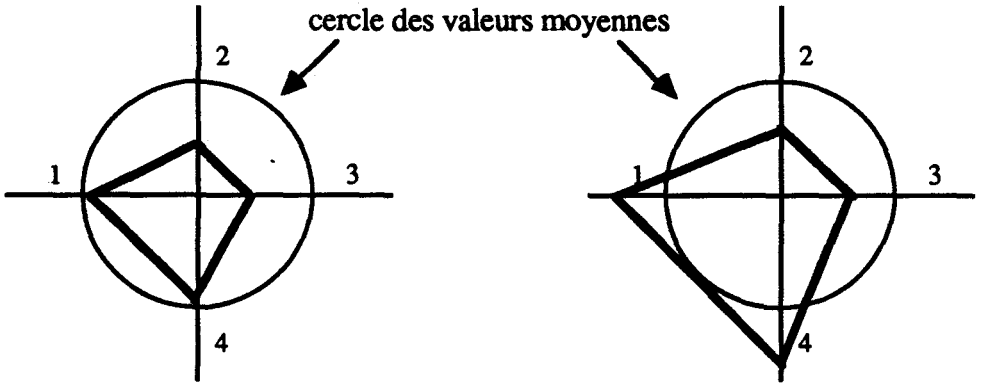


Figure I.7 : Polygone des variables, cas de 4 variables

Pour représenter globalement et simultanément les modalités choisies sur un questionnaire à choix multiple et les "certitudes" associées à leur choix (cf I.1-3), une représentation par histogramme avec nuances peut être adoptée, figure I.8.

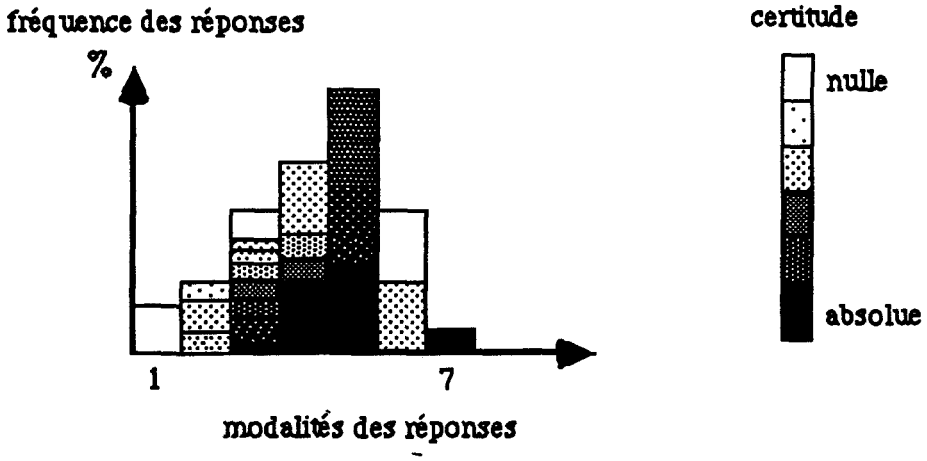


Figure I.8 : Histogramme intégrant les certitudes associées aux choix des modalités

Il existe bien entendu beaucoup d'autres modes de représentation, citons par exemple les méthodes strictement graphiques de BERTIN /BERTIN 77/ ou d'autres telles que la méthode d'Andrews ou les visages de Chernof citées dans /DIDAY 82/.

I.2-4 Opérateurs de comparaison

Ce sont des opérateurs qui permettent d'exprimer numériquement un lien existant entre les individus ou entre les variables. Le premier qui correspond au minimum de propriétés est l'indice de similarité.

Soit s un tel indice sur un ensemble Ω, il doit réunir les trois propriétés suivantes :

(1) - non-négativité

s est une application de $\Omega \times \Omega$ dans $\Re^+$

(2) - symétrie

$\forall (\omega,\omega') \in \Omega \times \Omega \quad s(\omega,\omega') = s(\omega',\omega)$

(3) - normalisation

$\forall (\omega,\omega') \in \Omega \times \Omega \quad \omega \neq \omega' \quad s(\omega,\omega) = s(\omega',\omega') > s(\omega,\omega')$

Pour un indice de dissimilarité il suffit de remplacer (3) par :

(4) - $\forall \omega \in \Omega \quad s(\omega,\omega) = 0$

Pour un indice de distance il convient de rajouter à l'indice de dissimilarité la propriété suivante:

(5) - $\forall (\omega,\omega') \in \Omega \times \Omega \quad s(\omega,\omega') = 0 \Leftrightarrow \omega = \omega'$

et pour obtenir une distance il suffit de reprendre l'indice de distance (propriétés (1),(2),(4) et (5)) et d'y adjoindre :

(6) - inégalité triangulaire

$\forall (\omega,\omega',\omega'') \in \Omega^3 \quad s(\omega,\omega') \leq s(\omega,\omega'') + s(\omega'',\omega')$

Enfin, si (6) est remplacée par :

(7) - inégalité ultramétrique

$\forall (\omega,\omega',\omega'') \in \Omega^3 \quad s(\omega,\omega') \leq \max (s(\omega,\omega''),s(\omega'',\omega'))$

on parle de distance ultramétrique.

Les indices les plus couramment utilisés sont la corrélation et la covariance. Avant de définir les distances les plus importantes les notations utilisées sont données ci-après.

n : nombre d'observations ω_i

p : nombre de variables v_j

$X = (x_i^j)$: matrice p x n où $x_i^j = v_j(\omega_i)$ est la valeur de la variable v_j prise au cours de l'observation ω_i.

$$x_i = \begin{pmatrix} x_i^1 \\ \cdots \\ x_i^p \end{pmatrix}$$: la $i^{ème}$ ligne de X qui correspond à ω_i

$$x^j = \begin{pmatrix} x_1^j \\ \dots \\ x_n^j \end{pmatrix}$$: la $j^{ème}$ colonne de X qui correspond à v_j

Les distances les plus couramment utilisées sont alors :

- la métrique euclidienne classique

$$d^2(\omega_i,\omega_k) = \sum_{j=1}^{p} (x_i^j - x_k^j)^2$$

- la distance du "CHI-DEUX" principalement utilisée en analyse factorielle

$$d^2(\omega_i,\omega_k) = \sum_{j=1}^{p} \frac{1}{x_\bullet^j} \left(\frac{x_i^j}{x_i^\bullet} - \frac{x_k^j}{x_k^\bullet}\right)$$

où : $x_\bullet^j = \sum_{i=1}^{n} x_i^j \qquad x_i^\bullet = \sum_{j=1}^{p} x_i^j$

- la distance de MAHANALOBIS

$$d^2(\omega_i,\omega_k) = \sum_{j=1}^{p} \sum_{l=1}^{p} w_{jl} (x_i^j - x_k^j)(x_i^l - x_k^l)$$

où w_{jl} est l'élément de l'inverse de la matrice de covariance.

- distance rectangulaire ou "City Block"

$$d(\omega_i,\omega_k) = \sum_{j=1}^{p} p_j \mid x_i^j - x_k^j \mid$$

et avec $p_j = 1$ distance dite de HAMMING

- distance de CHEBYCHEV

$$d(\omega_i,\omega_k) = \max_j (x_i^j - x_k^j)$$

La phase d'évaluation - choix des outils de mesure et des différents opérateurs - étant achevée, il s'agit alors de déterminer quelles méthodes mettre en oeuvre pour caractériser le comportement du système ; en d'autres termes, construire des modèles exprimés sous des formes différentes - verbale, graphique ou mathématique - qui permettent une représentation adéquate du système.

I.3 - LA MODELISATION DES SYSTEMES

La perception d'un système concret n'est possible qu'au travers de modèles représentatifs que l'Homme se forme individuellement, image mentale, ou qu'il explicite de manière plus ou moins formelle. La construction d'un modèle pose alors le problème du choix de son mode de formalisation. Ce choix est un compromis entre la volonté de prendre en compte une partie plus ou moins grande de l'ensemble des relations entre les propriétés du système et du modèle et les différentes situations envisagées quant à l'utilisation du modèle. Tout modèle peut se présenter de façon plus ou moins abstraite et plusieurs voies sont possibles pour atteindre un niveau de formalisation donné.

I.3-1 Les différents niveaux de formalisation d'un modèle

La mise en relation d'un modèle et d'un système relève des mêmes principes que ceux intervenant dans la mise en correspondance des séries empirique et formelle de l'échelle de mesure. Les conditions pour obtenir un isomorphisme entre le modèle et le système sont difficiles à remplir : les propriétés de l'un ne sont pas toujours vérifiables par l'autre. De plus, il existe une infinité de modèles d'un même système car il est possible de mettre l'accent sur telle ou telle propriétés du système. En pratique tous les modèles sont alors imparfaits et incomplets.

La formalisation d'un modèle peut s'envisager selon des niveaux différents d'abstraction. En passant du moins abstrait au plus abstrait il faut distinguer les trois niveaux suivants :

- le premier niveau pour aborder la modélisation est **l'approche verbale.** Elle s'appuie sur les langues usuelles mais peut également faire appel à des langages "artificiels" - langages informatiques. Les phrases suivantes, exprimant une prémisse et une relation de cause-conséquense, en sont des exemples :

"La pression est proche de 25 bars "
"Si la pression est proche de 25 bars alors arrêter la machine "

Avec les développements de l'intelligence artificielle, les systèmes experts, par exemple, il est possible d'intégrer et d'inférer sur de nombreuses expressions de ce type. Il faut ici déjà insister sur la difficulté de prendre en compte les notions d'imprécision et d'incertitude au cours des inférences successives.

- Le second niveau est constitué de **symboles graphiques**. Il faut distinguer les icônes, par exemple les sigles iconiques représentatifs d'un sport aux jeux olympiques, des tableaux ou diagrammes issus de traitements simples - graphiques statistiques.

- L'**approche mathématique** constitue le niveau d'abstraction le plus élevé. Il peut faire appel à la théorie des ensembles, notion d'appartenance, calcul des probabilités..., ou à des systèmes d'équations analytiques.

Un système étant caractérisé par l'ensemble de ses variables d'entrées et de sorties, une notion importante dans sa modélisation est celle de leur mise en relation. De nombreux modèles verbaux, graphiques ou mathématiques permettent en fait de traduire les réseaux de relations entre variables caractérisant des situations particulières. Si l'on suppose qu'il existe une relation entre la variables V0 et les variables V1,V2 et V3, il existe essentiellement les relations de causalité et de dépendance /WALLISER 77/.

Les relations de causalité traduisent un effet et une action. on distingue :

- les relations de condition : V1 et V2 et V3 ======> V0
- les relations de conséquence : V1 ou V2 ou V3 ======> V0

Les relations de dépendance traduisent essentiellement :

- des similarités, parenté entre les comportements de V0 et V1,
- des influences, par exemple V1 agit sur V0.

Si les variables sont construites à partir d'échelles quantitatives, les relations de causalité et de dépendance peuvent être traduites par une relation fonctionnelle. A ce niveau, il importe de savoir si les variables sont aléatoires ou non. Par exemple, dans le cas de variables aléatoires, le théorème de BAYES fait intervenir un rapport de vraisemblance entre une probabilité à priori et une probabilité à posteriori.

I.3-2 Les différentes formes d'un modèle

Une première façon de présenter un modèle est un **graphe** figurant aux noeuds les variables ou les objets et par des arcs les relations. Les arcs peuvent être orientés ou non et valués ou non. Deux catégories de graphes sont souvent utilisés : les arbres hiérarchiques et les treillis, figure I.9. Ils traduisent une structure d'héritage simple ou multiple entre les variables ou les objets.

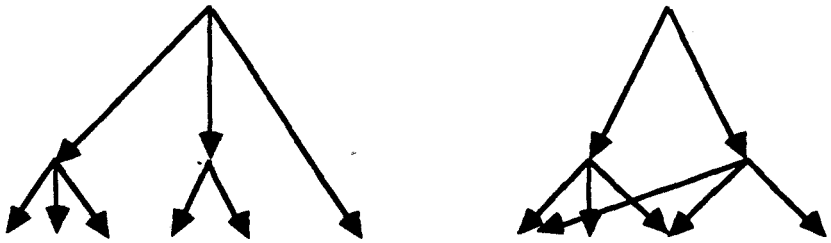


figure I.9 : exemples particuliers de graphes

Dans l'analyse des systèmes, on est souvent amené à caractériser un ensemble S de situations - des individus, des objets, des instants - par un ensemble V de variables. L'application de l'ensemble produit SxV dans un ensemble X est caractérisé par une **matrice.** En fonction de la nature de ces trois ensembles, la matrice peut traduire, par exemple, une notion de proximité - ressemblance entre individus, corrélation entre variables... -, de probabilité d'apparitions d'occurences - probabilité que la variable V0 prenne une valeur donnée pour une situation donnée, table cause-symptômes...

Une autre façon de caractériser le comportement d'un système est de faire appel aux **modèles analytiques.** Ils sont construits à partir d'ensemble d'équations liant les différentes variables mesurables entre elles. Les modèles les plus couramment utilisés sont les modèles différentiels, systèmes du premier ordre et du second ordre,...

Le choix d'une forme de modèle, choix de sa "syntaxe", introduit inévitablement une idée de "ressemblance" entre le modèle et le système. Cette ressemblance doit tenir compte de la signification du modèle par rapport au système (aspect sémantique du modèle). C'est la connaissance de cette ressemblance qui permet d'élargir le champ de validité du modèle choisi.

I.3-3 La généralisation du modèle

Dans la pratique, même s'il a fallu faire appel à des hypothèses concernant la construction du modèle, celui-ci est déterminé dans un cadre bien défini autour de situations expérimentées. Mais les objectifs étant a priori ou a posteriori plus vastes que ceux ayant initialement conduits à la phase expérimentale, la généralisation du modèle à des situations non étudiées est très délicate, bien qu'elle soit une phase indispensable à son utilisation - la réalité n'est jamais identique à la situation expérimentale. Si une nouvelle situation est intermédiaire à celles expérimentées, il est alors possible de faire appel à l'interpolation du modèle, si elle est en dehors du domaine expérimenté, il faut recourir à l'extrapolation. Ceci nécessite de vérifier l'adéquation des nouvelles situations aux situations de bases et aux

hypothèses sur lesquelles le modèle a été fondé. C'est *"le modèle qui doit suivre les données et non l'inverse"* /BENZECRI 73/.

I.3-4 Les différents outils

Après avoir choisi le niveau d'abstraction et la forme syntaxique du modèle il faut recueillir et traiter les données. Ceci implique d'estimer les variables les plus pertinentes et de construire des relations entre celles-ci. Dans le cadre de l'analyse des systèmes homme-machine, le besoin de prendre en compte simultanément des données de nature objective et subjective nécessite de faire appel à des méthodes complexes.

En fonction du type de résultat que l'on attend du traitement des données, il est possible de distinguer trois groupes de méthodes de traitements.

- Les méthodes de classification.

Elles consistent à construire des partitions ou des hiérarchies de partitions des situations selon des critères de ressemblance et de dissemblance particulier. Le partitionnement peut se faire par divisions ou par agrégations successives, classification ascendante hiérarchique, par exemple. Chaque élément peut appartenir à une classe et une seule (arbre) ou à plusieurs classes simultanément (pyramide).

- Les méthodes de réduction

Leur objectif est de résumer "au mieux" l'information contenue dans plusieurs variables en gardant celles d'entre elles qui apportent le plus d'informations - arbre de décision - ou des combinaison de celles-ci - méthodes d'analyse factorielle.

- Les méthodes de mises en relation

Elles visent à mettre en évidence l'influence de certains facteurs sur les variables. Par exemple, l'analyse de la variance permet de rechercher si un facteur expérimental à une influence significative ou non, à un seuil d'erreur donné, sur une variable particulière. Les méthodes factorielles - correspondances simple, multiples, canoniques - permettent de faire apparaître sur les plans factoriels les relations entre les variables, les proximités entre les observations et les correspondances entre les individus et les variables.

En pratique, ces méthodes sont plus ou moins complexes et, de par leur aspect multivariable, posent le problème de l'homogénéisation et du codage des données. Ces

méthodes conduisent à des modèles verbaux - interprétation des plans factoriels -, graphiques - arbres - ou mathématiques - modèles linéaires issus de la régression multiple et analyse canonique ou non linéaires si l'on procède à des changement de variable pour se ramener au cas linéaire. Ne répondant pas aux mêmes objectifs, ces méthodes sont souvent utilisées conjointement.

I.4 - CONCLUSION

La modélisation des systèmes repose sur l'utilisation des fonctions de représentation et d'opérateur. Si la construction d'une échelle de mesure pour caractériser des phénomènes de nature "physique" pose peu de problèmes, il en est tout autrement lorsqu'on souhaite faire appel à des jugements de valeur. En pratique, l'obtention de ces informations se heurte à la construction d'échelle subjectives puis à la formulation des questionnaires proprement dits.

La fonction d'opérateur peut faire appel à des outils mathématiques très différents et il importe alors de les adapter à la nature des variables prises en compte. A ce niveau, il faut alors distinguer les opérations proposées par la théorie des ensembles "classiques" avec celles proposées par la théorie des sous-ensembles-flous.

Remarquons alors, que dans la construction d'un modèle, les relations entre les entrées et les sorties doivent être un juste compromis entre une définition trop "souple" - pour obtenir une sortie répondant de façon suffisamment précise avec l'entrée - et une définition trop "rigide" qui supposerai de faire appel à des hypothèses ne "cadrant" pas à la réalité.

Avec la théorie des sous-ensembles flous, ZADEH a justement tenté de faire mieux "cadrer" l'outil mathématique avec les données subjectives qui sont, par essence, imprécises et incertaines. Le chapitre suivant présente, dans ce cadre, le traitement de données subjectives à l'aide des sous-ensembles flous.

CHAPITRE II

ANALYSE D'IMPRESSIONS SUBJECTIVES A L'AIDE DES SOUS-ENSEMBLES ALEATOIRES FLOUS

L'introduction des concepts flous peut difficilement se concevoir dans un contexte de mesures objectives, issues de capteurs physiques. Par contre, dans les systèmes Homme-Machine, où l'Homme peut jouer le rôle de "capteur", le recueil des données est indispensable par son intermédiaire - données nécessairement incertaines et imprécises - et l'approche floue semble indiquée, voire nécessaire.

S'il existe déjà des méthodes multidimensionnelles bien adaptées permettant le traitement de certaines données subjectives, issues de questionnaires par exemple, ce chapitre montre avant tout que le fait d'utiliser les sous-ensembles flous et plus particulièrement les sous-ensembles aléatoires flous, ramène à considérer le problème de l'analyse d'une manière peu "orthodoxe" par l'emploi de fonctions de répartition. D'autre part il montre également qu'il est nécessaire de faire appel, principalement pour la classification, aux méthodes classiques d'analyse des données, ces aspects étant encore limités dans la théorie des sous-ensembles flous.

II.1 - DOMAINE DE L'ANALYSE ET ANALYSE DES DONNEES

II.1-1 Recueil d'impressions subjectives

Si pour l'évaluation des connaissances ou pour des sondages d'opinion on fait en général appel à des questionnaires et plus particulièrement à des questionnaires à choix multiples, il est des domaines où le mode de recueil est très diversifié et où il est difficile même de se fixer sur un choix. Par exemple, les modes de recueil sont des plus divers dans l'extraction de connaissances pour un système à base de connaissances, citons entre autres : l'interview, les questionnaires, fermés ou ouverts, l'analyse de grilles de classification et l'analyse de protocoles /NASSIET 87/.

Parmi ces différentes méthodes, il reste à définir sur quels types de données appliquer un traitement utilisant les mathématiques floues. Des méthodes telles que l'entretien ou l'analyse de protocoles, qui permettent de dégager des ensembles de notions importantes et sont en général préalables à la réalisation de questionnaires, ne peuvent faire l'objet d'un traitement, les questions étant ouvertes et laissant trop de "liberté" au sujet ou à l'expert.

Le seul mode de recueil permettant une complète standardisation est le questionnaire fermé dont les différents types ont été évoqués au chapitre I en parallèle avec les échelles de mesure. Parmi ceux-ci les questionnaires à choix multiples, par le caractère forcé de la réponse, ne laissent que peu de place aux aspects subjectifs et sont donc peu intéressants. Il reste alors les méthodes laissant une place importante à la subjectivité à savoir : les questionnaires utilisant des différenciateurs sémantiques continus, les questionnaires à choix multiples utilisant un degré de certitude et éventuellement les grilles d'évaluation utilisant une échelle continue.

Le domaine d'application ayant été défini, il s'agit de dégager les différentes étapes de l'analyse permettant le traitement des données, et pour ce faire, introduire rapidement la démarche générale d'une analyse des données.

II.1-2 Les étapes d'une analyse des données

D'une manière générale, du point de vue du statisticien, les étapes nécessaires à une analyse de données sont résumées figure II.1. Elle fait apparaître la nécessité, dûe aux aspects multivariés des méthodes multidimensionnelles, de passer par une phase de codage de façon à homogénéiser l'ensemble des variables (cf chapitre I). Ce codage effectué, le choix d'une mesure de ressemblance permet alors de situer les observations les unes par rapport aux autres. Enfin, il convient de déterminer quelles méthodes sont à mettre en oeuvre et comment visualiser graphiquement les résultats.

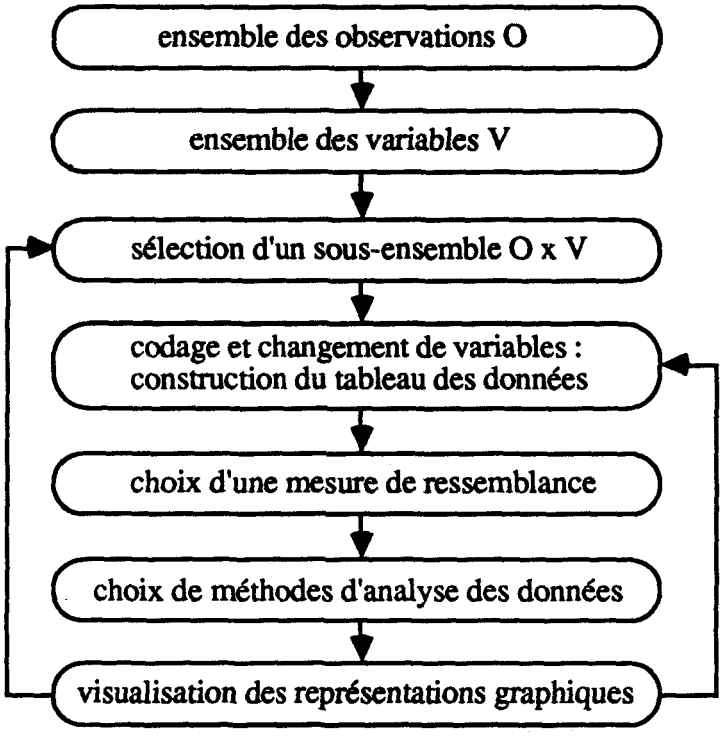


Figure II.1 : Etapes d'une analyse de données

Notons la nécessité de réaliser des bouclages :

- la procédure codage-analyse s'effectuant plusieurs fois sur un ensemble de données,
- les résultats de l'analyse permettant de détecter des données aberrantes et de sélectionner les variables les plus discriminantes, il s'agit alors d'effectuer une nouvelle analyse en tenant compte des résultats de la première.

L'analyse proposée se base sur une telle démarche. Le point de départ étant les sous-ensembles aléatoires flous définis dans le prochain paragraphe.

II.2 - LES SOUS-ENSEMBLES ALEATOIRES FLOUS

II.2-1 Définition et exemples

Le concept de sous-ensemble aléatoire flou, les différentes propriétés associées ont été principalement introduits par FERON /FERON 76/ et HIROTA /HIROTA 81/. Le concept est basé sur les théories des probabilités et des sous-ensembles flous.

Définition : Soient E un référentiel fini et A un sous-ensemble flou de E, supposons que la fonction d'appartenance μ_A soit une variable aléatoire prenant ses valeurs dans [0,1] et dont la densité est donnée :

$\forall x \in E \quad f(\mu_A(x) = \alpha)$. On dira que A est un sous-ensemble aléatoire flou.

Un sous-ensemble aléatoire flou (SEAF) de $\Re$ où pour tout x la densité $f(\mu(x))$ est donnée est représenté figure II.2.

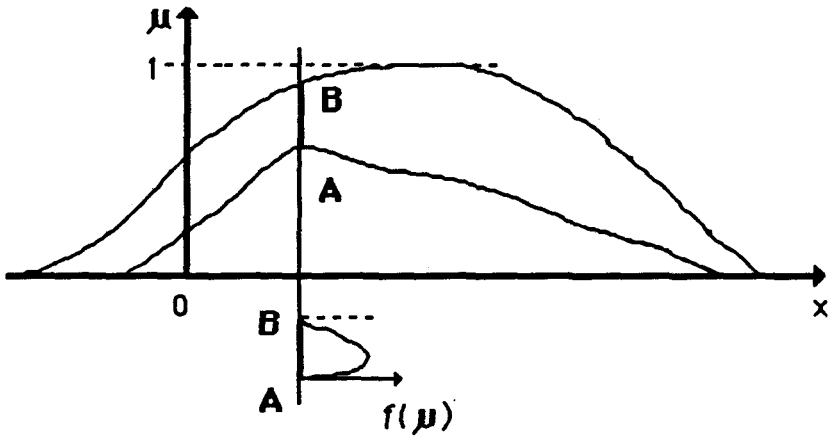


Figure II.2 : Représentation d'un sous-ensemble aléatoire flou de $\Re$

En discret, il est alors possible de présenter un sous-ensemble aléatoire flou à l'aide de fonctions de répartition complémentaire /KAUFMANN 83/.

Pour plus de compréhension, un exemple en discret est exposé.

Soit le référentiel $E = \{ x_1, x_2, x_3, x_4, x_5 \}$ et des occurences de $\mu(x_i)$ dans [0,1]. Le tableau figure II.3 donne un exemple de sous-ensemble aléatoire flou mis sous forme non cumulée.

E	x_1	x_2	x_3	x_4	x_5
	0,3	1	0,4	0,1	1
occurences	0,1	0,9	0,4	0	1
de $\mu(x_i)$	0,4	0,8	0,4	0	0,8
	0,6	0,7	0,2	0,2	1

Figure II.3 : Exemple de SEAF sous forme non cumulée

Un simple comptage sur chaque élément de E permet la construction des fonctions de répartition complémentaires sur chaque variable, renormées sur [0,1]. Cet ensemble de fonctions de répartition complémentaires donne alors le sous-ensemble aléatoire flou sous forme cumulée figure II.4.

x_1		x_2		x_3		x_4		x_5	
Occ.	Cum	Occ.	Cum	Occ.	Cum	Occ.	Cum	Occ.	Cum.
0	1	0	1	0	1	0	1	0	1
0,1	1	0,7	1	0,2	1	0,1	0,5	0,8	1
0,3	0,75	0,8	0,75	0,4	0,75	0,2	0,25	1	0,75
0,4	0,5	0,9	0,5	1	0	1	0		
0,6	0,25	1	0,25						
1	0								

Occ.=occurences
Cum.=cumul

Figure II.4 : Exemple de SEAF sous forme cumulée

Ce qui se traduit sur un graphique, avec $\mu \in [0,1]$ représentant les occurrences et $F(\mu) \in [0,1]$ les valeurs cumulées, par :

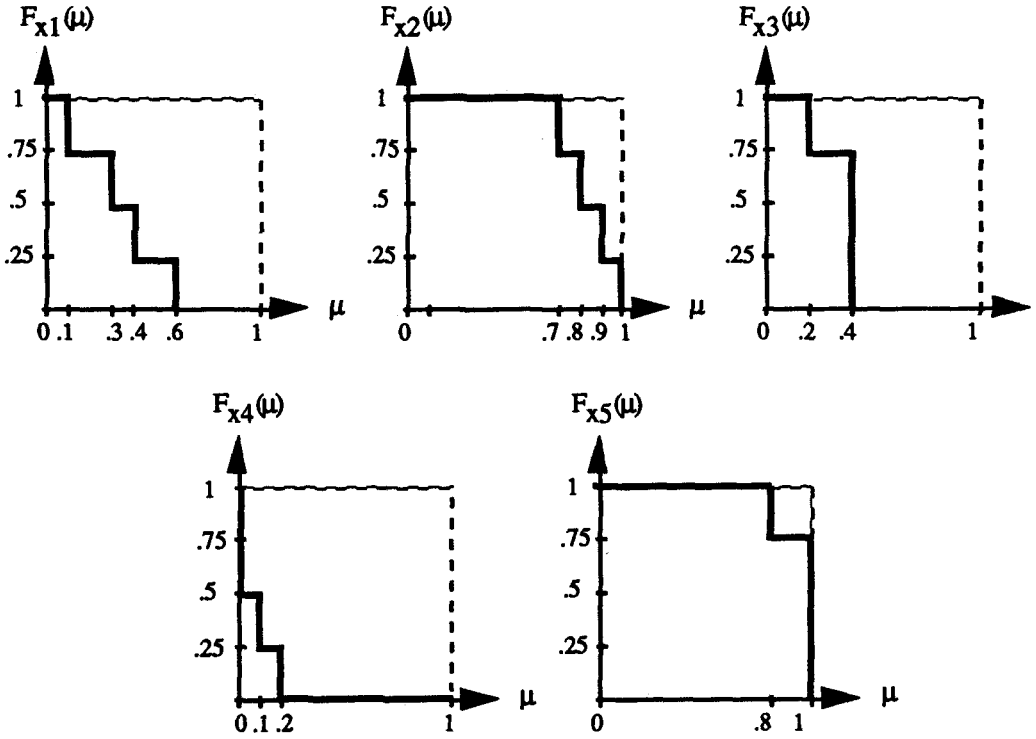


Figure II.5 : Représentation du SEAF sous forme cumulée

II.2-2 Comparaison de deux sous-ensembles aléatoires flous

Pour comparer deux sous-ensembles aléatoires flous d'un même référentiel, le concept de distance entre deux fonctions de répartitions complémentaires $F(\mu)$ et $F'(\mu)$ ($\mu \in [0,1]$) avec $F(\mu)$, $F'(\mu) \in [0,1]$ est nécessaire. Cette distance est obtenue en prenant l'aire qui sépare les deux fonctions de répartitions figure II.6, et il vient :

$$d(F(\mu),F'(\mu)) = \int_0^1 | F(\mu) - F'(\mu) | \, d\mu$$

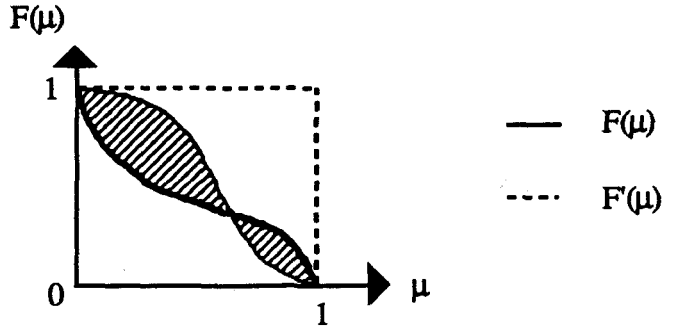


Figure II.6 : Distance entre 2 fonctions de répartition complémentaires

Les axiomes de distance sont aisément vérifiables, cette distance est dite de Hamming généralisée.

Dans le cas de fonctions de répartitions complémentaires obtenues par niveaux discrets,comme l'exemple figures II.4, II.5, la distance est donnée par :

$$d(F(\mu),F'(\mu)) = \sum_{i=1}^{k} | F(N(i)) - F'(N(i)) | \, (N(i) - N(i-1))$$

où :

K : ensemble des occurences de F, K' : ensemble des occurences de F'

k : nombre d'occurences de $K \cup K'$

N : $N \rightarrow K \cup K'$, $N(i) = \mu_i \quad 0 \leq i \leq k$

avec $\mu_i \in K \cup K'$ $(0 \leq i \leq k)$ occurence de F et/ou de F'

Cette distance reste bien entendu dans [0,1].

Soient alors deux sous-ensembles aléatoires flous A et A' d'un même référentiel E = [a,b], la comparaison entre A et A' se fait à l'aide de la distance suivante :

$$\partial (A,A') = \frac{1}{b-a} \int_a^b d(F(\mu,x),F'(\mu,x)) \, dx \qquad /KAUFMANN\ 84/$$

et en discret, la comparaison revient au calcul de la moyenne sur les éléments du référentiel, $E = \{x_i, 1 \leq i \leq n\}$, des distances entre chaque fonction de répartition prise pour chaque élément, c'est à dire :

$$\partial (A,A') = \frac{1}{n} \sum_{i=1}^{n} d(F(\mu,x_i),F'(\mu,x_i))$$

L'utilisation de l'aire séparant deux fonctions de répartition complémentaires semble naturelle. Malheureusement elle ne présente pas que des avantages. En effet, pour un x_i donné considérons les 3 cas suivants, *figure II.7 :*

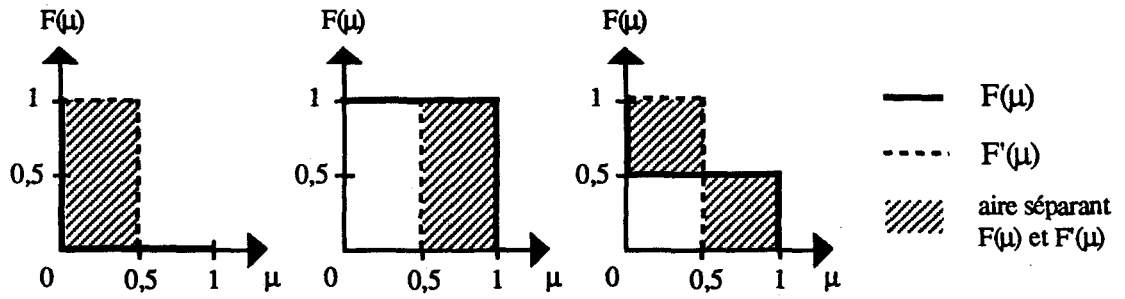


Figure II.7 : 3 cas d'aires identiques pour deux fonctions de répartition complémentaires

L'aire séparant les deux fonctions de répartitions complémentaires dans chaque cas est identique, $d(F(\mu),F'(\mu)) = 0.5$, alors que les trois situations sont très différentes. Pour pallier ce problème, il a été décidé de prendre en compte les aires supérieure et inférieure entre les deux fonctions.

Soient donc F et F', 2 fonctions de répartition complémentaires. La comparaison, appelée indice et notée I, associe à chaque couple de fonctions deux valeurs IS, aire supérieure, et II, aire inférieure, obtenues de la manière suivante :

$$IS(F(\mu),F'(\mu)) = \int_0^1 (F/_S(\mu) - F'/_S(\mu))\, d\mu$$

avec : $F/_S(\mu) = F(\mu)$ $F'/_S(\mu) = F'(\mu)$ si $F(\mu) > F'(\mu)$

$F/_S(\mu) = F'/_S(\mu) = 0$ si $F(\mu) \leq F'(\mu)$

$$II(F(\mu),F'(\mu)) = \int_0^1 (F'/_I(\mu) - F/_I(\mu))\, d\mu$$

avec : $F/_I(\mu) = F(\mu)$ $F'/_I(\mu) = F'(\mu)$ si $F(\mu) < F'(\mu)$

$F/_I(\mu) = F'/_I(\mu) = 0$ si $F(\mu) \geq F'(\mu)$

et bien entendu :

$$d(F(\mu),F'(\mu)) = IS(F(\mu),F'(\mu)) + II(F(\mu),F'(\mu))$$

Pour l'exemple figure II.7 on a alors respectivement pour les 3 cas :

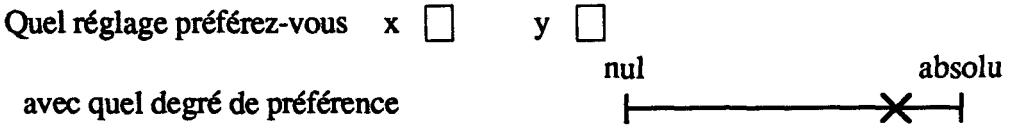
$I(F(\mu),F'(\mu)) = (0 ; 0,5)$, $(0,5 ; 0)$ et $(0,25 ; 0,25)$

Les différents rappels sur les SEAF viennent d'être cités. Avant de passer aux méthodes d'analyse, le paragraphe suivant montre comment les réponses à un questionnaire peuvent être synthétisées sous forme de SEAF.

II.2-3 Le passage des réponses à un questionnaire à des sous-ensembles aléatoires flous

La composante commune aux différents modes de recueil d'impressions subjectives retenus est la possession d'un segment continu dans la formulation de la question, la réponse se fait alors en plaçant une croix sur ce segment. Pour un différenciateur sémantique continu par exemple :

Pour une question utilisant une notion de certitude ou de confiance :



Les réponses de x experts à q questions peuvent donc être considérées comme x sous-ensembles flous d'un référentiel E composé de q éléments. Par exemple en reprenant le tableau figure II.3, il peut représenter la réponse de 4 experts à un questionnaire composé de 5 questions. Il peut donc être considéré comme un sous-ensemble aléatoire flou et en conserve alors toutes les propriétés, figure II.8.

Q = questions
E = experts

E\Q	x_1	x_2	x_3	x_4	x_5
1	0,3	1	0,4	0,1	1
2	0,1	0,9	0,4	0	1
3	0,4	0,8	0,4	0	0,8
4	0,6	0,7	0,2	0,2	1

Figure II.8 : SEAF correspondant à 4 experts répondant à 5 questions

Enfin, l'ensemble des réponses d'un expert étant représenté par un SEF peut être mis sous la forme d'un sous-ensemble aléatoire flou par extension. Par exemple l'ensemble des réponses de l'expert 3 sera représenté par le sous-ensemble aléatoire flou suivant :

x_1		x_2		x_3		x_4		x_5	
0	1	0	1	0	1	0	1	0	1
0,4	1	0,8	1	0,4	1	1	0	0,8	1
1	0	1	0	1	0			1	0

Figure II.9 : SEAF pour un expert répondant à 5 questions

En conclusion, ce paragraphe montre la possibilité de synthétiser les données à partir de SEAF. La mesure de ressemblance est choisie comme étant la distance entre deux fonctions de répartition complémentaires. A ce stade, il faut déterminer quelles méthodes peuvent être mises en oeuvre pour traiter les données et quelles sont les représentations qui permettent de visualiser les résultats de l'analyse.

II.3 - ANALYSE DES QUESTIONNAIRES

Afin de situer les objets les uns par rapport aux autres, le premier traitement à réaliser est une classification qui est *"la base de toute connaissance"* /DIDAY 82/. Pour déterminer d'autre part l'importance relative de chaque paramètre, une représentation par plan issue directement de la prise en compte des données a été mise en oeuvre.

II.3-1 Classification

Le but d'une classification est d'obtenir sur une population donnée des classes à l'aide d'un nombre fini de paramètres. Il est alors possible de réaliser une classification en travaillant directement sur les données brutes, ou de partir d'une matrice de proximité. Celle-ci est construite à partir d'une comparaison entre les différents éléments. Les données étant prises en compte sous forme de fonctions de répartition complémentaires, il a été décidé de retenir cette deuxième approche fondée sur la matrice de proximité.

Soit un ensemble A de sous-ensembles aléatoires flous définis sur le même référentiel E, une classification sur ceux-ci, par exemple sur des experts, peut être obtenue en considérant la matrice de proximité obtenue par :

$\forall A_i, A_j \in A$	$d(A_i,A_j) = d(A_j,A_i)$	symétrie
$\forall i$	$d(A_i,A_i) = 0$	antiréflexivité

Cette matrice est dite de dissemblance.

Le problème est alors de pouvoir exploiter au mieux les informations contenues dans une telle matrice. En considérant la théorie des sous-ensembles flous, la décomposition en

sous-relation maximale de similitude est une méthode qui reste très limitée quant à l'extraction des résultats. Pour pallier ce problème, il est nécessaire de faire appel aux méthodes classiques d'analyse des données. Dans ce contexte, seules les méthodes arborescentes, hiérarchies, pyramides... ont été utilisées. Enfin, une méthode nouvelle a été mise en oeuvre : l'arbre à liaisons incomplètes.

a. Sous-relations maximales

La matrice ne possédant pas la propriété de transitivité, la recherche des sous-relations maximales, ou classes, va en général permettre l'obtention de classes empiétantes.

Pour décomposer une relation de dissemblance $\Re$, des relations de dissemblance $\Re_\alpha$ sont utilisées. Elles sont obtenues pour un seuil α de la manière suivante :

$$\text{si} \quad \Re_\alpha = (r_{ij}) \quad \begin{array}{lll} r_{ij} = 1 & \text{si} & r_{ij} \leq \alpha \\ r_{ij} = 0 & \text{sinon} & \end{array}$$

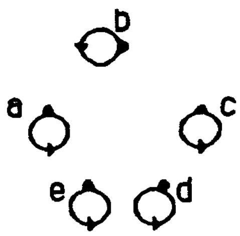
Il existe déjà des algorithmes de décomposition en sous-relations maximales de similitude, par exemple ceux de MALGRANGE ou de PICHAT cités par KAUFMANN /KAUFMANN 77/, mais l'intérêt d'en réaliser un qui soit plus adapté aux systèmes informatiques est apparu. Cet algorithme est présenté dans l'annexe I. Notons également que dans le cas où la matrice est rectangulaire l'analyse peut se faire à partir de treillis de Galois.

Nous allons illustrer sur un exemple la décomposition d'une matrice de dissemblance. Soit la relation de dissemblance $\Re$ définie comme suit :

$$\Re = \begin{pmatrix} 0 & 0{,}1 & 0{,}2 & 0{,}5 & 0{,}4 \\ 0{,}1 & 0 & 0{,}2 & 0{,}9 & 0{,}4 \\ 0{,}2 & 0{,}2 & 0 & 0{,}8 & 0{,}8 \\ 0{,}5 & 0{,}9 & 0{,}8 & 0 & 0{,}5 \\ 0{,}4 & 0{,}4 & 0{,}8 & 0{,}5 & 0 \end{pmatrix}$$

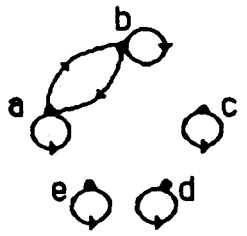
En désignant par a, b, c, d, e les colonnes de la relation, au seuil 0 la relation suivante est obtenue ainsi que son graphe associé :

$$\Re_0 = \begin{pmatrix} 1 & & & & \\ & 1 & & 0 & \\ & & 1 & & \\ & 0 & & 1 & \\ & & & & 1 \end{pmatrix}$$



c'est à dire 5 classes singletons: {a} {b} {c} {d} {e}. Au seuil 0,1 il est obtenu :

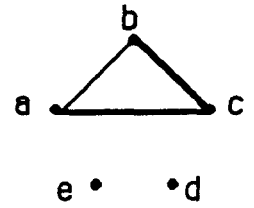
$$\mathfrak{R}_{.1} = \begin{pmatrix} 1 & 1 & & & \\ 1 & 1 & & 0 & \\ & & 1 & & \\ & 0 & & 1 & \\ & & & & 1 \end{pmatrix}$$



donc 4 classes : {a,b} {c} {d} {e}.

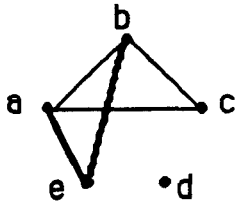
Dans la suite, pour plus de lisibilité des graphes associés aux matrices, les boucles de réflexivité et la symétrie ne sont représentées que par un trait simple. D'autre part, toutes les nouvelles liaisons qui se forment par seuil sont tracées en gras. Il vient donc au seuil .2 :

$$\mathfrak{R}_{.2} = \begin{pmatrix} 1 & 1 & 1 & & \\ 1 & 1 & 1 & 0 & \\ 1 & 1 & 1 & & \\ & 0 & & 1 & \\ & & & & 1 \end{pmatrix}.$$



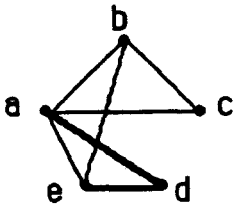
et 3 classes: {a,b,c} {d} {e}. Au seuil .4 :

$$\mathfrak{R}_{.4} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$



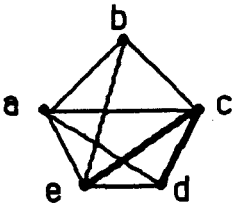
et 3 classes {a,b,c} {a,b,e} {d}. Au seuil .5 :

$$\mathfrak{R}_{.5} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$



et 3 classes {a,b,c} {a,b,e} {a,d,e}. Au seuil .8 :

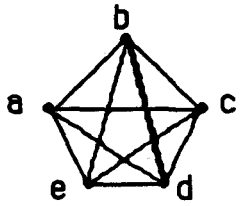
$$\mathfrak{R}_{.8} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$



et 2 classes {a,b,c,e} {a,c,d,e}

Enfin au seuil .9 on retrouve bien évidemment la classe unique {a,b,c,d,e} :

$\mathfrak{R}_{.9} = (1)$



Si cette représentation par graphes permet de bien assimiler la détermination des classes empiétantes, elle reste malheureusement "pauvre" pour permettre de dégager des résultats, surtout si le nombre d'éléments de la matrice devient important.

b. Les méthodes classiques

A partir d'un tableau de distances de nombreuses méthodes existent permettant d'extraire des résultats. Par exemple, une méthode visuelle /BERTIN 77/ est fondée sur la permutation des lignes et des colonnes de la matrice. Seules les méthodes arborescentes où il s'agit de représenter les ressemblances entre les observations en les classant seront traitées. Dans le cas d'une classification ascendante, elles sont d'abord considérées comme des singletons puis agrégées en fonction de leurs ressemblances. Deux méthodes peuvent alors être distinguées : les hiérarchies faisant intervenir des partitions emboîtées et les pyramides faisant intervenir des recouvrements.

* Les hiérarchies indicées

D'un point de vue matriciel les méthodes hiérarchiques cherchent à*"transformer la matrice des proximités en une nouvelle matrice dans laquelle les groupes d'objets sont plus apparents que dans la matrice des proximités initiales"* /CHANDON et PINSON 81/ citant JARDINE et SIBSON.

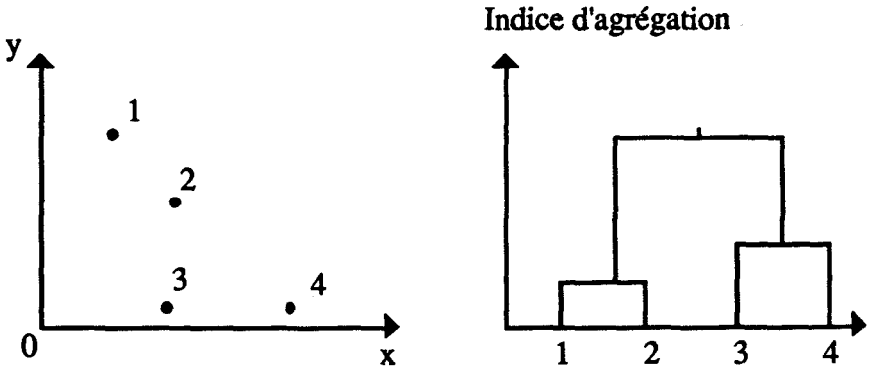


Figure II.10 : *hiérarchie indicée*

Les hiérarchies cherchent à produire un ensemble de parties hiérarchiquement emboîtées, figure II.10, et l'indiçage de la hiérarchie permet alors d'affiner la partition. Une hiérarchie peut s'obtenir selon deux stratégies opposées : par agrégations successives, classification ascendante, et par divisions successives, classification descendante.

Le choix du critère d'agrégation ou de division est fondamental et est dicté par le type de données à traiter et l'objectif fixé par l'analyse. Il en existe beaucoup, par exemple : le lien minimum, le lien maximum, la moyenne des distances, pour ne citer qu'eux.

D'autre part, on montre qu'il y a équivalence entre l'ensemble des hiérarchies indicées et un ensemble de distances appelées ultramétriques /DIDAY 82/ (cf I.2-4) et qu'il existe un ordre qui transforme une matrice ultramétrique en une matrice particulière dite de ROBINSON, c'est à dire une matrice dont les termes des lignes et des colonnes sont croissants à partir de chaque terme de la diagonale. Par exemple :

$$M = \begin{pmatrix} 0 & 1 & 2 & 6 \\ 1 & 0 & 2 & 6 \\ 2 & 2 & 0 & 2 \\ 6 & 6 & 2 & 0 \end{pmatrix}$$

D'un point de vue matriciel, une matrice de proximité issue de données expérimentales n'est jamais de ce type, et pour se ramener à une telle matrice qui permet une représentation hiérarchique il est nécessaire de réaliser des concessions. Enfin, le problème de ce type de méthode est de trouver un "bon" indice d'agrégation qui ne donne pas de résultats artificiels.

* Les pyramides /DIDAY 86/

Si une hiérarchie est formée d'une suite de partitions emboîtées, DIDAY montre alors que *"l'ensemble des hiérarchies peut être plongé dans un ensemble plus vaste donnant des représentations (appelées pyramides), plus riches d'informations, plus proches des données initiales, faisant apparaître des recouvrements emboîtés au lieu de partitions"* et qu'en supprimant les contraintes dûes à l'inégalité ultramétrique, il est possible de déterminer une classe d'indices particuliers appelés *"indices pyramidaux"* qui inclut l'ensemble des ultramétriques.

D'un point de vue matriciel cela revient à travailler sur les matrices suivantes :

- matrices de Robinson qui ont été définies plus haut,

• matrices SDR (sur-diagonale "rectangle") : une matrice est dite SDR si et seulement si chaque terme de la sur-diagonale est inférieur aux termes du rectangle qui lui est associé.

exemple : $\begin{pmatrix} 0 & 2 & 6 & 4 \\ 2 & 0 & 3 & 9 \\ 6 & 3 & 0 & 4 \\ 4 & 9 & 4 & 0 \end{pmatrix}$ et $3 \leq \begin{pmatrix} 6 & 4 \\ 3 & 9 \end{pmatrix}$

• matrices SDD (sur-diagonale "dominée"): une matrice est dite SDD si et seulement si les termes des lignes et des colonnes de sa partie triangulaire supérieure sont plus grands que le terme de la sur-diagonale qu'elles contiennent.

exemple : $\begin{pmatrix} 0 & 2 & 5 & 3 \\ 2 & 0 & 4 & 6 \\ 5 & 4 & 0 & 1 \\ 3 & 6 & 1 & 0 \end{pmatrix}$ et $4 \leq 5$ et $4 \leq 6$

La visualisation d'une pyramide à partir d'une matrice quelconque nécessite de déterminer un ordre θ, pas toujours unique, sur les singletons qui soit compatible avec la pyramide.

Exemple : Soit la matrice de Robinson M précédente, la pyramide induite par cette matrice est la suivante :

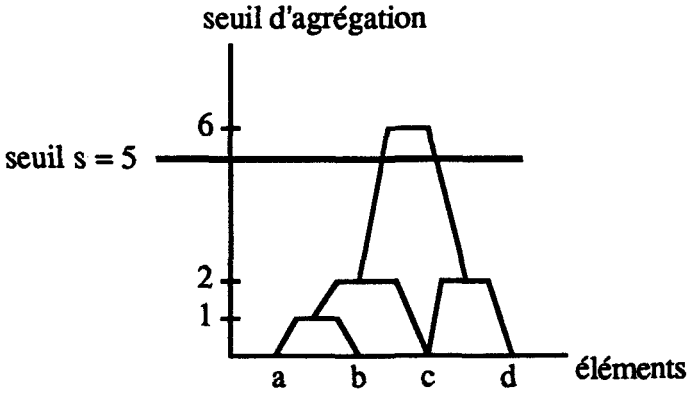


Figure II.11 : exemple de pyramide

Pour "lire" cette pyramide il suffit de "couper" à un seuil s donné et les classes existantes à ce seuil sont tous les paliers existants directement inférieurs. Par exemple en coupant à s = 5 sur cette pyramide les classes obtenues sont les suivantes :

{a,b,c} et {c,d}

L'intérêt des pyramides est la possibilité pour chaque palier d'avoir deux prédécesseurs et donc de pouvoir représenter des classes empiétantes, comme le montre la figure II.11. Cette méthode est une des seules à essayer de prendre en compte le problème de la représentation visuelle des classes empiétantes, même si elle est limitée par le fait qu'un successeur ne peut avoir plus de deux prédécesseurs. A partir du moment ou un élément appartient à plus de deux classes la représentation présente donc des "concessions" par rapport à la matrice des proximités initiale. Une seule autre méthode, non arborescente, à notre connaissance, permet de résoudre en partie le problème des classes empiétantes à partir de diagrammes de Hasse modifiés en diagrammes à anneaux (Ring diagrams) /REGGIANI et MARCHETTI 88/.

En résumé, les méthodes existantes permettent une représentation visuelle de classes à partir d'une matrice de dissemblance. Cependant, le problème consiste à déterminer pour les hiérarchies un indice d'agrégation I et pour les pyramides un ordre θ, I et θ devant assurer une représentation arborescente sans croisements fidèle aux données.

Une nouvelle méthode a alors été mise en oeuvre pour dégager une représentation permettant la prise en compte des classes empiétantes.

c. Arbre à liaisons incomplètes

D'après les travaux de DIDAY et BROSSIER /BROSSIER 86/, il apparaît qu'une représentation par arbres qui cherche à garder le maximum de liaisons possibles nécessite toujours un certain nombre de contraintes. Une représentation arborescente d'une classification est un compromis entre le nombre de classes à représenter et le nombre de liaisons joignant ces classes. Pour les deux méthodes qui ont été décrites, la classification à partir d'une matrice quelconque, revient à transformer la matrice initiale en une matrice particulière, de Robinson pour les hiérarchies, de Robinson, SDD ou SDR pour les pyramides, qui permette de représenter une arborescence sans croisements qui reste fidèle aux données.

Pour pouvoir prendre en compte toutes les classes issues de la matrice initiale, c'est à dire ne réaliser aucune "concession" et obtenir une représentation visuelle de ces classes sans croisements, le principe de l'arbre à liaisons incomplètes est de ne pas représenter toutes les liaisons joignant ces classes. Les seules liaisons représentées sont telles que si C_1 et C_2 sont deux classes telles que C_1 est une feuille, $C_1 \subset C_2$ et Ω l'ensemble des classes on a :

$$\forall i \in \Omega \quad \nexists C_i \quad C_1 \subset C_i \subset C_2$$

Exemple :

Au seuil s_i supposons qu'il y ait trois classes : {ab} {ac} et {bd}
et au seuil s_{i+1} deux : {abc} {bcd}. La représentation est alors la suivante :

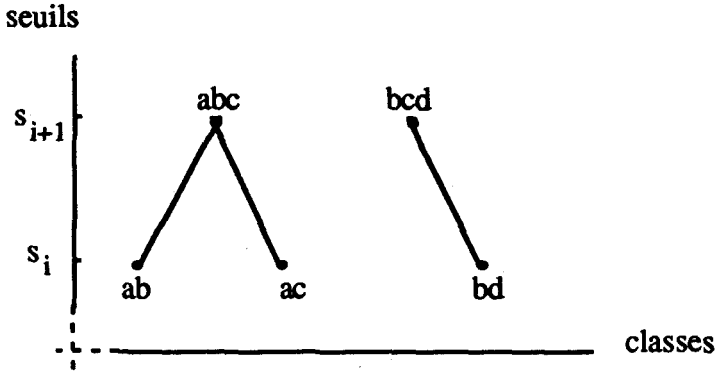


Figure II.12 : exemple de représentation de classes empiétantes

On remarque que {bcd} se créant au seuil s_{i+1} élimine {bd} mais n'intervient ni avec {ab}, ni avec {abc} alors que b appartient à ces deux classes.

L'algorithme de construction de l'arbre à liaisons incomplètes ainsi que des exemples et une comparaison avec les hiérarchies et les pyramides sont présentés annexe II. Celle-ci montre que cette méthode est difficilement exploitable, quant à l'extraction des résultats, pour des matrices dépassant 10 éléments, mais qu'il semble nécessaire de l'utiliser pour, tout au moins, valider les résultats issus des autres méthodes arborescentes.

En conclusion, on peut dire que l'interprétation issue des différentes méthodes arborescentes est d'autant plus facile que les contraintes imposées à la matrice initiale sont importantes, les hiérarchies sont les plus "faciles à lire", mais, que plus on se rapproche d'une représentation fidèle aux données, matrice sans contraintes, et plus celle-ci est difficile à interpréter.

II.3-2 Représentation par plan

En analyse des données il est souvent fait appel à des représentations par plan permettant de donner une vue synthétique du contenu d'un tableau de données. Citons les exemples de l'analyse en composante principale ou l'analyse factorielle des correspondances. A partir de cette constatation l'idée de construire un plan issu directement de la formulation du questionnaire et des réponses à celui-ci est apparue /LOSLEVER 88/.

A cette fin deux SEAF particuliers doivent être définis.

a. Construction de deux SEAF particuliers

* un SEAF optimal

La formulation du questionnaire permet en général la détermination d'un SEAF optimal constitué des réponses "idéales" à chaque question. En considérant la question illustrée figure II.13, dans le cadre d'une expérience cherchant à tester la qualité d'un écran par exemple, la réponse "idéale" est "pas du tout" et la fonction de répartition complémentaire associée est alors celle présentée.

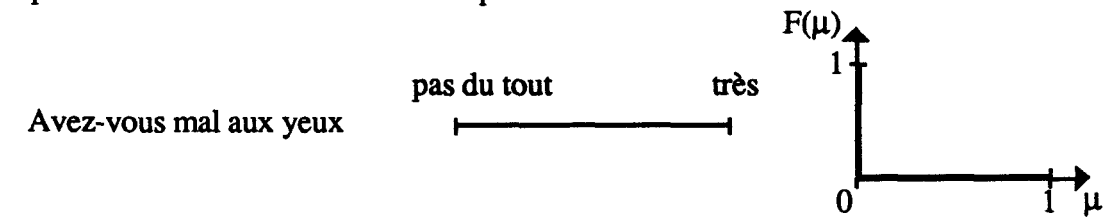


Figure II.13 : fonction de répartition complémentaire optimale associée à une question

Si dans ce cas, et dans de nombreux cas la réponse idéale se trouve à une des extrémités du différenciateur sémantique continu, il est possible d'imaginer des questions qui ne suivent pas cette règle, par exemple à la question de la figure II.14 la réponse "idéale" peut se trouver au centre du différenciateur sémantique.

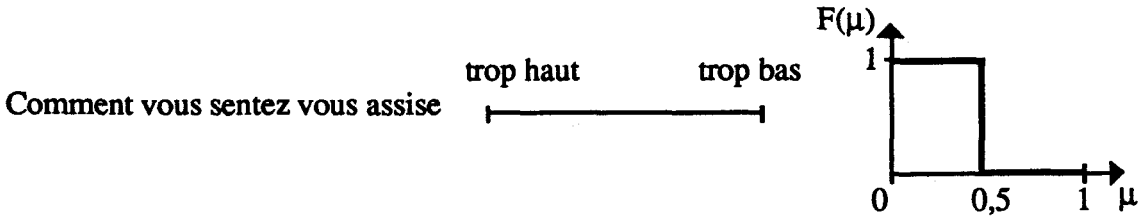


Figure II.14 : fonction de répartition optimale complémentaire associée à une question

* un SEAF type

Le deuxième SEAF est représentatif de l'ensemble des SEAF considérés. Il est possible d'imaginer de nombreux SEAF représentatifs, par exemple obtenus à l'aide d'une moyenne ou liés à l'écart-type. Mais, le fait d'utiliser des fonctions de répartition a amené tout naturellement à considérer pour chaque question la fonction de répartition obtenue par concaténation de toutes les autres, figure II.15, comme fonction de répartition représentative. Le SEAF obtenu à l'aide de ces fonctions de répartitions représentatives à été nommé SEAF type. L'intérêt de ce SEAF par rapport à d'autres qu'il aurait été possible de choisir comme représentatifs, réside dans le fait de ne perdre que peu d'informations par rapport aux données initiales.

Effectivement, comme le montre figure II.15, si la première fonction de répartition complémentaire représente la réponse d'un expert à deux moments différents à une question et la deuxième celle d'un autre expert à la même question, le fait de concaténer les deux permet de retrouver les 4 réponses à la question, l'information étant perdue sur la réponse des experts aux questions.

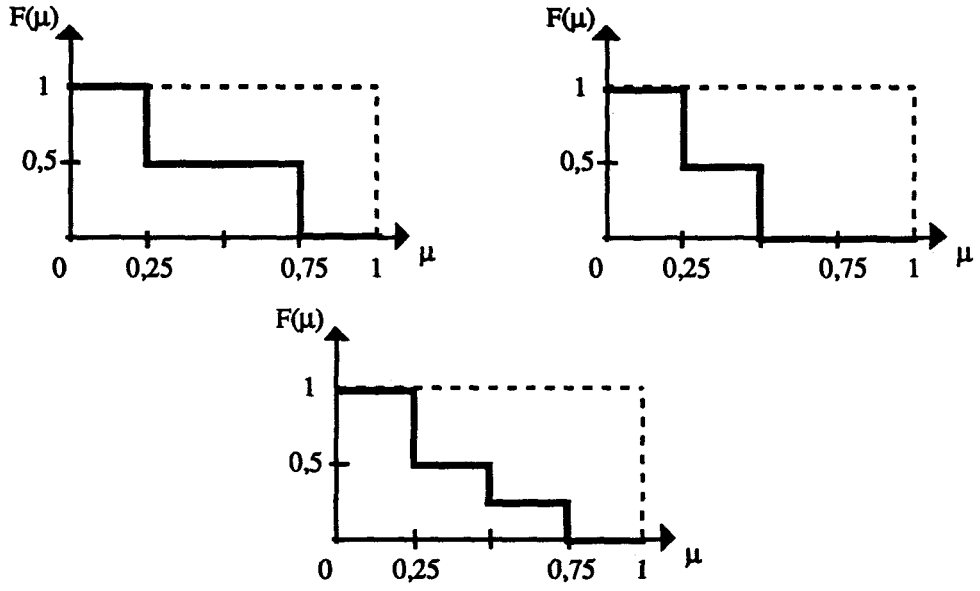


Figure II.15 : concaténation de 2 fonctions de répartition complémentaires

Deux SEAF particuliers ayant été définis, le SEAF optimal qui est lié à la formulation du questionnaire et le SEAF type qui varie en fonction des variables prises en compte, il est alors possible de comparer tous les SEAF à ces 2 SEAF particuliers.

b. Construction du plan

* utilisation de la distance

Dans un premier temps, en utilisant la distance, définie au II.2-2, comme moyen de comparaison chaque SEAF A_i peut être défini par 2 distances :

$d(A_i$,SEAF optimal)
$d(A_i$,SEAF type)

Ces 2 distances peuvent être considérées comme des coordonnées et les SEAF A_i peuvent alors être projetés dans un plan de la manière suivante, figure II.16 :

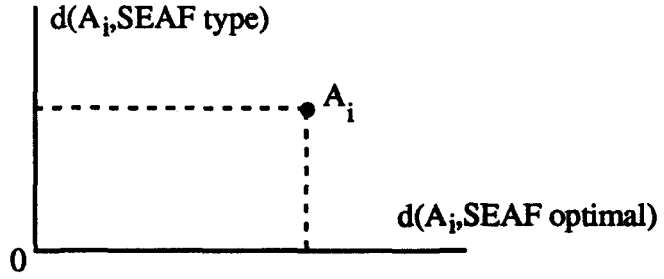


Figure II.16 : projection d'un point dans le plan avec la distance

En conservant l'exemple de la figure II.15 et en considérant que l'optimal pour la question prise en compte se trouve au centre du différenciateur, figure II.14, les distances sont calculées entre les deux experts et les deux SEAF particuliers. La représentation figure II.17 illustre le résultat obtenu.

$d(A_1,\text{SEAF optimal}) = 0,25$ $\quad d(A_1,\text{SEAF type}) = 0,0625$
$d(A_2,\text{SEAF optimal}) = 0,125$ $\quad d(A_2,\text{SEAF type}) = 0,0625$

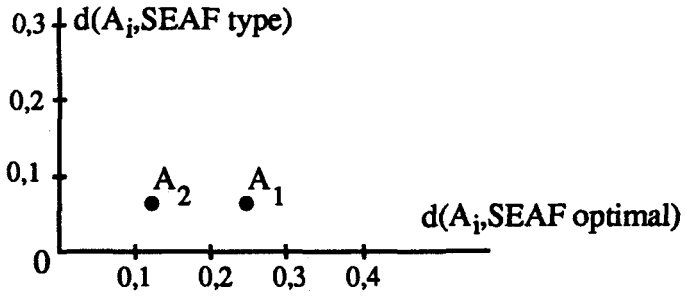


Figure II.17 : exemple de plan en utilisant la distance comme moyen de comparaison

* utilisation de l'indice

En utilisant à présent l'indice défini au II.2-2 comme moyen de comparaison chaque SEAF A_i peut être défini par 4 valeurs que l'on peut également représenter dans un plan, figure II.18. Les 4 valeurs IS_0, II_0, IS_t, II_t sont définies à l'aide de l'indice de la manière suivante :

$I(A_i,\text{SEAF optimal}) = (IS_0,II_0)$
$I(A_i,\text{SEAF type}) = (IS_t,II_t)$

(IS_0,IS_t) et (II_0,II_t) peuvent alors représenter 2 points correspondant au SEAF A_i.

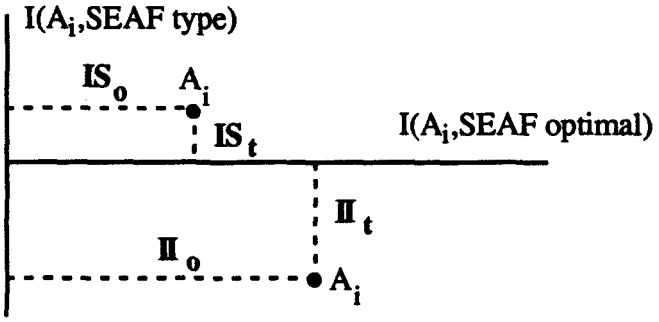


Figure II.18 : projection d'un point dans le plan avec l'indice

En reprenant l'exemple de la figure II.15 le plan obtenu est présenté figure II.19. L'exemple étant très simple, A_1 est toujours au dessous du SEAF type ce qui entraîne $II_t = 0$ et un des représentants de A_1 sur l'axe, et A_2 toujours au dessus ce qui entraîne $IS_t = 0$ et un des représentants de A_2 sur l'axe également.

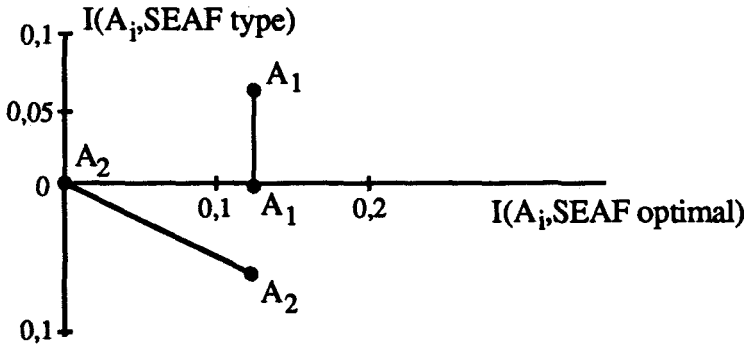


Figure II.19 : exemple de plan en utilisant l'indice comme moyen de comparaison

Il convient à présent d'expliquer comment dégager des résultats de ces plans et pour ce faire passer à la phase de l'interprétation.

c. Interprétation

Si en général, en analyse des données, par exemple en analyse factorielle des correspondances, l'interprétation des plans est relativement complexe, faisant intervenir de nombreux paramètres /VOLLE 81/, dans le cas présent elle est plus aisée. Malheureusement nous ne pourrons pas approfondir ici l'interprétation car elle est trop dépendante des questions et du choix de l'optimal. Seules les grandes lignes sont données, le chapitre IV montre sur un exemple, comment affiner l'interprétation.

Dans un premier temps il convient de déterminer quand 2 points peuvent être considérés comme significativement distincts sur les plans. Effectivement, comme il a été dit

précédemment les distances utilisées restent dans [0,1], donc si tous les points projetés se retrouvent par exemple dans un carré de coté 0,05 il s'agit de savoir si on a le **droit** d'interpréter. Pour cela il est nécessaire d'effectuer un calcul d'erreur dépendant, au minimum, de la longueur du segment utilisé et d'une hypothèse issue de l'expérience ; à partir de quand deux réponses peuvent être considérées comme différentes.

Les grandes lignes de l'interprétation sont alors les suivantes :

- par rapport au SEAF type

Les SEAF projetés dans le plan permettent de déterminer ceux qui sont les plus proches du SEAF type correspondant alors à un comportement "normal", et inversement les SEAF les plus éloignés déterminant ceux qui sont globalement opposés aux autres SEAF. Les résultats peuvent être affinés si on utilise l'indice I comme moyen de comparaison, dans la mesure où chaque SEAF a deux représentants dans le plan.

- par rapport au SEAF optimal

En supposant que le SEAF optimal correspond à un objectif à atteindre, il convient de rechercher quels SEAF sont les plus proches - respectivement éloignés - déterminant ainsi ceux qui répondent le mieux - respectivement le plus mal - aux critères souhaités. En utilisant l'indice I comme moyen de comparaison il est alors possible de dégager les tendances de réponses par rapport au SEAF optimal.

- Enfin, en regardant simultanément par rapport aux deux SEAF particuliers type et optimal, une tendance globale peut être dégagée. Par exemple, un SEAF globalement opposé aux autres, éloigné par rapport au SEAF type, qui se trouve être le plus proche de l'optimal indique ainsi que par son comportement "anormal" par rapport aux autres SEAF, il présente un comportement "satisfaisant" par rapport à l'expérience.

II.4 - CONCLUSION

Le domaine d'application de l'analyse a été restreint aux questionnaires laissant le plus de liberté possible au sujet, des questionnaires plus restrictifs, questionnaires à choix multiples par exemple, pourraient être traités de la même façon mais cela semble peu intéressant car : d'une part, pour ces types de questionnaires des méthodes simples et efficaces existent déjà, et d'autre part le fait d'utiliser des fonctions de répartition libère du choix d'un nombre de modalités pour permettre de travailler avec une échelle continue.

L'utilisation des sous-ensembles aléatoires flous a amené une différence fondamentale avec les méthodes multidimensionnelles d'analyse de données - analyse factorielle par exemple - quant à la prise en compte des données : fonctions de répartition par rapport aux données brutes utilisées pour ces méthodes. Si la partie classification a du être traitée par des méthodes d'analyse de données classiques une méthode descriptive a pu être dégagée permettant de donner une vue synthétique du contenu du tableau analysé. Enfin, l'utilisation du formalisme des sous-ensembles flous permet le recours à des techniques propres à ces derniers telles que l'inférence déductive. Cette technique est alors utilisés comme base à une mise en relation entre données comme l'illustre le chapitre suivant.

CHAPITRE III

MISE EN RELATION DE DONNEES

De nombreux domaines expérimentaux nécessitent le recueil de données dites objectives et subjectives. Le but est d'extraire des résultats à l'aide de différentes méthodes compte tenu de ces deux types de données de nature différente. Pour aller plus loin dans l'analyse il est important également de vérifier l'adéquation entre ceux-ci, et d'étudier s'il existe une relation stable entre ces deux types de données expérimentales.

Le choix du formalisme des mathématiques floues pour l'analyse des impressions subjectives a permis d'élaborer une méthode de mise en relation, à l'aide de l'inférence déductive et du Modus Ponens généralisé. Le début du chapitre rappelle alors les différents principes issus de la théorie des sous-ensembles flous qui permettent la mise en oeuvre de cette méthode.

III.1 - L'INFERENCE DEDUCTIVE

III.1-1 Implication

Définition : on appelle "implication" une opération logique définie de la façon suivante : P et Q étant 2 propositions, V(P) et V(Q) leur valeur logique, on associe à l'application, notée P → Q, l'opérateur V(¬P)∨V(Q) où ¬P = non P et ∨ représente le maximum.

La table de vérité est alors la suivante :

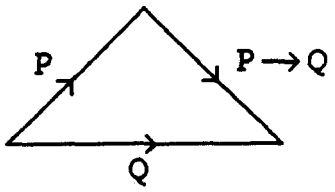
V(P) \ V(Q)	0	1
0	1	1
1	0	1

L'implication peut alors être utilisée de 2 manières différentes :

- modus ponens

V(P) = 1	prémisse
V(P → Q) = 1	prémisse
V(Q) = 1	conclusion

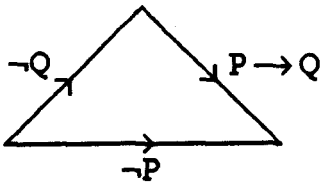
qui peut se schématiser de la manière suivante :



- modus tollens

V(Q) = 0	prémisse
V(P → Q) = 1	prémisse
V(P) = 0	conclusion

qui peut se schématiser de la manière suivante :



III.1-2 méta-implication

Une méta-implication n'est pas une opération booléenne, elle est notée P ⇒ Q et sa table de vérité est la suivante :

V(P) \ V(Q)	0	1
0	?	?
1	0	1

c'est à dire si V(P) = 0 on ne doit rien pouvoir déduire. Une méta-implication ne peut donc être utilisée, au contraire d'une implication, que si la valeur de vérité de P est égale à 1.

Or pour $P \rightarrow Q$ comme pour $P \Rightarrow Q$ on peut dire: "si P est vrai alors Q est vrai". C'est pourquoi KAUFMANN /KAUFMANN 87/ suggère *"d'employer :*

pour $P \Rightarrow Q$ *: si P alors Q*

pour $P \rightarrow Q$ *: P implique (ou infère) Q"*

Dans beaucoup de contextes, il s'avère nécessaire de modéliser des assertions imprécises. Ce problème peut être pris en compte à l'aide de principes issus des théories des sous-ensembles flous et des possibilités. Dans ce cadre, et en prenant en compte le formalisme utilisé par d'autres auteurs /DUBOIS et PRADE 84/ /MOREAU 87/ quelques définitions vont être rappelées :

Si pour "X est A" la variable X est définie par sa distribution Π_X, A par sa fonction caractéristique μ_A il est possible d'écrire /ZADEH 81/ :

$$\Pi_X = \mu_A$$

Une méta-implication du type "si X est A alors Y est B" sera représentée par une distribution de possibilité conditionnelle notée $\Pi_{Y/X}$. La valeur $\Pi_{Y/X}(x,y)$ est alors la possibilité que Y prenne la valeur y sachant que X vaut x.

$\Pi_{Y/X}(x,y)$ s'écrira alors en fonction des distributions Π_X et Π_Y.

Beaucoup d'auteurs ont alors proposé des distributions $\Pi_{Y/X}(x,y)$ et ont effectué des comparaisons dans différents contextes /KISZKA 85/ /CAO et KANDEL 89/ par exemple. Citons quelque unes de ces distributions :

- ZADEH /ZADEH 75/

$$\Pi_{Y/X}(x,y) = \min [\, 1 \,,\, 1 - \Pi_X(x) + \Pi_Y(y) \,]$$

$$\Pi_{Y/X}(x,y) = \max [\, 1 - \Pi_X(x) \,,\, \min (\Pi_X(x) \,,\, \Pi_Y(y)) \,]$$

- MIZUMOTO et ZIMMERMANN /MIZUMOTO et ZIMMERMANN 82/

$$\Pi_{Y/X}(x,y) = 1 - \Pi_X(x) + \Pi_X(x).\Pi_Y(y)$$

- MAMDANI /MAMDANI 77/

$$\Pi_{Y/X}(x,y) = \min [\, \Pi_X(x) \,,\, \Pi_Y(y) \,]$$

III.1-3 L'inférence déductive

Si une règle du type "si X est A alors Y est B" est intéressante elle ne peut être utilisée que si "X est A" existe. Or dans beaucoup de cas, on cherche à déduire des résultats avec un A' qui diffère de A.

Par exemple si l'on considère "la réponse de l'expert est l'optimal", il semble nécessaire de pouvoir déduire un résultat avec "la réponse de l'expert est proche de l'optimal". ZADEH /ZADEH 77/ définit ce type d'inférence comme le modus ponens généralisé qui correspond au schéma syllogistique suivant :

si X est A alors Y est B
X est A'

Y est B'

La modélisation de la composition inférentielle est alors possible en utilisant la formule suivante /DUBOIS et PRADE 84/ qui fournit un résultat concernant la variable Y sous la forme d'une distribution Π_y :

$$\Pi_y(y) = \sup_x [\ \Pi_x(x) \Lambda \Pi_{y/x}(x,y)\ ]$$

où l'opérateur Λ désigne une fonction de [0,1] x [0,1] dans [0,1] qui vérifie :

$$\begin{cases} \Lambda \text{ est monotone} \\ 1 \Lambda 1 = 1 \\ 1 \Lambda 0 = 0 \Lambda 1 = 0 \\ 1 \Lambda a = 0 \Rightarrow a = 0 \\ a \Lambda 1 = 0 \Rightarrow a = 0 \end{cases}$$

L'extension de cette formule permet de modéliser le modus ponens généralisé de ZADEH /ZADEH 77/ :

$$\mu_{B'}(y) = \sup_x [\ \mu_{A'}(x) \Lambda \Pi_{y/x}(x,y)\ ]$$

D'autres travaux, /YAGER 80/ /MAGREZ 85/ proposent d'autres modélisations. Seule celle de DUBOIS et PRADE est utilisée dans la suite de notre travail.

Le résultat de l'inférence est alors déterminé par un couple $(\Pi_{y/x}, \Lambda)$. A partir des distributions de possibilité proposées, chapitre II.1-2, par exemple, il faut trouver un opérateur Λ qui permette de réaliser une inférence qui ne donne pas de résultats contra-intuitifs. L'association de l'opérateur de composition et d'une distribution de possibilité conditionnelle doit alors respecter les deux critères cités, chapitre III.1-1, à savoir :

- le modus ponens

si X est A alors Y est B

X est A

Y est B

ce qui se traduit par :

$$\mu_B(y) = \sup_x [\, \mu_A(x) \wedge \Pi_{y/x}(x,y) \,]$$

- le modus tollens

si X est A alors Y est B

non (Y est B)

---

non (X est A)

ce qui revient à appliquer le modus ponens à la règle "si non(Y est B) alors non(X est A)" et qui s'écrit, c représentant une négation forte :

$$c[\mu_A(x)] = \sup_y [\, c[\mu_B(y)] \wedge \Pi_{\text{non } x/\text{non } y}(x,y) \,]$$

Comme $\Pi_{\text{non } x/\text{non } y}(x,y) = \Pi_{y/x}(x,y)$ par symétrie contrapositive, il vient :

$$c[\mu_A(x)] = \sup_y [\, c[\mu_B(y)] \wedge \Pi_{y/x}(x,y) \,]$$

Beaucoup de tels couples ont été mis au point, vérifiant les deux critères ou seulement l'un d'entre eux, chacun dépendant plus ou moins de l'utilisation requise. Citons quelques exemples de couples vérifiant ces deux critères :

$$\Pi_{y/x}(x,y) = \min [1, 1 - \Pi_x(x) + \Pi_y(y)] \quad \text{et} \quad a \wedge b = \max (0, a + b - 1)$$

$$\Pi_{y/x}(x,y) = \min [1 - \Pi_x(x), \Pi_y(y)] \quad \text{et} \quad a \wedge b = \begin{cases} 0 & \text{si } a+b \le 1 \\ b & \text{sinon} \end{cases}$$

$$\Pi_{y/x}(x,y) = 1 - \Pi_x(x) + \Pi_x(x).\Pi_y(y) \quad \text{et} \quad a \wedge b = \begin{cases} 0 & \text{si } a = 0 \\ \max [\, 0, \frac{a+b-1}{a} \,] & \text{sinon} \end{cases}$$

$$\Pi_{y/x}(x,y) = \min [\, (\Pi_x(x) \; S \; \Pi_y(y)) \,, (1 - \Pi_x(x)) \; G \; (1 - \Pi_y(y)) \,]$$

avec :

$$S \begin{cases} 1 & \text{si } \Pi_x(x) \le \Pi_y(y) \\ 0 & \text{sinon} \end{cases} \quad \text{et} \quad G \begin{cases} 1 & \text{si } \Pi_x(x) \le \Pi_y(y) \\ \Pi_y(y) & \text{sinon} \end{cases}$$

l'opérateur de composition est défini par : $a \Lambda b = \min [a,b]$

et sous la condition suivante : $\forall x \quad \exists y \ / \ \Pi_y(y) = \Pi_x(x)$ /MIZUMOTO 82/

D'autres couples ne vérifient que le modus ponens, par exemple :

$$\Pi_{y/x}(x,y) = \begin{cases} 1 & \text{si } \Pi_x(x) = 0 \\ \min [1 , \dfrac{\Pi_y(y)}{\Pi_x(x)}] & \text{sinon} \end{cases} \qquad \text{et} \quad a \Lambda b = a.b$$

$$\Pi_{y/x}(x,y) = \begin{cases} 1 & \text{si } \Pi_x(x) \leq \Pi_y(y) \\ \Pi_y(y) & \text{sinon} \end{cases} \qquad \text{et} \quad a \Lambda b = \min [a , b]$$

Enfin, des modèles existent vérifiant les deux critères mais ne pouvant se mettre sous la forme d'un couple $(\Pi_{y/x},\Lambda)$, par exemple celui de MOREAU /MOREAU 87/ issu de celui de ZADEH :

$$\mu_{B'}(y) = \sup_x [\min (\mu_{A'}(x) , 1 - \mu_A(x) + \mu_B(y))]$$

et transformé de la manière suivante :

$$\mu_{B'}(y) = 2 \sup_x [\min (\mu_{A'}(x) , 1 - \mu_A(x) + \mu_B(y))] - 1$$

sous la condition :

$$\forall \alpha \in \{ \beta \in [0,1] / \exists y, \mu_B(y) = \beta \} \quad \exists x \ / \ \mu_A(x) = \frac{1+\alpha}{2}$$

Après avoir rappelé les différentes définitions relative à l'inférence déductive, la suite du chapitre est consacrée à la mise en oeuvre de la méthode de mise en relation de données.

III.2 - MISE EN RELATION DE DONNEES

Ce paragraphe expose dans un premier temps la méthodologie permettant de déterminer s'il existe une adéquation entre deux groupes de données, avant de présenter au travers d'un exemple les divers problèmes liés à cette méthode /GUERRA et ROGER 91/.

Le début du paragraphe présente alors les notations qui seront utilisées par la suite pour les ensembles et les variables.

III.2-1 Méthodologie de mise en relation de deux ensembles

Les ensembles employés sont les suivants :

E : ensemble des r experts,

S : ensemble des p variables subjectives,

O : ensemble des q variables objectives.

La méthodologie proposée se décompose alors en quatre étapes :

a. Création de la relation R

La mise en relation des deux ensembles doit s'effectuer sur un certain nombre de variables et d'experts. Si dans l'étape préalable de traitement séparé des ensembles il apparait des experts ayant un comportement "incohérent" ou, des variables peu discriminantes, l'analyse peut alors être réduite à des sous-ensembles excluant ces experts et ces variables. Dans ce contexte, il est nécessaire, pour cette première étape de déterminer les sous-ensembles des experts et variables choisis pour réaliser la mise en relation. Soient alors les ensembles :

Ex ($\subset$ E) : un sous-ensemble de e experts ($1 \leq e \leq r$),
Su ($\subset$ S) : un sous-ensemble de s variables subjectives ($1 \leq s \leq p$),
Ob ($\subset$ O) : un sous-ensemble de o variables objectives ($1 \leq o \leq q$).

Supposons que l'on désire créer une relation sur ces e experts entre les s variables subjectives et les o variables objectives, il s'agit alors de construire les 2 fois e sous-ensembles aléatoires flous suivants, figure III.5 :

- S_i ($1 \leq i \leq e$) : SEAF du $i^{ème}$ expert construit sur l'ensemble des variables subjectives Su. Il est constitué de s fonctions de répartition et représente l'avis de l'expert i associé aux variables subjectives Su,
- O_i ($1 \leq i \leq e$) : SEAF du $i^{ème}$ expert construit sur l'ensemble des variables objectives Ob. Il est constitué de o fonctions de répartition et représente pour l'expert i les données recueillies sur les variables objectives Ob.

Les données ayant été synthétisées sous forme de SEAF, la création de la relation R nécessite alors la construction de deux SEAF particuliers, représentatifs des 2 sous-ensembles considérés. Le choix de ce SEAF particulier s'est porté sur le SEAF type, chapitre II paragraphe 3.

Ces deux SEAF particuliers, figure III.1b, notés TYPS et TYPO, permettent la création de la relation R en modélisant :

"si X est le SEAF type subjectif alors Y est le SEAF type objectif"
ou : "si X est le SEAF type objectif alors Y est le SEAF type subjectif".

La règle présentée sous-entend que les SEAF types ne sont pas mis sous forme cumulée, chapitre II paragraphe2. La relation créée correspond en fait à la concaténation de e x n R_{ij} ($1 \leq i \leq e$, $1 \leq j \leq n$) relations créées à partir de n règles pour chaque expert. Une autre façon de présenter ce raisonnement est :

$$\begin{cases} \text{expert 1} \begin{cases} \text{Si X est } A_{11} \text{ alors Y est } B_{11} \\ \dots \quad \dots \\ \text{Si X est } A_{1n} \text{ alors Y est } B_{1n} \end{cases} \\ \dots \quad \dots \\ \text{expert e} \begin{cases} \text{Si X est } A_{e1} \text{ alors Y est } B_{e1} \\ \dots \quad \dots \\ \text{Si X est } A_{en} \text{ alors Y est } B_{en} \end{cases} \end{cases}$$

où les A_{ij} et B_{ij} représentent respectivement des SEF subjectif et objectif associés à chaque expert i dans une situation j donnée. Pour chaque règle une relation R_{ij} peut être dégagée et la relation R correspond alors à la concaténation de ces règles.

Dans la suite on gardera le formalisme :

"Si X est TYPS alors Y est TYPO"

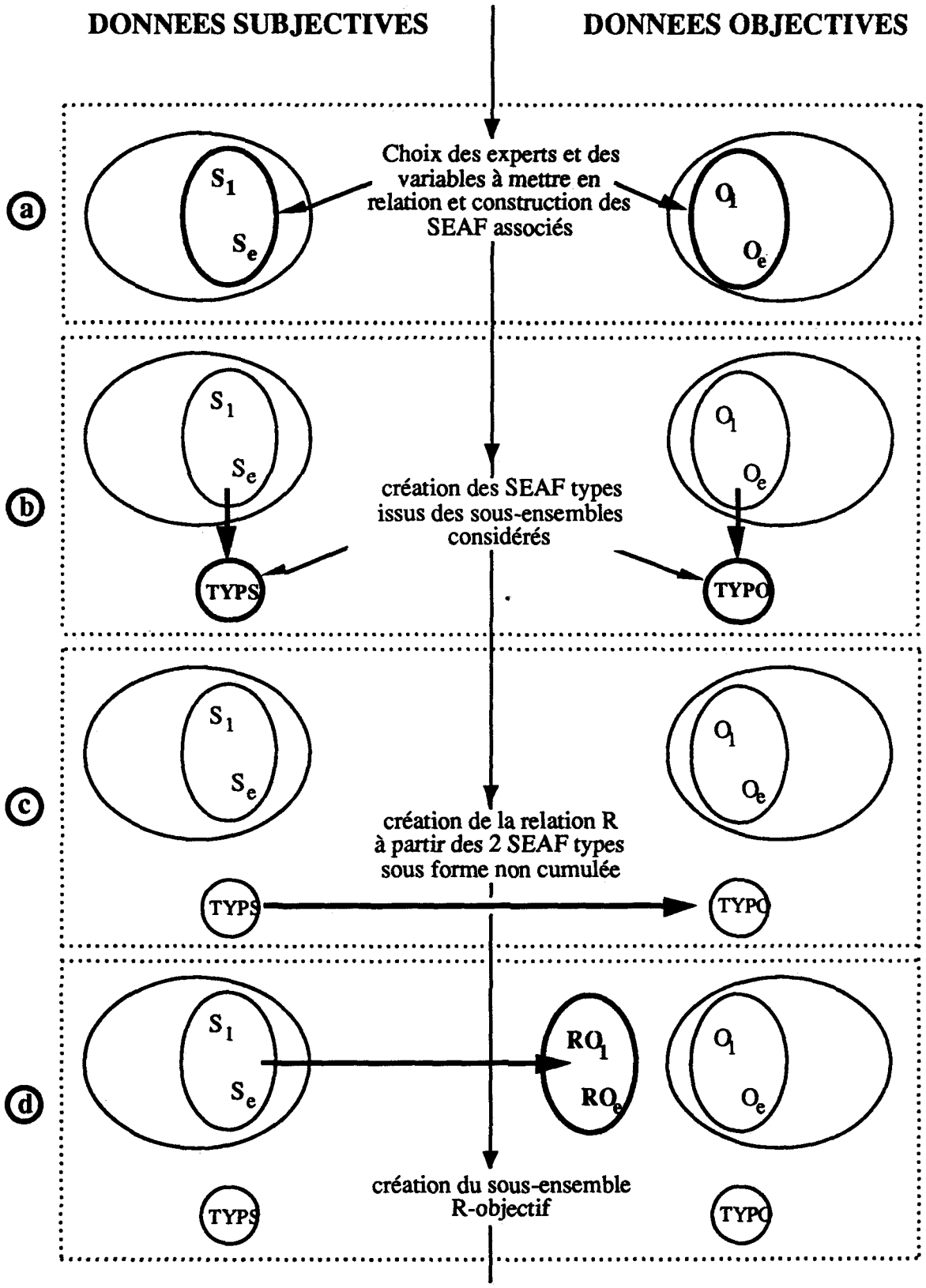
Cette relation R peut alors se représenter par une distribution de possibilité conditionnelle $\Pi_{y/x}$ entre Su et Ob composée de s x o fonctions de répartitions. Le choix de $\Pi_{y/x}$ doit se faire en liaison avec l'opérateur de composition Λ.

Une telle création est TOUJOURS possible, il s'agit maintenant de savoir si le couple utilisé convient et dans ce cas si la relation créée a réellement un sens.

b. Création d'un ensemble R-objectif ou R-subjectif

Dans toute la suite, la relation R est supposée créée en modélisant "si X est TYPS alors Y est TYPO" c'est à dire en considérant l'ensemble des données subjectives comme ensemble de départ.

Cette étape consiste à créer un ensemble particulier appelé R-objectif à partir des SEAF S_i ($1 \leq i \leq e$) construits sur les variables Su, et du couple ($\Pi_{y/x}$,Λ).



a, b, c : création de la relation R

d : création du sous-ensemble R-objectif

Figure III.1 : Les deux premières étapes de la mise en relation

Remarque : si la relation est créée en modélisant "si X est TYPO alors Y est TYPS" le nom donné à l'ensemble créé est alors R-subjectif.

La création de l'ensemble R-objectif se fait en utilisant le modus ponens généralisé :

si X est TYPS alors Y est TYPO
X est S_i

Y est RO_i

Le résultat est alors e SEAF, notés RO_i composés de o fonctions de répartitions. Cet ensemble représente un ensemble fictif de données objectives issu des impressions subjectives.

Ces deux premières étapes de la mise en relation sont présentées figure III.1.

c. Vérification de la relation R

Cette troisième étape consiste à vérifier si le couple $(\Pi_{y/x},\Lambda)$ utilisé ne donne pas de résultats contra-intuitifs par rapport aux données traitées. Dans ce but, il est nécessaire dans un premier temps, de calculer les valeurs des dérives par modus ponens et modus tollens. Ce problème est bien entendu dépendant des couples $(\Pi_{y/x},\Lambda)$ utilisés et des données prises en compte. En conséquence il est traité dans l'exemple du III.2.2 suivant la mise en oeuvre de la méthode.

Si ces deux critères sont nécessaires pour la vérification du couple $(\Pi_{y/x},\Lambda)$ utilisé, ils ne sont pas suffisants pour permettre de conclure que la relation R est adaptée aux données à traiter. Un autre critère est alors mis en oeuvre se basant sur l'utilisation du modus ponens généralisé :

si X est TYPS alors Y est TYPO
X est S_i

Y est RO_i

Cette formulation permet de penser que :

si S_i est "proche" de TYPS alors RO_i est "proche" de TYPO

mais surtout que :

si S_j est "plus loin" de TYPS que S_i alors RO_j est "plus loin" de TYPO que RO_i

La vérification de la relation R est alors liée à cette constatation. Le nouveau critère mis en oeuvre nécessite le calcul, à l'aide de la distance d définie chapitre II, de $d(TYPS,S_i)$ et $d(TYPO,RO_i)$ ($\forall i \in \{1,\ldots,e\}$). En considérant alors la fonction V :

$$\forall i \in \{1,\ldots,e\} \quad V\left(d(TYPO,RO_i)\right) = d(TYPS,S_i)$$

le critère d'une "bonne" relation est :

Critère I : la fonction V est croissante

ou $\forall i,j\ (i \neq j) \in \{1,\ldots,e\}\ d(TYPO,RO_i) \geq d(TYPO,RO_j) \Leftrightarrow d(TYPS,S_i) \geq d(TYPS,S_j)$

En considérant chaque expert comme un point de coordonnées $(d(TYPO,RO_i),d(TYPS,S_i))$ et en projetant tous ces points dans un plan, le critère énoncé peut alors s'exprimer par :

> La relation R est vérifiée si tous les points du plan sont situés de façon croissante suivant les deux axes.

Un exemple d'un tel plan est donné figure III.2.

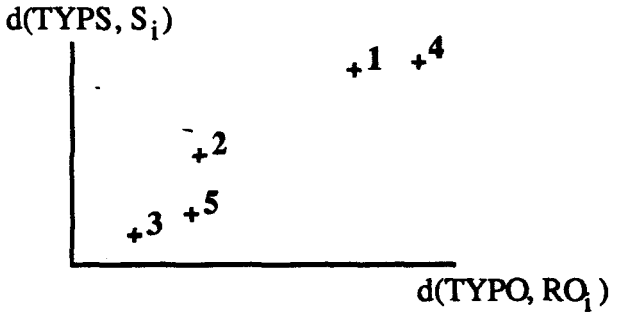


Figure III.2 : Exemple de plan de vérification de la relation R

d. Validation de la relation

Cette dernière étape, figure III.4, a pour but de déterminer s'il y a adéquation entre les deux groupes de données ou non. Elle ne peut être mise en oeuvre que SI la relation est VERIFIEE, c'est à dire que le couple $(\Pi_{y/x},\Lambda)$ est adapté aux données à traiter. Remarquons enfin que cette étape ne fait intervenir que les variables objectives, les variables subjectives étant présentes au travers du sous-ensemble R-objectif, figure III.5.

Il s'agit alors de trouver des critères qui permettent de valider la relation. Deux critères sont mis en oeuvre qui se basent sur la constatation suivante :

s'il existe une relation stable entre les deux groupes de données, l'intérêt de la relation est de pouvoir déduire pour un nouvel expert e+1, à partir uniquement de son avis S_{e+1}, un résultat sur l'ensemble des données objectives Ob.

En supposant alors que S_{e+1} corresponde exactement à un des avis d'expert S_i, et en appliquant la relation R à S_{e+1}, un SEAF RO_{e+1} sera obtenu. Le minimum que l'on puisse attendre de la relation est alors que RO_{e+1} soit le plus proche de O_i. Partant de cette constatation le critère d'une "bonne" relation est donc d'avoir O_i et RO_i les plus proches possibles, figure III.4.

Cette constatation permet de proposer les deux critères suivants :

Critère II :

$\forall i \in \{1,\ldots,e\} \quad |d(TYPO,O_i) - d(TYPO,RO_i)| < \varepsilon \quad \varepsilon$ étant le plus petit possible

Critère III :

$\forall i \in \{1,\ldots,e\} \quad d(O_i,RO_i) < \varepsilon' \qquad \varepsilon'$ étant le plus petit possible

Ces deux critères peuvent alors être présentés graphiquement. En effet, en considérant chaque expert comme un point de coordonnées $(d(TYPO,O_i),d(TYPO,RO_i))$ et comme le centre d'un cercle de diamètre $d(O_i,RO_i)$, la représentation graphique des points et de leur cercle associé dans un plan permet de dégager le critère graphique suivant :

Le critère visuel d'une "bonne" relation est alors d'avoir les points distribués les plus près de la bissectrice avec des cercles de diamètre les plus petits possibles.

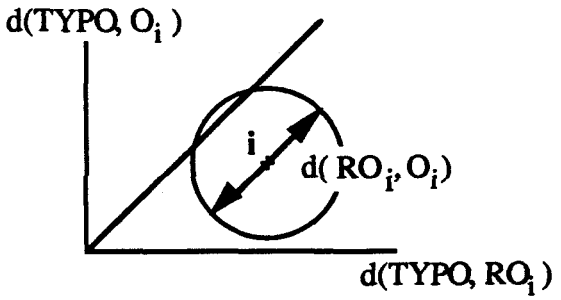


Figure III.3 : Plan de validation de la relation R

Ces deux dernières étapes de vérification et de validation de la mise en relation entre deux groupes de données sont résumées figure III.4.

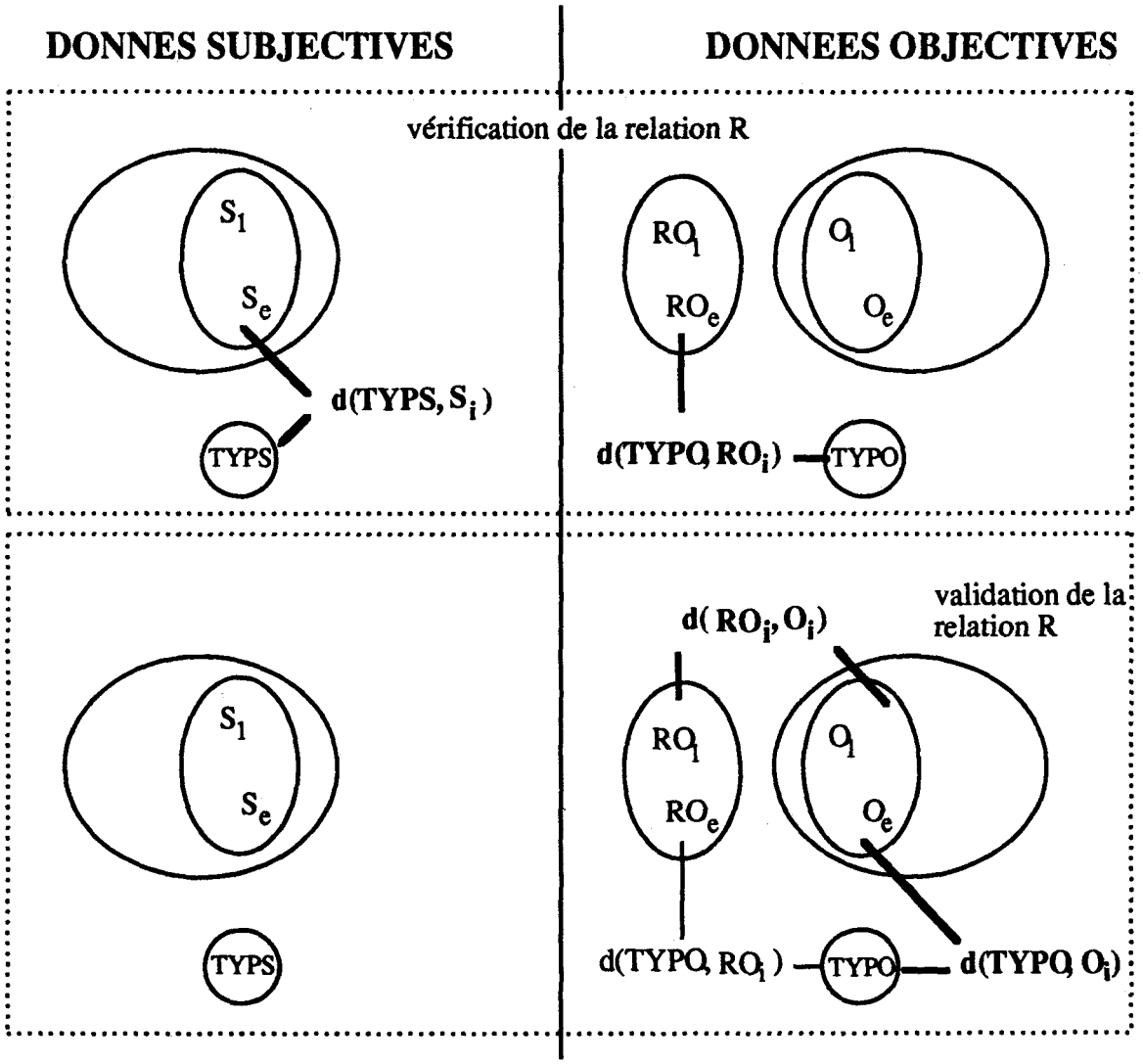


Figure III.4 : Etapes de vérification et de validation de la relation R

Enfin, la méthodologie de la mise en relation entre deux ensembles est présentée figure III.5 faisant apparaître les différents retours nécessaires. Si la relation n'est pas vérifiée, il est nécessaire de tester d'autres couples ($\Pi_{y/x}$,Λ) ou de choisir d'autres sous-ensembles Ex, Su et Ob pour créer la relation. Quand la relation est vérifiée et qu'il est donc possible de passer à la validation de la relation R, les mêmes retours sont nécessaires en cas d'échec.

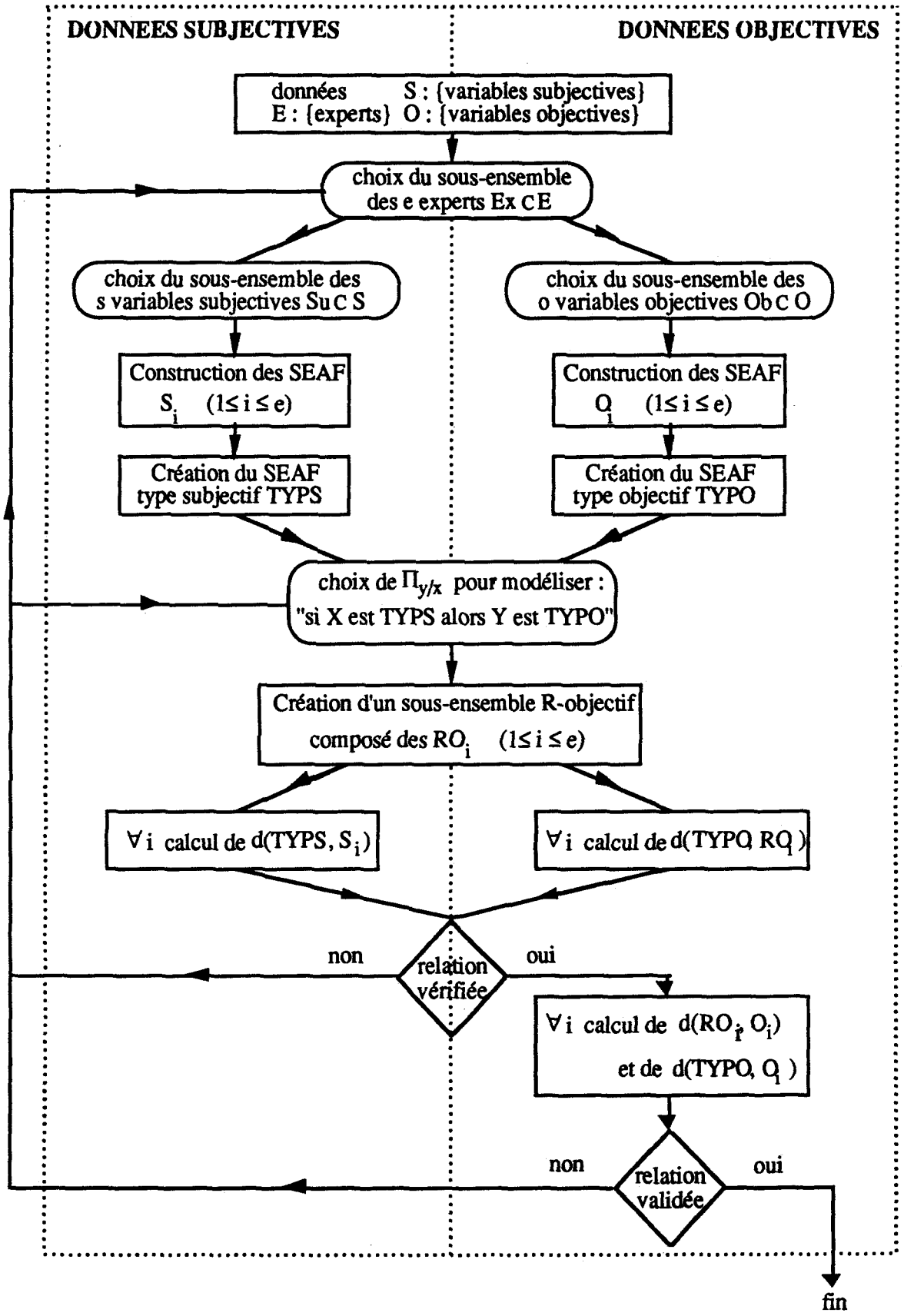


Figure III.5 : Méthodologie de mise en relation de deux groupes de données

III.2-2 Présentation d'un exemple et problèmes liés à la méthode

Comme il a été signalé précédemment, l'exemple présenté n'a pour but que de préciser toutes les étapes de la démarche de mise en relation entre deux groupes de données. Les divers problèmes liés à la méthodologie sont également discutés au travers de cet exemple. Le lecteur trouvera aussi dans l'annexe III un deuxième exemple traité.

a. Prise en compte des données et choix des couples

Soient les deux ensembles notés respectivement SUB et OBJ, construits respectivement sur 8 variables qui seront appelées "subjectives" et 6 variables "objectives".

Ex \ Su	1	2	3	4	5	6	7	8
1	1	0	1	1	0	1	0	0,7
2	1	0	0,9	0,9	0	0,9	0,1	0,7
3	1	0	0,9	0,5	0	0,3	0,6	0,5
4	1	0,1	0,9	0,5	0	0,4	0,6	0,6
5	1	0,1	0,8	0,4	0	0,4	0,5	0,7
6	1	0,1	0,9	0,4	0	0,5	0,5	0,7
7	1	0,1	0,9	0,4	0	0,6	0,4	0,7
8	1	0,1	0,9	0,6	0	0,7	0,3	0,8
9	1	0	0,9	0,5	0	0,5	0,4	0,6
10	0,8	0	1	1	0,2	1	0	1
11	1	0	0,9	0,5	0	0,4	0,5	0,6
12	0,4	0,4	0,2	0,2	0,4	0,2	0,2	0,1
13	0,6	0,6	0,5	0,5	0,4	0,4	0,6	0,2
14	0,3	0,4	0,2	0,3	0,3	0,3	0,1	0,1
15	1	0,1	0,2	0,9	0,9	0,1	0,9	0,9

Ex \ Ob	1	2	3	4	5	6
1	1	1	0,3	1	1	1
2	0,9	0,9	0,3	1	1	1
3	0,4	0,4	0,5	1	1	1
4	0,4	0,5	0,5	0,9	1	1
5	0,5	0,3	0,3	0,9	1	1
6	0,5	0,4	0,4	0,9	1	1
7	0,6	0,4	0,3	0,9	1	1
8	0,7	0,6	0,2	0,9	1	1
9	0,6	0,5	0,4	1	1	1
10	1	1	0,2	1	0,9	0,8
11	0,5	0,5	0,4	1	1	1
12	0,4	0,2	0,1	0,2	0,4	0,4
13	0,4	0,5	0,8	0,5	0,6	0,6
14	1	0,2	0,9	0,8	1	0,8
15	0,1	0,1	0,1	0,1	0,9	0,2

La recherche d'une ou de plusieurs relations stables entre ces deux ensembles de données est effectuée globalement sur 15 experts (e = 15). En considérant l'ensemble des 8 variables "subjectives" (s = 8) et des 6 "objectives" (o = 6), il vient donc :

$Ex = \{ i, 1 \leq i \leq 15 \}$ $Su = \{ j, 1 \leq j \leq 8 \}$ $Ob = \{ k, 1 \leq k \leq 6 \}$

Il convient de choisir des distributions de possibilité conditionnelles $\Pi_{y/x}$, en liaison avec un opérateur Λ permettant la modélisation de la méta-implication :

"si X est le SEAF type subjectif alors Y est le SEAF type objectif"

Les couples permettant cette modélisation sont nombreux, nous en avons retenus 19 issus principalement de /MIZUMOTO 82,85/ /MOREAU 87/. Sur ces 19 couples, seuls 6 couples représentatifs des 19 sont étudiés dans ce paragraphe. Les résultats obtenus pour les 19 couples sont présentés dans l'annexe III ainsi qu'un deuxième exemple. Soient les 6 couples suivants :

n° des couples	$\Pi_{y/x}$	opérateur Λ
1	$\min [1,1 - \Pi_x(x) + \Pi_y(y)]$	$\max (0,a + b - 1)$
2	$\begin{cases} 1 \quad \text{si } \Pi_x(x) \leq \Pi_y(y) \\ \Pi_y(y) \text{ sinon} \end{cases}$	$\min [a , b]$
3	$\begin{cases} 1 \quad \text{si } \Pi_y(y) = 1 \\ \min [1 , \frac{1 - \Pi_x(x)}{1 - \Pi_y(y)}] \text{ sinon} \end{cases}$	$a.b$
4	$\min [\Pi_x(x),\Pi_y(y)]$	$\min [a , b]$
5	$\min [(\Pi_x(x) \, S \, \Pi_y(y)) , (1 - \Pi_x(x)) \, G \, (1 - \Pi_y(y))]$	$\min [a , b]$
6	$\begin{cases} 1 \quad \text{si } \Pi_y(y) = 1 \\ \min [1 , \frac{1 - \Pi_x(x)}{1 - \Pi_y(y)}] \text{ sinon} \end{cases}$	$\begin{cases} 0 \quad \text{si } b = 0 \\ \max [0, \frac{a+b-1}{b}] \text{ sinon} \end{cases}$

En fonction du couple utilisé un sous-ensemble R-objectif est alors créé. Il s'agit de rechercher, dans un premier temps, si le couple choisi est adapté aux données à traiter, et dans un deuxième temps, déterminer s'il existe effectivement une relation stable entre les deux groupes de données.

b. Vérification et validation de la relation

Les problèmes liés à cette mise en relation sont les choix des différents paramètres qui permettent d'accepter ou de refuser la validation de la relation. Ces paramètres sont au nombre de 4 :

- dérive par modus ponens,
- dérive par modus tollens,
- ε et ε' seuils au dessous desquels les distances seront jugées acceptables et donc la relation validée selon la procédure présentée précédemment.

La première vérification des couples se situe au niveau du calcul des deux dérives. Si une dérive nulle est, bien entendu, une très bonne valeur, ou une dérive de 0.5 une très mauvaise, la connaissance d'une dérive correcte est basée sur les données et l'expérimentation.

Sur l'exemple traité, ces deux dérives ont été calculées pour les 6 couples choisis et les résultats sont présentés figure III.6.

couple	1	2	3	4	5	6
modus ponens	0,081	0,063	0,081	0,038	0,152	0,097
modus tollens	0,181	0,187	0,123	0,326	0,169	0,172

Figure III.6 : dérives par modus ponens et modus tollens calculées sur les 6 couples

Le tableau figure III.6 révèle la différence des résultats selon les couples utilisés. Les valeurs des dérives par modus ponens étant en dessous de 0,1 à l'exception du couple 5 et les valeurs des dérives par modus tollens étant similaires sauf pour le couple 4. Néanmoins, avant de prendre en compte ces valeurs, il faut déterminer si des experts ne doivent pas être éliminés de l'analyse au vu des critères mis en oeuvre pour vérifier et valider la relation. De tels experts peuvant créer des perturbations importantes et ainsi fausser l'interprètation des valeurs obtenues.

Pour ce faire, le critère graphique explicité au paragraphe III.2.1 permettant la vérification de la relation a été utilisé. Rappelons que ce critère graphique s'exprime par :

> la relation R est vérifiée si tous les points du plan sont situés de façon croissante suivant les deux axes.

L'utilisation de ce critère permet de montrer qu'il y a trois tendances principales pour les 6 couples. On retrouve le même comportement pour les couples 1, 2, 3 et 4, figure III.7a, un comportement différent mais proche du précédent pour le couple 5, figure III.7b et un comportement tout à fait différent pour le couple 6, figure III.8a.

Cette première étape de vérification fait ressortir que pour tous les couples hormis le 6, l'expert 15 présente une inversion importante et ne permet pas au critère I d'être vérifié. Pour ces couples il est alors nécessaire d'éliminer cet expert avant de passer à l'étape de validation de la relation.

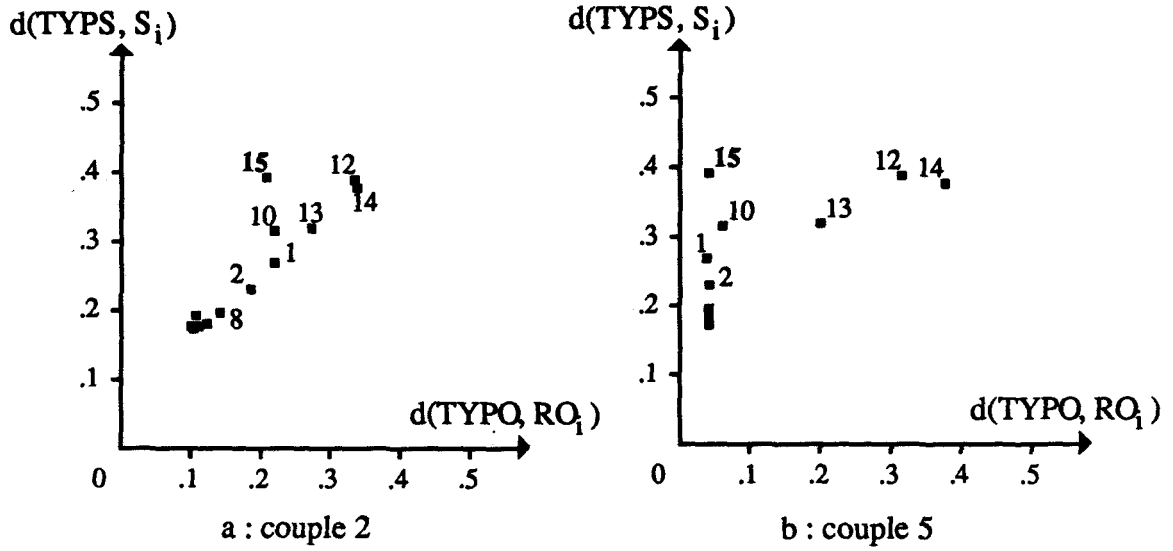


Figure III.7 : Plans de vérification

En ce qui concerne le couple 6, le critère I étant vérifié, le deuxième critère graphique regroupant les critères II et III est mis en oeuvre figure III.8b. Ce critère graphique, rappelons le s'exprime par :

> la relation est validée si les points sont distribués les plus près de la bissectrice et si les cercles ont de petits diamètres.

Ce critère étant loin d'être vérifié, le couple 6 ne peut convenir pour la mise en relation des deux ensembles. Il est exclu du reste de l'analyse.

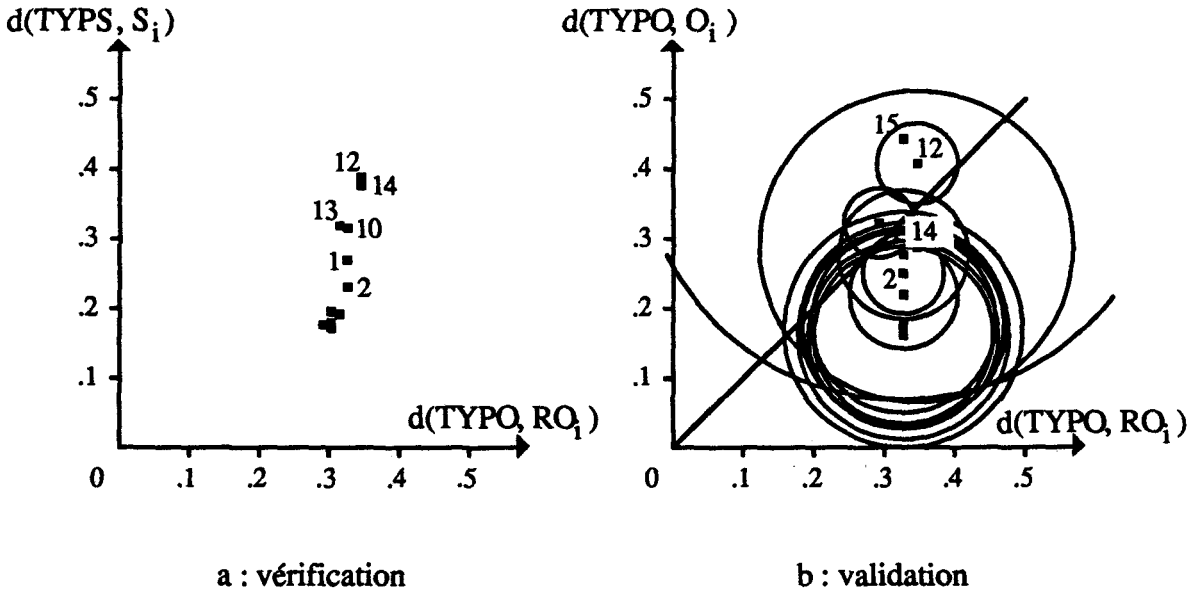


Figure III.8 : Plans de vérification et validation présentés pour le couple 6

Pour les couples 1 à 5 la relation est redéterminée à partir d'un ensemble d'experts Ex1 = { i, $1 \leq i \leq 14$ } = Ex - {15} toujours sur les ensembles de variables Su et Ob. Le critère graphique est à nouveau utilisé et la relation est alors vérifiée sur les 14 experts pris en compte, figure III.9a.

Pour déterminer s'il existe une relation stable entre les deux groupes de données réduits, les critères II et III doivent être vérifiés, le problème étant, en fonction de ε et ε' de déterminer la frontière entre les experts à conserver ou à rejeter. On sait qu'une distance $d(O_i,RO_i)$ nulle est idéale et une distance de 0.5 est à rejeter et, avant de déterminer si la relation est validée, il s'agit d'abord de regarder si un ou des experts ne présentent pas de comportements inadéquats, c'est à dire une distance $d(O_i,RO_i)$ trop grande - des cercles trop importants - et/ou une valeur $|d(TYPO,O_i) - d(TYPO,RO_i)|$ trop élevée - des points trop éloignés de la bissectrice.

A l'aide de l'ensemble d'experts Ex1 et pour tous les couples il apparait que l'expert 14 ne vérifie pas les deux critères, figure III.9b, il présente une distance $d(O_i,RO_i)$ trop importante (> 0.4), il convient donc de l'éliminer de l'analyse et de recommencer celle-ci sur un ensemble d'experts plus réduit : Ex2 = { i, $1 \leq i \leq 13$ } = Ex1 - {14}.

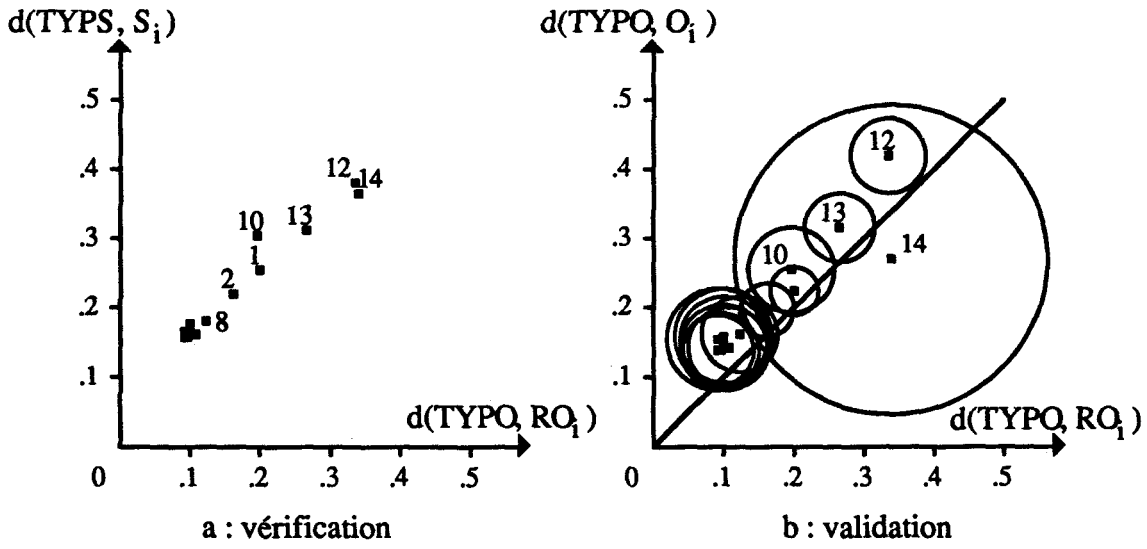


Figure III.9 : Plans de vérification et validation présentés pour le couple 2 sur Ex1

L'expert 14 étant éliminé de l'analyse, le critère I reste vérifié sur Ex2 pour les 5 couples restant. Les plans de validation donnent alors des résultats différents suivant les couples. Les couples 1, 2 et 3 permettent de valider la relation, figure III.10a de façon sensiblement identique. Le couple 4 quant à lui ne le permet pas à cause de l'expert 10, figure III.10b. Enfin le plan de validation associé au couple 5 présente des cercles trop

grands, c'est à dire des distances $d(O_i,RO_i)$ trop importantes (de l'ordre de 0,2) pour permettre de valider une relation.

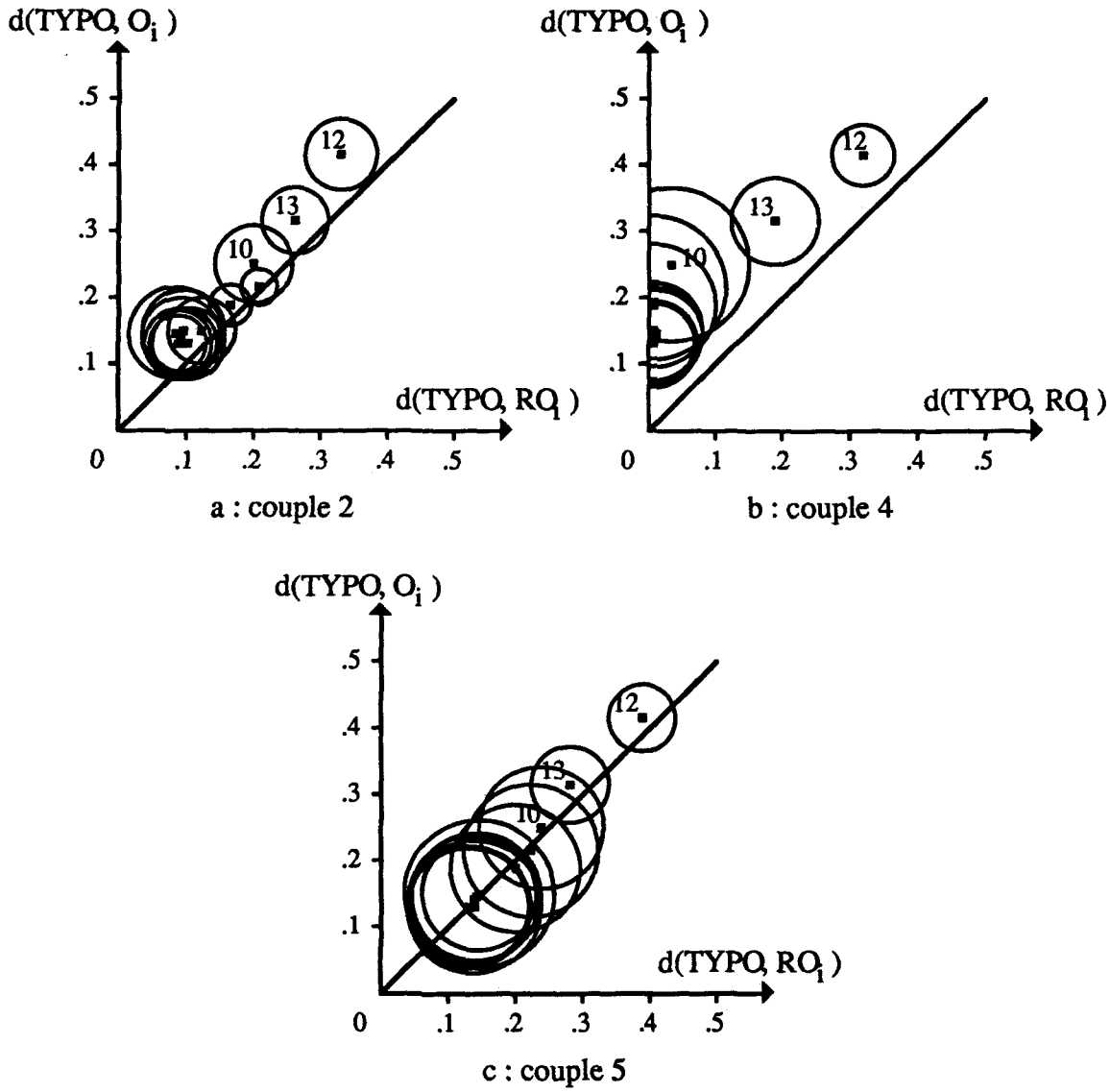


Figure III.10 : Plans de validation déterminés sur Ex2

Avant de conclure, un tableau résumant les différents résultats est présenté figure III.11. Les valeurs obtenues sur Ex pour les deux derniers paramètres d1 et d2, représentent le maximum de ces paramètres pour les experts qui seront conservés dans l'analyse c'est à dire les 13 premiers. Ceci étant réalisé afin d'avoir une base de comparaison entre Ex et Ex2. Par exemple, pour le couple 5 la valeur de $d(O_i,RO_i)$ maximum est supérieure à 0.5 sur Ex, c'est à dire pour les 15 experts, mais les experts 14 et 15 étant éliminés dans la suite de l'analyse, cette distance maximum, prise sur les 13 premiers experts, est de 0,29.

couple	1		2		3		4		5		6	
ensemble	Ex	Ex2	Ex	Ex2	Ex	Ex2	Ex	Ex2	Ex	Ex2	Ex	Ex2
d_{mp}	0,08	0,07	0,06	0,05	0,08	0,06	0,04	0,01	0,15	0,12	0,10	0,09
d_{mt}	0,18	0,20	0,19	0,20	0,12	0,13	0,33	0,35	0,17	0,17	0,17	0,20
d1	0,21	0,20	0,15	0,14	0,20	0,17	0,28	0,23	0,29	0,23	0,34	0,34
d2	0,06	0,08	0,07	0,08	0,06	0,08	0,21	0,22	0,05	0,04	0,16	0,17

d_{mp} : dérive par modus ponens d_{mt} : dérive par modus tollens

$$d1 = \max_{Ex2} d(O_i,RO_i) \qquad d2 = \max_{Ex2} \left| d(TYPO,O_i) - d(TYPO,RO_i) \right|$$

Figure III.11 : Tableau récapitulatif des 4 paramètres utilisés

c. Résultats

En premier lieu, il est à noter que les valeurs des 4 paramètres de la figure III.11 restent sensiblement égales lorsque les experts présentant des comportements particuliers sont éliminés de l'analyse. Les paramètres 1 et 3 baissent légérement, 2 et 4 augmentent légèrement.

Au vu de ces valeurs et des plans de vérification et de validation la relation est alors considérée comme validée à l'aide des 3 premiers couples sur les 13 premiers experts.

Les autres couples mis en oeuvre ne permettent pas de valider la relation entre les deux ensembles, le couple le moins adapté semble être le couple 6.

En conclusion, les couples 1, 2, 3 donnent des résultats corrects, et le choix d'un couple pour mettre en relation les deux ensembles se porte sur les couples 2 et 3 ayant des valeurs plus faibles au niveau de d1, figure III.11, que le couple 1.

Enfin, des résultats peuvent également être fournis au niveau des experts. En considérant les couples 2 et 3 comme les plus adaptés à la mise en relation, la figure III.7a montre que l'expert 15 présente une inversion importante, ce qui signifie qu'il a un comportement particulier par rapport aux autres. De même, mais d'une manière différente, en se référant à la figure III.9b, il se dégage que l'expert 14 présente un comportement "incohérent" : il présente en fait de faibles valeurs pour les variables subjectives et des grandes pour celles objectives. Enfin, remarquons que l'expert 12 est situé loin des autres mais de façon sensiblement identique pour les deux groupes de données ce qui lui permet de rester dans l'analyse.

III.3 - CONCLUSION

La méthode de mise en relation de deux groupes de données basée sur les SEAF est dépendante d'un certain nombre de choix qui ne sont pas aisés, choix des couples et des quatre paramètres. Mais, comme le souligne /VOLLE 81/ *"le dernier mot de l'interprétation appartient au flair du praticien, à sa culture aussi qui lui permettra de trouver une expression littéraire des résultats permettant leur communication"*. C'est pourquoi le choix de la limite acceptable des différents paramètres doit se faire en se basant sur l'expérimentation mais aussi en "regardant" les données, les valeurs de ces limites ne pouvant être fixées "une fois pour toute".

D'autre part, dans le cas de grands tableaux, quelques centaines de variables et d'experts, le choix du sous-ensemble d'experts à mettre en relation peut être déterminé par une première classification sur chacun des ensembles, mais il est impossible dans l'état actuel de déterminer précisément les variables à mettre en relation. Une solution pourrait être d'utiliser une méthode classique d'analyse des données, analyse factorielle des correspondances par exemple, préalablement à la méthode, pour permettre de dégager les variables les plus discrimantes par exemple.

Après avoir présenté la méthode de mise en relation de deux ensembles ainsi qu'un exemple sur des données fictives, le chapitre suivant propose d'appliquer cette méthode et les méthodes de traitement d'impressions subjectives présentées chapitre II, à des données réelles issues d'une étude ergonomique du poste de travail bureautique.

CHAPITRE IV

ANALYSE DE DONNEES PROVENANT D'UNE EVALUATION OBJECTIVO-SUBJECTIVE D'UN POSTE DE TRAVAIL BUREAUTIQUE

Ce quatrième chapitre propose une application des méthodes décrites aux chapitres précédents à l'analyse de données issues d'une étude ergonomique d'un poste de travail bureautique. Au cours de cette étude, effectuée en laboratoire, des données objectives et subjectives sont recueillies. Le grand nombre de variables utilisées nécessite de faire appel à des méthodes multidimensionnelles pour leur traitement.

En premier lieu, les données subjectives étant recueillies à l'aide de questionnaires utilisant des différenciateurs sémantiques continus, les méthodes proposées au chapitre II sont appliquées pour leur traitement, de la classification à la détermination d'une configuration optimale de réglage du poste de travail.

En second lieu, disposant également d'un ensemble de données objectives recueillies par capteurs, la méthode de mise en relation entre deux ensembles de données élaborée au chapitre précédent est utilisée pour déterminer l'existence de relations entre les deux groupes de données.

IV.1 PRESENTATION DE L'ETUDE EXPERIMENTALE

IV.1-1 Le protocole expérimental

L'objectif de cette étude expérimentale consiste à évaluer l'influence de trois facteurs importants du poste de travail bureautique concernant le mobilier /LOSLEVER 88a/ :

- facteur assise : hauteur du plan de l'assise par rapport au sol
- facteur table : éloignement de la table par rapport à l'assise
- facteur dossier : éloignement du dossier par rapport au clavier

La tâche demandée à l'opérateur consiste à recopier un texte, présenté sur un support papier, en utilisant une machine à écrire électrique. La population expérimentale est constituée de 12 sujets.

Pour évaluer l'influence de ces 3 facteurs, un ensemble de variables objectives sont prises en compte au niveau du rachis, et un ensemble de variables subjectives sont recueillies par l'intermédiaire d'un questionnaire.

Le schéma présenté sur la figure IV.1 résume les principaux éléments du protocole expérimental

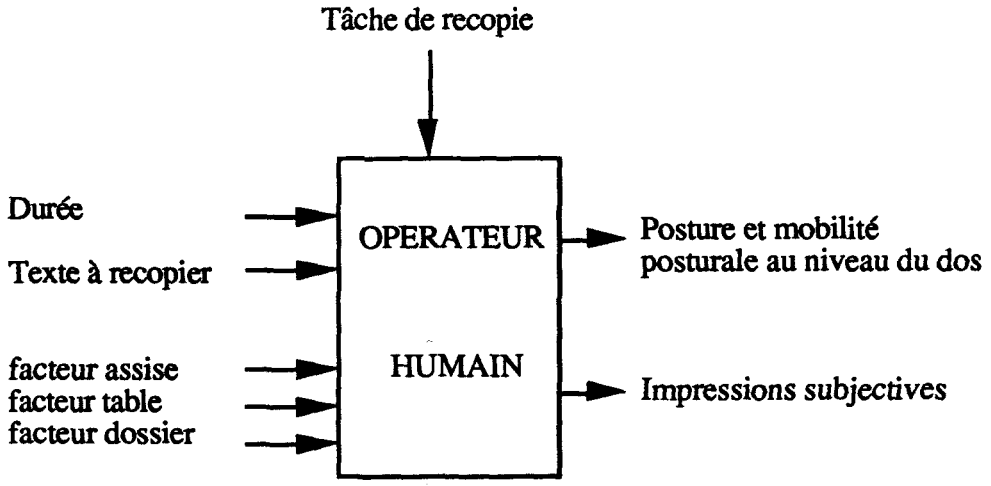


Figure IV.1 : présentation schématique du protocole et du dispositif expérimental

Chacun des 3 facteurs étudiés est caractérisé par 3 modalités de réglage, deux extrêmes et une intermédiaire, qui ont été déterminées par des manipulations préalables suivant des critères ergonomiques /LEPOUTRE et ROGER 82/. Le plan d'expérience retenu est alors le suivant :

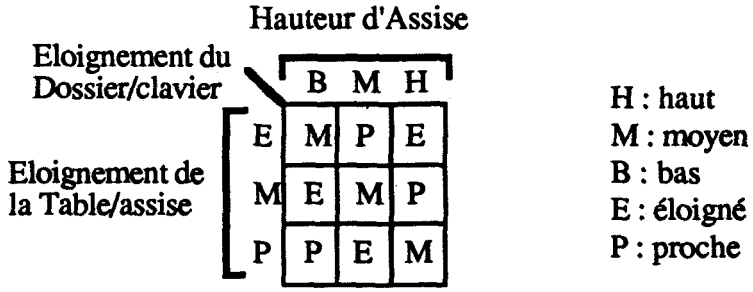
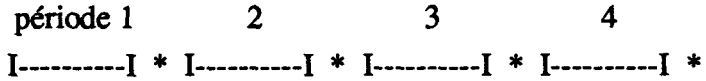


Figure IV.2 : plan d'expérience

Chaque case définit une expérience. Par exemple, la case en bas à gauche correspond à la combinaison "BPP" c'est à dire : l'assise Basse, la table Proche de l'assise et le dossier Proche du clavier.

Selon les auteurs de l'expérience, celle-ci est divisée en 4 périodes d'environ 34 minutes de frappe effective entre lesquelles sont introduites des pauses de quelques minutes /LEPOUTRE et ROGER 82/. Durant ces pauses, le sujet répond à un questionnaire. L'organisation temporelle d'une expérience est présentée figure IV.3.



* : temps de pause durant lequel le sujet répond au questionnaire
I---I : le sujet recopie un texte

Figure IV.3 : Organisation temporelle d'une expérience

IV.1-2 Ensemble des mesures expérimentales

L'atome expérimental de base est alors la période. A partir de celle-ci, l'ensemble O des observations est construit en considérant toutes les combinaisons des triplets (s,e,p) étudiées au cours de la phase expérimentale où :

- s désigne un sujet de la population observée S = {a,b,...,k,l}, Card(S) = 12,
- e est une expérience de l'ensemble E des différentes "cases" du plan d'expérience, card(E) = 9,
- p désigne une période expérimentale de P, card(P) = 4.

Le nombre total de combinaisons réalisées est :

$$card(O) = card (S \times E \times P) = 432$$

Chaque observation (s,e,p) de O est caractérisée par un ensemble V de variables de nature objective et subjective.

Variables objectives :

Pour les variables objectives, les informations prises en compte portent sur le comportement postural au niveau du rachis.

Concernant ce comportement postural, quatre courbures de dos et l'inclinaison de la base du tronc par rapport à la verticale sont enregistrées en continu par l'intermédiaire d'un inclino-courbomètre, figure IV.4 /LEPOUTRE 79/. Un tel dispositif permet des mesures relativement précises en statique et en dynamique et est peu gênant pour le sujet /GUERRA 85/ /LEPOUTRE 85/ /LEPOUTRE et GUERRA 86/. En plus de ces cinq mesures, la force d'appui exercée sur le dossier est également enregistrée. Les six voies de mesure sont :

- voie 1 : la courbure C1 au niveau de la première vertèbre dorsale,
- voie 2 : la courbure C2 au niveau de la sixième vertèbre dorsale,
- voie 3 : la courbure C3 au niveau de la dixième vertèbre dorsale,
- voie 4 : la courbure C4 au niveau de la troisième vertèbre lombaire,
- voie 5 : l'inclinaison de la base du dos par rapport à la verticale,
- voie 6 : la force d'appui sur le dossier.

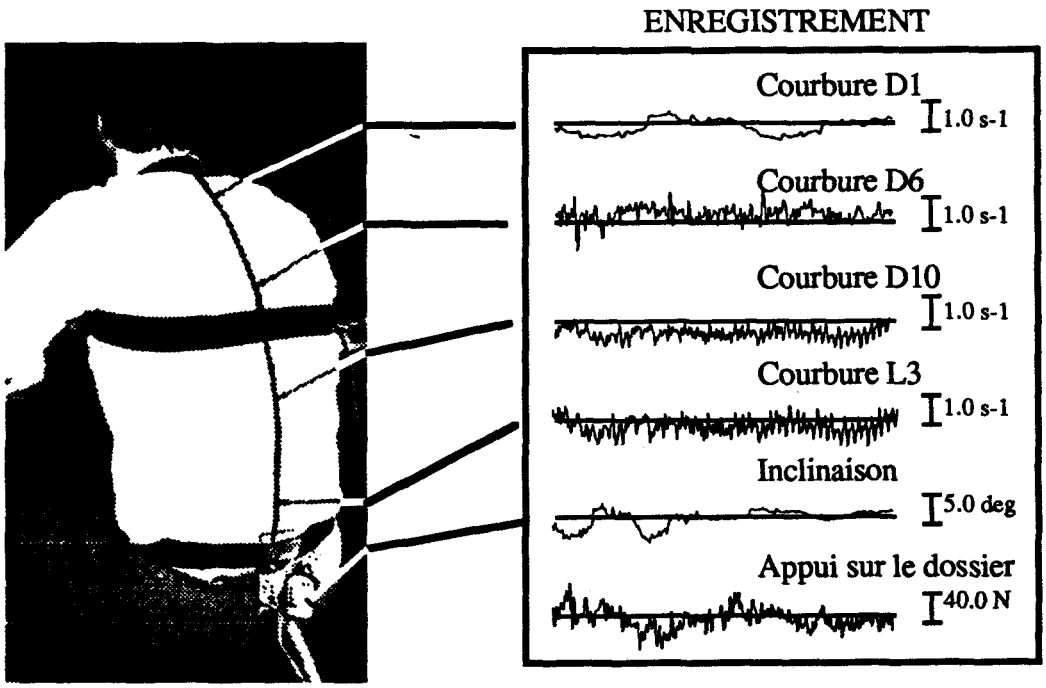


Figure IV.4 : Evaluation de la forme du dos par un inclino-courbomètre

Les informations enregistrées sur chacune des voies des capteurs au cours d'une période de travail sont synthétisées sous la forme d'une moyenne MOY, d'un écart-type ECT et d'un indice de variation IV de posture correspondant à la valeur efficace de la dérivée du signal. Les variables posturales sont résumées par :

VOBJ= { MOY^1, ECT^1, IV^1,..., MOY^6, ECT^6, IV^6 }, l'exposant indique le numéro de la voie.

Variables subjectives :

Les informations subjectives prises en compte portent sur les gênes perçues et sur les appréciations du poste de travail. Elles sont obtenues à partir d'un questionnaire comportant 17 questions - numérotées 1, 2, 3, 4a ... 4g, 5 ... 11 - dont les différenciateurs sémantiques sont continus, figure IV.6. Les réponses fournies rendent compte :

- de la satisfaction du sujet sur le plan général, questions 1, 2, 3 ;
- des gênes locales, questions 4a, 4b,..., 4f, 4g ;
- de l'intérêt porté à la tâche, questions 5, 6 ;
- de l'ambiance lumineuse et sonore, questions 7, 8 ;
- des réglages des 3 facteurs étudiés, questions 9, 10, 11.

Chaque variable subjective est obtenue à partir de la mesure de la position de la réponse sur le différenciateur sémantique.

L'ensemble VSUBJ contiendra les 17 variables :

VSUBJ = { 1, 2, 3, 4a, ..., 4f, 5,..., 11 }

En résumé la base de données générée à partir de O x (VSUBJ ∪ VOBJ) a la structure suivante :

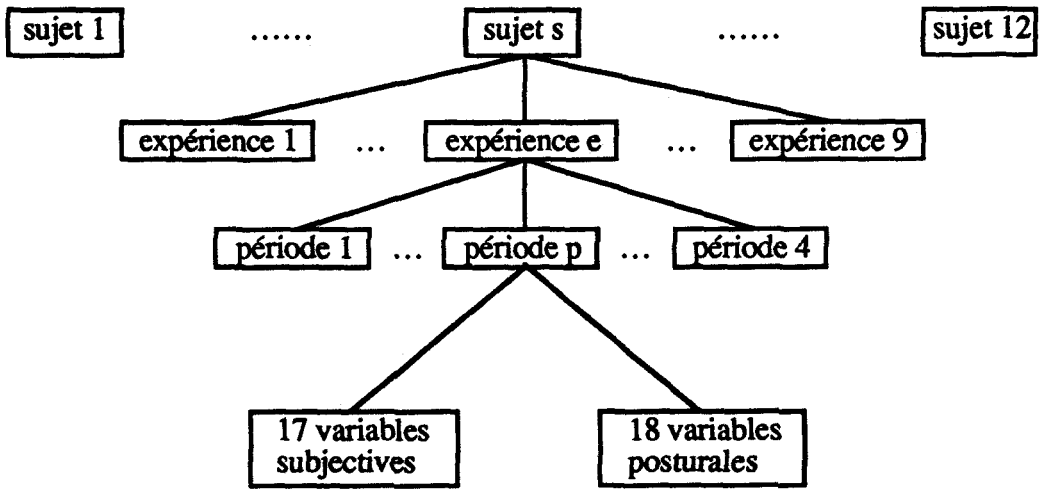


Figure IV.5 : Structure de données

		beaucoup	pas du tout
1.	Avez-vous envie de bouger	⊢———	———⊣
2.	Avez-vous mal à la tête	⊢———	———⊣
3.	Etes-vous fatiguée	⊢———	———⊣
4	Avez-vous mal aux jambes	⊢———	———⊣
	aux mains	⊢———	———⊣
	aux bras	⊢———	———⊣
	aux épaules	⊢———	———⊣
	aux yeux	⊢———	———⊣
	à la nuque	⊢———	———⊣
	au dos	⊢———	———⊣
5.	Avez-vous des difficultés à concentrer votre attention	⊢———	———⊣
6.	Trouvez-vous le texte intéressant	très ⊢———	———⊣ pas du tout
7.	Trouvez-vous l'ambiance sonore	beaucoup trop forte ⊢———	———⊣ beaucoup trop faible
8.	Trouvez-vous l'ambiance lumineuse	⊢———	———⊣
9.	Comment vous sentez-vous assise	beaucoup trop haut ⊢———	———⊣ beaucoup trop bas
10.	Comment trouvez-vous la table	⊢———	———⊣
11.	Comment trouvez-vous le dossier	beaucoup trop loin ⊢———	———⊣ beaucoup trop près

12. Autres remarques : _____

_____

Figure IV.6 : questionnaire utilisé, le sujet répond en plaçant une croix sur le segment

La structure des données ayant été précisée, les objectifs de l'étude sont les suivants.

Analyser les données subjectives. Dans un premier temps, il faut dégager l'existence de classes d'appréciations subjectives pour les différents sujets. Dans une deuxième phase, il est nécessaire de classer les combinaisons de réglages les mieux et les moins bien perçues. Remarquons que la formulation du questionnaire permet de proposer une réponse optimale pour chaque question et par conséquent d'appliquer la méthode développée au chapitre II.

A l'opposé, concernant les données objectives, il n'est pas possible de fournir, a priori, un modèle de liaison entre un indice postural et le confort. Dans ces conditions, il est nécessaire de déterminer l'existence des relations stables entre les données subjectives et les données objectives, ceci permettant, éventuellement, de créer de tels modèles.

IV.2 ANALYSE DES DONNEES SUBJECTIVES

IV.2-1 Attitude des sujets face au différenciateur sémantique

Chaque sujet, face à une représentation continue de l'axe a un comportement voire même une "stratégie" de réponse qui lui est propre /LOSLEVER 88b/. L'un des problèmes posés est alors la valeur connotative associée aux expressions antinomiques liées au segment du différenciateur. De plus il est probable que certains sujets appréhendent de façons différentes l'axe continu. Certains sujets peuvent, en effet, utiliser le segment dans sa totalité et d'autres se cantonner à une partie seulement.

Concrètement, il s'agit de vérifier quand cela est possible, la cohérence des réponses avec certains faits connus de l'expérimentateur seul, réglages, intensité lumineuse et sonore...

Dans cette perspective, il est nécessaire d'étudier l'histogramme des réponses associées à chaque question. La fonction de densité a été construite en découpant l'axe en 7 modalités, car le fait d'utiliser un nombre impair permet de tenir compte de la symétrie de l'axe des différenciateurs sémantiques des 5 dernières questions. La première analyse s'intéresse à la façon dont l'ensemble des sujets "apprécie" les trois réglages d'un facteur donné. Par exemple pour la hauteur de table, question 10, les trois histogrammes correspondant aux trois positions différentes de la table proche, moyenne, éloignée, tous autres réglages confondus sont construits. A partir de chaque histogramme un modèle de la fonction de densité est proposé figure IV.7.

La disposition le long de l'axe et la symétrie par rapport au centre de cet axe est remarquable. En effet, les appréciations sont données "dans le sens des réglages" pour l'ensemble des sujets.

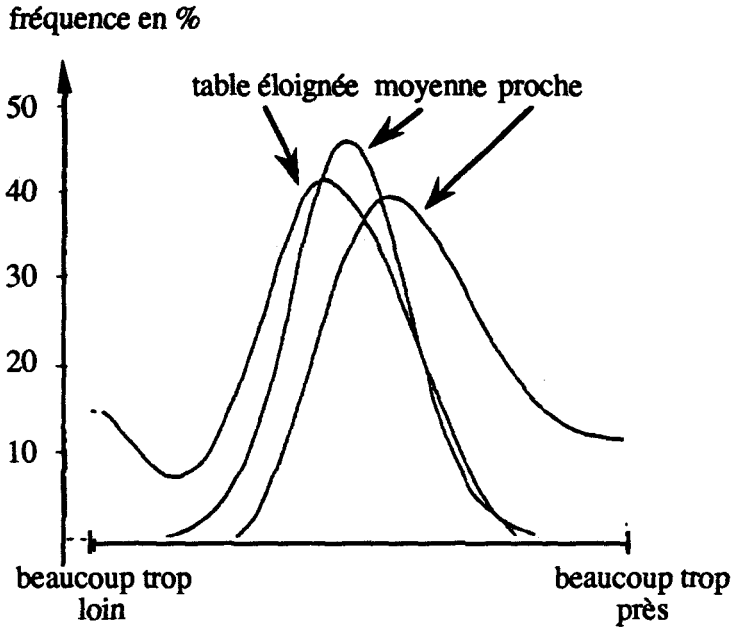


Figure IV.7 : Appréciations du facteur table par l'ensemble des sujets

Les histogrammes ont également été construits tous réglages confondus pour les 6 premières questions et leur allure générale est présentée figure IV.8. Il apparait que le champ d'utilisation de l'axe dialectique est relativement faible pour les rubriques de la question 4, hormis, pour 4g - douleur dans le dos. Concernant la question 6, intérêt porté au texte, les appréciations sont plus mitigées et, si certains sujets ont trouvé le texte absolument inintéressant, ce qui parait une réponse franche, un d'entre eux l'a trouvé très intéressant, ce qui est plutôt une réponse de "complaisance".

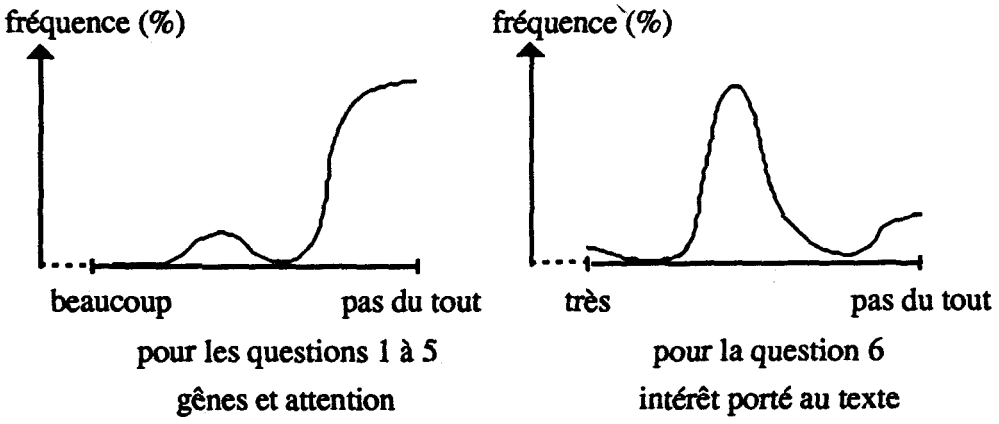


Figure IV.8 : Type d'histogramme pour les réponses 1 à 6

IV.2-2 Classification des sujets

L'utilisation des sous-ensembles aléatoires flous (SEAF) impose de ramener toutes les données entre 0 et 1. Le choix adopté est un codage linéaire, le zéro correspondant à la borne gauche du différenciateur, le un à la droite. D'autres codages peuvent être envisagés, min-max ou centrage réduction avec la moyenne et l'écart-type, mais, au vu des histogrammes construits sur les questions, il semble nécessaire de ne pas dilater artificiellement la plage de réponses, des sujets se "cantonant" dans une partie relativement restreinte de l'axe dialectique.

La réponse à un questionnaire est considéré comme un SEAF à 17 variables. Le nombre d'observations pris en compte pour chaque variable dépend alors de la combinaison des triplets (s,e,p) étudiée - s pour sujet, e pour expérience et p pour période. Par exemple, une classification sur les sujets pour une période et tous réglages confondus entraîne 9x1 observations par variable.

L'analyse a été effectuée à l'aide de hiérarchies, de pyramides et d'arbre à liaisons incomplètes. Il n'est pas possible de présenter l'ensemble des résultats fournis par l'analyse étant donné la combinatoire élevée des observations prises en compte. Néanmoins, les figures IV.9 à IV.12 présentent des exemples de résultats obtenus en fonction des paramètres étudiés qui sont détaillés ci-après.

L'analyse typologique des réponses des 12 sujets au questionnaire est illustrée figure IV.9. Chaque sujet est caractérisé par les 17 réponses fournies pour les 9 expériences et les 4 périodes. Les résultats issus de cet arbre sont principalement l'opposition de 4 sujets, i, h, g, l par rapport aux autres. Ce fait est souligné par la présence d'une classe abcdefjk au seuil 0,12. Il faut remarquer que ce sont principalement les sujets g et h qui présentent un comportement particulier, restant des singletons dans la classification jusqu'au seuil 0,12 pour h et 0,14 pour g, alors que le sujet i fait partie d'une classe de 4 éléments à 0,10 et l d'une classe à 5 éléments à 0,11.

Une classification pyramidale est illustrée figure IV.10, prenant en compte les mêmes paramètres qu'auparavant en excluant les sujets g et h ceux-ci présentant un comportement trop particulier par rapport à l'ensemble des sujets. Le comportement de l est bien retrouvé et on remarquera la bonne homogénéité des classes.

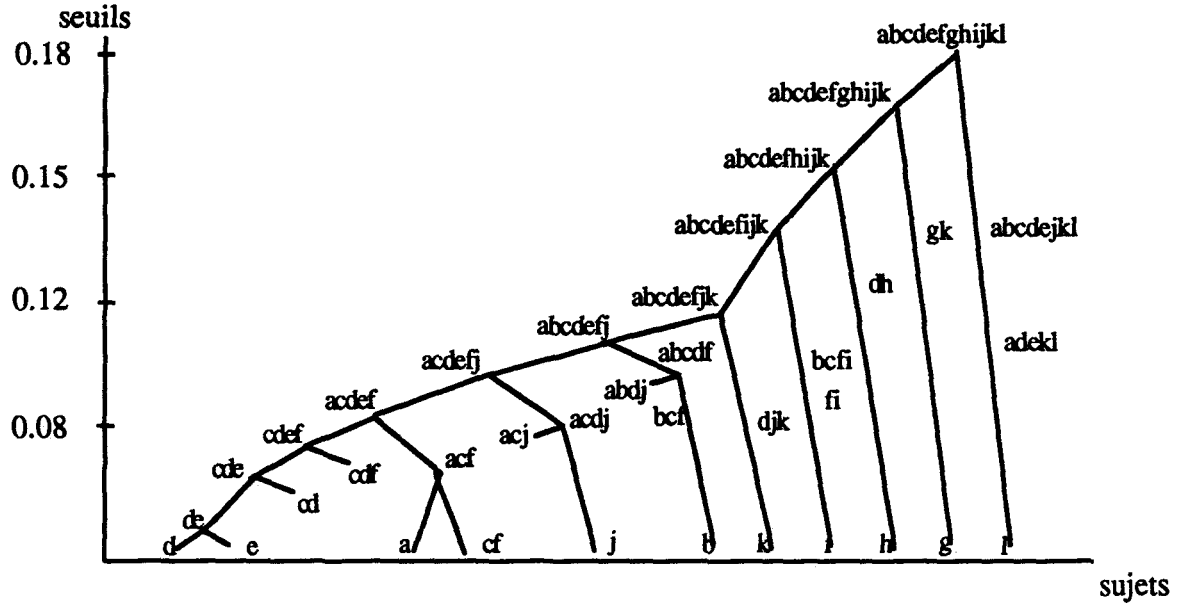


Figure IV.9 : Représentation à l'aide de l'arbre à liaisons incomplètes :
classification des 12 sujets en considérant 17 questions pour les 9 réglages et les 4 périodes

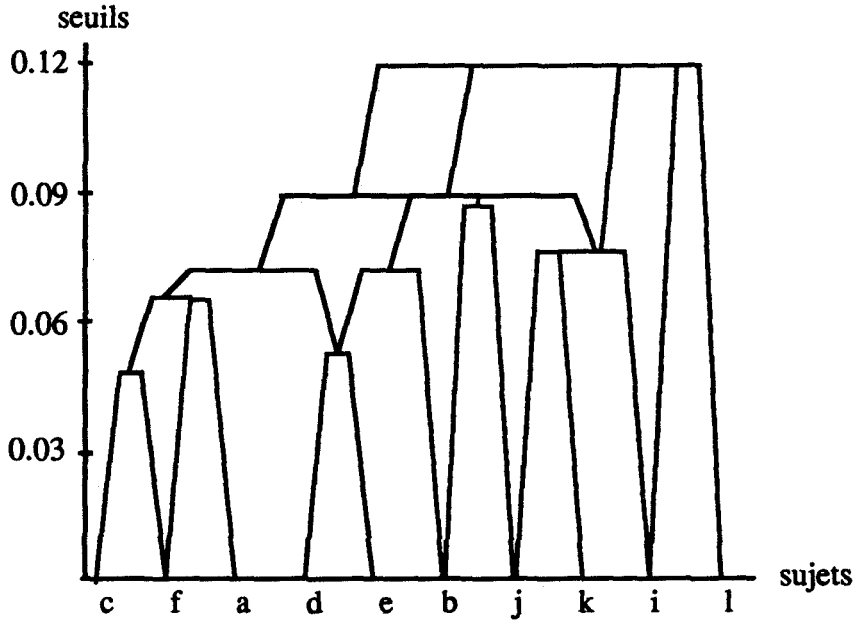


Figure IV.10 : Représentation à l'aide d'une pyramide :
classification de 10 sujets (sans g et h), 17 questions, 9 réglages et 4 périodes

Une classification des sujets par arbre à liaisons incomplètes est présentée figure IV.11. Chaque sujet est caractérisé par les 17 réponses fournies au cours des 4 périodes pour le seul réglage HMP, assise Haute, table Moyenne et dossier Proche. Si, par rapport à la figure IV.9 on retrouve bien les sujets g et l avec un comportement particulier, il est à noter que les sujets h et i s'agrègent eux, très bien.

Enfin, une classification des sujets hiérarchique est proposée figure IV.12. Chaque sujet est caractérisé par ses réponses données aux deux questions 4e, relative aux yeux, et 8, relative à l'ambiance lumineuse pour les 9 réglages et les 4 périodes. On remarque que deux sujets sont à l'opposé du comportement général, et correspondent en fait à deux personnes ayant ressenti des problèmes visuels, ce qui confirme encore la cohérence des réponses des sujets.

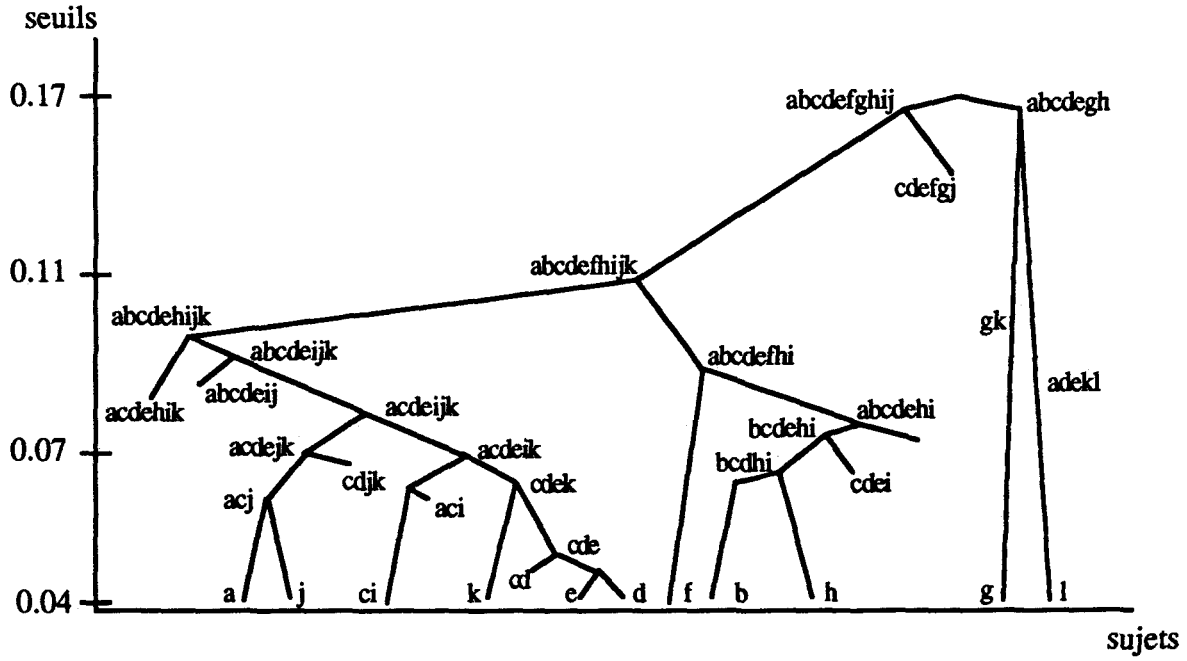


Figure IV.11 : Représentation à l'aide de l'arbre à liaisons incomplètes : classification des 12 sujets pour 17 réponses correspondant au réglage HMP pour les 4 périodes

Les principaux résultats se dégageant alors de l'analyse sont :

- au niveau de la stratégie employée par les sujets vis à vis de la réponse portée sur le différenciateur sémantique, il apparait que 10 sujets sur 12 ont un comportement analogue qui correspond en fait à une utilisation très limitée de l'axe dialectique. Les

deux derniers experts, g et h, sont ceux qui ont le plus largement utilisé la totalité de l'échelle.

- les rubriques de la question 4 exceptée 4g, douleur dans le dos, sont très peu discriminantes, le champ d'utilisation de l'axe étant très faible,
- la question 6, *intérêt porté au texte*, est la question où les classes sont les plus dispersées et où l'agrégation est la plus "mauvaise".

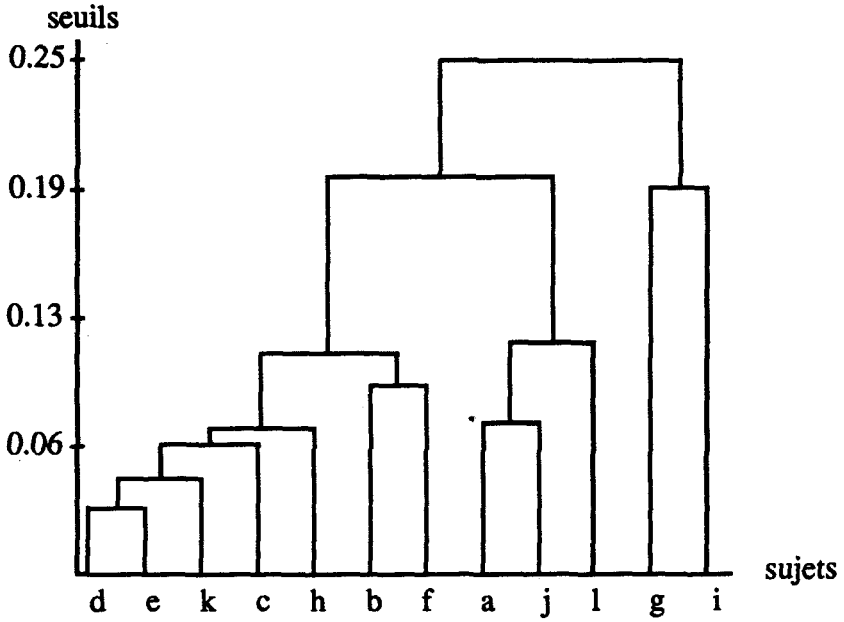


Figure IV.12 : Représentation à l'aide d'une hiérarchie :
classification des 12 sujets, pour 2 réponses données aux questions 4e relative aux yeux et 8 relative à l'ambiance lumineuse correspondant aux 9 réglages et aux 4 périodes

Enfin, quelques remarques peuvent être faites concernant les réglages : il est à noter que hormis 2 réglages HMP, assise Haute, table Moyenne, dossier Proche et BME, assise Basse, table Moyenne, dossier Eloigné, les classifications sont analogues. Pour HMP figure IV.11, il semble que la classification soit plus homogène que pour les autres réglages, quant à BME, l'agrégation des classes est beaucoup plus tardive que pour les autres réglages. Le dernier résultat important à relever est une dégradation de la cohérence des experts, sur le plan temporel - passage de la période 1 à la période 4 - et ce, quel que soit le réglage.

IV.2-3 Comparaison des appréciations des différents réglages

Il s'agit également de déterminer s'il existe une ou plusieurs configuration de réglage mieux ou moins bien perçues par les sujets. Dans cette partie les éléments à classer sont les réglages. Dans ce but, la représentation par plan présentée chapitre II, paragraphe 3.2, est utilisée.

La première étape consiste à construire les deux SEAF particuliers :

le SEAF optimal : celui-ci correspond à une expérience "optimale au sens du confort" construit à l'aide des réponses "idéales" associées à chaque question, chapitre II paragraphe 3.2, il est représenté figure IV.13.

questions : 1 à 5 6 7 à 11

Figure IV.13 : représentation du SEAF optimal

le SEAF type : celui-ci est obtenu pour chaque question par concaténation de toutes les fonctions de répartition complémentaires suivant le nombre d'observations prises en compte. Par exemple, pour représenter les réglages en intégrant sur toutes les périodes, chaque réglage correspondra à un SEAF composé de 17 fonctions de répartition complémentaires construites chacune sur 12 x 4 observations.

Ces deux SEAF particuliers définis ont pour but, comme le montre le paragraphe 3.2 du chapitre II, la construction des plans permettant de représenter visuellement les résultats. Comme le souligne ce même paragraphe, avant d'interpréter les résultats issus de ces plans, il convient de déterminer la précision des points projetés dans le plan. Pour ce faire, un calcul d'erreur est mis en oeuvre déterminant à partir de quelle distance deux points projetés dans le plan peuvent être considérés comme significativement distincts.

a. Calcul d'erreur

Le calcul d'erreur est dépendant du questionnaire utilisé, dans le cas présent celui de la figure IV.6. Le SEAF optimal est imposé par celui-ci, il correspond à celui de la figure IV.13.

Le calcul d'erreur est dépendant également du nombre d'observations prises en compte. Le nombre de ces observations peut être différent suivant l'analyse effectuée, il varie en fonction du nombre de sujets de réglages et de périodes pris en compte. Il est possible d'avoir 12 observations, analyse effectuée pour les 12 sujets en considérant leurs réponses sur un réglage et une période, ou 48 - 12 sujets, 1 réglage, 4 périodes - ou 108 - 12 sujets, 9 réglages, 1 période - ou 432 - 12 sujets, 9 réglages et 4 périodes.

C'est sur le minimum, c'est à dire 12 observations, que le calcul d'erreur est effectué car le fait de travailler à l'aide de fonctions de répartitions entraîne que plus le nombre d'observations est petit, c'est à dire le nombre de paliers de ces fonctions est petit, plus l'erreur a des chances d'être grande.

Les sujets ayant répondu sur un segment continu mesurant 50 mm, la base de départ du calcul d'erreur est la précision de cette réponse qui a été estimée à 2 mm.

A partir de cette estimation, le calcul d'erreur se décompose en deux étapes :

- soient les S_i $i \in \{1...12\}$ 12 SEAF construits à l'aide des 17 réponses au questionnaire des sujets. Le SEAF optimal étant déterminé et le SEAF type calculé sur ces 12 sujets, chaque sujet est alors projeté dans un plan (d(S_i,SEAF optimal),d(S_i,SEAF type)).

Ces 12 points projetés sont les "références", il reste à faire varier les réponses autour de celles données et projeter les nouveaux SEAF obtenus en observant quelle est la dispersion obtenue autour de ces points de référence.

- Pour ce faire, les réponses de chaque sujet i sont tirées aléatoirement entre la réponse mesurée - 2 mm et la réponse + 2mm. On obtient ainsi pour chaque sujet i un nouveau SEAF noté Sa_i. Les distances entre ce SEAF et le SEAF optimal et le SEAF type sont recalculées et les points obtenus projetés dans le plan. Il s'agit de faire très attention au fait que les réponses ayant variées le SEAF type doit être recalculé à chaque fois. Pour chaque sujet 5000 SEAF Sa_i sont déterminés puis projetés dans le plan, figure IV.14.

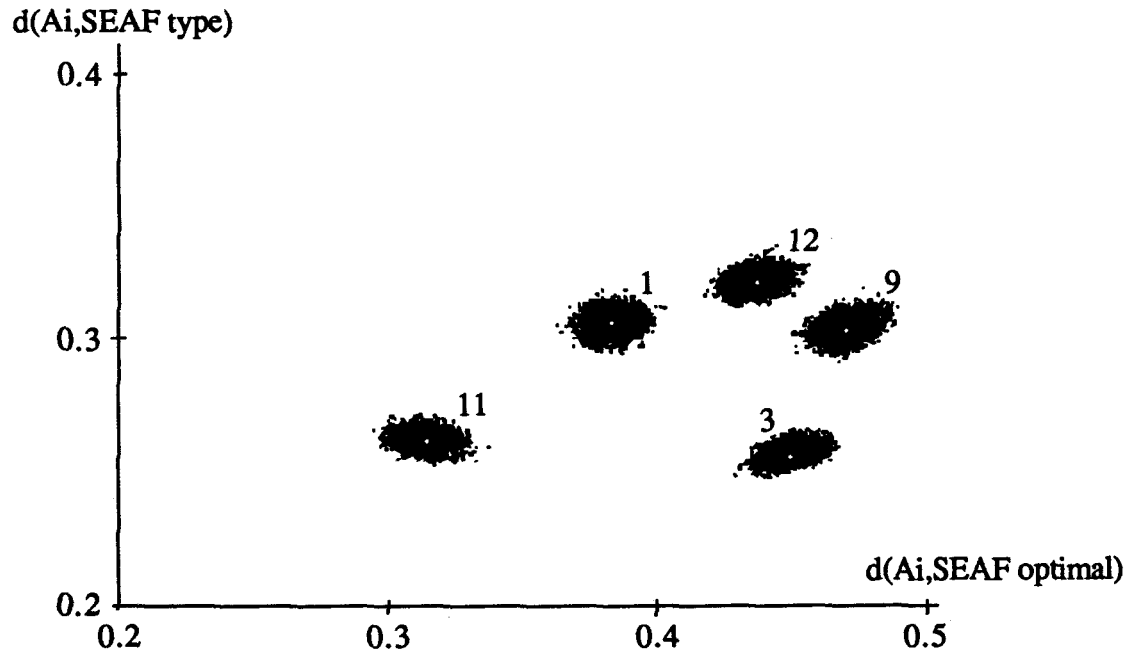


Figure IV.14 : Projection de 5 SEAF S_i et des 5000 Sa_i correspondants dans le plan

Le plan présenté figure IV.14 n'est composé que de 5 sujets pour montrer des nuages de points distincts. La première constatation est que les nuages sont environ moitié moins hauts que larges ce qui indique que l'erreur par rapport au SEAF type est plus faible que celle par rapport au SEAF optimal. Ceci est dû au fait que que le SEAF type, variant en fonction des réponses alétaoires Sa_i, provoque des compensations. Ce phénomène n'apparait pas pour le SEAF optimal car il est indépendant des réponses des sujets. L'erreur maximale a été considérée comme la distance séparant les deux points d'un même nuage les plus éloignés par rapport à l'un des axes. Il vient ainsi en notant E_{opt} l'erreur par rapport au SEAF optimal et E_{typ} celle par rapport au SEAF type :

$E_{opt} = 2,5\ \%$ et $E_{typ} = 1,5\ \%$

Deux points sont donc considérés comme significativement distincts par rapport au SEAF type s'ils sont distants de 0,015 et par rapport au SEAF optimal s'ils sont distincts de 0,025.

Le calcul d'erreur étant effectué, l'analyse, ici présentée, s'intéresse à l'évolution intra et interexpérimentale des appréciations de l'ensemble des sujets. Dans un premier temps la distance définie, chapitre II, est utilisée comme moyen de comparaison.

b. Comparaison à l'aide de la distance

Les résultats sont présentés par des points projetés dans un plan :

$d(A_i, \text{SEAF optimal})$; $d(A_i, \text{SEAF type})$

La première analyse représentée dans le plan figure IV.15, s'intéresse à l'étude inter-expérience, le plan est construit à l'aide des réponses des 12 sujets aux 17 questions en intégrant les 4 périodes. Une étude inter-période, pour déterminer quelle est l'influence du temps sur les réponses des sujets, fait l'objet de la seconde analyse. Deux plans sont construits sur l'ensemble des sujets dans ce but. Le premier figure IV.16, pour déterminer si la douleur augmente avec le temps, est construit avec les réponses des sujets aux questions 1 à 6. Le second figure IV.17, pour déterminer si le temps influe sur l'appréciation des réglages et de l'environnement extérieur, est construit à l'aide des réponses des sujets aux questions 7 à 11.

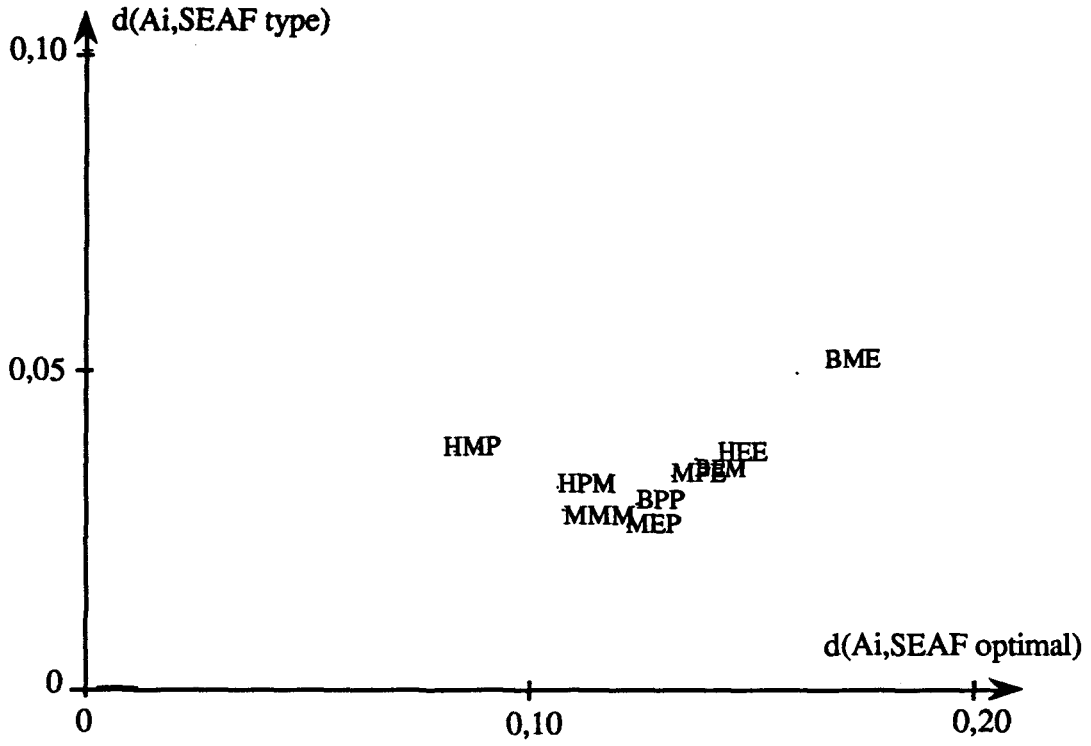


Figure IV.15 : Plan construit pour les réponses des 12 sujets aux 17 questions en intégrant les 4 périodes

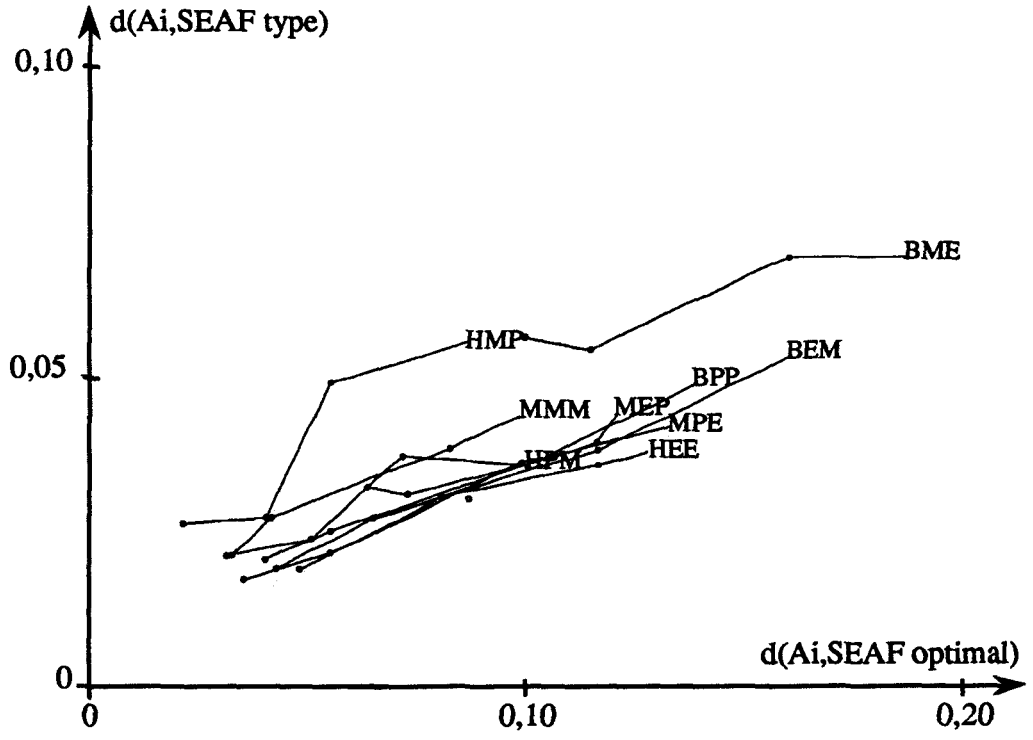


Figure IV.16 : Plan construit pour les réponses des 12 sujets aux questions 1 à 6 en séparant les périodes. Les points correspondants à un même réglage sont reliés dans le sens croissant de la période 1 à la période 4

Le plan figure IV.15, fait apparaitre le réglage assise Haute, table Moyenne, dossier Proche, HMP, comme celui étant le mieux perçu par l'ensemble des experts à l'opposé du réglage assise Basse, table Moyenne, dossier Eloigné, BME. Notons également qu'en dehors de BME tous les réglages sont sensiblement proches du SEAF type.

Les évolutions "intra expérience" sont illustrées sur le plan figure IV.16. Ce plan fait apparaître, pour les 6 premières questions, que le passage de la première à la quatrième période s'accompagne toujours d'une dégradation plus ou moins importante du niveau de confort, éloignement de l'axe vertical, ainsi qu'un éloignement du SEAF type. Ce dernier point indique que quel que soit le réglage, le comportement global des sujets est identique.

A l'opposé de la figure IV.16, la figure IV.17 montre que pour les questions relatives à l'environnement extérieur, 7 et 8, et aux réglages du poste de travail, 9, 10 et 11, il n'y a pas d'influence du temps sur l'appréciation des sujets ce qui corrobore la cohérence de leurs réponses.

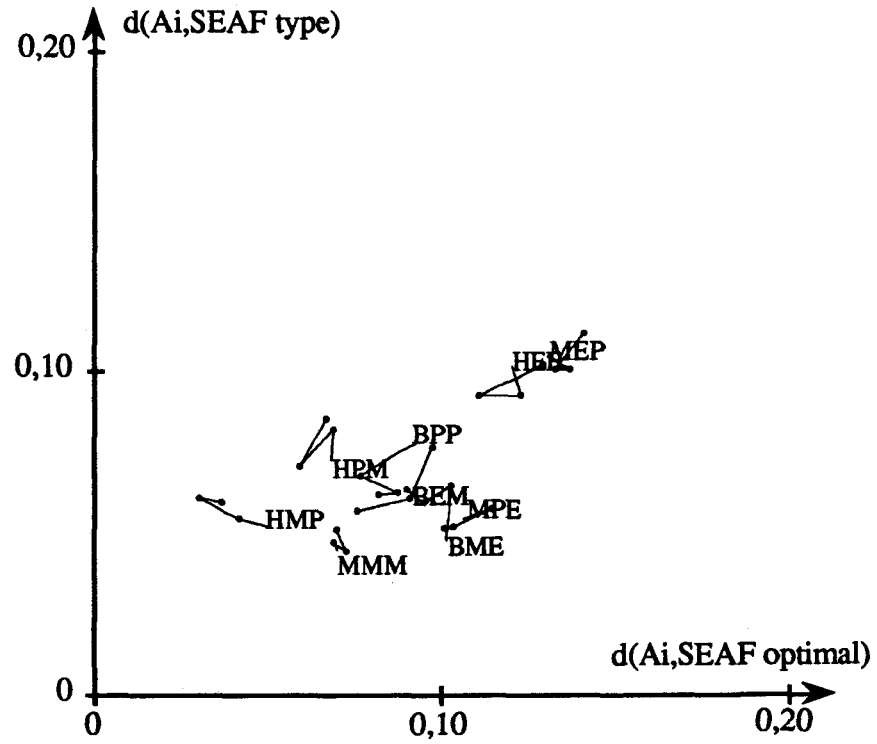


Figure IV.17 : Plan construit pour les réponses des 12 sujets aux questions 7 à 11 en séparant les périodes. Les points correspondants à un même réglage sont reliés dans le sens croissant, de la période 1 à la période 4

c - Utilisation de l'indice I comme moyen de comparaison

Comme il est indiqué chapitre II, paragraphe 3.2, le résultat de la comparaison à l'aide de l'indice I donne quatre valeurs IS_0, II_0, IS_t et II_t qui correspondent aux aires supérieures et inférieures calculées par rapport au SEAF optimal et au SEAF type :

$I(A_i, \text{SEAF optimal}) = (IS_0, II_0)$ et $I(A_i, \text{SEAF type}) = (IS_t, II_t)$

La réponse "idéale" pour les questions 1 à 6 se trouve à une des extrémités du différenciateur sémantique continu, figure IV.13, ce qui entraîne qu'une des valeurs issue de la comparaison au SEAF optimal est obligatoirement nulle : pour les questions 1 à 5, $II_0=0$ et pour la 6, $IS_0=0$. En conséquence, pour ces questions le fait de représenter un plan avec l'indice I n'apporte pas d'informations supplémentaires. L'utilisation de l'indice n'est donc présentée que pour les questions 7 à 11 présentant une réponse "idéale" au centre du différenciateur sémantique continu.

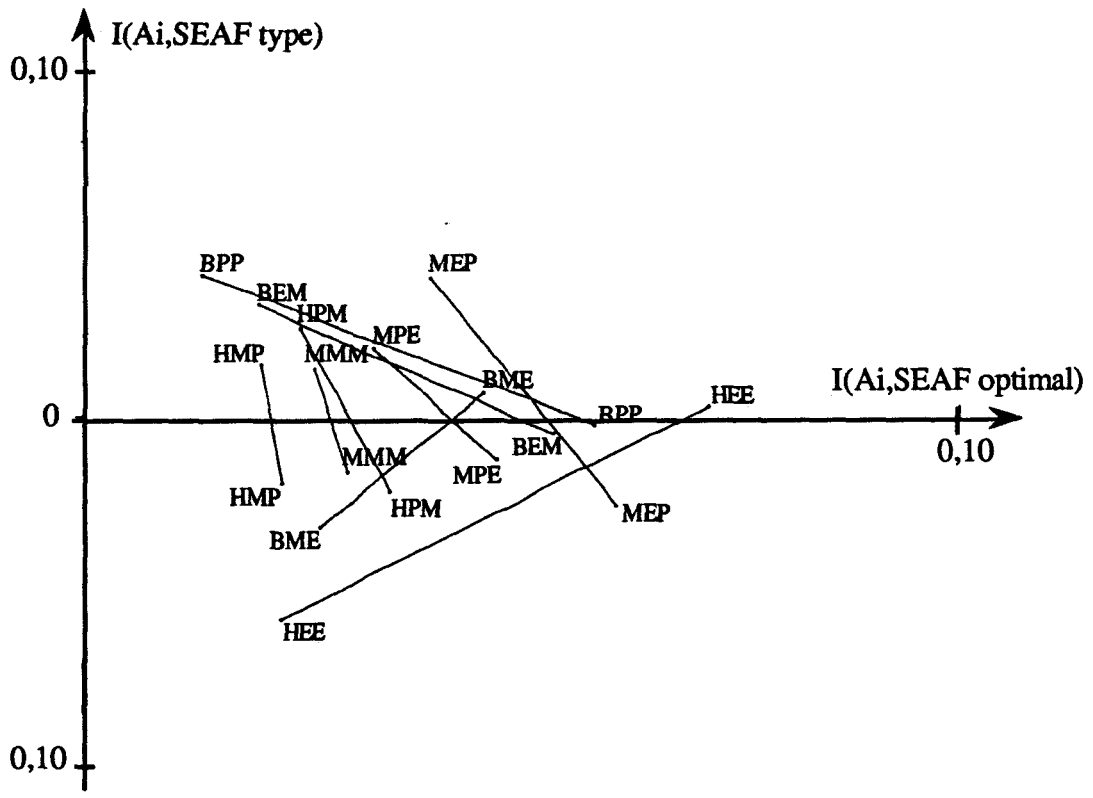


Figure IV.18 : Plan construit pour les réponses des 12 sujets aux questions 7 à 11 en intégrant les périodes. Un même réglage est relié entre les 2 parties du plan

En conséquence, le plan présenté figure IV.18 a été construit en considérant les réponses des 12 sujets aux cinq dernières questions en intégrant les 4 périodes. Les réglages

ont été reliés entre les deux parties du plan. Les trois réglages qui présentent une bonne distribution des aires supérieures et inférieures par rapport au SEAF type, IS_t de l'ordre de II_t, et au SEAF optimal, IS_0 de l'ordre de II_0, sont HMP, assise Haute, table Moyenne et dossier Proche, MMM correspondant aux trois modalités moyennes et HPM, assise Haute, table Proche et dossier Moyen. Par ailleurs, en considérant les aires supérieures, c'est à dire la partie "haute" du plan, le réglage le plus proche de l'optimal est BPP, assise Basse, table Proche et dossier Proche, il présente par contre un comportement opposé au regard de l'aire inférieure.

La comparaison à l'aide de l'indice ne donne pas de résultats différents de ceux précédemment cités, elle ne fait que les affiner. Une des constatations principales pour les 5 dernières questions, est que ce sont les réglages HMP, MMM et HPM, dont les fonctions de répartitions présentent une aire inférieure du même ordre que l'aire supérieure par rapport aux fonctions de répartition du SEAF type qui sont les plus proches de l'optimal.

IV.2-4 Conclusion

L'analyse des questionnaires a permis de :

- faire apparaitre que la façon d'appréhender l'axe continu du différenciateur joue un rôle important, certains experts se cantonnant dans les zones extrêmes et dans la partie centrale, d'autres utilisant volontiers le segment dans sa totalité.
- montrer la relative cohérence des réponses des experts sur les appréciations concernant les réglages et la dégradation dû niveau de confort dans le temps ;
- dégager une configuration qui semble être la mieux perçue par l'ensemble des sujets : HMP, assise Haute, table Moyenne, dossier Proche ;

IV.3 - MISE EN RELATION DES DONNEES OBJECTIVES ET DES DONNEES SUBJECTIVES

Disposant de données objectives et subjectives, il est intéressant de déterminer s'il existe une réelle adéquation entre ces deux groupes. Pour ce faire, la méthode décrite au chapitre III est mise en oeuvre. Néanmoins, avant de passer à la méthode proprement dite, il est nécessaire de rendre les données homogènes, dans le cas présent, réaliser un codage des variables tant objectives que subjectives permettant de ramener les données dans l'intervalle [0,1].

IV.3-1 Codage et choix des variables

L'utilisation d'un différenciateur continu a permis de ramener les données subjectives dans l'intervalle [0,1] en considérant 0 et 1 à chaque extrémité du segment. Par contre, pour les données objectives, il est nécessaire, vu l'hétérogénéité des variables, de passer par une étape préalable de codage qui permette la comparaison de celles-ci. Ce codage devant de plus ramener les données dans l'intervalle [0,1], il a été décidé d'utiliser un codage par transformation linéaire :

pour une variable v_i : $v_i \rightarrow \frac{v_i - a}{b}$

avec $a = \min(v_i)$ et $b = \max(v_i) - \min(v_i)$.

Notons de suite que cette normalisation n'est pas faite "une fois pour toutes", mais qu'elle varie en fonction des variables et des observations choisies. Par exemple, le fait de travailler sur 10 sujets au lieu de 12 sujets peut entrainer un changement du $\min(v_i)$ et/ou du $\max(v_i)$ ce qui par conséquent, entraîne un nouveau codage à partir de ces 10 sujets seuls pour les mêmes observations.

Le codage étant réalisé, préalablement au traitement, il s'agit comme le souligne, le paragraphe 2.1.a du chapitre III, de déterminer les sous-ensembles des variables à mettre en relation. Etant donné le nombre important de variables prises en compte, 17 variables subjectives et 18 objectives il a été décidé de ne retenir dans l'analyse que les variables les plus significatives.

La détermination de ces variables significatives pour les données subjectives est basée sur les résultats de l'analyse effectuée précédemment. Celle-ci fait ressortir que certaines questions sont très peu discriminantes, les sujets se sont cantonés dans des zones très restreintes du différentiateur. Ces questions ont été éliminées de la mise en relation. Les variables subjectives restantes correspondent alors aux questions relatives :

- à l'envie de bouger, question 1,
- à la fatigue perçue, question 3,
- au niveau de douleur dans le dos, question 4g,
- aux appréciations des réglages de l'assise, de la table, et du dossier, questions 9, 10 et 11

Les données objectives ont été traitées par des méthodes multidimensionnelles d'analyse des données /LOSLEVER 88a/, analyse factorielle des correspondances multiples par exemple, et ont fait ressortir parmi les 18 variables objectives les variables les plus significatives. Celles-ci correspondent :

- aux six moyennes, MOY^1 à MOY^6 ;
- aux six indices de variation, IV^1 à IV^6.

IV.3-2 Résultats de la mise en relation

Les couples $(\Pi_{y/x},\Lambda)$ choisis pour la mise en relation sont au nombre de 6, ce sont ceux cités au chapitre III, paragraphe 2.2.a, utilisés pour l'exemple.

Comme pour les données de l'exemple du chapitre précédent, les couples $(\Pi_{y/x},\Lambda)$ utilisés induisent des résultats différents. A nouveau, il semble que ce soit les couples 2 et 3 qui présentent la meilleure adéquation avec les données. Le couple 1 ayant un comportement similaire à ces couples mais avec des valeurs des paramètres plus élevés. Les autres couples ne permettent pas, quant à eux de dégager des relations. Les résultats ne sont alors présentés que pour le deuxième couple $(\Pi_{y/x},\Lambda)$.

Les deux seuls résultats obtenus sur la mise en relation sont :

- une relation entre la question 1 relative à l'envie de bouger et les 6 indices de variation de posture, figures IV.19, pour 8 experts sur 12, tous réglages et toutes périodes pris en compte.

Cette relation semble naturelle, elle rend compte de l'adéquation de la réponse des sujets sur leur envie de bouger et leur variation réelle de posture.

- une relation entre d'une part la question 11 correspondant à l'appréciation sur le dossier et d'autre part la moyenne sur la voie 4 - courbure au niveau de la troisième vertèbre lombaire - et l'indice de variation de la voie 6 - appui sur le dossier - figures IV.20.

Cette relation traduit l'influence du facteur dossier, sa position est naturellement perçue comme telle, l'appui sur le dossier est alors plus faible pour le dossier éloigné que lorsque le dossier est proche.

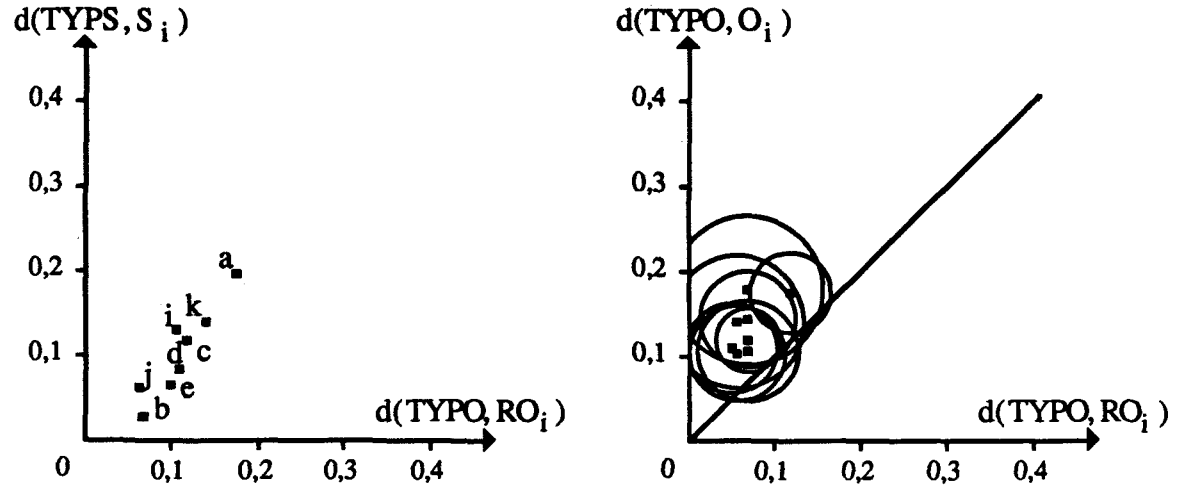


Figure IV.19 : plans de vérification et de validation construits entre la question 1, envie de bouger, et les 6 indices de variation pour 8 experts sur 12 sur les 9 réglages en intégrant les périodes

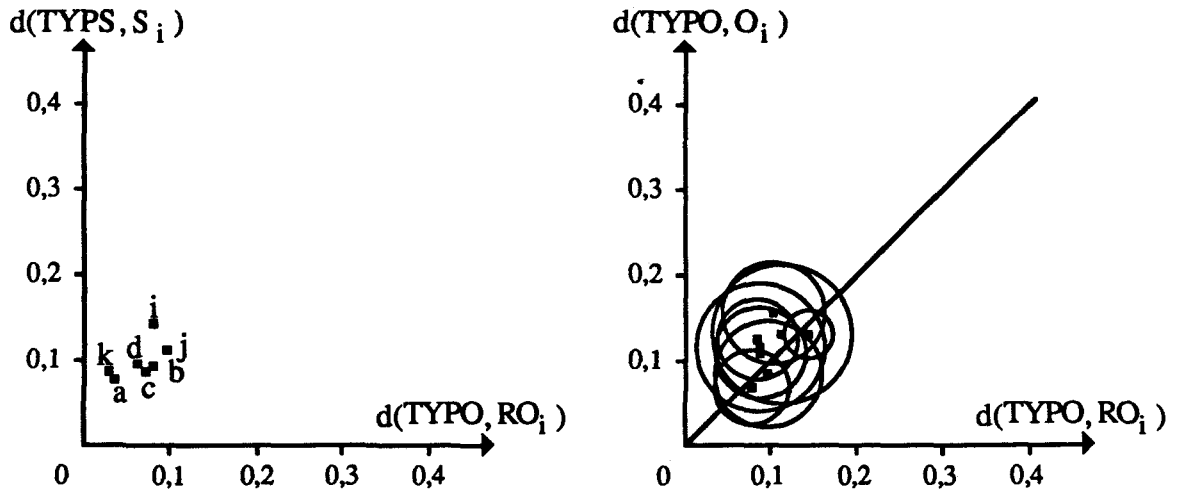


Figure IV.20 : plans de vérification et de validation construits entre la question 17, appréciation du dossier, et la moyenne 4 et l'indice de variation 6 pour 8 experts sur 12 sur les 9 réglages en intégrant les périodes

Le fait qu'il y ait peu de relations entre les deux groupes de données, objectifs et subjectifs, résultat confirmé par l'analyse factorielle des correspondances multiples, montre que le recueil d'un seul type de données, subjectives ou objectives, ne peut suffire dans un tel domaine d'étude.

Ce manque d'adéquation entre les deux groupes de données peut être dû principalement à deux causes :

- en premier lieu, les sujets semblent présenter une bonne cohérence dans leur réponse par exemple :

le niveau de douleur augmente lors du passage de la première à la quatrième période, ou, lorsqu'un réglage est modifié, le sujet perçoit cette modification ou encore la perception du réglage ne varie pas en fonction du temps. Malgré cette cohérence, il leur est sans doute difficile de quantifier subjectivement les aspects de fatigue ou de gênes liées à la manipulation, ces aspects étant liés également à des facteurs indépendants de celle-ci - problèmes de dos, énervement, problèmes familiaux par exemple.

- en second lieu, les variables objectives choisies pour synthétiser les signaux enregistrés sur chaque voie de mesure, la moyenne l'écart-type et l'indice de variation, ne sont peut être pas une représentation adéquate des informations enregistrées.

IV.4 - CONCLUSION

Notons tout d'abord, que tous les résultats trouvés par les méthodes utilisées confirment ceux obtenus par des méthodes d'analyse des données multidimensionnelles utilisées par un autre chercheur du laboratoire /LOSLEVER 88a/. Sans rentrer dans les détails des différences méthodologiques ou des différences de présentation et d'interprétation des résultats, les résultats concernant les données subjectives sont similaires et le peu de résultats de la mise en relation des données objectives et subjectives se retrouve également au niveau de la précédente étude.

L'exemple traité, concernant l'étude ergonomique d'un poste de travail, a été choisi pour valider les méthodes développées. Si l'analyse des données subjectives issues de questionnaires a permis de dégager des résultats, la mise en relation des données objectives et subjectives n'a pas donné de réelle adéquation entre les deux ensembles, seules deux relations très ponctuelles ont pu être dégagées. Il nous semble nécessaire de poursuivre l'application de la méthode de mise en relation à d'autres exemples, avec des données moins hétérogènes, car son originalité réside dans le fait de mettre en évidence des relations globales.

Dans ce contexte, une des perspectives à envisager est l'extrapolation de la relation par rapport à un nouvel expert intervenant dans l'analyse. Ce dernier point est traité dans le chapitre suivant.

CHAPITRE V

TRAITEMENT DE DONNEES
SYSTEMES EXPERTS ET PERSPECTIVES

Après avoir appliqué les méthodes développées aux chapitre II et III à un exemple concret, ce dernier chapitre présente leurs perspectives et tente de les situer par rapport au problème général de l'analyse et de la modélisation des systèmes.

La première partie de ce présent chapitre donne une extension à la mise en relation : il s'agit de déterminer dans quelle mesure une relation construite sur un ensemble de sujets peut-être extrapolée par rapport à un nouveau sujet intervenant dans l'analyse.

La deuxième partie pose le problème de l'extraction des connaissances en général. Dans des domaines, tels que le diagnostic technique, où, en général, une expertise existe, l'approche par les systèmes à base de connaissance se justifie mais elle ne peut se généraliser à tous les domaines. Il est alors nécessaire pour des situations complexes où l'expertise est inexistante, voire impossible, de suivre une démarche expérimentale pour acquérir des données in situ ou en laboratoire, puis d'analyser les données multidimensionnelles afin de mettre en évidence des classes ou des relations existant entre objets.

V.1 EXTRAPOLATION DE LA RELATION

Dans la suite de ce paragraphe une relation R sera supposé créée entre des variables d'un ensemble Su et des variables d'un ensemble Ob sur un ensemble E de sujets. Il y a deux façons de considérer l'extrapolation de la méthode. La première est de considérer qu'un nouveau sujet intervient dans la mise en relation ce qui permet de déterminer un nouvel ensemble de règles. La deuxième est de considérer que pour ce nouveau sujet on ne dispose que d'un seul sous-ensemble déterminé sur l'ensemble des variables de départ Su.

Dans ces deux cas il est nécessaire de savoir si la relation déterminée sur l'ensemble E des sujets peut s'étendre au nouveau sujet. Néanmoins les deux cas étant différents sur le plan méthodologique, ils sont étudiés à part.

V.1-1 Premier cas : introduction d'un ensemble de règles par un sujet

Le premier cas peut se résumer par la formulation suivante :

la relation est supposée créée à partir des règles suivantes, voir chapitre III, sur e sujets :

$$\begin{cases} \text{expert 1} \begin{cases} \text{Si X est } A_{11} \text{ alors Y est } B_{11} \\ \dots \quad \dots \\ \text{Si X est } A_{1n} \text{ alors Y est } B_{1n} \end{cases} \\ \dots \qquad\qquad \dots \\ \text{expert e} \begin{cases} \text{Si X est } A_{e1} \text{ alors Y est } B_{e1} \\ \dots \quad \dots \\ \text{Si X est } A_{en} \text{ alors Y est } B_{en} \end{cases} \end{cases}$$

Un nouveau sujet intervient, introduisant un ensemble de n règles :

$$\text{expert e+1} \begin{cases} \text{Si X est } A_{(e+1)1} \text{ alors Y est } B_{(e+1)1} \\ \dots \quad \dots \\ \text{Si X est } A_{(e+1)n} \text{ alors Y est } B_{(e+1)n} \end{cases}$$

Il convient de déterminer si ce nouveau sujet peut être inclus dans l'analyse. Ce premier cas semble relativement simple à résoudre et deux étapes sont nécessaires et suffisantes.

On rappelle que la relation R a été créée à partir des sous-ensembles aléatoires flous (SEAF) S_i construits à partir des A_{ij} et des SEAF O_i construits à partir des B_{ij}. De la même façon, pour le nouveau sujet deux SEAF S_{e+1} et O_{e+1} sont construits à partir respectivement des $A_{(e+1)i}$ et $B_{(e+1)i}$.

La première étape consiste à appliquer la relation R obtenue sur les e sujets à ce nouveau sujet, le résultat étant l'obtention du SEAF RO_{e+1}. Ce SEAF est comparé au SEAF O_{e+1}. Les distances calculées pour ce nouveau sujet permettent alors de le projeter dans les plans de vérification et de validation de la relation comme donnée "supplémentaire". Ce terme étant employé dans le sens où ce sujet ne participe pas à la construction de la relation R. Dans le cas où il présente un comportement adéquat par rapport aux critères I, II et III définis au chapitre III paragraphe 2, il peut être considéré comme "entrant" dans l'analyse et cette première étape peut être suffisante. Effectivement, en considérant que ce nouveau sujet n'apporte que peu d'informations supplémentaires il n'est pas absolument nécessaire de l'introduire dans l'analyse. D'un autre coté il est possible de construire une nouvelle relation R' sur les n+1 sujets pour renforcer la relation R, figure V.1.

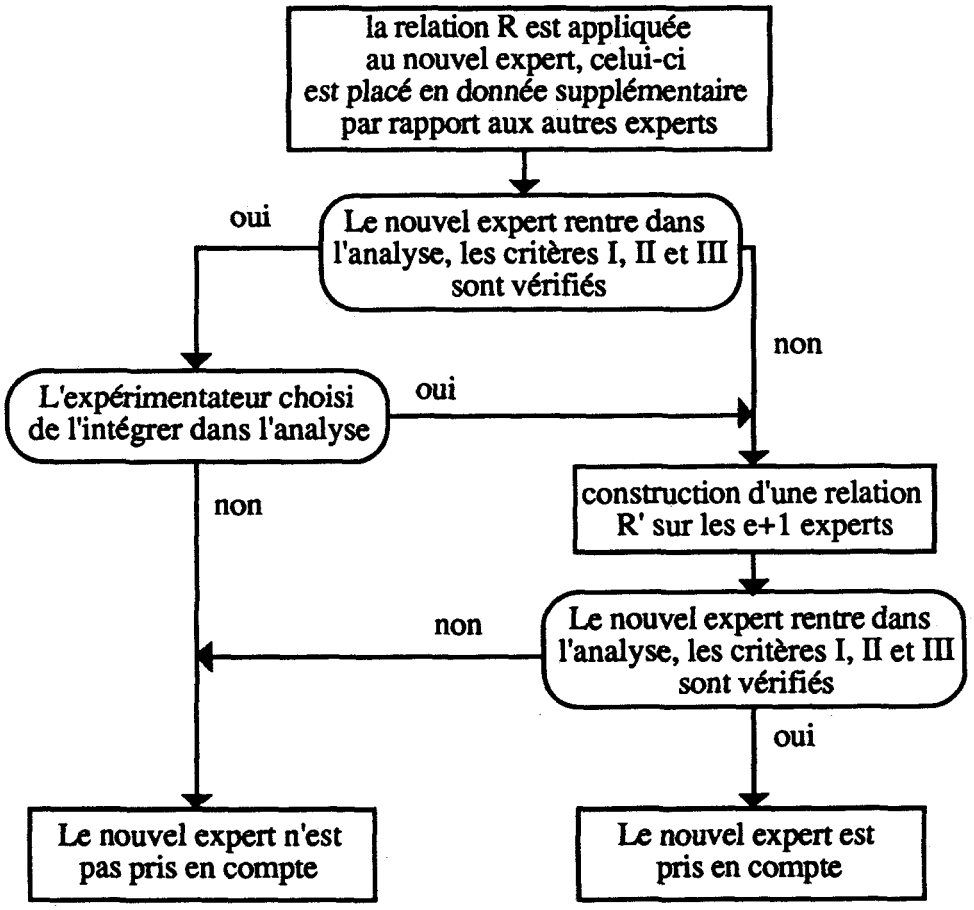


Figure V.1 : Extrapolation de la relation dans le cas d'un nouvel ensemble règles introduit

Dans le cas où le nouveau sujet ne vérifie pas les critères I, II ou III il est alors nécessaire de déterminer si une nouvelle relation peut être dégagée sur les e+1 sujets. Cette deuxième étape consiste alors à construire une relation R' entre les ensembles de variables Su et Ob sur e+1 sujets de la même façon que la relation R et à appliquer les critères I, II et III de vérification et de validation à la relation R'. Si R' est validée le sujet e+1 "rentre" dans

l'analyse, et il est nécessaire de le garder car il apporte une extension de la relation à un cas qui n'était pas pris en compte dans la relation R. Dans le cas contraire il est exclus de l'analyse. Ces deux étapes sont résumées figure V.1.

V.1-2 Deuxième cas : les données ne sont recueillies pour un nouveau sujet que sur l'ensemble des variables de départ

Ce deuxième cas correspond à la formulation suivante :

la relation R est supposée créée à partir des règles suivantes sur e sujets :

$$\begin{cases} \text{expert 1} \begin{cases} \text{Si X est } A_{11} \text{ alors Y est } B_{11} \\ \dots \quad \dots \\ \text{Si X est } A_{1n} \text{ alors Y est } B_{1n} \end{cases} \\ \dots \qquad\qquad \dots \\ \text{expert e} \begin{cases} \text{Si X est } A_{e1} \text{ alors Y est } B_{e1} \\ \dots \quad \dots \\ \text{Si X est } A_{en} \text{ alors Y est } B_{en} \end{cases} \end{cases}$$

le nouveau sujet n'intervenant que sur les variables de Su il vient :

$$\text{expert e+1} \begin{cases} \text{X est } A_{(e+1)1} \\ \dots \\ \text{X est } A_{(e+1)n} \end{cases}$$

La question posée est de savoir dans quels cas la relation peut être appliquée à ce nouveau sujet. En reprenant l'exemple du chapitre IV et en supposant créée une relation R, il s'agit de déterminer, à partir des réponses d'un nouveau sujet à un questionnaire, et à partir de ses réponses seulement, si l'on peut déduire le comportement de ce dernier par rapport aux variables objectives à l'aide de la relation R.

Dans cette optique, il est nécessaire de déterminer le comportement du nouveau sujet introduit par rapport aux autres. *Une classification, hiérarchie, pyramide...*, est parfaitement appropriée à cela. Il est alors possible de considérer que si ce sujet présente un comportement trop particulier par rapport aux autres - il s'agrège très mal - il ne peut entrer dans l'analyse, celle-ci ne pouvant être valable que pour les comportements similaires à ceux étudiés sur les e sujets.

Si la classification donne de "bons" résultats sur ce sujet une deuxième étape consiste alors à appliquer la relation à ce nouveau sujet. Le résultat obtenu est un SEAF RO_{e+1}, mais le fait de ne pas disposer des $B_{i(e+1)}$, et par voie de conséquence de O_{e+1}, ne permet pas

d'utiliser les critères II et III. Le critère I, de vérification de la relation est seul appliqué. Si le critère reste vérifié pour cette donnée supplémentaire on peut alors considérer que la relation est applicable à ce nouveau sujet. L'extrapolation de la relation est résumée figure V.2.

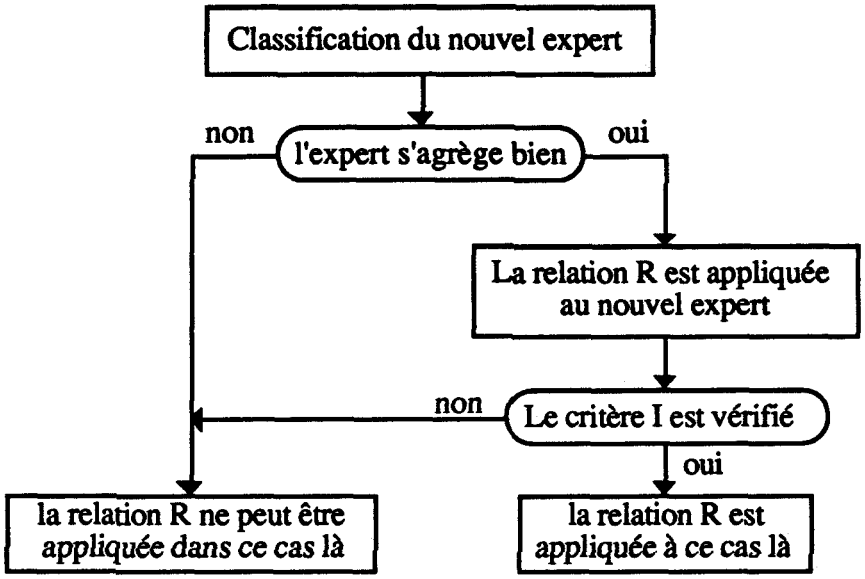


Figure V.2 : Extrapolation de la relation dans le cas où le nouveau sujet n'intervient que sur l'ensemble Su

Ce deuxième cas revient à connaître le domaine d'applicabilité de la méthode. Une perspective est de déterminer la robustesse de la relation par rapport aux couples utilisés pour la créer, pour dégager ce domaine d'applicabilité. Cette robustesse peut se déterminer en faisant varier des données en entrée et/ou en sortie et en regardant au travers des différents critères la "dérive" de la relation par rapport à ces variations.

Le problème lié à l'extrapolation étant traité, la partie suivante est relative au problème plus général de l'extraction de connaissances. Elle permet de montrer les difficultés rencontrées dans ce domaine et de différencier les approches utilisées pour les systèmes à base de connaissance de celles basées sur l'expérimentation et l'analyse des données. Il ne s'agit pas dans ce présent chapitre de réaliser une étude bibliographique des domaines étudiés mais, en guise de conclusion, de bien situer et différencier les deux approches.

V.2 EXTRACTION DE CONNAISSANCES

L'extraction de connaissances est un terme couramment employé dans le domaine de l'intelligence artificielle et plus particulièrement pour les systèmes à base de connaissances. Dans ces derniers, extraire des connaissances revient à *"acquérir les connaissances qui composent le savoir-faire de l'expert et les structurer dans une forme exploitable par une*

machine" /NASSIET 87/. Néanmoins, il est souvent des domaines où l'expertise elle-même fait défaut, le domaine étant trop complexe pour être résumé à des formes déclaratives ou procédurales. Reprenons simplement l'exemple du chapitre IV sur l'influence des réglages d'un poste de travail bureautique. Il n'existe pas d'expert, ou un groupe d'experts, ayant les connaissances suffisantes permettant de dégager pour une personne donnée, dans une situation donnée, les réglages du siège assurant un confort "optimal". Le nombre de facteurs intervenant tant subjectifs - personnalité du sujet, état de santé... - qu'objectifs - données anthropométriques, âge, problèmes physiques du sujet...- est trop grand. L'extraction de la connaissance dans de tels cas nécessite le recours à des méthodes expérimentales, qui permettent d'observer le comportement de plusieurs sujets dans une situation précise, et l'utilisation d'une approche d'analyse de données multidimensionnelles permettant de dégager des résultats.

La première approche traitée est celle liée à l'extraction des connaissances dans les systèmes à base de connaissances dont les différentes méthodes sont rappelées succinctement.

V.2-1 Approches utilisées pour les systèmes à base de connaissance

L'extraction des connaissances dans le domaine des systèmes à base de connaissances est certainement la phase la plus importante et la plus complexe de développement de tels systèmes. Etant donné l'absence d'outils permettant une extraction automatique, le processus est extrêmement laborieux, certains auteurs n'hésitent pas à parler de goulot d'étranglement /NASSIET 87/ pour décrire cette phase.

Les principales techniques utilisées pour l'extraction de la connaissance peuvent être classées en deux familles distinctes /OLSON et RUETER 87/ :

- les techniques directes : les connaissances que l'expert est capable de verbaliser sont directement recueillies ;
- les techniques indirectes : l'expert utilise de manière implicite certaines connaissances qu'il est incapable de restituer directement, ces techniques visent alors à collecter ces connaissances.

Ces deux types de techniques vont être abordées rapidement. D'autres auteurs, /NASSIET 87/ /BENKIRANE 91/ ont développé plus longuement cette phase. Notons simplement que la richesse de la connaissance humaine requiert, pour extraire la

connaissance, la combinaison de plusieurs de ces techniques. Le cogniticien doit déterminer quelles techniques utiliser en fonction des connaissances à extraire.

a. Les techniques directes

Il existe quatre grandes techniques appartenant à ce type :

* l'interview : c'est certainement la plus utilisée des techniques d'extraction de la connaissance. Le cogniticien obtient la connaissance de l'expert par interaction directe avec ce dernier ;

* les questionnaires : au contraire de l'interview, les questionnaires permettent le recueil d'informations précises et structurées, ils peuvent être de différents types, voir chapitre I, ouverts, fermés, avec ou sans notion de certitude...

* l'analyse de protocoles : elle consiste à observer l'expert résolvant une situation réelle à "haute voix". Un enregistrement sonore et/ou vidéo de la séance étant effectué afin de saisir les connaissances procédurales impossibles à communiquer verbalement ;

* le tri conceptuel : il consiste à proposer à l'expert une fiche d'identification pour chaque concept. L'expert trie ces fiches et les classe en sous-groupes, et recommence itérativement le processus sur les sous-groupes jusqu'à l'obtention d'une hiérarchie complète des concepts pris en compte.

b. Les techniques indirectes

Plus complexes que les techniques précédentes les techniques indirectes peuvent nécessiter le recours à des méthodes d'analyse des données pour extraire les résultats. Citons entre autres :

* l'évaluation multidimensionnelle : l'expert fourni un coefficient de ressemblance pour chaque paire d'objets du domaine étudié. Le résultat obtenu, en tenant compte de la symétrie des coefficients, est une matrice de ressemblance. Il est alors nécessaire d'utiliser des techniques d'analyse des données pour permettre l'obtention de résultats sur une telle matrice ;

* l'analyse de grilles de classification : l'expert doit proposer, à partir d'objets du domaine, une caractéristique qui permette de séparer ces objets en deux sous-

ensembles distincts. Cette caractéristique et son opposé, chaud/froid par exemple, sont appelées *constructions mentales*. Une matrice croisant en ligne ces constructions mentales et en colonnes les objets est alors établie et l'expert note tous les objets sur une échelle arbitraire, de 1 à 5 par exemple. Cette matrice, appelée grille de classification permet alors d'obtenir des règles de production.

La représentation des connaissances peut alors se faire de façons différentes, règles de production, arbres de décision par exemple. Dans notre cadre seule la représentation à l'aide de règles de production sera discutée.

L'obtention de ces règles au travers des différentes techniques qui viennent d'être évoquées est déjà difficile dans le cas où il existe un expert du domaine. En règle générale on essaye, en plus, d'obtenir des connaissances précises malgré la difficulté de l'expert à exprimer de telles connaissances /FARRENY 85/.

Dans les systèmes à base de connaissance, il est nécessaire de souligner la place prise actuellement par la *théorie des possibilités pour traiter* le problème de l'imprécis et de l'incertain, certains systèmes tels que SPII-2 permettant de prendre en compte ces deux aspects /LEBAILLY 85/. Les connaissances actuelles permettent de prendre en compte les règles *dont les antécédents et les conséquences* sont imprécis, cet aspect a été traité chapitre III. La prise en compte de l'incertain se fait au travers de mesures de possibilité/nécessité /DUBOIS et PRADE 87/ qui ne seront pas développées ici. Notons également le besoin d'aller plus en avant que des règles de production simples du type : si X est A alors Y et B, et dans ce sens des travaux effectués sur le raisonnement flou multidimensionnel basé sur le schéma suivant /SUGENO et TAKAGI 83/ :

$$\frac{\text{Si } X_1 \text{ est } A_1,\ X_2 \text{ est } A_2,\ldots,\ X_n \text{ est } A_n \text{ alors } Y \text{ est } B \qquad X_1 \text{ est } A'_1,\ X_2 \text{ est } A'_2,\ldots,\ X_n \text{ est } A'_n}{Y \text{ est } B' = ?}$$

ou encore /MINGSHENG 88/ :

$$\frac{\begin{cases} \text{Si } X \text{ est } A_1 \text{ alors } Y \text{ est } B_1 \\ \text{Si } X \text{ est } A_2 \text{ alors } Y \text{ est } B_2 \\ \quad \ldots \qquad\qquad \ldots \\ \text{Si } X \text{ est } A_n \text{ alors } Y \text{ est } B_n \end{cases} \qquad X \text{ est } A}{Y \text{ est } B = ?}$$

Cette dernière représentation s'apparente aux travaux exposés dans ce mémoire. La méthode développée par MINGSHENG ne permet pas de déterminer s'il existe une relation stable entre les A_i et les B_i. Elle consiste, les A_i et les B_i étant définis à obtenir des résultats sur B en fonction de A qui ne soient pas contra-intuitifs.

Enfin, un autre problème à prendre en compte dans ces systèmes est la présence de plusieurs experts qui introduit une difficulté supplémentaire dans le processus. En effet, le cogniticien *"doit obtenir une expertise de synthèse, fruit d'un consensus entre les experts et non pas une expertise de compromis, dans laquelle chaque expert s'y retrouve un peu mais qui en totalité ne satisfait personne"* /BENKIRANE 91/.

Dans le cas où la modélisation d'une expertise humaine est impossible, il est alors nécessaire de faire appel à une approche plus traditionnelle, basée autour d'une approche statistique.

V.2-2 Approche par l'analyse des données multidimensionnelles

L'approche appelée analyse des données multidimensionnelles est en premier lieu, une approche basée sur l'expérimentation. Elle est justifiée lorsque le système est complexe, faisant intervenir un nombre important de variables et d'observations, ainsi que de nombreux couplages entre ces variables /RICHETIN et DUFOUR 79/.

Dans ce contexte, des données sont recueillies qualifiées d'objectives si elles sont issues de capteurs, et de subjectives si elles sont recueillies par l'intermédiaire de sujets. L'emploi du terme sujet, et non plus expert, est délibéré pour ne pas créer d'ambiguïté avec les systèmes à base de connaissances, car dans cette présente approche il n'est plus possible de parler d'expert du domaine.

Il est alors nécessaire de pouvoir structurer les données et, dans ce sens, d'avoir recours à des méthodes statistiques permettant de dégager des paramètres liant variables ou groupes de variables, ou permettant la représentation du système sous forme graphique.

De telles méthodes sont nombreuses, rappelons simplement les méthodes purement descriptives telles que l'analyse en composantes principales ou l'analyse des correspondances, les méthodes répondant à un problème plus structuré telles que l'analyse canonique ou la régression linéaire ou encore, pour permettre la mise en évidence des relations non linéaires entre variables, l'approche basée sur la théorie de l'information /RICHETIN et DUFOUR 79/.

Notre travail s'insère dans ce contexte. Le développement de méthodes nouvelles basées sur la théorie des sous-ensembles flous semble prendre une part de plus en plus importante dans le contexte de données incertaines ou imprécises. En premier lieu, le

traitement des questionnaires évoqué chapitre II s'apparente à une méthode descriptive permettant de dégager des résultats sur les variables étudiées. En second lieu, la mise en oeuvre d'une méthode basée sur l'utilisation des techniques d'inférences propres à la théorie des sous-ensembles flous permet la mise en évidence de relations entre variables.

Néanmoins, le but d'une analyse de système est de pouvoir dégager des commandes permettant d'agir sur le système. En gardant une approche systèmes à base de connaissance, il est alors nécessaire de pouvoir exploiter les résultats obtenus, le problème étant de pouvoir représenter les connaissances acquises sous forme de règles de production du type "si... alors...". Ce problème se ramène au délicat passage des données à une méthode de traitement dont on attend qu'elle fournisse des résultats sous cette forme /LOSLEVER 88a/.

Les résultats obtenus peuvent alors permettre dans certains cas, d'élaborer des règles de production. En reprenant l'exemple du poste bureautique du chapitre IV, les résultats issus de l'analyse des données subjectives peuvent permettre la formulation de règles. Deux exemples de règles peuvent être :

Exemple I :

SI le réglage correspond à HMP,
une assise haute et une table moyenne et un dossier proche du clavier

ALORS les sujets disent :
- qu'ils ont peu envie de bouger
- qu'ils sont peu fatigués
- qu'ils ont peu mal au dos
- que l'assise n'est ni trop haute ni trop basse
- que la table n'est ni trop haute ni trop basse
- que le dossier n'est ni trop loin ni trop proche

l'influence de la durée de l'expérience est faible

Exemple II :

SI la durée de la tâche augmente

ALORS les sujets disent que :
- leur envie de bouger augmente
- leur fatigue augmente
- leurs douleurs dans le dos augmentent

La mise en relation entre les variables ne permet pas, quant à elle, de dégager des règles de production. Néanmoins, dans une perspective plus lointaine, il est possible d'imaginer qu'une telle méthode puisse servir dans le cas de plusieurs experts du même domaine déterminant chacun une règle de production vague sur une même notion. La méthode développée permettrait de déterminer leur cohérence, voire de rejeter les experts considérés comme ayant un comportement particulier.

V.3 CONCLUSION

Ce dernier chapitre présente les perspectives envisagées pour la mise en relation entre variables. En premier lieu, la nécessité de dégager la robustesse d'une relation pour permettre l'extrapolation de cette dernière, en second lieu l'application pour les systèmes experts au cas de plusieurs experts du domaine.

Il présente enfin un bref aperçu de deux approches permettant dans des contextes et avec des outils différents d'analyser des systèmes complexes. Ces deux approches, les systèmes à base de connaissance et l'analyse des données ne répondent pas, en général, aux mêmes objectifs. La première se base sur des systèmes où l'expertise existe et est fiable, permettant à l'aide d'un, ou plusieurs, experts du domaine de dégager des connaissances et de commander le système. La deuxième tente dans des domaines trop complexes pour permettre une quelconque expertise de dégager des paramètres permettant la liaison de variables ou permettant la représentation graphique du système.

CONCLUSION GENERALE

L'objectif du travail présenté était l'analyse de données objectivo-subjectives fondée sur la théorie des sous-ensembles flous et plus particulièrement celle des sous-ensembles aléatoires flous. L'importance croissante des sous-ensembles flous pour la prise en compte de données incertaines et imprécises nous a amené, dans le contexte de système à composante humaine, à utiliser les outils mis en place par cette théorie.

Le premier chapitre rappelle alors les différents aspects de l'évaluation et de la modélisation des systèmes. Dans le cadre des systèmes à composante humaine le recueil de données subjectives pose alors le double problème du mode de recueil et de traitement de telles données.

Dans ce contexte, le deuxième chapitre fait appel à la théorie des sous-ensembles aléatoires flous pour le traitement de questionnaires dont les réponses sont portées sur un différenciateur sémantique permettant de laisser au répondeur une liberté importante. Une méthode descriptive de traitement de tels questionnaires a ainsi été élaborée.

La composante humaine du système requiert le recueil de données de nature subjectives et objectives. Il importe dans ce contexte de rechercher l'existence de relations éventuelles liant ces deux ensembles. Cette préoccupation fait l'objet du chapitre III où, à l'aide de l'inférence déductive et du modus ponens généralisé, une méthode de mise en relation a été élaborée.

Les méthodes mises au point dans les deux chapitres précédents ont alors été testées dans le cadre d'une évaluation d'un poste de travail bureautique. L'application de la méthodologie a permis de mettre en évidence l'efficacité de cette nouvelle approche.

Dans le dernier chapitre, les perspectives liées aux méthodes développées sont discutées avant de replacer l'approche d'analyse des données par rapport aux systèmes à base de connaissances dans le contexte général de l'extraction de connaissances.

Enfin, il nous semble important de poursuivre la recherche sur la mise en relation de données à l'aide des sous-ensembles aléatoires flous, l'originalité d'une telle mise en relation étant l'obtention d'une relation globale entre deux groupes de données. Dans ce contexte il s'agit d'une part, d'élargir l'application de la méthodologie à d'autres groupes de données et d'autre part, comme le souligne le dernier chapitre, de déterminer le domaine d'applicabilité et la robustesse des relations obtenues.

BIBLIOGRAPHIE

/BEHRAKIS et NICOLAPOULOS 87/ - T. BEHRAKIS, I. NICOLAPOULOS
Une typologie des forces politiques en Grèce
in Journal International pour l'analyse des grands tableaux et données d'enquête
T. Aluja Banet Marti Recorder éditeurs - Blanes

/BELLON 89/ - C. BELLON, P. BOSC, H. PRADE
Le "boum" du flou au Japon
Rapport I.N.R.I.A.

/BEN ZINEB et BOUZGHAYA 90/ - S. BEN ZINEB, E. BOUZGHAYA
Un essai de représentation des classes empiétantes : l'arbre à liaisons incomplètes
Rapport de projet - Ecole Nationale des Sciences de l'Informatique - Tunis février 1990

/BENKIRANE 91/ - M. BENKIRANE
Contribution à la méthodologie d'extraction des connaissances dans le domaine du diagnostic technique
Thèse de Doctorat - L.A.I.H. - Université de Valenciennes - 18 mars 1991

/BENZECRI 73/ - J.P. BENZECRI
Traité d'analyse des données : l'Analyse des Correspondances
DUNOD éditeur

/BERTIN 77/ - BERTIN
Le graphique et le traitement graphique de l'information
FLAMMARION éditeur

/BROSSIER 86/ - G. BROSSIER
Problèmes de représentation de données par des arbres
Thèse d'Etat - Université de Haute-Bretagne - Rennes 2

/CAO et KANDEL 89/ - Z. CAO, A. KANDEL
On the applicability of some fuzzy implication operators
Fuzzy Sets and Systems - n°31 - pp 151,186

/CHANDON et PINSON 81/ - J.L. CHANDON, S. PINSON
Analyse typologique : théories et applications
MASSON éditeur

/DEWITTE 86/ - P. DEWITTE
Analyse comparative de la perception selon la nature du support : papier ou écran
Rapport de D.E.A. - L.A.I.H. - Université de Valenciennes

/DIDAY 82/ - E. DIDAY, J. LEMAIRE, J. POUJET, F. TESTU
Eléments d'analyse des données
DUNOD éditeur

/DIDAY 86/ - E. DIDAY
Une représentation visuelle des classes empiétantes : les pyramides
R.A.I.R.O. APII - ATP spécial - vol 20 - n°5 - pp 475,526

/DUBOIS et PRADE 84/ - D. DUBOIS, H. PRADE
Fuzzy logics and the generalized modus ponens revisited
Cybernetics and Systems - vol 15 - pp 293,331

/DUBOIS et PRADE 87/ - D. DUBOIS, H. PRADE
Théorie des possibilités : application à la représentation des connaissances en informatique
MASSON éditeur - 2ème édition

/FABRE 80/ - J.M. FABRE
Jugement de certitude
P. Lang éditeur

/FARRENY 85/ - H. FARRENY
Les systèmes experts - principes et exemples
CEPADUES éditions

/FERON 76/ - F. FERON
Théorie des probabilités - Ensembles aléatoires flous
Compte Rendu Académie des Sciences Paris - tome 282 - pp 903,906

/GALLEGO 82/ - F.J. GALLEGO
Codage flou en analyse des correspondances
Les Cahiers de l'Analyse des Données - vol. VII - n°4 - pp 413,430

/GUERRA 85/ - T.M. GUERRA
Conception d'un système d'acquisition micro-informatique embarquable de mesures de signaux posturaux sur engin routier
Rapport de DEA - LAIH - Université de Valenciennes

/GUERRA 88a/ - T.M. GUERRA, P. DEWITTE, P. LOSLEVER, D. ROGER
Introduction de la subjectivité dans la formulation et le traitement des questionnaires fermés
ESIEE : Formation, Evaluation, Sélection par questionnaires fermés - pp 214,237

/GUERRA 88b/ - T.M. GUERRA, P. LOSLEVER, D. ROGER
Mise en relation de données objectives et subjectives
4th International Symposium on Applied Stochastic Models and Data Analysis - Nancy

/GUERRA et ROGER 91/ - T.M. GUERRA, D. ROGER
A method for getting in relation two data sets
International Fuzzy Systems Association - IFSA'91 - Brussels - 7,12 juillet 1991

/HIROTA 81/ - K. HIROTA
Concepts of probabilistic sets
Fuzzy Sets and Systems - 5 - pp 31,46

/KAMOUN 89/ - KAMOUN
Contribution à la répartition dynamique des tâches entre opérateur et calculateur pour la supervision des procédés automatisés
Thèse de Doctorat - L.A.I.H. - Université de Valenciennes - 17 avril 1989

/KAUFMANN 77/ - A. KAUFMANN
Introduction à la théorie des sous-ensembles flous - Tome I : éléments théoriques de base
MASSON éditeur - 2ème édition

/KAUFMANN 83/ - A. KAUFMANN
Sous-ensembles aléatoires flous et possibilité aléatoire
Note de travail n°105 - édité par l'auteur

/KAUFMANN 84/ - A. KAUFMANN
Avis d'experts et "expert systems" utilisant les sous-ensembles aléatoires flous
Note de travail n°118 - édité par l'auteur

/KAUFMANN 87/ - A. KAUFMANN
Nouvelles logiques pour l'intelligence artificielle
HERMES éditeur

/KISZKA 85/ - J. B. KISZKA, M.E. KOCHANSKA, D.S. SLIWINSKA
The influence of some fuzzy implication operators on the accuracy of a fuzzy model part I and part II
Fuzzy Sets and Systems - n°15 - pp 111,128 et pp 223,240

/LEBAILLY 85/ - J. LEBAILLY, R. MARTIN-CLOUAIRE, A. PRADE
Use of Fuzzy Logic in a Rule-bases System in Petroleum Geology
Computers in Earth Sciences for Natural Resources Characterization. Nancy

/LEPOUTRE 79/ - F.X. LEPOUTRE
Analyse et traitement des courbures du dos d'un sujet, dans le plan sagittal, modélisation et application biomécanique
Thèse de Docteur Ingénieur - LAIH - Université de Valenciennes

/LEPOUTRE et ROGER 82/ - F.X. LEPOUTRE, D. ROGER
Système vision-posture : ergonomie des sièges de bureau
Rapport final CETIM - LAIH - Université de Valenciennes

/LEPOUTRE 85/ - F.X. LEPOUTRE, P. CLOUP, T.M. GUERRA
Posture and dorsal shape at a sitted workstation
Biostereometrics'85 - Cannes - 3/6 décembre 1985

/LEPOUTRE et GUERRA 86/ - F.X. LEPOUTRE, T.M. GUERRA
Recherche de base sur l'autocar : ergonomie des postes de conduite, autocar de ligne. Système d'analyse de la posture dorsale
Rapport final INRETS - LAIH - Université de Valenciennes

/LOSLEVER 88a/ - P. LOSLEVER
Etude ergonomique du poste bureautique : approche par les méthodes multidimensionnelles d'analyse des données
Thèse de Doctorat - L.A.I.H. - Université de Valenciennes

/LOSLEVER 88b/ - P. LOSLEVER, T.M. GUERRA, D. ROGER
Analyse des questionnaires en ergonomie : l'appréciation des réglages d'un poste de travail
Les Cahiers de l'Analyse des Données - vol. XIII - n°2 - pp 175,187

/MAGREZ 85/ - P. MAGREZ
Modèles de raisonnement approché dans le cadre des systèmes experts médicaux
Thèse d'agrégation de l'enseignement supérieur - Bruxelles

/MAMDANI 77/ - E.H. MAMDANI
Application of fuzzy logic to approximate reasoning using linguistic synthesis
IEEE - Transactions on Computers - vol 26 - n°12

/MINGSHENG 88/ - Y. MINGSHENG
Some notes on multidimensional fuzzy reasoning
Cybernetics and Systems - n°19 - pp 281,293

/MIZUMOTO 82/ - M. MIZUMOTO
Fuzzy conditional inference under Max- composition
Information Sciences - n°27 - pp183,209

/MIZUMOTO et ZIMMERMANN 82/ - M. MIZUMOTO, H.J. ZIMMERMANN
Comparison of fuzzy reasoning methods
Fuzzy Sets and Systems - vol 8 - n°3 - pp 253,283

/MIZUMOTO 85/ - M. MIZUMOTO
Extended fuzzy reasoning
Approximate Reasoning in Expert Systems - M. GUPTA and al. editors
Elsevier Science Publishers B.V. (North-Holland) - 1985 - pp 71,85

/MOREAU 87/ - A. MOREAU
Contribution au traitement des informations subjectives dans les systèmes experts
Thèse de Doctorat - L.A.I.H. - Université de Valenciennes

/NASSIET 87/ - D. NASSIET
Contribution à la méthodologie de développement des Systèmes Experts : application au domaine du diagnostic technique
Thèse de Docteur Ingénieur - LAIH - Université de Valenciennes

/OLSON et RUETER 87/ - J.R. OLSON, H.H. RUETER
Extracting expertise from experts : methods for knowledge acquisition
Expert Systems - n°4 - pp 152,168 - Août 1987

/REGGIANI et MARCHETTI 88/ - M.G. REGGIANI, F.E. MARCHETTI
A proposed method for representing hierarchies
IEEE Transactions on Systems, Man, and Cybernetics - vol 18 - n°1 - pp 2,8

/RICHETIN et DUFOUR 79/ - M. RICHETIN, J. DUFOUR
Analyse structurale des systèmes complexes par l'analyse des données
dans A. TITLI et al. - *Analyse et commande des systèmes complexes* - pp 57,74
CEPADUES éditions

/STEVENS 74/ - S.S. STEVENS
Scaling : a sourcebook for behavorial scientists
Aldine Publishing Co., Chicago

/SUGENO et TAKAGI 83/ - M. SUGENO, T. TAKAGI
Multi-dimensional fuzzy reasoning
Fuzzy Sets and Systems - n°9 - pp 313,325

/SUGENO M., NISHIDA 85/ - M. SUGENO, M. NISHIDA
Fuzzy control of a model car
Fuzzy Sets and Systems - 16 - pp 103,113

/THOLE 79/ - U. THOLE, H.J. ZIMMERMANN, P. ZYSNO
On the suitability of minimum and product operators for the intersection of fuzzy sets
Fuzzy Sets and System - n°2 - pp 167,180

/VOLLE 81/ - M. VOLLE
Analyse des données
Economica éditeur - 2ème édition

/WALLISER 77/ - WALLISER
Systèmes et modèles. Introduction critique à l'analyse de systèmes
Editions du Seuil

/YAGER 80/ - R.R. YAGER
An approach to inference in approximate reasoning
International Journal of Man-Machine Studies - Vol. n°13 - pp 323,338

/ZADEH 65/ - L.A. ZADEH
Fuzzy sets
Information and Control - n°8 - pp 338,353

/ZADEH 75/ - L.A. ZADEH
Calculus of fuzzy restrictions
Fuzzy sets and their applications to Cognitive and Decision Processes - pp 1,40
Zadeh, Fu, Tanaka, Shimura editors

/ZADEH 77/ - L.A. ZADEH
A theory of approximate reasoning
Memorandum UCB/ERL - M77/58

/ZADEH 78/ - L.A. ZADEH
Fuzzy sets as a basis for a theory of possibility
Fuzzy Sets and Systems - n°1 - pp 3,28

/ZADEH 81/ - L.A. ZADEH
Test score semantics for natural languages and meaning representation via PRUF
Technical note n°247 - SRI - Menlo Park

ANNEXE I

ALGORITHME DE DECOMPOSITION EN SOUS-RELATIONS

MAXIMALES DE SIMILITUDE

L'algorithme de détermination des classes empiétantes (ou sous-relation maximales) issues d'une matrice **M** quelconque se base sur des remarques élémentaires issues des graphes associés. On remarque en effet qu'il n'y a que 4 façons possibles de construire de nouvelles classes à partir des classes précédentes.

Au préalable à la présentation de l'algorithme il est utile de donner quelques notations préliminaires.

A.1 - NOTATIONS

a - Seuils

Soit une matrice de dissemblance $\mathbf{M} = (m_{ij})$, l'ensemble des seuils S_k définis par la matrice **M** correspond à toutes les valeurs m_{ij} différentes ordonnées par ordre croissant.

Exemple :

$$\mathbf{M} = \begin{pmatrix} 0 & 0,1 & 0,3 & 0,2 \\ 0,1 & 0 & 0,1 & 0,2 \\ 0,3 & 0,1 & 0 & 0,3 \\ 0,2 & 0,2 & 0,3 & 0 \end{pmatrix} \quad \text{et} \quad S_k = \{ 0 ; 0,1 ; 0,2 ; 0,3 \}$$

b - Liaison

Les lignes et les colonnes de la matrice **M** étant notées α_i ($1 \leq i \leq \dim(\mathbf{M})$), une liaison $\alpha_i\alpha_j$ correspond à élément m_{ij} de la matrice, où α_i correspond à la ième ligne et α_j la jième colonne.

A un seuil donné peuvent correspondre une ou plusieurs liaisons.

Exemple : En reprenant l'exemple précédent, il vient :

au seuil 0.1 : les liaisons $\alpha_1\ \alpha_2$; $\alpha_2\ \alpha_3$
au seuil 0.2 : les liaisons $\alpha_1\ \alpha_4$; $\alpha_2\ \alpha_4$
au seuil 0.3 : les liaisons $\alpha_1\ \alpha_3$; $\alpha_3\ \alpha_4$

A2 - L'ALGORITHME DE CONSTRUCTION DES CLASSES EMPIETANTES

Soit une matrice de dissemblance $\mathbf{M} = (m_{ij})$ $i,j \in \{1,...,n\}$ où "n" est la dimension de la matrice et les α_k ($k \in \{1,...,n\}$), ses colonnes ou lignes, puisque la matrice est symétrique, mais aussi ses éléments à classer.

Les règles sur lesquelles est basé l'algorithme sont les suivantes :

Au départ, l'ensemble des classes est constitué par tous les singletons $\{\alpha_i\}$ ($i \in \{1,...,n\}$). En se plaçant à un seuil s_p, soit $C = \{C_z, z \in \{1,...,k\}\}$, l'ensemble des classes existantes à ce seuil où k représente le nombre de classes. Au seuil s_{p+1}, soit $Z = \{z_t = \alpha_i\alpha_j, t \in \{1,...,r\}\}$ l'ensemble des nouvelles liaisons créées. Le traitement effectué sur chaque liaison de Z est le suivant :

- la recherche de toutes les classes du seuil s_p qui contiennent d'une part α_i, soit alors $X_i = \{ C_z \in C / \alpha_i \in C_z \}$ l'ensemble de ces classes. D'autre part la recherche de toutes les classes du seuil s_p qui contiennent α_j, soit alors $X_j = \{ C_z \in C / \alpha_j \in C_z \}$ l'ensemble de ces classes. Les autres classes n'interviennent pas pour la liaison $\alpha_i\alpha_j$ considérée,

- la première étape est de vérifier si X_i ou X_j contient respectivement le singleton $\{\alpha_i\}$ ou $\{\alpha_j\}$. Dans ce cas, une seule nouvelle classe sera créée : $\{\alpha_i\alpha_j\}$ et la, ou les, classes singletons sont éliminées,

- dans le cas contraire, pour chacune des classes de X_i et X_j prises deux à deux les étapes sont les suivantes :

soient les deux classes : $C_{u1} \in X_i$ et $C_{u2} \in X_j$, elles sont telles que : $\alpha_i \in C_{u1}$ et $\alpha_j \in C_{u2}$ notons alors :

$$C'_{u1} = C_{u1} - \{\alpha_i\}$$
$$C'_{u2} = C_{u2} - \{\alpha_j\}$$

les nouvelles classes induites par la liaison $\alpha_i\alpha_j$ sont déterminées par l'un des quatre cas suivants :

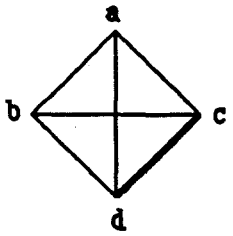
- 1er cas :

Si $C'_{u1} \cap C'_{u2} = \varnothing$ alors la nouvelle liaison n'a pas d'importance pour les deux classes C_{u1} et C_{u2} prises ensemble et l'investigation est poursuivie avec les autres classes.

- 2ème cas :

Si $C'_{u1} = C'_{u2}$ alors une nouvelle classe est créée au seuil s_{p+1} :
$CN = C'_{u1} + \{\alpha_i,\alpha_j\}$ et les 2 classes C_{u1} et C_{u2} disparaissent à ce seuil.

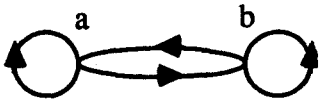
Exemple : $C_{u1} = \{a,b,c\}$ $C_{u2} = \{a,b,d\}$ et la nouvelle liaison $\alpha_i\alpha_j = cd$.



C = {a,b,c,d}

et les deux classes {a,b,c} et {a,b,d} n'ont plus lieu d'exister.

Remarque : Les traits reliant les éléments dans les graphes utilisés pour les exemples ne font pas apparaître la réflexivité et la symétrie pour permettre une meilleure lisibilité. En fait un trait :

a _____ b représente 

Les nouvelles liaisons qui apparaissent sont quant à elles tracées en trait plus foncé.

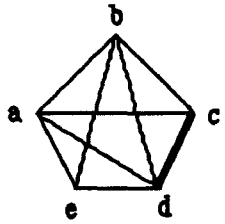
- 3ème cas :

Si $C'_{u1} \neq C'_{u2}$ et $C'_{u1} \cap C'_{u2} = C'_{u1}$ (resp. C'_{u2}) alors une nouvelle classe est créée au seuil s_{p+1} :

$CN = C'_{u1} + \{\alpha_i,\alpha_j\}$ (resp. $CN = C'_{u2} + \{\alpha_i,\alpha_j\}$)

et la classe C_{u1} (resp. C_{u2}) disparait à ce seuil.

Exemple : $C_{u1} = \{a,b,c\}$ $C_{u2} = \{a,b,d,e\}$ et la nouvelle liaison $\alpha_i\alpha_j = cd$



C = {a,b,c,d}

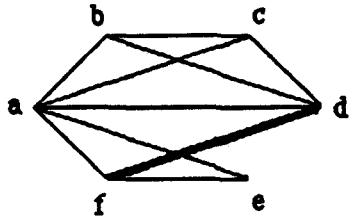
et {a,b,c} disparait alors que {a,b,d,e} reste.

- 4ème cas :

Si $C'_{u1} \neq C'_{u2}$ et $C'_{u1} \cap C'_{u2} \neq C'_{u1}$ et $\neq C'_{u2}$ mais $C'_{u1} \cap C'_{u2} \neq \varnothing$ alors une nouvelle classe est créée au seuil s_{p+1} :

$CN = C'_{u1} \cap C'_{u2} + \{\alpha_i,\alpha_j\}$ et les deux classes C_{u1} et C_{u2} restent à ce seuil.

Exemple : $C_{u1} = \{a,b,c,d\}$ $C_{u2} = \{a,e,f\}$ et la nouvelle liaison $\alpha_i\alpha_j = df$



C = {a,d,f}

et les deux classes {a,b,c,d} et {a,e,f} restent.

Enfin, de toutes les classes CN_x créées pour une même liaison, seules celles telles que:

$\forall$ i CN_x C CN_i sont conservées.

Exemple :

Soit pour un seuil s_p donné les classes existantes {abd}, {cdf} et {acd}, si au seuil s_{p+1} la liaison bc apparait, les nouvelles classes créées seront :

- une classe {bcd} créée à partir des classes {abd} et {cdf}
- une classe {abcd} créée à partir de {abd} et {acd} (ces 2 dernières disparaissant)

{bcd} étant incluse dans {abcd} les classes du seuil s_{p+1} seront : {cdf} et {abcd}.

En résumé, un organigramme de l'algorithme est présenté figure A1.

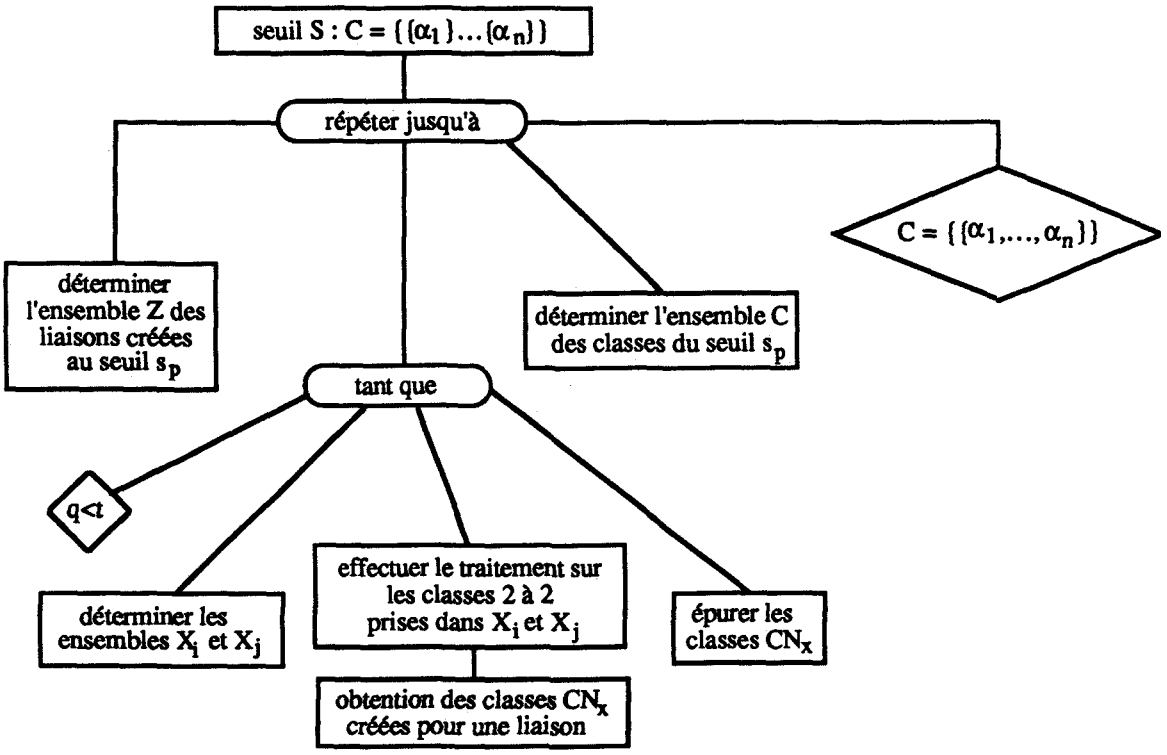


Figure A1 : organigramme général de l'algorithme

ANNEXE II

UN ESSAI DE REPRESENTATION DE CLASSES EMPIETANTES :

L'ARBRE A LIAISONS INCOMPLETES

Comme il a été dit chapitre II, on cherche à représenter une arborescence où toutes les classes issues de l'algorithme de l'annexe I sont présentes. Ω étant l'ensemble des classes, les seules liaisons apparentes sont celles qui lient deux classes C_1 et C_2 telles que si $C_1 \subset C_2$ et C_1 est une feuille :

$$\forall C_i \in \Omega \quad \nexists C_i \quad C_1 \subset C_i \subset C_2$$

Le problème est de pouvoir construire cet arbre de manière à ce qu'il ne puisse jamais y avoir de croisements et que les classes les plus "ressemblantes" - leur intersection est la plus grande - soient les plus proches possible dans l'arborescence.

Pour ce faire, un algorithme descendant a été mis au point.

B.1 - ALGORITHME DE CONSTRUCTION DE L'ARBRE A LIAISONS INCOMPLETES

a - Principe

La racine de l'arbre contiend la classe unique correspondant au seuil le plus élevé, 1er niveau. Les notations utilisées par la suite sont résumées figure B1.

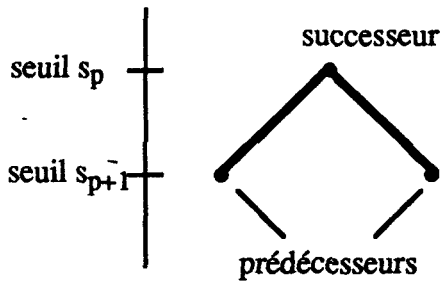


Figure B.1 : Notations utilisées

Etant donné les classes du seuil s_p, le placement d'une classe du seuil s_{p+1}, s'effectue par la recherche de la première classe du seuil précédent (s_p) qui la contient. Ce qui revient à rechercher pour une classe donnée de s_p les prédécesseurs qui ont été directement éliminés par la création de cette classe.

Après une première détermination du placement de chacune des classes d'un même niveau, ces dernières sont replacées en fonction de leur proximité à des classes ayant le plus d'éléments en commun. Ce "replacement" est effectué afin de faciliter la "lecture" de l'arbre.

b - Représentation de l'arbre

L'algorithme étant descendant, la racine de l'arbre est placée en premier. Comme celle-ci a obligatoirement un fils droit et un fils gauche elle se traite de la même façon qu'un noeud ayant deux prédécesseurs.

A un paramètre près, le traitement d'un sous-arbre droit ou gauche est identique. Seul le placement des noeuds et liaisons pour la partie "gauche" est alors expliqué.

La première étape de l'algorithme consiste à déterminer le nombre de feuilles de l'arbre pour procéder à un "découpage" de l'écran en "pas" constants. Ce "pas" sera déterminé par :

$$\text{"Pas"} = \frac{\text{taille de l'écran}}{\text{nombre de feuilles}}$$

La distance entre deux feuilles consécutives est donc toujours d'un "pas".

Si le calcul de l'ordonnée d'un noeud est directement déterminé à partir du seuil qui lui est associé le calcul de l'abcisse d'un noeud "i" se fait en distinguant trois types de noeuds :

- noeud ayant un fils gauche et un fils droit,
- noeud apparaissant dans une "cascade",
- une feuille.

** Noeud possédant un fils droit et un fils gauche*

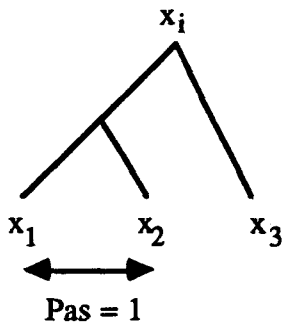
Soit N un tel noeud, et le sous-arbre issu de ce noeud. La première étape consiste à déterminer le nombre "n" de feuilles que contient le sous-arbre issu de N. Soit f_g le nombre de noeuds contenus dans le sous-arbre gauche issu de N et f_d le nombre de noeuds contenus dans le sous-arbre droit. Le noeud N est alors placé au barycentre de l'abcisse de la première feuille affectée du poids f_g et de la dernière affectée du poids f_d. De cette façon le sous-arbre droit ou gauche issu du noeud N le plus important - comportant le plus de feuilles - a le "plus de place".

$$x = (x_m + x_1)\frac{x_1 f_d + x_m f_g}{f_d + f_g} = \frac{x_m f_d + x_1 f_g}{f_d + f_g} = \left((n-1)\frac{f_g}{f_d + f_g} + 1\right) * \text{"Pas"}$$

où :

- x_1 est l'abcisse de la 1ère feuille du sous-arbre dont la racine est le noeud "i",
- x_m est l'abcisse de la dernière feuille du sous-arbre dont la racine est le noeud "i",
- n est le nombre de feuilles du sous-arbre issu du noeud "N".

Exemple :



"Pas" = 1
1ère feuille $x_1 = 1$
Dernière feuille $x_3 = 3$

Placement du $i^{ème}$ noeud x_i : $x_i = (2 * \frac{f_g}{f_d + f_g} + 1) * 1 = \frac{2 \times 2}{3} + 1 = \frac{7}{3}$

** Noeud d'une "cascade"*

C'est un noeud dont plusieurs successeurs consécutifs n'ont qu'un seul fils droit (respectivement gauche) s'il se trouve du côté droit (respectivement gauche).

Exemples :

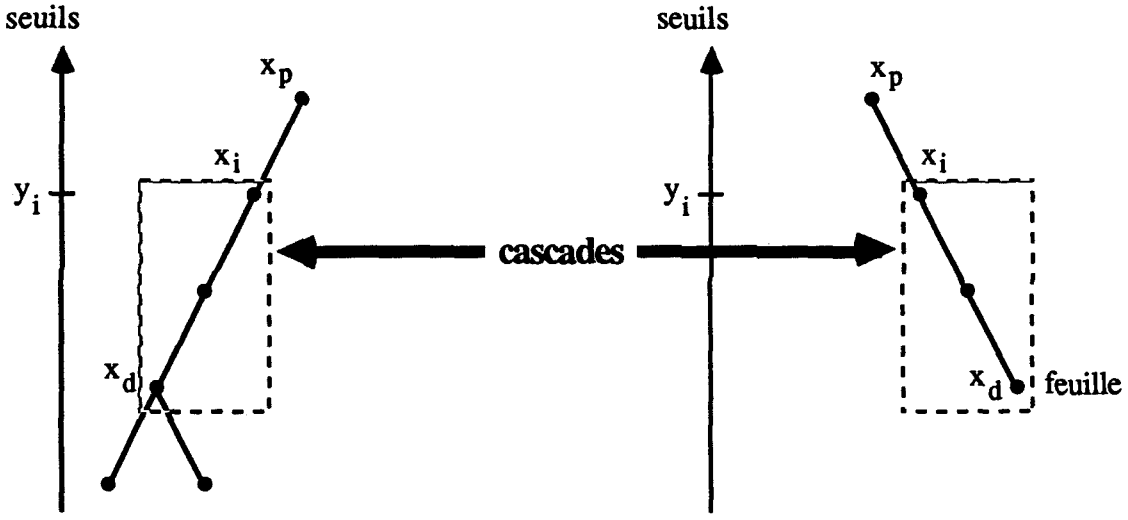


Figure B.2 : Les deux types de cascades

Il s'agit de calculer l'abcisse x_d du dernier noeud de la cascade - premier noeud qui a deux fils ou qui est une feuille. L'abcisse du noeud "i", sera alors, l'intersection de la droite passant par le point x_p - abcisse du père - et x_d et la droite y_i ordonnée du noeud "i".

$y = ax + b$

$y = y_i \quad \Rightarrow \quad a = \frac{y_p - y_d}{x_p - x_d}$ et $\quad b = y_p - \frac{y_p - y_d}{x_p - x_d} x_p = \frac{y_d x_p - y_p x_d}{x_p - x_d}$

donc :

$$y_i = \frac{y_p - y_d}{x_p - x_d} x + \frac{y_d x_p - y_p x_d}{x_p - x_d}$$

$$x = \frac{y_i (x_p - x_d) - y_d x_p + y_p x_d}{y_p - y_d}$$

** Le noeud "i" est une feuille*

S'il se trouve du côté gauche : $x = E(\frac{x_p}{\text{“Pas”}}) * \text{“Pas”}$

S'il se trouve du côté droit : $x = E(\frac{x_p}{\text{“Pas”}} + 1) * \text{“Pas”}$

c - Exemple de construction

On considère la matrice suivante :

$$\mathbf{M} = \begin{pmatrix} 0 & 0,1 & 0,2 & 0,5 & 0,5 \\ 0,1 & 0 & 0,1 & 0,2 & 0,4 \\ 0,2 & 0,1 & 0 & 0,5 & 0,5 \\ 0,5 & 0,2 & 0,5 & 0 & 0,1 \\ 0,5 & 0,4 & 0,5 & 0,1 & 0 \end{pmatrix}$$

L'algorithme de recherche des classes empiétantes donne les classes suivantes où a, b, c, d et e représentent les lignes ou colonnes de la matrice :

au seuil 0 : {a}, {b}, {c}, {d}, {e}
au seuil .1 : {ab}, {bc}, {de}
au seuil .2 : {abc}, {bd}, {de}
au seuil .4 : {abc}, {bde}
au seuil .5 : {abcde}

Avec un "Pas" = 1 il vient alors :

- Abcisse du noeud 1 : {abcde}

$n = 4, f_g = 2, f_d = 2 \qquad x = ((n-1) \frac{f_g}{f_d + f_g} + 1) * \text{“Pas”} = 3 * \frac{2}{4} + 1 = \frac{5}{2}$

- Abcisse du noeud 2 : {abc}

$y_p = 5, y_i = 4, x_d = \frac{3}{2}, y_d = 2, x_p = \frac{5}{2} \qquad x = \frac{4\left(\frac{5}{2} - \frac{3}{2}\right) - 2 * \frac{5}{2} + 5 * \frac{3}{2}}{3} = \frac{13}{6}$

- Abcisse du noeud 3 : {abc}

$n = 2, f_g = 1, f_d = 1 \qquad x = 1 * \frac{1}{2} + 1 = \frac{3}{2}$

- Abcisse de la feuille 4 : {ab}

$x_p = \frac{3}{2}$ $\qquad\qquad$ $x = 1 * 1 = 1$

- Abcisse de la feuille 5 : {bc}

$x_p = \frac{3}{2}$ $\qquad\qquad$ $x = (1 + 1) * 1 = 2$

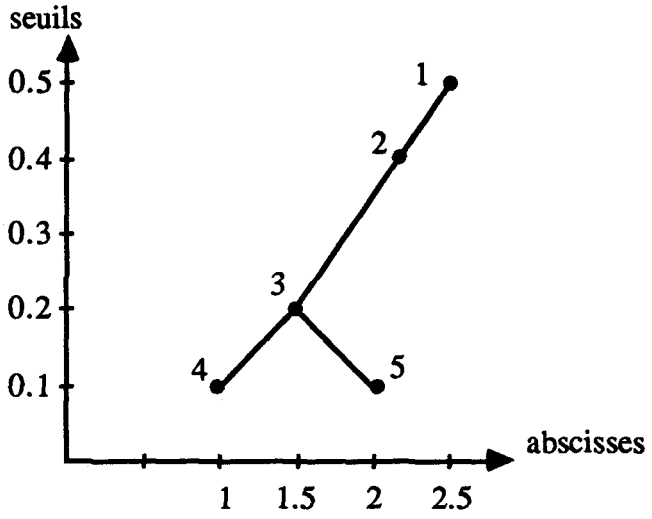


Figure B.3 : Construction du sous-arbre gauche

De la même manière est construit le sous-arbre droit et il vient alors la représentation suivante :

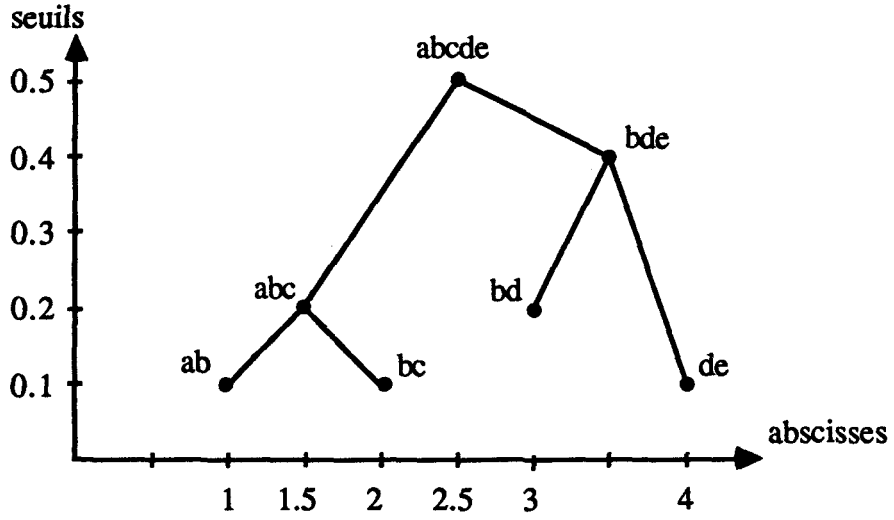


Figure B.4 : Exemple d'arbre à liaisons incomplètes

d. Interprétation

L'arbre à liaisons incomplètes étant représenté, sa "lecture", contrairement aux hiérarchies et aux pyramides, n'est pas aisée, ne faisant pas apparaître directement des partitions ou des recouvrements.

Pour déterminer à un seuil donné les classes, il suffit de "couper" l'arbre à ce seuil et de regarder quelles sont les classes directement précédentes. Par exemple, figure B.4, en "*coupant*" au seuil 0,3 les classes sont : {a,b,c} {b,d} et {d,e}.

L'interprétation d'un tel arbre se fait en effectuant une "lecture" simultanée horizontale et verticale :

- la première "lecture" est une lecture horizontale permettant de dégager la disparité des classes de départ - pour les premiers seuils -, c'est à dire déterminer si un élément appartient à une ou plusieurs classes et à un ou plusieurs sous-arbres. Si plusieurs classes, dans des sous-arbres différents contiennent le même élément ce dernier peut être considéré comme s'agrégeant bien, inversement un élément n'appartenant qu'à une classe, donc un seul sous-arbre, voire deux classes d'un même sous-arbre, ou de sous-arbres proches c'est à dire issus d'un même noeud, présente un comportement globalement opposé aux autres.

En reprenant l'exemple traité précédemment il vient :

a appartient à la classe {a,b} et au sous-arbre issu du noeud abc,

b appartient aux classes {a,b} {b,c} et {b,d} et aux sous-arbres issus des noeuds abc et bde,

c appartient à la classe {b,c} et au sous-arbre issu du noeud abc,

d appartient aux classes {b,d} et {d,e} et au sous-arbre issu du noeud bde,

e appartient à la classe {d,e} et au sous-arbre issu du noeud bde,

c'est à dire que b est l'élément s'agrégeant le mieux, a c et e les plus mal.

- cette première étape réalisée, une lecture verticale de l'arbre permet d'observer comment se regroupent les éléments et quels sont les sauts ($S_{p+1}-S_p$) importants qui permettent de déterminer les seuils où une autre "lecture" horizontale est nécessaire.

Dans l'exemple figure B.4 les deux sauts les plus importants ($S_{p+1}-S_p = 0.3$) permettent de passer respectivement des classes {a,b,c} et {d,e} aux classes {a,b,c,d,e} et {b,d,e}, les résultats de la classification sur cet exemple peuvent alors s'exprimer ainsi :

b est l'élément qui s'agrège le mieux, e est l'élément qui est globalement le plus loin des autres,

on peut constater que deux classes disjointes s'opposent {a,b,c} et {d,e}

Un logiciel de représentation de l'arbre à liaisons incomplètes a été mis au point /BEN ZINEB et BOUZGHAYA 90/ permettant à l'aide de trois fonctions de faciliter l'interprétation de l'arbre :

- un zoom permettant de visualiser un sous-arbre à partir de la sélection d'un noeud,
- une représentation de chemins permettant, à partir d'un singleton ou d'une classe de visualiser toutes les classes dont le singleton est un élément - dont la classe est une sous-classe - et toutes les liaisons liant ces différentes classes,
- une fonction permettant de visualiser la partie, inférieure ou supérieure, à un seuil donné.

Il est nécessaire de déterminer l'apport d'une telle méthode en la comparant avec les deux méthodes arborescentes qui ont été décrites chapitre II, hiérarchie et pyramide. Pour une classification entraînant des classes faiblement empiétantes, la matrice est proche d'une Robinson, des méthodes dites "classiques" se révèlent tout à fait adaptées et, les résultats de l'arbre à liaisons incomplètes sur de tels exemples concordent tout à fait.

Il est par contre intéressant de regarder quels sont les résultats pour des matrices entraînant des classes fortement empiétantes. Pour cela deux exemples particuliers ont été construits.

B2 - COMPARAISON AVEC LES HIERARCHIES ET LES PYRAMIDES

a - Premier exemple

Soit la matrice suivante : $\mathbf{M} =$

$$\begin{pmatrix} 0 & 0,1 & 0,1 & 0,1 & 0,1 & 0,3 \\ 0,1 & 0 & 0,4 & 0,8 & 0,7 & 0,8 \\ 0,1 & 0,4 & 0 & 0,3 & 0,5 & 0,4 \\ 0,1 & 0,8 & 0,3 & 0 & 0,2 & 0,3 \\ 0,1 & 0,7 & 0,5 & 0,2 & 0 & 0,6 \\ 0,3 & 0,8 & 0,4 & 0,3 & 0,6 & 0 \end{pmatrix}$$

Les hiérarchies et les pyramides déterminées à l'aide de trois indices d'agrégation sont présentées figures B5 à B7.

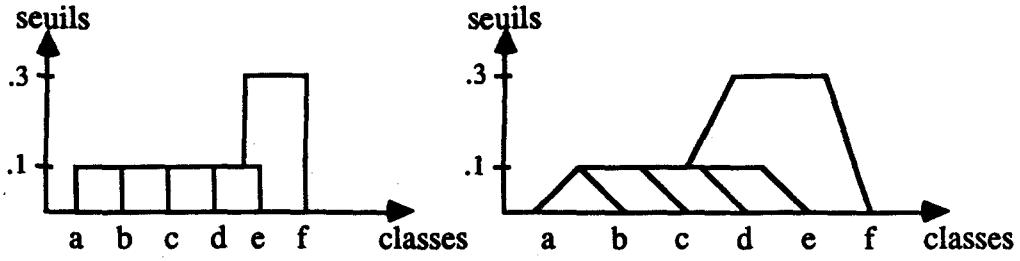


Figure B.5 : Hiérarchie et pyramide obtenues avec l'indice du saut minimum

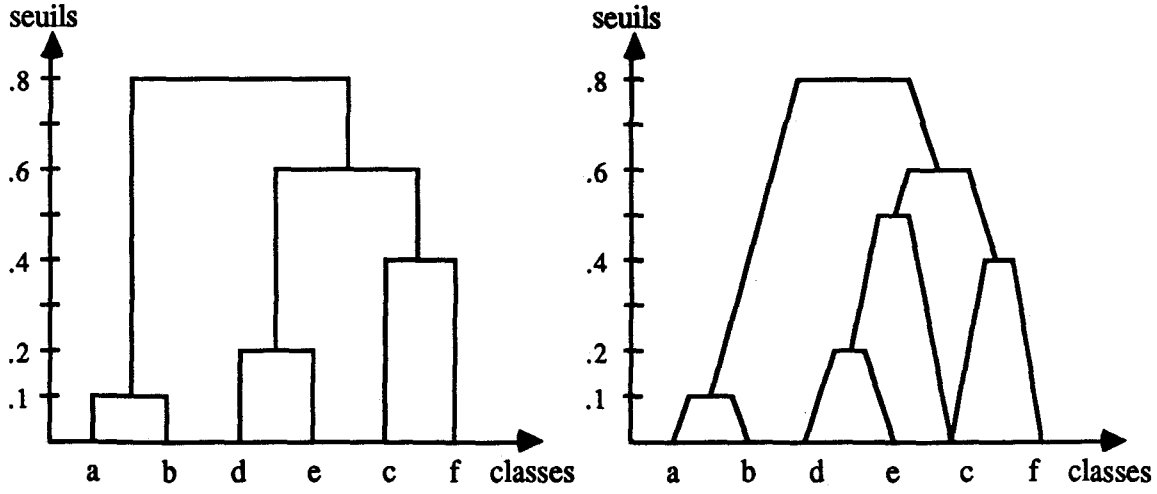


Figure B.6 : Hiérarchie et pyramide obtenues avec l'indice du saut maximum

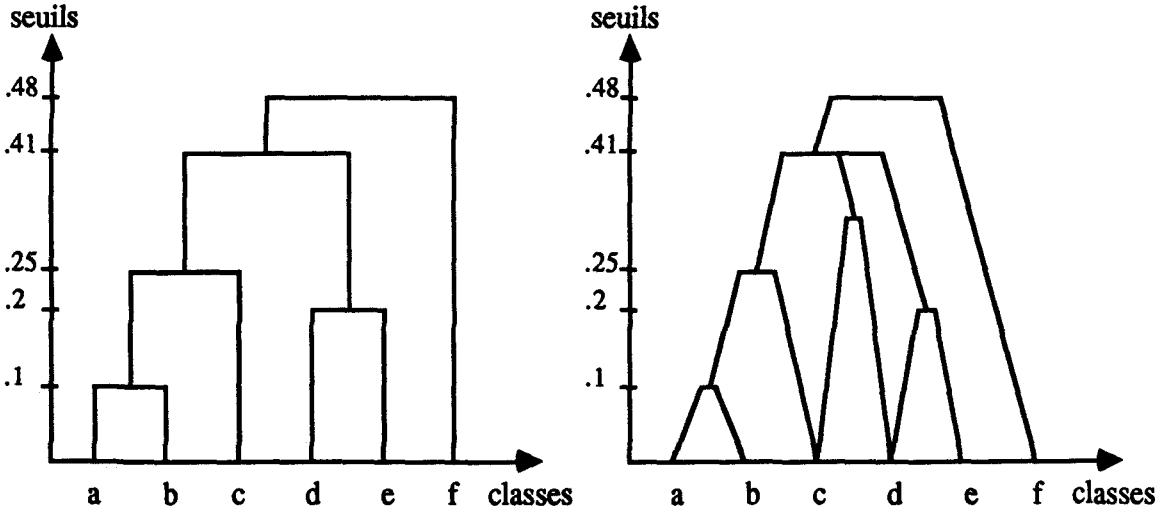


Figure B.7 : Hiérarchie et pyramide obtenues avec l'indice de la moyenne des distances

Remarquons tout d'abord que les trois indices donnent trois interprétations qui ne concordent pas. Effectivement en éliminant le saut minimum, peu adapté pour un tel type de matrice, il ressort les résultats suivants :

- pour les hiérarchies :

avec l'indice du lien maximum trois classes : {ab}, {de}, {cf}
avec la moyenne des distances trois classes : {abc}, {de}, {f}

- pour les pyramides :

avec l'indice du lien maximum : deux comportements, une classe {ab} et un groupement, plutôt qu'une classe, decf,

avec l'indice de la moyenne des distances : il ressort principalement que le singleton {f} est loin des autres éléments.

L'arbre à liaisons incomplètes issu de la matrice a été représenté la figure B.8

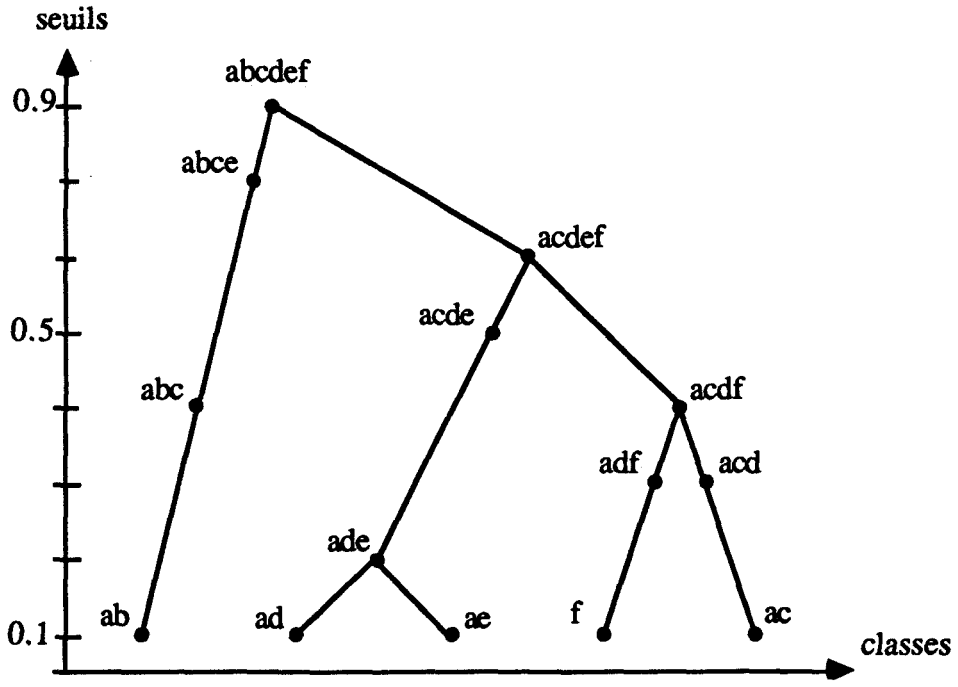


Figure B.8 : Arbre à liaisons incomplètes

L'interprétation issue de l'arbre à liaison minimale fait ressortir que les résultats trouvés à l'aide des 2 autres méthodes sont inexacts. En ce qui concerne la classe {ab}, celle-ci est bien présente dans l'arbre mais au même seuil que {ad}, {ae}, {ac} ce qui prouve que a est un élément qui s'agrège très bien. De même, si on peut remarquer que {f} est effectivement un singleton qui s'agrège "mal", il est important de noter que b est l'élément qui est le plus loin des autres. Ce dernier résultat n'a pu être mis en évidence par les autres représentations.

b - Deuxième exemple

Soit la matrice suivante : $\mathbf{M} = \begin{pmatrix} 0 & 0,1 & 0,8 & 0,15 & 0,9 & 0,1 \\ 0,1 & 0 & 0,1 & 0,85 & 0,1 & 0,8 \\ 0,8 & 0,1 & 0 & 0,1 & 0,7 & 0,15 \\ 0,15 & 0,85 & 0,1 & 0 & 0,1 & 0,7 \\ 0,9 & 0,1 & 0,7 & 0,1 & 0 & 0,8 \\ 0,1 & 0,8 & 0,15 & 0,7 & 0,8 & 0 \end{pmatrix}$

Les hiérarchies et les pyramides déterminées à l'aide de trois indices d'agrégation sont présentées figures B9 à B11.

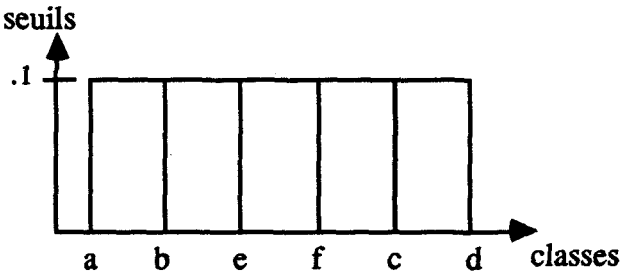


Figure B.9 : Hiérarchie obtenue à l'aide de l'indice du saut minimum

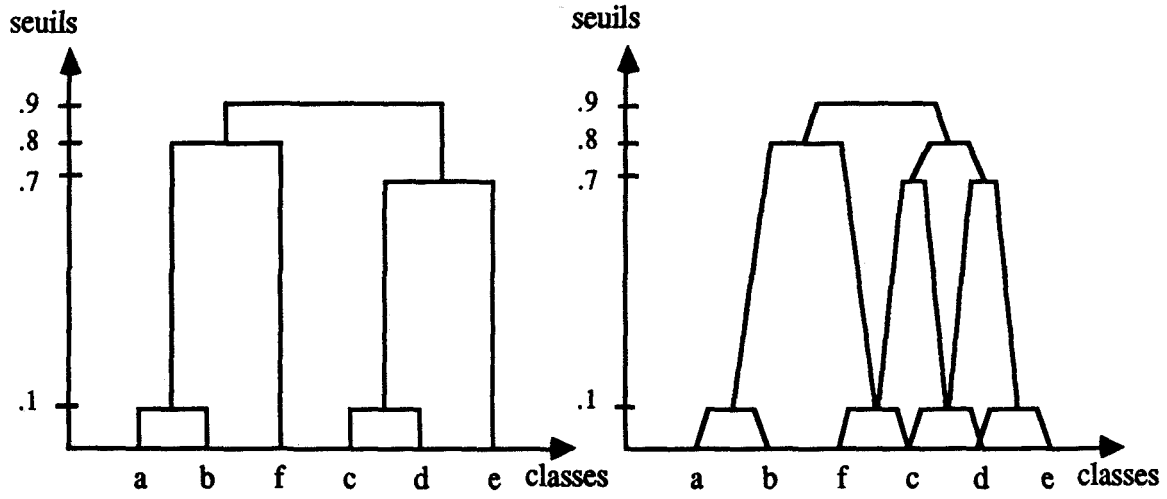


Figure B.10 : Hiérarchie et pyramide obtenues à l'aide de l'indice du saut maximum

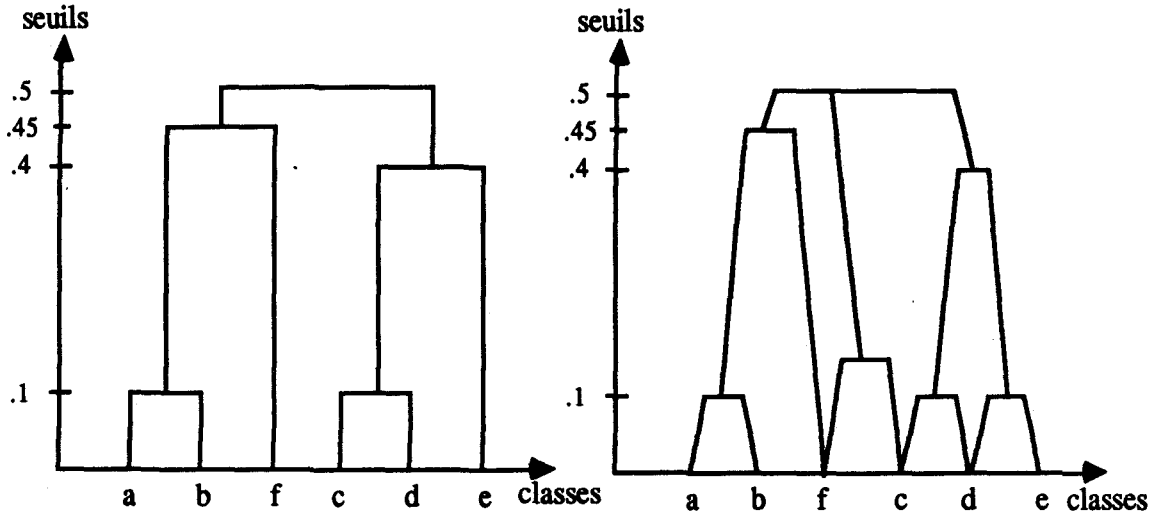


Figure B.11 : Hiérarchie et pyramide obtenues à l'aide de l'indice de la moyenne des distances

Le saut minimum ne pouvant convenir, l'interprétation des hiérarchies donne 3 types de comportements différents :
{a,b}, {c,d} et les singletons {f} et {e}. Constatons également que c'est l'agrégation de deux groupes de trois éléments qui forment la classe unique {abcdef}.

Pour les pyramides, le résultat est affiné, car si on retrouve bien {ab} on remarque qu'il n'y a pas à vrai dire un comportement {ce} mais plutôt un groupe cdef.

Le résultat principal issu de ces 2 méthodes est qu'il existe une classe {ab} qui "s'oppose" aux autres éléments.

L'arbre à liaisons incomplètes issu de la matrice a été représenté figure B.12

Sur cet arbre, on voit surtout "qu'il n'y a rien à voir", car si une classe {ab} existe bien, aux mêmes seuils les classes {af}, {ad} et {bc}, {be} sont également présentes. Vu la disparité des classes, il est important de remarquer que la classification ne peut donner aucun résultat. Effectivement, chaque élément aux deux premiers seuils, 0,1 et 0,15, appartient au moins à deux classes différentes, celles-ci appartenant de même à deux sous-arbres différents. D'autre part il faut remarquer l'importance des sauts qui permettent de passer des classes de deux éléments aux classes de trois indiquant ainsi que toutes les agrégations se font tardivement (≥ 0.7). La classification sur une telle matrice ne peut donc pas donner de résultats.

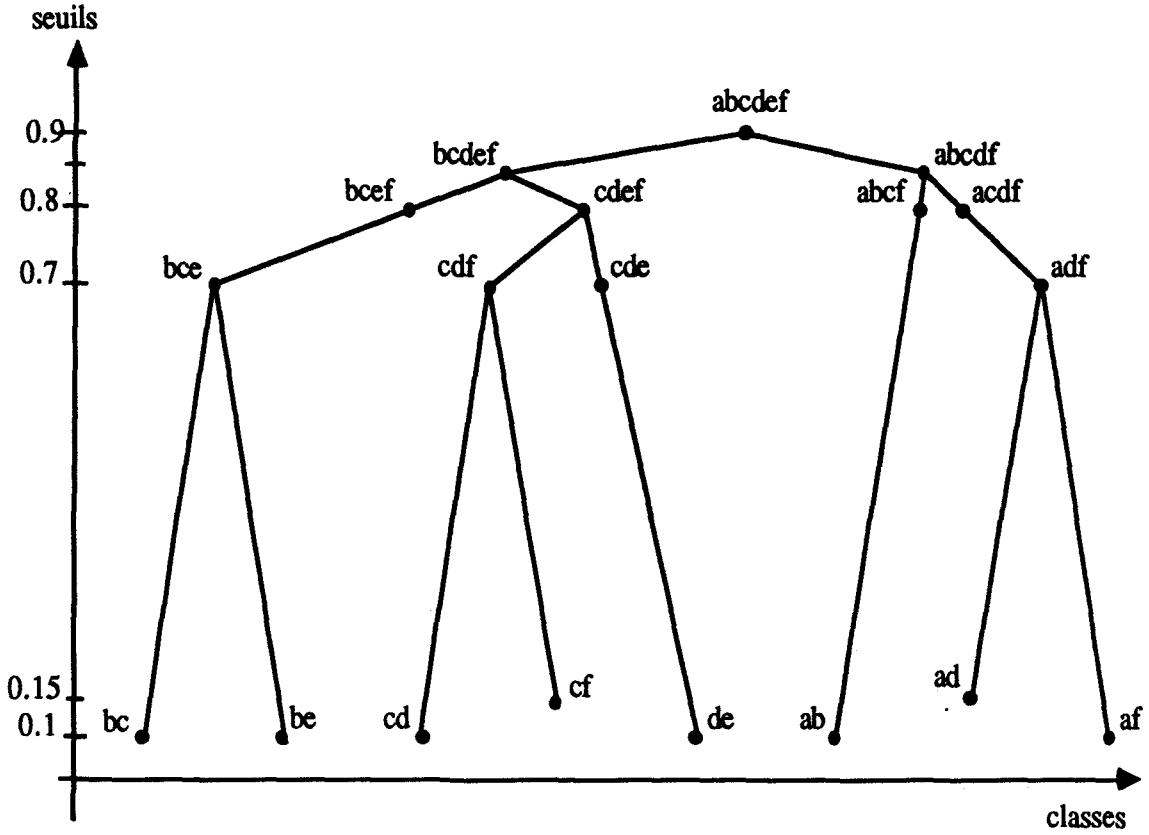


Figure B.12 : Arbre à liaisons incomplètes

c - Conclusion

La comparaison entre l'arbre à liaisons incomplètes et les hiérarchies et pyramides vient d'être traitée sur deux exemples. Il n'est pas dit qu'il n'existe pas d'indice d'agrégation pour les hiérarchies ou d'ordre pour les pyramides qui permettent d'obtenir des résultats similaires. L'avantage que présente un tel arbre est de représenter fidèlement la matrice de départ. Il n'y a donc pas de "tatonnement" préalable à l'utilisation d'un tel arbre alors que pour les hiérarchies surtout, il s'agit de justifier l'indice d'agrégation utilisé ce qui n'est pas toujours aisé.

En contrepartie, si hiérarchies et pyramides sont "faciles" à interpréter, il n'en va pas de même de l'arbre à liaisons incomplètes surtout quand le nombre d'éléments à classer dépasse 10. Il s'agit alors dans ces cas-là d'utiliser cette méthode comme appoint aux hiérarchies et pyramides pour valider, ou non, les résultats issus de ces méthodes.

Enfin, plus que les trois fonctions d'aide à la lecture de l'arbre qui ont été introduites dans le logiciel, une perspective serait de déterminer des règles - basées sur la disparité des classes et les sauts importants par exemple - qui permettent de guider l'utilisateur dans la lecture et l'interprétation de l'arbre.

ANNEXE III

DEUX EXEMPLES DE MISE EN RELATION DE DONNEES

Cette dernière annexe se rapporte à la méthodologie de mise en relation de deux groupes de données étudiée dans le chapitre III. Dans un premier temps, elle complète et étend les résultats obtenus pour l'exemple traité dans le chapitre III paragraphe 2. Dans un deuxième temps, la méthode est appliquée à un deuxième exemple.

Nous allons commencer par donner les 19 couples utilisés.

C.1 - COUPLES UTILISES

Les couples utilisés sont au nombre de 19, ils ont été choisis principalement dans /MIZUMOTO 82,85/ /MOREAU 87/. Le tableau suivant donne 9 premiers couples :

n° des couples	$\Pi_{y/x}$	opérateur Λ
1	$\min [1, 1 - \Pi_x(x) + \Pi_y(y)]$	$\max (0, a + b - 1)$
2	$\begin{cases} 1 & \text{si } \Pi_x(x) = 0 \\ \min [1, \frac{\Pi_y(y)}{\Pi_x(x)}] & \text{sinon} \end{cases}$	a.b
3	$\max [1 - \Pi_x(x), \Pi_y(y)]$	$\begin{cases} 0 & \text{si } a+b \leq 1 \\ b & \text{sinon} \end{cases}$
4	$1 - \Pi_x(x) + \Pi_x(x).\Pi_y(y)$	$\begin{cases} 0 & \text{si } a = 0 \\ \max [0, \frac{a+b-1}{a}] & \text{sinon} \end{cases}$
5	$\begin{cases} 1 & \text{si } \Pi_x(x) \leq \Pi_y(y) \\ 1 - \Pi_x(x) & \text{sinon} \end{cases}$	$\begin{cases} 0 & \text{si } a+b \leq 1 \\ a & \text{sinon} \end{cases}$
6	$\begin{cases} 1 & \text{si } \Pi_y(y) = 1 \\ \min [1, \frac{1 - \Pi_x(x)}{1 - \Pi_y(y)}] & \text{sinon} \end{cases}$	$\begin{cases} 0 & \text{si } b = 0 \\ \max [0, \frac{a+b-1}{b}] & \text{sinon} \end{cases}$
7	$\begin{cases} 1 & \text{si } \Pi_x(x) = 0 \\ \min [1, \frac{\Pi_y(y)}{\Pi_x(x)}] & \text{sinon} \end{cases}$	$\begin{cases} 0 & \text{si } b = 0 \\ \max [0, \frac{a+b-1}{b}] & \text{sinon} \end{cases}$
8	$\begin{cases} 1 & \text{si } \Pi_x(x) \leq \Pi_y(y) \\ \Pi_y(y) & \text{sinon} \end{cases}$	$\begin{cases} 0 & \text{si } a+b \leq 1 \\ a & \text{sinon} \end{cases}$
9	$\begin{cases} 1 & \text{si } \Pi_y(y) = 1 \\ \min [1, \frac{1 - \Pi_x(x)}{1 - \Pi_y(y)}] & \text{sinon} \end{cases}$	a.b

Les 19 couples choisis peuvent se scinder en plusieurs classes notamment par rapport à l'opérateur de *composition. Les 6 couples choisis dans* le chapitre III faisant partie de ces différentes classes, et par souci de numérotation cohérente, les couples numérotés 1 à 6 dans le chapitre III ne portent plus leur numéro. Ils sont renumérotés respectivement 1, 17, 9, 15, 10 et 6.

Le deuxième tableau donne toutes les distributions de possibilités conditionnelles $\Pi_{y/x}$ qui sont utilisées avec l'opérateur de composition min :

n° des couples	$\Pi_{y/x}$
10	$\min [(\Pi_x(x)\ S\ \Pi_y(y)) , (1 - \Pi_x(x))\ G\ (1 - \Pi_y(y))]$
11	$\min [(\Pi_x(x)\ G\ \Pi_y(y)) , (1 - \Pi_x(x))\ G\ (1 - \Pi_y(y))]$
12	$\begin{cases} 1 & \text{si } \Pi_x(x) \le \Pi_y(y) \\ 1 - \Pi_x(x) & \text{sinon} \end{cases}$
13	$\min [1, 1 - \Pi_x(x) + \Pi_y(y)]$
14	$\max [\min (\Pi_x(x) , \Pi_y(y)) , 1 - \Pi_x(x)]$
15	$\min [\Pi_x(x) , \Pi_y(y)]$
16	$\begin{cases} 1 & \text{si } \Pi_x(x) \le \Pi_y(y) \\ 0 & \text{sinon} \end{cases}$
17	$\begin{cases} 1 & \text{si } \Pi_x(x) \le \Pi_y(y) \\ \Pi_y(y) & \text{sinon} \end{cases}$
18	$\max [1 - \Pi_x(x), \Pi_y(y)]$
19	$\begin{cases} 1 & \text{si } \Pi_x(x) \le \Pi_y(y) \\ \dfrac{\Pi_x(x)}{\Pi_y(y)} & \text{sinon} \end{cases}$

Les 19 couples étant définis, l'exemple traité au chapitre III paragraphe 2 est repris sur ces différents couples.

C.2 - PREMIER EXEMPLE

L'exemple traité dans le chapitre III permet sur les 6 couples testés de dégager 4 comportements différents que nous allons classer :

classe I :

Le couple utilisé ne peut en aucun cas permettre la validation de la relation - cas du couple 6.

classe II :

le couple utilisé ne permet pas, après exclusion des experts 14 et 15 de valider la relation, les valeurs de $d(O_i,RO_i)$ étant jugées trop grandes pour plusieurs experts - cas du couple 10 (5 dans le chapitre III).

classe III :

le couple utilisé ne permet pas, après exclusion des experts 14 et 15 de valider la relation, les valeurs de $d(O_i,RO_i)$ et de $|d(TYPO,O_i) - d(TYPO,RO_i)|$ étant jugées trop grandes pour l'expert 10 - cas du couple 15 (4 dans le chapitre III).

classe IV :

le couple utilisé permet la validation la relation après exclusion des experts 14 et 15 - cas des couples 1, 17 et 9 (1, 2 et 3 dans le chapitre III).

n° couple	d_{mp}	d_{mt}	d1	d2	classe	validation
1	0,071	0,202	0,204	0,076	V	sans 5
2	0,063	0,203	0,188	0,078	V	sans 5
3	0,122	0,287	0,206	0,149	III	non
4	0,164	0,355	0,195	0,177	V	sans 5
5	0,137	0,231	0,342	0,172	I	non
6	0,091	0,205	0,342	0,172	I	non
7	0,100	0,224	0,342	0,172	I	non
8	0,128	0,234	0,342	0,172	I	non
9	0,058	0,133	0,176	0,076	IV	oui
10	0,123	0,172	0,228	0,035	II	non
11	0,055	0,214	0,155	0,081	II	oui
12	0,059	0,137	0,141	0,076	IV	oui
13	0,104	0,248	0,204	0,076	V	sans 5
14	0,053	0,197	0,202	0,139	III	non
15	0,006	0,350	0,230	0,219	III	non
16	0,076	0,153	0,176	0,081	IV	oui
17	0,052	0,205	0,139	0,085	IV	oui
18	0,053	0,201	0,206	0,149	III	non
19	0,092	0,230	0,188	0,078	V	sans 5

Figure C.1 : tableau récapitulatif des 4 paramètres utilisés pour les 19 couples

Les 19 couples se séparent alors suivant ces 4 classes. Le tableau figure C.1 résume alors sur tous les couples les résultats obtenus pour l'ensemble des experts Ex2 = Ex - {14,15}. La signification des 4 paramètres est rappelée ci-dessous :

d_{mp} : dérive par modus ponens $\quad$ d_{mt} : dérive par modus tollens

$$d1 = \max_{Ex2} d(O_i,RO_i) \qquad d2 = \max_{Ex2} \left| d(TYPO,O_i) - d(TYPO,RO_i) \right|$$

En dehors des remarques faites au chapitre III sur les 6 couples étudiés, remarques qui s'étendent aux 19 couples, une cinquième classe de couples se dégage. Celle-ci correspond à des valeurs trop élevées pour d1 et/ou d2 pour l'expert 5. Dans ce cas là, la relation est reconstruite sur 12 experts, en éliminant l'expert 5, et elle est validée. Ce cas apparait dans le tableau figure C.1 dans la colonne validation par "sans 5". Les valeurs des 4 paramètres ne sont pas redonnées dans ce dernier cas.

Après l'extension de la méthodologie sur l'exemple traité dans le chapitre III à 19 couples, un nouvel exemple est abordé.

C.3 - DEUXIEME EXEMPLE

Soient les deux ensembles suivants construits respectivement sur 8 variables qui seront dites "subjectives" et 6 variables "objectives".

Ex \ Su	1	2	3	4	5	6	7	8
1	1	0	1	1	0	1	0	0,7
2	0,6	0,2	0,6	0,7	0,3	0,7	0,3	0,3
3	0,4	0	1	0,5	0,5	0,6	0,4	0,2
4	0,5	0,2	0,8	0,8	0,5	0,6	0,3	0,8
5	1	0,4	1	0,9	0	1	0	0,2
6	1	0,2	0,7	0,8	0,1	1	0	0,7
7	1	0	0,9	0,5	0	0,5	0,4	0,6
8	0,8	0	1	1	0,2	1	0	1
9	1	0	0,9	0,7	0	0,4	0,5	0,6
10	0,4	0,4	0,2	0,2	0,4	0,2	0,2	0,1
11	0,4	0,6	0,5	0,5	0,4	0,4	0,6	0,2
12	1	0,4	1	0,3	0,3	0,9	0,1	0,9
13	1	0	0,2	0,9	0,3	1	0,9	0,9
14	1	0,1	0,1	1	0,1	0,4	0,2	0,1
15	0	1	0	0	1	0,8	0,2	0,4

Ex \ Ob	1	2	3	4	5	6
1	1	0,5	0,7	1	0,9	1
2	0,7	0,3	0,3	0,7	0,7	0,7
3	1	0,5	0,7	1	1	1
4	0,8	0,7	0,8	0,7	0,8	0,8
5	1	0,5	0,6	1	0,8	1
6	1	0,5	0,7	1	0,9	1
7	1	0,4	0,6	1	1	1
8	1	0,8	1	1	1	0,9
9	1	0,5	0,6	0,9	1	1
10	0,4	0,2	0,4	0,3	0,4	0,2
11	0,6	0,5	0,6	0,4	0,3	0,6
12	1	0,7	0,9	1	1	1
13	1	0,9	0,9	1	0,8	0,8
14	0	0,5	1	0	0,7	1
15	0	0,2	0,4	0	0,7	0

Les couples utilisés pour rechercher une relation entre ces deux ensembles sont bien entendu les 19 cités plus haut.

Avant d'aller plus loin dans l'analyse donnons de suite les couples inadaptés à toute mise en relation - correspondant aux classes I et II définies plus en avant -, il s'agit des mêmes que pour l'exemple I, à savoir les couples 5, 6, 7 et 8 pour la classe I et 10 et 11 pour la classe II. Ces 6 couples ne seront plus traités dans la suite.

Les plans de vérification et de validation ne sont présentés que pour les couples 1 et 15. 2 premiers plans de vérification sont présentés figure C.2

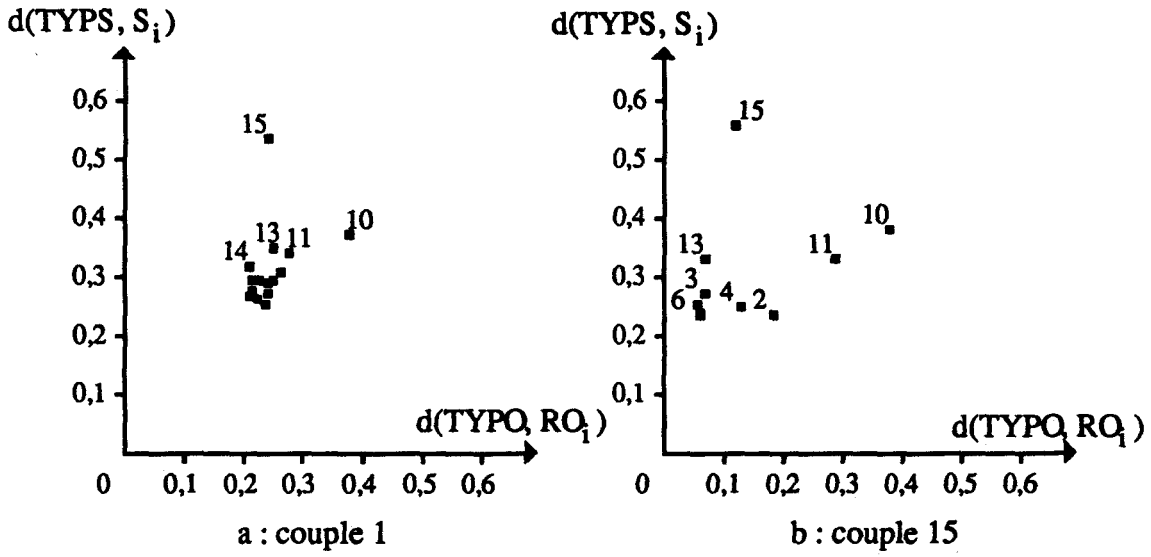


Figure C.2 : plans de vérification

Remarquons que le critère I - tous les points doivent être situés de façon croissante suivant les deux axes - est quelque peu "malmené". Si l'expert 15 présente une inversion importante - il devra être éliminé de l'analyse - un groupe d'experts présente également des inversions. Ceci est dû à la grande hétérogénéité des données. Effectivement, pour confirmer ce dernier point, le tableau figure C.3 donne les distances de chaque expert à TYPS et TYPO. Dans le cas de données aussi hétérogènes, le SEAF type n'est plus à proprement parler un bon représentant des données prises en compte. L'idéal serait de lui substituer un SEAF fixe par rapport aux données.

Néanmoins, le critère I reste globalement vérifié sur 14 experts pour le couple 1, les experts 10, 11 et 13 ayant les plus grandes distances par rapport à TYPS se retrouvant les plus isolés sur le plan figure C.2a. Quant au couple 15, il semble que l'expert 13 présente lui aussi une inversion importante. Il est conservé dans l'analyse, la suite de la mise en relation déterminera son adéquation aux autres experts .

expert	d(TYPS, S_i)	d(TYPO,O_i)
1	0,2900	0,1944
2	0,2550	0,2844
3	0,2958	0,2000
4	0,2733	0,2333
5	0,2933	0,1944
6	0,2650	0,1944
7	0,2683	0,2089
8	0,3075	0,2544
9	0,2767	0,2022
10	0,3733	0,4389
11	0,3408	0,3244
12	0,2967	0,2289
13	0,3508	0,2533
14	0,3192	0,3911
15	0,5342	0,5167

Figure C.3 : tableau des distances de chaque expert par rapport aux SEAF TYPS et TYPO

L'expert 15 ayant été éliminé, l'analyse est reprise sur Ex1 = Ex - {15}. Les plans de vérification et de validation pour les couples 1 et 15 sont présentés respectivement figures C.4 et C.5.

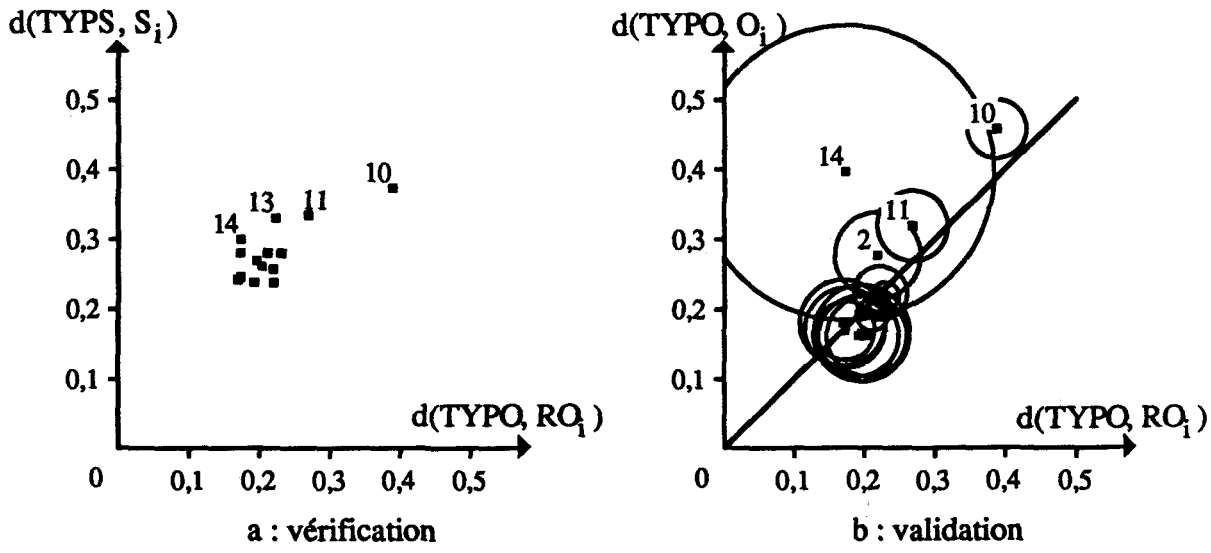


Figure C.4 : plans de vérification et de validation présentés pour le couple 1 sur Ex1

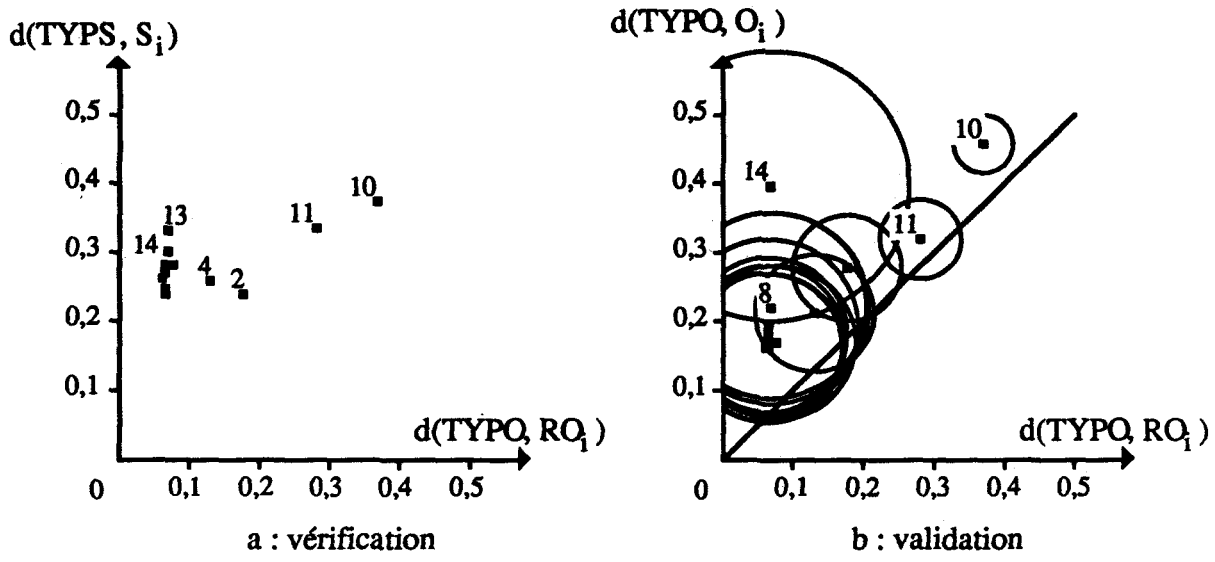


Figure C.5 : plans de vérification et de validation présentés pour le couple 15 sur Ex1

L'expert 14 est bien entendu retiré de l'analyse pour le couple 1, figure C.4b, et le couple 15 n'a que peu de chances de valider la mise en relation au vu du plan de validation figure C.5b.

En conséquence, deux ultimes plans sont présentés figure C.6, correspondant au couple 1. La relation est considérée comme validée dans ce cas là.

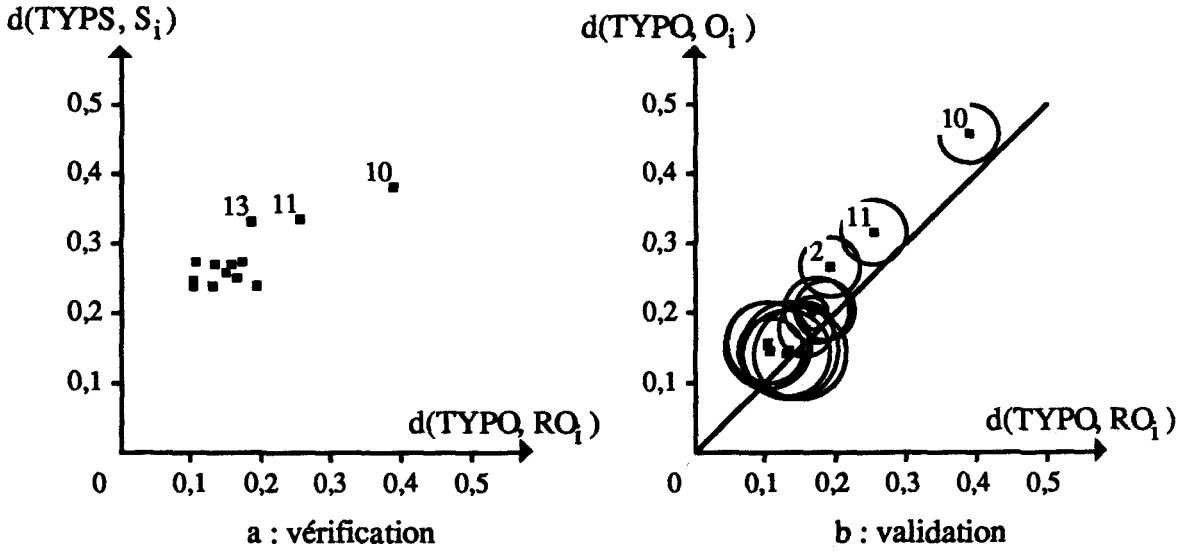


Figure C.6 : plans de vérification et de validation présentés pour le couple 1

Enfin, le tableau figure C.7 résume les résultats pour 13 couples sur la mise en relation des 13 premiers experts, 6 ayant été éliminés dès le départ de l'analyse au vu de leurs réultats négatifs.

n° couple	d_{mp}	d_{mt}	d1	d2	validation
1	0,076	0,231	0,135	0,068	oui
2	0,081	0,231	0,131	0,068	oui
3	0,094	0,310	0,131	0,113	oui
4	0,140	0,371	0,128	0,068	oui
9	0,112	0,149	0,139	0,084	oui
12	0,117	0,156	0,141	0,072	oui
13	0,100	0,232	0,135	0,068	oui
14	0,073	0,232	0,156	0,115	oui
15	0,055	0,312	0,256	0,141	non
16	0,138	0,148	0,164	0,065	oui
17	0,087	0,232	0,126	0,077	oui
18	0,067	0,232	0,131	0,113	oui
19	0,099	0,232	0,131	0,068	oui

Figure C.7 : tableau récapitulatif des 4 paramètres utilisés pour 13 couples

CONCLUSION

Les deux exemples traités, bien que différents amènent tous les deux à un résultat de mise en relation. Un ou plusieurs couples peuvent alors être choisis pour mettre en relation les données. Il semble également que certains couples ne sont pas adaptés, soit à la méthodologie, soit qu'un exemple n'ait pu être trouvé permettant de valider leur utilisation. Enfin, notons que certains couples sont adaptés aux deux exemples, alors que certains ne sont adaptés qu'au premier.

Les résultats obtenus sont très encourageants pour continuer la recherche à l'aide de cette méthodologie.