Méthodes statistiques et informatiques pour le traitement des données manquantes

Weila Vila Gu Co

THESE

présentée en vue de l'obtention du DOCTORAT du

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

PARIS

Spécialité

INFORMATIQUE

par **CO Vila**

Sujet de la thèse

METHODES STATISTIQUES ET INFORMATIQUES POUR LE TRAITEMENT DES DONNEES MANQUANTES

Présentée le 8 mars 1997 devant le jury composé de :

M. Ludovic **LEBART,** Rapporteur, C.N.R.S., E.N.S.T., Paris

M. Michel **LEJEUNE,** Professeur, C.N.A.M., Paris

M. Jan L.A. Van **RIJCKEVORSEL,** Rapporteur, TNO, Leiden

M. Gilles **SANTINI,** Président Directeur Général IMS, Paris

M. Gilbert **SAPORTA,** Directeur de Recherche, C.N.A.M., Paris

M. Jean **SOUSSELIER,** Président Directeur Général STATIRO, Paris

REMERCIEMENTS

A Monsieur le Professeur Gilbert SAPORTA, mon directeur de thèse, sans qui cette thèse n'aurait certainement jamais vu le jour, je ne saurais jamais assez le remercier de m'avoir aidé à mener à bien ce travail par sa disponibilité, ses encouragements, ses conseils avisés et sa gentillesse.

A Monsieur le Professeur Michel LEJEUNE pour son attention portée à la lecture et de m'avoir fait bénéficier de ses judicieuses remarques et d'avoir accepté de faire partie du jury de cette thèse.

A Messieurs le Professeur Ludovic LEBART et le Docteur Jan L.A. Van RIJCKEVORSEL pour l'honneur qu'ils m'ont fait en acceptant d'être rapporteurs et pour leurs remarques judicieuses qui m'ont été d'une très grande utilité.

A Messieurs Jean SOUSSELIER et Gilles SANTINI pour les discussions que nous avons eues, leurs nombreux conseils et suggestions et pour leurs participations à ce jury de thèse.

A Monsieur le Professeur Robert CLEROUX pour les discussions que nous avons eues lors de ses séjours en France.

A toute ma famille pour son soutien et sa présence à mes côtés.

Au C.N.A.M. de Paris pour son soutien matériel, au Département de Mathématiques et particulièrement aux membres de la Chaire de Statistiques qui m'ont accueillie et encouragée.

A STATIRO qui m'a permis de préparer cette thèse dans le cadre d'une convention de recherche CIFRE.

A Ruilin REN, Béatrice DUCRE, Jean-Jacques LOCHE, Chaochau LOCHE, Ndèye NIANG, Luan JAUPI et à toutes celles et ceux qui m'ont apporté leur appui.

RESUME

La présence des données manquantes est très fréquente en pratique En l'absence du traitement approprié, des statistiques qui n'en tiendraient pas compte seraient fortement biaisées.

Cette thèse est consacrée à des méthodes d'estimation de données manquantes, en particulier qualitatives.

Nous nous intéressons à la méthode de l'analyse homogène développée par des chercheurs néerlandais qui peut reconstituer des données manquantes qualitatives du type non complètement aléatoires.

Comme l'Analyse en Composantes Principales (A.C.P.) et la classification automatique sont deux techniques fréquemment employées dans le dépouillement d'enquête, nous avons mis au point une méthode pour l'A.C.P. et une pour la classification automatique avec données incomplètes. Les résultats sont validés par reconstitution de données connues mais cachées, dans des cas réels et simulés.

La deuxième partie de la thèse est consacrée à la fusion de fichiers. C'est un outil indispensable pour rassembler des informations provenant de différentes sources. La méthode de l'analyse homogène y est importée et adaptée à la fusion de fichiers. L'évaluation de la méthode est faite simultanément sur données réelles et simulées.

Table des matières

PARTIE II : LA FUSION DE FICHIERS

PARTIE I

TRAITEMENT DES DONNEES MANQUANTES

INTRODUCTION

En analyse de données multidimensionnelles, on rencontre fréquemment des données manquantes : les non-réponses dans une enquête, les pannes apparues au cours d'un processus expérimental industriel, la perte de document ou une bande de données abîmée etc. Cependant si on supprime les unités ayant des données manquantes, la population retenue diffère alors de celle de l'étude et en conséquence, les résultats peuvent être biaisés. De plus, si le taux de réponses manquantes est élevé, le fichier résultat peut être très petit. Pour éviter cette élimination, on doit procéder à un traitement des données manquantes. Pour cela, le statisticien devra préciser le mécanisme qui engendre les données manquantes et savoir si elles peuvent être ignorées ou non pour pouvoir choisir un traitement approprié. Nous étudierons des méthodes relativement robustes de traitement des données manquantes. Selon la nature et le taux des données manquantes (noté *D.M.*), on a deux approches différentes :

1. Estimer les *D.M.* en recourant à des modèles implicites ou explicites.
2. Adapter la technique d'analyse à des données incomplètes, autrement dit s'en accommoder.

De nombreux travaux traitent seulement le cas où les données manquantes sont complètement aléatoires. Les techniques classiques de traitement des données manquantes supposent souvent que les données suivent une loi normale. Peu d'ouvrages traitent les données manquantes qualitatives qui ne sont pas complètement aléatoires.

L'absence de réponses est préjudiciable à la qualité des résultats mais malgré les avantages que procure le traitement des données manquantes, de nombreux praticiens en ignorent l'importance. Comme le dit J. Meulman dans sa thèse (1982). "... Although statisticians have long appreciated that the existence of such missing information can change an ordinarily simple statistical analysis into a complex one... and responded to this challenge by producing enormous amounts of literature... there is little indication that survey researchers have paid much attention to the literature".

Les objectifs de cette thèse sont de décrire un ensemble de techniques bien adaptées à certains cas particuliers de traitement des données manquantes, en y intégrant des travaux épars réalisés par des équipes Françaises (Nora, Escofier, Benali et Caron) et Néerlandaises (De Leeuw, Gifi, Meulman, Van Buuren et Van Rijckevorsel). Si les français parlent d'analyse des correspondances, les néerlandais utilisent plus volontairement le terme d'analyse homogène ("homogeneity analysis") qui conduit à la même méthode mais dont le critère permet d'incorporer plus aisément les données manquantes. Nous présenterons également une technique d'analyse en composantes principales (ACP) adaptée et une technique de classification automatique applicable à des données incomplètes.

Le Premier Chapitre expose l'origine et la nature des données manquantes, l'historique et le développement du traitement, notamment des notions générales et des notations fondamentales employées dans le domaine du traitement des données manquantes.

Le Second Chapitre présente les principales méthodes du traitement des données manquantes. En particulier, on s'intéresse à l'analyse homogène qui consiste à estimer les *D.M.* qualitatives dont le type n'est pas complètement aléatoire. L'avantage de cette méthode est qu'elle ne nécessite aucune hypothèse distributionnelle des données. Le principe de la méthode sera détaillé et un exemple y sera illustré.

Le Troisième Chapitre est consacré à l'étude de méthodes classiques de l'analyse des données dans le cas de données incomplètes. Nous présenterons l'analyse en composantes principales et la classification automatique et pour chacune, une technique appropriée. Nous effectuerons parallèlement deux analyses sur des tableaux de données. Nous testerons et comparerons l'efficacité des méthodes proposées avec les techniques existantes sur la base de simulations. Les avantages et les inconvénients de chacune seront aussi expliqués. Nous exposerons également des méthodes d'analyse factorielle des correspondances avec données incomplètes.

CHAPITRE I: GENERALITES SUR LE TRAITEMENT DES DONNEES MANQUANTES

1.1. ORIGINE DES DONNEES MANQUANTES

On distingue les données manquantes partielles où manque une partie des réponses d'un individu, des données manquantes totales (observation entièrement absente).

Les données manquantes partielles :

- les données manquantes partielles dans une enquête ; les non réponses proviennent de plusieurs sources (Rubin et Little, 1986) :
  - le refus de répondre à une question (par exemple : les revenus),
  - l'incompréhension ou l'impossibilité de répondre à une question (par exemple, les étrangers, les personnes âgées, une question mal passée).
  - l'incohérence ou l'invalidation de réponse (par exemple, une personne de 13 ans retraitée, ou le total des dépenses dépassant le revenu),
  - la mauvaise qualité du travail de l'enquêteur,

- dans une expérimentation industrielle, certains résultats sont mauvais à cause des incidents survenus au cours d'un processus expérimental.

Les données manquantes totales :

- le refus de répondre ou l'abandon en cours d'une enquête nécessitant plusieurs visites,
- l'incapacité de répondre (par exemple, des étrangers où personnes âges),
- la négligence du répondant (la plupart des enquêtes auprès des entreprises sont postales, le rendement spontané est environ de 40 à 60%),
- le défaut du processus de production (la perte ou vol de documents, l'effacement de données),
- des documents inexploitables (bande de données trop ancienne ou abîmée),
- l'absence de la personne à interroger.

Traitement pratique des données manquantes (*D.M.*) :

1. Les méthodes préméditées pour éviter d'avoir des D.M. (Deville J.C., 1995)

- un plan de sondage préventif (un panel surreprésentant les catégories présumées mal répondantes),
- une pratique de rattrapage,
- l'utilisation de techniques d'enquêtes appropriées (face à face papier, moyens télématiques, etc.),
- la rédaction rigoureuse des questionnaires.

2. Procéder à une technique statistique sur les données manquantes

En général, le logiciel utilise un code pour identifier ce type de non-réponses du genre "ne sait pas", "ne répond pas", "rebut" etc. Les statisticiens excluent souvent les unités ayant des réponses manquantes ou bien, ils utilisent une technique de "pairwise" qui consiste à employer des paires de valeurs disponibles en même temps sur deux variables. Mais ces stratégies sont généralement non appropriées parce que les enquêteurs s'intéressent à la population entière plutôt qu'à la partie de la population qui leur procure des réponses à toutes les variables (Little R.J.A. et Rubin D.B. 1986).

Pour certaines questions sensibles, souvent on n'obtient pas la vraie réponse. Par exemple, "Quelle sorte de drogue utilisez-vous ?". Il s'agit de convaincre l'enquêté que l'anonymat de sa réponse à une question sensible sera totalement protégé. Dans ce cas la méthode "des réponses aléatoires" peut être utile ; l'enquêté répond à une question non sensible A ou à une question sensible B en fonction de la boule retirée au hasard d'une urne sans la montrer à l'enquêteur.

En effet, on doit savoir si le résultat réel est masqué et influencé par les non-réponses, et si les non-réponses sont liées aux variables d'intérêt. C'est à dire on tente d'identifier le mécanisme qui génère les données manquantes et sa modélisation ; si les données manquantes sont ignorables ou non. La procédure consiste en :

- une première étape consiste à effectuer une analyse descriptive complète des non-réponses et d'essayer de trouver la cause et les cofacteurs de *D.M.* Autrement dit, il s'agit d'identifier le mécanisme qui génère les données manquantes et sa modélisation (Deville J.C., 1995).

- une deuxième étape consiste à élaborer une méthode de compensation pour la non-réponse qui débouche sur un estimateur corrigé de celle-ci.

- une troisième étape prend en compte les non-réponses au niveau de l'estimation de précision.

Le traitement, quant à lui, se regroupe en deux parties :

I. Le traitement des données manquantes totales :

1. Repondérer les répondants de façon à assurer une représentativité selon des variables redéfinies (calages sur marges, méthode RAS),
2 Pratiquer une de fusion de fichiers pour reconstituer une observation entière s'il existe des informations auxiliaires et prédictives.

II. On distingue deux approches pour le traitement des données manquantes partielles :

1. Estimer les données manquantes en recourant à des modèles implicites (l'imputation) ou explicites (la régression par exemple) : la donnée manquante est remplacée par une valeur estimée sans se préoccuper de son usage. On peut utiliser d'autres variables qui peuvent aider à prédire ou compléter les *D.M.*. On peut aussi prendre en compte d'autres sources d'informations : les données disponibles sur le sujet d'autres enquêtes ou quelques fois des informations externes. Par exemple, la méthode du "Hot-Deck" repose sur l'hypothèse que les observations ayant des réponses similaires sont ressemblantes sur les autres variables non-observées, donc on impute ainsi les données manquantes.

2. Chercher à s'en accommoder : c'est à dire adapter les techniques de calcul aux cas des données incomplètes ; tenter de se placer au sein d'un modèle (de l'analyse de données) et chercher à estimer directement les paramètres à partir des données incomplètes sans estimer explicitement les données manquantes.

Avantages provenant du traitement de données manquantes :

- la réduction des aberrations dues aux données manquantes,
- la production de données propres et complètes,
- la simplification pour l'analyse statistique puisqu'il existe déjà des modèles pour les données complètes. Donc, il est possible de produire des calculs sur les variables observées complètement et d'améliorer ainsi les qualités des paramètres estimés,
- la préservation de la distribution de la population et des variables, donc la moyenne, la variance et d'autres paramètres seront plus proches de la situation réelle.

En conclusion, le traitement des données manquantes permet d'avoir plus d'informations satisfaisantes. Parfois les informations cachées deviennent disponibles grâce à une méthode d'estimation et certaines informations dites confidentielles ou masquées peuvent être explicitées.

Inconvénients et dangers : cependant on ne devrait pas oublier que des données reconstituées ne sont pas des données observées, donc des écarts inévitables entre les deux existent toujours. La nature des écarts varie en fonction de chaque méthode d'estimation. Plus le taux des D.M. dans un tableau est grand, moins fiable est le résultat.

1.2. HISTORIQUE DU DEVELOPPEMENT DU TRAITEMENT DES DONNEES MANQUANTES

Wilks (1932) a étudié un cas de données bivariées en se servant d'un modèle de maximum de vraisemblance sans estimer explicitement les données manquantes. Il est la première personne à remarquer que la régression de y sur x (avec des *D.M.* sur x) en utilisant seulement les observations complètes de x, est équivalente à une substitution de la moyenne de x pour les *D.M.*.

Yates (1933), en utilisant une suggestion de Fisher, fut le premier à utiliser la méthode d'estimation classique des moindres carrés : cette méthode consiste à compléter d'abord des données par l'imputation à la *D.M.* d'une valeur, telle que la somme des carrés entre les

valeurs de la variable dépendante (ayant D.M.) et ses valeurs estimées, soit minimale. L'estimation des paramètres β dans le modèle linéaire, $y = \beta X + \varepsilon$, ($y$ - variable dépendante (continue), X - vecteur des valeurs du plan d'expérience (valeur 1 ou 0)), est donnée par la formule standard. Les deux objectifs de la méthode sont d'obtenir une estimation correcte des moindres carrés des paramètres et d'obtenir une somme des carrés des résidus correcte.

Les procédures servant à estimer les *D.M.* peuvent être classées en deux catégories itération et non-itération. Federspiel (1959) et Buck (1960) ont introduit des méthodes de régression itérative. La variable ayant des *D.M.* est traitée comme une variable dépendante dans une régression, toutes les autres variables comme des variables explicatives avec l'imputation de la valeur manquante. Par ces équations, on impute une nouvelle valeur estimée à chaque valeur manquante. L'opération est faite successivement pour chaque variable ayant des *D.M.*.

Tim (1970), Deer (1959), Christofferson (1965), Wold (1966) et Gleason et Staelin (1975) travaillent sur les procédures qui décomposent la matrice des données en deux parties les valeurs complètes et les valeurs incomplètes et ensuite utilisent les premières composantes principales en se servant de la formule de reconstitution pour estimer les éléments inconnus.

La méthode du maximum de vraisemblance est illustrée par Hartley et Hocking (1971) qui traitent le problème des D.M. comme une estimation des paramètres sur des données observées. Les données sont supposées être générées à partir d'une distribution connue (il s'agit très souvent de la loi multinormale).

L'algorithme utilisé par Orchard et Woodbury (1972) ainsi que par Beale et Little (1975) est un exemple de l'algorithme *EM* (Espérance Maximisation) : la première étape consiste à estimer l'espérance des statistiques et la seconde à appliquer l'algorithme, maximum de vraisemblance, sur les données estimées complètes pour obtenir les estimateurs de paramètres.

Nora (1975) a proposé une reconstitution des *D.M.* dans un tableau de contingence grâce à la formule de reconstitution de l'analyse des correspondances.

Press et Scott (1976) s'intéressent au problème des *D.M.* dans un contexte de régression linéaire avec un point de vue bayésien. Ils définissent une distribution simultanée a posteriori des paramètres de la régression et des *D.M.* Ils maximisent cette distribution a posteriori par rapport aux paramètres et aux *D.M.* et obtiennent ainsi les estimateurs bayésiens. L'estimation des *D.M.* sert à l'imputation.

Rubin (1976) s'interroge sur le processus à l'origine des données manquantes alors qu'auparavant bien des auteurs se contentaient d'hypothèses floues sur la notion de *D.M.* Il présente le modèle- probabiliste formel du problème des *D.M.* en général et définit les conditions par lesquelles on peut ignorer le processus à l'origine des *D.M.* lorsque l'on fait de l'inférence statistique.

Kim et Cury (1978) comparent les méthodes "listwise" et "pairwise". La première méthode utilise seulement les observations complètes : un sérieux inconvénient à cette stratégie est que le nombre des observations complètes diminue lorsque les *D.M* sont très aléatoires et dispersées. La deuxième méthode, quant à elle consiste, par exemple, à utiliser des paires de données complètes pour calculer les corrélations entre les différentes variables. Mais elle aussi a un désavantage, en effet, les corrélations ou les covariances entre variables ne sont pas homogènes et la matrice de corrélation n'est alors plus nécessairement semi-définie positive : ses valeurs propres peuvent être négatives, ce qui peut conduire à des corrélations multiples supérieures à 1 ou négatives. Haitovsky (1968), Kim et Curry (1978) sont tombés d'accord : quand le nombre de données manquantes est grand ou que les D.M. ne sont pas aléatoires, il est plus intéressant d'utiliser les méthodes d'imputation des D M..

Rubin (1978) propose l'imputation multiple. On pense qu'une seule valeur pour *D.M.* ne permet pas d'exprimer l'incertitude que l'on a sur sa valeur. L'imputation multiple consiste à donner plusieurs valeurs d'estimation au lieu d'une.

Fèvre (1979) travaille sur la classification automatique des données quantitatives incomplètes en mettant une distance qui ignore les *D.M.*.

Rubin (1981) définit une méthode de bootstrap bayésien qui simule la distribution a posteriori du paramètre. En 1987, le même auteur propose une approximation du bootstrap

bayésien et applique cette technique à l'imputation des *D.M.* dans le domaine des sondages. La méthode se déroule comme suit, par exemple, si n_1 individus ont répondu pour la variable Y et que n_0 valeurs sont manquantes :

1) On effectue n_1 tirages aléatoires avec remise dans $Y=(Y_1, Y_2, ..., Y_{n1})$, ils forment ainsi l'ensemble Y_{obs}

2) On effectue n_0 tirages aléatoires avec remise dans Y_{obs} pour effectuer les n_0 imputations.

L'analyse homogène a été développée par une équipe de chercheurs mathématiciens et psychologues au sein de l'université de Leiden aux Pays-Bas : De Leeuw (1973), Gifi (1980, 1981, 1990), Meulman (1982), Van Buuren et Van Rijckevorsel (1991). Il s'agit d'une méthode de transformation (quantification) optimale des variables nominales équivalente à une analyse des correspondances multiples, qui a été étendue ensuite à des données incomplètes par (Meulman J., 1982). Van Buuren et Van Rijckevorsel (1991, 1992) présentent une technique d'estimation des *D.M.* basée sur ce modèle. Elle permet d'estimer des données manquantes non complètement aléatoires.

Little et Schluchter (1985) combinent la méthode du maximum de vraisemblance sur données continues (multinormal) et sur données catégorielles (modèle de Poisson ou multinomial) pour engendrer une méthode d'analyse des données multivariées avec des variables continues et catégorielles contenant des données manquantes.

Escofier (1981, 1987) et Benali (1985, 1987) ont traité les données manquantes dans le cadre de l'analyse des correspondances multiples.

Crettaz de Rotten (1991, 1993, 1995) présente une imputation de données manquantes à l'aide du coefficient RV pour les variables quantitatives. Cette méthode repose sur la minimisation d'une distance entre deux nuages de points ou de manière équivalente sur le lien entre deux groupes de variables. Le coefficient de corrélation vectorielle empirique RV est par définition une mesure de similitude entre deux nuages de points, cette méthode impute par la valeur qui minimise la distance entre les deux nuages de points ou de manière équivalente qui maximise ce coefficient.

Deville J C. et Sàrndal C.E. (1994) travaillent sur l'estimation des variances valides où l'imputation des *D.M.* est faite à l'aide d'un modèle de régression multiple.

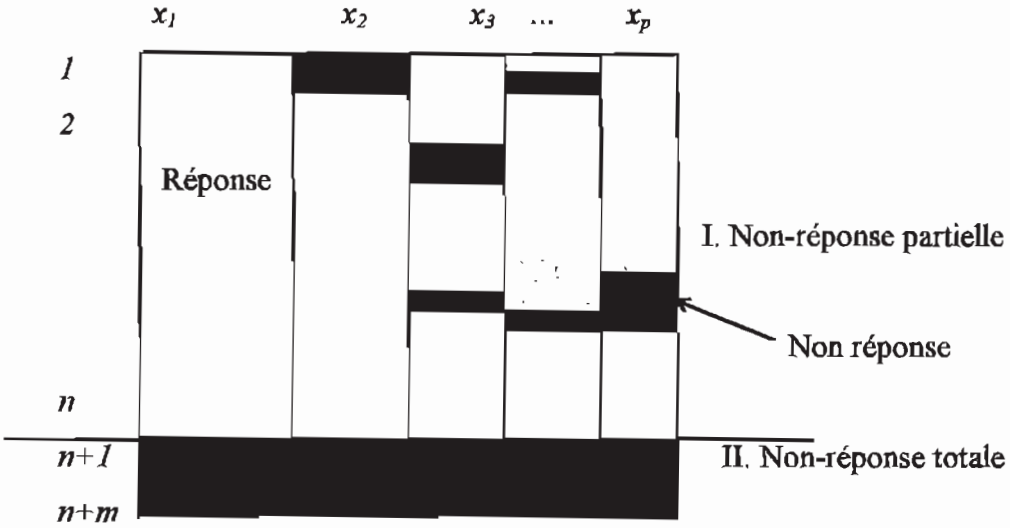
Caron N. (1996) donne une synthèse des méthodes de traitement des unités manquantes totales.

1.3. NOTIONS PRINCIPALES SUR LES DONNEES MANQUANTES

Les données statistiques sont généralement présentées sous forme d'un tableau à n lignes, les données étant recueillies sur n unités-observations. Si nous observons pour chaque unité, les p caractères-variables, le tableau a la forme d'une matrice à n lignes et p colonnes :

$$T = \begin{array}{c} \\ 1 \\ i \\ n \end{array} \begin{array}{c} \begin{array}{ccc} x_1, & \cdots, x_j, & \cdots, x_p \end{array} \\ \begin{bmatrix} x_{11} & x_{1j} & x_{1p} \\ & x_{ij} & \\ x_{n1} & & x_{np} \end{bmatrix} \end{array} \qquad (1)$$

Avec la présence de données manquantes, (1) pourrait être explicité, par exemple, de la façon suivante :



Si on considère les observations de l'unité *1* à l'unité *n* :

- x_1 est une variable à réponse totale,

- x_2 *à* x_p sont des variables entachées de données manquantes. Il s'agit ici de "*données manquantes partielles*"

Si on considère les observations de l'unité $n+1$ à l'unité $n+m$, elles sont du type "unités manquantes" ou "non-réponse totale".

Suivant les relations des données manquantes d'une variable avec celles observées et avec des données d'autres variables, nous regroupons les **"données manquantes partielles"** en trois types différents, *D.M.C.A.*, *D.M.A.*, *D.M.N.A.* (Rubin, 1976).

- ***D.M.C.A.*** (Données Manquantes Complètement Aléatoires) : si la probabilité qu'une variable x_j soit manquante, est indépendante des valeurs de tous les x_k $(k = 1,\cdots,p)$, y compris x_j, les données manquantes de x_j sont dites "complètement aléatoires". Car les données manquantes de x_j sont réparties de façon complètement aléatoire dans l'échantillon d'origine et les valeurs observées sont un sous-ensemble aléatoire de l'ensemble d'origine. Les données manquantes du type "D.M.C.A." sont ignorables dans toutes les circonstances.

Exemple 1 - Il existe deux variables X_1 = âge et X_2 = revenu dans un tableau de données. Si la probabilité de réponse pour le revenu est indépendante du montant de celui-ci et de celle de l'âge, les données manquantes sur le revenu sont donc dites complètement aléatoires.

- ***D.M.A.*** (Données Manquantes Aléatoires) : si la probabilité qu'une variable x_i soit manquante, est indépendante des valeurs de x_i sans être indépendante des valeurs des autres variables x_j $(j=1, \ldots, p \text{ et } j \neq i)$, les données manquantes de x_i sont dites aléatoires mais non complètement aléatoires.

Exemple 2 - Dans le même cadre que celui de l'exemple 1, si la probabilité de réponse pour le revenu ne dépend pas du montant de celui-ci, mais dépend de l'âge, alors les données manquantes sur le revenu sont aléatoires.

- ***D.M.N.A.*** (Données Manquantes Non Aléatoires) : si la probabilité qu'une variable x_i soit manquante, dépend des valeurs de x_i et (ou) des autres variables x_j, les données manquantes sur x_i sont dites non aléatoires.

1.4. TYPES DE DONNEES MANQUANTES ET CHOIX DE LEUR TRAITEMENT

Différent type de données manquantes :

Il est très délicat de vérifier quel est le type des *D.M.*. Il n'existe pas de méthode générale pour détecter des types de *D.M.* et la méthodologie varie selon la nature de chaque problème concret et dépend des informations disponibles.

Lorsque des données manquantes sont apparues dans une seule variable y, la procedure standard consiste à comparer les distributions des autres variables (complètement observées) sur des répondants de y et des non-répondants de y, au moyen d'un test pour la différence en moyenne.

Dans un contexte de régression, Simon et Simonoff proposent en 1986 des graphiques permettant de détecter si les données sont D.M.A., lorsqu'il manque des données dans une variable indépendante. Ils utilisent une limite inférieure et supérieure du vecteur des paramètres de régression comme une fonction de la mesure du caractère non-aléatoire du processus à l'origine des données manquantes. Pour ce faire, ils écrivent ces deux limites comme une fonction des données manquantes, qu'ils reportent dans un graphique, une mesure d'erreur sous l'hypothèse D.M C.A. Selon l'enveloppe définie par les deux courbes, ils évaluent l'importance du processus à l'origine des données manquantes et diagnostiquent si les données sont D.M.A..

En 1988, Little propose une statistique de test unique de l'hypothèse *D.M.C.A.*, pour les données multivariées ayant des données manquantes dans plusieurs variables, qui a la forme d'une distance de Mahalanobis entre le vecteur des moyennes et celui des estimateurs du maximum de vraisemblance des espérances. Il montre, de plus, que cette statistique est celle du

test de rapport des vraisemblances et que sous l'hypothèse *D.M.C.A*, la distribution est asymptotiquement celle d'un *Chi-deux*.

Crettaz de Rotten (1993) a cité que, en 1991, Heitjan et Rubin définissent un modèle statistique général pour des données incomplètes ("coarse data"), ce terme regroupe les données censurées, arrondies groupées, non-stochastiques (par exemple, les données groupées), d'autres sont de nature stochastique (par exemple, les données manquantes). Les résultats de Rubin (1976) sur l'ignorabilité du mécanisme à l'origine des données manquantes, sont généralisés pour les données incomplètes. Ils définissent sous quelles conditions on peut ignorer la nature stochastique des données et baser l'inférence sur le modèle adéquat pour les données observées.

Une autre solution serait de développer des méthodes d'imputation qui soient robustes à la violation de l'hypothèse de données D.M.C.A.. Dans le cas où il est impossible de savoir exactement le type de *D.M.* et si le mécanisme est "ignorable" ou "non ignorable", pour éviter le moindre risque, on pose toujours l'hypothèse du mécanisme de "non ignorable", et on applique une méthode de traitement destinée aux *D.M.* non ignorables. A notre avis, des méthodes adaptées aux *D.M.* du type "non ignorable" sont très précieuses.

Taux de données manquantes

Pour chaque type de *D.M.*, la quantité de *D.M.* a des influences différentes sur le résultat : dans le cas de *D.M.C.A.*, le taux de *D.M.* peut être relativement élevé sans causer trop de problème et cela dépend aussi du modèle de traitement correspondant. Par contre, une faible quantité de *D.M.* de type *D.M.A.*, surtout de *D.M.N.A.* peut nuire à la qualité du résultat.

Lorsque le taux de *D.M.* sur une observation est trop élevé et qu'on n'arrive pas à les estimer de façon satisfaisante, il est préférable de les exclure de l'analyse. Nous pourrons pour cela fixer un seuil au taux de données manquantes pour une catégorie, une variable, une observation en dessus duquel nous supprimons cette catégorie, variable, observation trop "abîmée".

Lorsqu'il existe un taux assez élevé de *D.M.* sur différentes variables et que nous traitons séparément variable par variable, il est préférable de faire les traitements sur les variables dont le taux de *D.M.* est le moins élevé.

Différents types de données manquantes et le choix de leurs traitements

Le cas des *D.M.N.A.* : la reconstitution des *D.M.* à l'aide du modèle d'estimation est recommandée. Lorsque des *D.M.* correspondent à une attitude particulière de données manquantes, soit le cas des *D.M.N.A.*, il n'est pas raisonnable d'ignorer ces *D.M.*. Si nous les traitons en supprimant simplement ces unités, il y aura non seulement perte d'informations, mais ça peut également générer des résultats statistiques aberrants. Donc, il faudrait reconstituer ces *D.M.* par les méthodes suivantes :

- imputation "*Hot-Deck*", du type de recherche des jumeaux ..,
- imputation par modèle probabiliste (la régression...),
- imputation par l'analyse homogène,
- méthode de maximum de vraisemblance.

Le cas des *D.M.A.* : leur prise en compte s'appuie sur l'objectif de l'analyse ; il existe deux cas possibles, les *D.M.* de type *D.M.A* peuvent être ignorées ou non ignorées ; par exemple, dans le tableau de données suivant, nous avons deux variables x et y :

$$\begin{matrix} X & Y \end{matrix}$$
$$\begin{bmatrix} x_1 & y_1 \\ x_2 & ? \\ x_3 & ? \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}$$

- si nous nous intéressons à la distribution conditionnelle de y sur x, $F(Y/X)$, les *D.M.* sur y peuvent être ignorées. Dans ce cas, la reconstitution des *D.M.* n'est pas obligatoire. On pourra travailler directement sur les données collectées,

- si nous nous intéressons au lien des variables, par exemple, la distribution conjointe $F(x, y)$, la distribution conditionnelle de x par rapport à y, ou à la distribution marginale de y, à sa moyenne ou son total. Les *D.M.* sur y ne peuvent pas être ignorées et il faut chercher à les reconstituer.

Le cas des *D.M.C.A.* : il existe deux possibilités de traitements, soit la reconstitution de données manquantes, soit l'application directe de méthodes statistiques "accommodées" pour des données incomplètes.

1. La reconstitution des données manquantes

Les *D.M.* reconstituées complètent les données observées. On peut les reconstituer soit :

- par l'imputation par la moyenne pour les données quantitatives, mais cette méthode, si elle garde constante la moyenne, produit une variance trop faible.
- par l'imputation aléatoire : les valeurs sont tirées au hasard parmi les données observées selon une probabilité proportionnelle à la fréquence marginale. Ce traitement conduit à une distribution raisonnable mais les résultats varie à chaque tirage,
- par l'imputation "homogène" : lorsque l'ensemble des données est assez homogène. Ce traitement conduit à une solution très satisfaisante.

De plus, les méthodes citées dans ce chapitre pour traiter les autres types de *D.M.* sont applicables pour traiter ce type des *D.M.*.

2. L'application directe des méthodes statistiques en s'accommodant des données manquantes

Lorsque le taux de données manquantes est assez faible, nous utilisons seulement les informations disponibles. Les techniques employées sont les suivantes :

- la technique "listwise" où l'analyse est basée sur les unités ayant des réponses complètes. L'avantage de cette analyse est la simplicité, les méthodes d'analyse de données peuvent être appliquées sans être modifiées mais elle engendre une perte importante d'unités lorsque la quantité de *D.M.* est grande,

- l'analyse basée sur toutes les données disponibles de chaque variable . il s'agit de la méthode utilisée dans la plupart des logiciels statistiques, l'analyse est basée sur toutes les données disponibles d'une variable pour calculer les fréquences marginales pour chaque variable qualitative et l'estimation de la moyenne pour chaque variable quantitative. L'avantage est l'utilisation d'un maximum d'informations, mais l'échantillon change d'une variable à l'autre.

- la technique "pairwise", l'analyse basée sur les paires de données disponibles de deux variables concernées ; par exemple, pour calculer la covariance de (y_j, y_k) ou la corrélation (y_j, y_k), nous nous basons sur les unités *i* dont les valeurs y_{ij} et y_{ik} sont toutes les deux présentes. Parfois, elle conduit à une situation contradictoire. Par exemple, sur le tableau de données suivant :

$$
\begin{array}{c}
y_1, y_2, y_3 \\
\begin{bmatrix}
1 & 1 & * \\
2 & 2 & * \\
3 & 3 & * \\
4 & 4 & * \\
1 & * & 4 \\
2 & * & 3 \\
3 & * & 2 \\
4 & * & 1 \\
* & 1 & 1 \\
* & 2 & 2 \\
* & 3 & 3 \\
* & 4 & 4
\end{bmatrix}
\end{array}
$$

On trouve en utilisant la méthode pairwise : $r_{12} = r_{23} = 1 \; et \; r_{13} = -1$ ce qui est absurde.

Lorsque les données incomplètes sont du genre "*D.M.C.A.*" et que la quantité des *D.M.* est faible, pour éviter tous les différents problèmes cités ci-dessus, nous pouvons modifier les méthodes de l'analyse des données pour qu'elles s'appliquent aux données incomplètes, comme celle de l'analyse des correspondances multiples pour les données incomplètes proposée par Benali H. & Escofier B. (1987). Nous proposerons dans le chapitre III une

méthode de classification automatique des données incomplètes pour des variables quantitatives en utilisant un algorithme du type "EM". Nous proposons également une méthode d'analyse en composantes principales pour des données incomplètes.

1.5. TRAITEMENT DES DONNEES MANQUANTES TOTALES

Dans une enquête ordinaire, les unités manquantes pourraient causer des problèmes de distorsion entre l'échantillon répondant et la représentativité de population et conduire à une mauvaise interprétation des résultats. Sauf si les non-répondants ont un comportement identique à celui des répondants, les estimations obtenues sur les répondants sont biaisées. De plus, en présence de non-réponse, l'estimateur est moins précis puisque c'est la taille de l'échantillon des répondants qui intervient dans les calculs de précision. Les non-réponses méritent donc d'être soigneusement étudiées. Peut-on distinguer des groupes de non-répondants ? Quelle est la cause des non-réponses ? Pour quels motifs principaux ? J.M Grosbras (1987) propose les techniques suivantes pour trouver les facteurs explicatifs de non-réponses.

Mise en évidence de facteurs explicatifs

D'abord, la relance de non-répondants peut fournir des informations supplémentaires pour la construction du modèle de non-répondant.

1. *L'analyse des données* permet de dégrossir le travail de description des non-répondants. On peut par exemple :

- faire une analyse des correspondances multiples sur l'échantillon global, en prenant pour variables actives celles du signalétique : catégorie de commune, type de logement, caractéristiques connues etc., et pour variable supplémentaire la variable 'réponse ou non-réponse'.

- faire une analyse des correspondances multiples sur le sous-échantillon des non-répondants, en portant cette fois en variables supplémentaire le critère 'refus motivé ou non' ainsi que les modalités de refus convenablement codifiées

- faire une analyse discriminante, la distinction à expliquer étant la séparation ' répondants ou non-répondants', les variables explicatives étant celles du signalétique

2. *Les techniques d'économétrie des variables qualitatives* (Gourieroux, 1984) peuvent être utilisées avec profit pour modéliser les variables 'réponse ou non-réponse'.

Soit Z

l'indicatrice définie par : $Z_i = \begin{cases} 1 \text{ si l'individu } i \text{ ne répond pas} \\ 0 \text{ s'il répond} \end{cases}$,

X_{1i} est la valeur signalétique X_1 pour l'individu i et X_{pi} est la valeur signalétique X_p pour l'individu i.

On suppose que la probabilité de non-réponse est une certaine fonction des variables X_1 , ..., X_p. $P(Z_i = 1) = F(X_{1i}, ..., X_{pi}) = F_i$, la vraisemblance de l'échantillon est : $\gamma = \Pi_{i=1}^{n} F_i^{Z_i} (1 - F_i)^{1-Z_i}$

Donc, on pourra estimer les paramètres de la fonction F par la méthode du maximum de vraisemblance. Les modèles se distinguent selon le choix des fonctions F.

- Le modèle LOGIT, F est défini par une loi logistique :

$$F_i = \left\{ 1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ji}) \right\}^{-1}$$

ou

$$\ln \frac{F_i}{1 - F_i} = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ji}$$

On estime donc les β_j par le maximum de vraisemblance. On démontre que la solution unique existe sauf si deux variables sont colinéaires. Lorsque le nombre d'observations augmente indéfiniment, les estimateurs tendent vers les valeurs vraies des coefficients et on peut estimer leurs variances asymptotiques, ce qui permet de tester l'égalité à 0 des coefficients

- Le modèle PROBIT, dans ce modèle, nous avons :

$$F_i = \Phi(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ji})$$

où Φ est la fonction de répartition de la loi normale centrée réduite.

Les lois logistique et normale sont assez voisines, et on obtient en général des résultats semblables par l'un ou l'autre des modèles.

Les méthodes de traitement des unités manquantes totales

Les principales méthodes de traitement de l'unité manquante sont basées sur un modèle de réponse et (ou) de l'utilisation des informations auxiliaires. La validité de l'inférence, c'est-à-dire l'extrapolation des résultats obtenus sur l'échantillon des répondants à la population, dépend de la validité du modèle choisi. N. Caron (1996) présente diverses méthodes de traitement des non-réponses totales et a appliqué ces méthodes dans des enquêtes au niveau national. Les méthodes principales de traitement de unités manquantes consistent en :

- la relance des non répondants,
- la substitution de l'individu manquant hors du panel d'enquête,
- la repondération des individus répondants,
- la post-stratification qui stratifie *a posteriori* les répondants. L'inférence concerne la taille de l'échantillon et des strates et le nombre des répondants dans chaque strate,
- le mécanisme du Raking ratio.

Relance des non-répondants (J.M. Grosbras, 1987) :

Cette méthode suppose que la population peut être classée en deux strates : celle de ceux qui répondent et celle de ceux qui ne répondent pas. Les n unités d'un premier échantillon révèlent n_1 répondants et n_2 non-répondants. Les premiers sont représentatifs de la première strate Il faudrait donc avoir au moins une estimation des éléments de la deuxième strate. Pour cela, on tire un sous échantillon de taille n_2' parmi les n_2 non-répondants au premier tour. Grâce aux bons enquêteurs et aux motivations appropriées, on réussit à obtenir les réponses de ces n_2' unités. On peut alors combiner les informations des deux strates :

$$\overline{Y}_h = \frac{n_1}{n}\overline{Y}_1 + \frac{n_2}{n}\overline{Y}_2'$$

où $\overline{Y}_1$ *et* $\overline{Y}_2'$ *sont les moyennes disponibles*

$\overline{Y}_h$ est un estimateur sans biais de $\overline{Y}$. Si on a décidé à l'avance le taux de sondage t au deuxième tour, nous avons :

$$V(\overline{Y}_H) = \frac{N-n}{N*n}S^2 + \frac{1}{n}\frac{N_2}{N}(\frac{1-t}{t})S_2^2$$

où N est l'effectif de l'ensemble de la population,

N_2 *est l'effectif de l'ensemble* des *non répondants de la population,*

$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2$ *et* $S_2^2 = \frac{1}{N_2-1}\sum_{i=1}^{N_2}(Y_{2i} - \overline{Y}_2)^2$, les variances respectives dans la population

Repondération des individus répondants :

L'idée des méthodes de repondération consiste à traiter la non-réponse en modifiant, plus exactement en augmentant, le poids de sondage des individus répondants. Souvent, elles se basent sur un choix de modèle de comportement des individus face à la non-réponse, c'est-à-dire, faire des hypothèses sur la distribution des non-répondants.

Dans le cas d'une enquête stratifiée, une unité sélectionnée avec une probabilité π, représente π_i^{-1} unités dans la population et donc π_i^{-1} est le poids dans l'estimation de la population totale. En l'absence des non répondants, le total T de la variable Y peut être estimé par l'estimateur de Horvitz-Thompson :

$$T = \sum_{i=1}^{N}\frac{y_i}{\pi_i}I_i, \quad \text{où } I_i = \begin{cases} 1 \textit{ répondant} \\ 0 \textit{ non répondant} \end{cases}$$

La moyenne de la population peut être estimée par

$$\overline{y}_w = \sum_{i=1}^{N} w_i y_i$$

$$\textit{où } w_i = \frac{I_i \pi_i^{-1}}{\sum_k I_k \pi_k^{-1}} \qquad (1)$$

dans le cas où présente des non répondants : le poids pour une variable y devient $(\pi_i \hat{p}_i)$, $\hat{p}_i$ est une estimation de la probabilité de réponse pour l'unité i. Bien entendu, cette technique

est basée sur l'hypothèse que l'on connaît préalablement les probabilités $\hat{p}_i$. Si tous les individus de la population ont la même probabilité de réponse et ont un comportement indépendant les uns des autres, l'estimateur de la moyenne Y par (1) reste valable et sans biais. Cela revient à calculer l'estimateur sur les répondants en ignorant les non-répondants. Ce modèle "ne rien faire pour traiter les non-réponses" revient à postuler un mécanisme de réponse uniforme vis à vis de la non-réponse (similaire au cas D.M.C.A.).

La repondération fondée sur la post-stratification :

La méthode la plus simple est fondée sur la stratification a posteriori. La population est divisée en sous-populations supposées homogènes au sens de la non-réponse. Elles sont constituées après réalisation de l'enquête en examinant en général le critère répond (ou ne répond pas) en fonction de variables connues pour les répondants et les non-répondants. On considère l'échantillon des répondants selon des critères pour lesquels on connaît la répartition dans la population parente. A partir de cette stratification, on calcule pour chaque sous-population un coefficient permettant de rétablir la part qu'elle représente dans l'ensemble de la population, en se basant sur les statistiques disponibles les plus proches. Les critères de stratification a posteriori doivent être le plus possible corrélés avec les variables de l'enquête. Cette méthode revient à remplacer les valeurs manquantes de l'échantillon par les valeurs moyennes des sous populations correspondantes. Cela revient à faire l'hypothèse implicite que les non-répondants des catégories ne se distinguent pas en moyenne de celles des répondants. Dans ce cas, l'estimateur du total Y appelé 'estimateur pondéré par classe' est sans biais :

$$\hat{Y} = \frac{N}{n} \sum_h n_h \bar{y}_{r_h}, \quad où \quad \bar{y}_{r_h} = \sum_{r_h} \frac{Y_i}{r_h},$$

où r_h, le nombre de répondants, supposé connu dans la sous-population h, et n_h, le nombre d'individus de l'échantillon de la sous-population h.

- si le nombre d'individus N_h dans la sous-population h est connu, l'estimateur est : $\hat{Y} = \sum_h N_h \bar{y}_{r_h}$.

- l'utilisation d'information auxiliaire externe permet de réduire le biais de l'estimateur obtenu dans le cas où le modèle de réponse n'est pas correct. Par exemple, si la variable X est connue au sein de l'échantillon complet s, nous obtenons l'estimateur par ratio par classe :

$$\hat{Y} = \frac{N}{n}\left(\sum_{s} x_i \frac{\sum_{h} n_h \bar{y}_{r_h}}{\sum_{h} n_h \bar{x}_{r_h}} \right)$$

Un compromis sur le nombre des classes reste à trouver entre deux positions contradictoires :

1. multiplier les classes pour réduire le biais lié au choix d'un modèle approximatif,

2. diminuer le nombre de classes pour obtenir une précision suffisante dans chaque classe.

L'inconvénient de cette technique est que l'on introduit artificiellement une concentration autour des valeurs moyennes. Les variances calculées sur l'échantillon redressé sous-estiment les variances véritables. Si on remplace les réponses manquantes par divers tirages à probabilités égales avec remise parmi les répondants, on peut éviter cet inconvénient

Enfin, dans le traitement des non-réponses totales, le mécanisme des unités manquantes mérite d'attirer notre attention.

- *Le mécanisme des unités manquantes* : d'abord nous devons bien connaître le mécanisme d'unité manquante pour pouvoir faire le choix des variables de redressement. Dans le redressement ordinaire, les échantillons sont souvent redressés par les variables démographiques et pas forcément par des variables, autres que les variables démographiques, qui régissent vraiment le mécanisme de non-réponse. Donc, il faudrait trouver des variables qui ont la plus grande différence entre répondants et non-répondants et identifier le mécanisme d'unité manquante.

D'ailleurs, on peut considérer le traitement de données manquantes totales comme un cas spécial de la fusion des fichiers que nous allons aborder dans la partie II.

CHAPITRE II : LES DIFFERENTES METHODES DE TRAITEMENTS DES DONNEES MANQUANTES PARTIELLES

Les méthodes d'imputation sont fréquemment utilisées dans le traitement des données manquantes partielles. Elles consistent à remplacer la donnée absente par une donnée 'plausible' qui est en général issue ou estimée à partir des répondants. L'avantage du traitement est qu'il procure des données complètes pour faciliter le calcul ensuite. Mais il n'est pas sans conséquence au niveau du résultat La dispersion au sein de la population n'est plus forcément valable en présence d'imputation. L'estimateur de la moyenne de la variable Y après imputation se présente sous la forme suivante :

$$\hat{\bar{Y}} = \frac{1}{n}\sum_{n}(Y_i\delta_i + (1-\delta_i)Y_i^*)$$

où Y_i^* représente une valeur imputée et δ_i vaut 1 si l'individu i est répondant et 0 sinon La distribution originelle des valeurs de Y est alors plus ou moins modifiée. De plus, l'estimation de la variance empirique, obtenue en traitant les données comme si elles étaient réelles, est modifiée par rapport à celle obtenue sur les répondants. En effet,

$$s_I^2 = \frac{r-1}{n-1}s_r^2 + \frac{n-r-1}{n-1}s_{nr}^2 + \frac{(n-r)r}{n(n-1)}(\bar{y}_r - \bar{y}_{nr})^2$$

où r-le nombre de répondants,

s_r^2 (s_{nr}^2) - représente la variance empirique modifiée pour les répondants (non-répondants).

2.1. ESTIMATION DES DONNEES MANQUANTES PAR DES MODELES EXPLICITES

2.1.1. Méthodes de régression

Les différentes méthodes de régression suivantes sont proposées depuis 1960 où la première méthode de ce type est présentée par Buck (1960) :

- la régression simple en prenant la variable la plus corrélée,

- la régression multiple en prenant le meilleur sous-ensemble de variables explicatives (Frane 1976, Seber 1984),

- la régression avec une technique stochastique en ajoutant à la prédiction un résidu, soit généré selon une loi normale, soit tiré au hasard parmi les résidus empiriques (Little et Rubin, 1987),

- la pondération d'une fonction de matrice de variance-covariance (Beale et Little, 1975).

Pour la régression, Frane (1976) montre qu'un trop grand nombre de prédicteurs produit de mauvaises imputations dans la pratique, pour des données manquantes

La prédiction par la moyenne peut être considérée comme un modèle de régression simple avec seulement la constante. Dans ce modèle, l'estimateur de la moyenne est sans biais. Mais la distribution des valeurs est fortement modifiée en ajoutant une proportion importante des valeurs égales à la moyenne. La variance empirique est sous-estimée en attribuant systématiquement la même valeur aux non-répondants. La sous-estimation est d'autant plus importante que le nombre de non-répondants est grand.

Puisque la **méthode de Buck** a sans cesse été commentée, évaluée et améliorée, nous la présenterons ici : Soit le cas où l'individu n a une donnée manquante sur la première variable notée x_{1n}, et où les variables explicatives $(x_{2n}, x_{3n}, \cdots, x_{pn})$ sont complètement observées.

La méthode de Buck suggère l'imputation de *D.M.* par :

$$E(x_{1n}|x_{2n}, x_{3n}, \cdots, x_{pn}) = \hat{\beta}_1 + \hat{\beta}_2 x_{2n} + \hat{\beta}_3 x_{3n} + \cdots + \hat{\beta}_p x_p$$

avec $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$, les estimateurs de moindres carrés des paramètres de la régression linéaire multiple évaluée sur les $(n\text{-}1)$ premiers individus (sans *D.M.*).

Soit $S_{n-1} = \begin{pmatrix} v_{11} & \alpha_1' \\ \alpha_1 & C \end{pmatrix}$ la matrice de variance-covariance évaluée sur *(n-1)* premiers individus.

l'estimation de la *D.M.* pour la variable x_j se fait par :

$$x_{1n} = \hat{\beta}_1 + \hat{\beta}_2 x_{2n} + \hat{\beta}_3 x_{3n} + \cdots + \hat{\beta}_{(p+q)} x_{(p+q)n} = \beta' X$$

$$avec\ X = (x_{2n}, x_{3n}, \cdots, x_{(p+q)n})\ et\ \beta' = \alpha_1' C^{-1}$$

$$\text{où } \beta = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p)'$$

Lorsque X_{1n} est un vecteur, la méthode de Buck utilise alors la régression multivariée.

Cette méthode équivaut à une minimisation de la distance de Mahalanobis estimée par :

$$D_1^2 = (X_{1n} - \overline{X}_{n-1})' S_{n-1}^{-1} (X_{1n} - \overline{X}_{n-1})$$

entre le vecteur individu ayant la *D.M.* X_{1n} et le vecteur moyenne $\overline{X}_{n-1}$.

Cette méthode peut être appliquée itérativement :

1. l'estimation initiale de la moyenne et de la matrice de variance-covariance et ensuite rèpéter les étapes 2 et 3 jusqu'à stabilisation des paramètres,

2. l'imputation par la méthode de Buck,

3. la réestimation de la moyenne et de la matrice de variance-covariance après imputation.

Dans le cas des données catégorielles :

Little et Rubin (1987) proposent des variantes de la méthode de Buck :

- lorsque la variable expliquée est dichotomique, à l'aide de la régression logistique on calcule la probabilité d'obtenir chacune des deux valeurs et on impute par la catégorie la plus probable,

- lorsque certaines des variables explicatives sont catégorielles, on utilise des variables indicatrices : pour une variable x_i *à* k_i catégories, on crée (k_i-1) variables indicatrices dont les paramètres s'utilisent comme ceux des variables continues pour imputer les données manquantes de la variable dépendante.

L'analyse de la méthode de la famille de Buck conduit aux conclusions suivantes (Crettaz de Rotten F, 1993) :

- la corrélation entre les variables doit être importante car la méthode n'est applicable que lorsque la structure de corrélation est assez forte,

- la matrice de variance-covariance après imputation est sous-estimée : cela peut être corrigé par l'élaboration de coefficients d'ajustement,

- les résultats de la méthode de Buck et ceux de l'approche du maximum de vraisemblance sont identiques sous l'hypothèse de multinormalité des données.

2.1.2. Méthodes du maximum de vraisemblance

Orchard et Woodbary (1972) avaient mis au point une méthode itérative qui estime les estimateurs du maximum de vraisemblance et impute les *D.M.* en même temps. Ils considèrent les *D.M.* comme des variables aléatoires à l'intérieur du modèle et donc des paramètres à estimer. Leur idée originelle consiste à estimer θ l'ensemble des paramètres du modèle par un point fixe de l'équation $\theta = \varphi(\theta)$ avec φ une transformation composée des équations de la maximisation de l'espérance du logarithme de la densité de Y_{obs} *et* Y_{mis} ayant θ comme ensemble des paramètres. On remplace donc un problème de maximisation par un problème de point fixe.

Dempster, Laird et Rubin (1977) introduisent le célèbre algorithme *EM*. Si la densité de $Y|\theta$ appartient à une famille exponentielle régulière $f(Y|\theta) = b(Y)\exp\{S(Y)\theta\} / \alpha(\theta)$, l'algorithme *EM* comporte deux étapes qu'on répète jusqu'à la stabilisation des estimateurs des paramètres :

E . estimer $t(Y)$, la statistique exhaustive de θ basée sur les données complètes, par l'espérance de $t(Y)|\ Y_{obs}$ et θ , dont implicitement il faut imputer les données manquantes par l'espérance conditionnelle de $Y_{mis}\ |Y_{obs}\ et\ \theta$. Si l'on fait l'hypothèse de la normalité multivariée, l'étape *E* impute par l'espérance conditionnelle de la *D.M.* d'individu, connaissant les valeurs inférentielles pour cet individu, c'est à dire par l'équation de la régression.

M : estimer les paramètres par maximum de vraisemblance ce qui est équivalent à chercher une solution de l'équation $E(t(Y|\theta)) = t$.

Par rapport aux autres algorithmes, l'algorithme *EM* ne nécessite pas de calcul ou d'approximation de dérivée seconde, ni d'inversion de matrice, il est simple, mais sa convergence est lente. Cet algorithme permet de traiter des *D.M.* dans beaucoup de contextes.

Celeux (1988), puis Celeux et Diebolt (1988) proposent une version stochastique de l'algorithme appelé *SEM* qui repose sur un principe d'attribution au hasard de la valeur des données manquantes.

- l'étape *E* est identique à celle de l'algorithme *EM*,

- l'étape *S* impute les *D.M.* par tirage aléatoire suivant la loi conditionnelle calculée à l'étape *E*,

- l'étape *M* calcule les estimateurs du maximum de vraisemblance sur l'échantillon complété de manière aléatoire suite à l'étape *S*.

Cet algorithme converge en distribution vers une loi de probabilité centrée sur les estimateurs du maximum de vraisemblance.

Mais l'algorithme *SEM* ou *EM* ne fournit pas de matrice de variance-covariance asymptotique des estimateurs du maximum de vraisemblance. Meng et Rubin (1991) proposent un algorithme *EM* supplémentaire qui comble cette lacune ; l'algorithme calcule la matrice de variance-covariance asymptotique à l'aide de la dérivée seconde du logarithme de la fonction de vraisemblance de données observées.

Cas des variables catégorielles :

On retrouve ici les mêmes idées que pour les modèles avec des données quantitatives. Little (1982) et Vardi, Shepp et Kaufman (1985) suggèrent que dans le cadre du modèle log-linéaire, l'algorithme *EM* soit utilisé comme suit :

- l'étape *E* calcule l'espérance conditionnelle du nombre d'individus par cellule à partir des estimations courantes des probabilités des cellules,

- l'étape *M* calcule les nouvelles estimations des paramètres.

Fuchs (1982) précise qu'à cause de la spécificité du modèle log-linéaire, l'estimation du maximum de vraisemblance n'est pas directe et l'étape *M* nécessite une itération.

Little et Schluter (1985) combinent les deux modèles, pour des données quantitatives et qualitatives. Pour générer une méthode d'analyse de données multivariées avec des variables continues et catégorielles contenant des données manquantes. L'ensemble des paramètres est . vecteur des probabilités des catégories, la moyenne et la matrice de variance-covariance des données continues. En présence de *D.M.*, le logarithme de la fonction de vraisemblance peut être maximisé par l'algorithme *EM* qui fournira les imputations et des estimateurs du maximum de vraisemblance des paramètres. Cette méthode permet de traiter la régression logistique et l'analyse discriminante avec les valeurs des variables explicatives incomplètes, la régression linéaire avec prédicteurs continus et catégoriques.

2.2. ESTIMATION DES DONNEES MANQUANTES PAR DES MODELES IMPLICITES

2.2.1 Méthode de type "Hot-deck"

L'imputation par le "hot-deck" regroupe plusieurs méthodes basées sur le concept de 'donneur'. Son principe consiste à remplacer les données manquantes d'un individu par les données réelles d'un donneur (autre individu complet). L'idée est de choisir une valeur à partir de la distribution empirique des répondants Cette méthode suppose que tous les individus de la population ont la même probabilité de répondre et ont un comportement indépendant les uns des autres. L'estimateur obtenu est sans biais. La valeur de l'aléa correspond à la déviation du répondant choisi comme donneur par rapport à la moyenne. En fait, cette méthode revient à repondérer les individus répondants ; c'est-à-dire à leur attribuer un poids aléatoire.

On distingue parmi la famille de type "hot-deck" :

- *Hot-deck aléatoire* avec ou sans remise (Caron N.,1996),
- *Hot-deck séquentiel* qui remplace la donnée manquante par la valeur correspondante de l'individu précédent selon un ordre choisi,
- *Hot-deck sur un critère* qui remplace la *D.M.* par la valeur de l'individu le plus proche au niveau du critère,
- *Hot-deck du plus proche voisin* qui remplace la *D.M.* par la valeur correspondante de l'individu ayant le minimum de distance en fonction des covariables (multicritère).

Hot-deck aléatoire avec remise :

La donnée manquante est remplacée par la valeur observée pour un individu répondant choisi au hasard avec remise. La variance de l'estimateur obtenu est :

$$V(\hat{\bar{Y}}|r) = (\frac{1}{r} - \frac{1}{N})S^2 + (1 - \frac{1}{r})(1 - \frac{r}{n})\frac{S^2}{n}$$

La variance se décompose en deux termes : le premier représente la variance des répondants et le second est induit par l'imputation. La variance obtenue est plus forte que celle obtenue par la méthode d'imputation par la moyenne ; elle présente cependant l'avantage de ne pas

déformer trop la distribution c'est-à-dire que la variance empirique modifiée calculée sur l'ensemble des données s_I^2 est proche de celle calculée à partir des répondants s_r^2.

Hot-deck aléatoire sans remise :

Le donneur est choisi (par exemple, lorsque $r \geq \frac{n}{2}$) par un tirage aléatoire sans remise.

La variance de l'estimateur obtenu est .

$$V(\hat{\bar{Y}}|r) = (\frac{1}{r} - \frac{1}{N})S^2 + \frac{(n-r)^2}{n^2}(1 - \frac{n-r}{r})\frac{S^2}{n-r}$$

où S^2 est la dispersion au sein de la population.

Nous avons la même conclusion que dans le cas avec remise. Mais la variance attribuable à l'imputation est plus petite dans le cas d'un hot-deck 'sans remise' que dans le cas 'avec remise'.

Hot-deck du plus proche voisin :

L'hypothèse de l'imputation est que les unités ayant la plupart de leurs profils similaires ont les mêmes valeurs sur les autres réponses manquantes. La donnée manquante est remplacée par une valeur observée pour l'individu le plus proche au sens d'une distance calculée en fonction de variables covariables renseignées pour les deux individus. La distance peut être choisie de façon à respecter la corrélation entre les variables et la variable d'intérêt en particulier en accordant plus d'importance aux variables les plus liées à la variable d'intérêt. Crettaz de Roten F. (1993) a montré que les méthodes de type "hot-deck" reviennent à minimiser des distances entre individus.

La méthode "Hot-deck" est souvent accompagnée d'une procédure de classification (Ford B.L., 1983) : Il s'agit de répartir l'ensemble des *n* individus en *m* groupes homogènes. Les valeurs correspondantes d'un individu dans le même groupe sont copiées pour remplacer les données manquantes d'un autre individu ou bien la donnée manquante est remplacée simplement par la moyenne observée des répondants de la classe. L'imputation de la moyenne par classe possède les mêmes désavantages que l'imputation de la moyenne globale. D'ailleurs,

toutes les méthodes présentées ci-dessus peuvent aussi être appliquées après la constitution de cellules homogènes et en recherchant le 'donneur' dans la même classe que la donnée manquante.

Choix de variables de la classification :

Les variables utilisées pour la classification doivent être bien corrélées avec le critère de choix du donneur et avec la variable ayant les données manquantes. Dans le cas contraire, le résultat ne sera pas très satisfaisant. C'est-à-dire que les variables servant à la classification doivent être des prédicteurs pertinents pour les variables ayant des données manquantes et le critère de choix du donneur. Ces variables ne devraient pas contenir de données manquantes : les variables auxiliaires sont toujours présentes. Les variables de classification peuvent être quantitatives ou qualitatives. Pour les variables qualitatives, cette notion de proximité est exprimée par une distance entre ces deux unités qui est une mesure de ressemblance entre les deux unités.

Un des avantages du hot-deck est de fournir des estimations réalistes des réponses manquantes (puisque ce sont de vraies réponses observées sur les répondants) et d'éviter ainsi des estimations incohérentes ou irréalistes.

2.2.2. Méthode basée sur le coefficient RV (Crettaz de Roten F., 1993)

Le dénominateur commun entre les différentes méthodes d'imputation des données manquantes, est la minimisation d'une distance basée sur une variable ou la maximisation du lien entre une variable et un groupe de variables. Cette méthode repose sur la minimisation d'une distance entre deux nuages de points ou de manière équivalente la maximisation du lien entre deux groupes de variables.

Soit $X=\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = (Y_1, Y_2)$ un vecteur aléatoire sur $(p+q)$ variables, partitionné en Y_1 *et* Y_2 et

$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ sa matrice de variance-covariance. Par la définition de Escoufier (1973), le coefficient de corrélation vectorielle entre Y_1 *et* Y_2 est :

$$\rho v = \rho v(Y_1, Y_2) = \frac{\operatorname{cov}(Y_1, Y_2)}{\sqrt{\operatorname{var}(Y_1) \cdot \operatorname{var}(Y_2)}} = \frac{tr(\sum_{12} \sum_{21})}{\sqrt{tr(\sum_{11}^2) tr(\sum_{22}^2)}}$$

ρv est une mesure de similarité entre Y_1 *et* Y_2.

En présence d'un échantillon $\hat{X}_1, \hat{X}_2, \cdots, \hat{X}_n$, on définit le coefficient de corrélation vectoriel empirique en remplaçant dans l'expression de ρv les paramètres par les estimateurs habituels :

$$RV = RV(\hat{Y}_1, \hat{Y}_2) = \frac{tr(S_{12} S_{21})}{\sqrt{tr(S_{11}) \cdot tr(S_{22})}}$$

$$où\ S_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (\hat{y}_{ik} - \overline{\hat{Y}_i})(\hat{y}_{jk} - \overline{\hat{Y}_j})' \quad i, j = 1, 2,$$

$\overline{\hat{Y}_1}$ *et* $\overline{\hat{Y}_2}$ désignent respectivement des moyennes empiriques des $\hat{Y}_{1k}$ *et* $\hat{Y}_{2k}$, $k = 1, \cdots, n$..

Supposons qu'une partie du premier groupe de variables du vecteur $\hat{X}_n$ contienne des données manquantes, nous noterons y_{1m} ce vecteur inconnu, y_{1o} la partie complète du premier groupe de variable et y_2 le deuxième groupe de variable. Ainsi nous pouvons écrire :

$$\hat{X}_n = (y_{1m}, y_{1o}, y_2)$$

Cette méthode consiste à remplacer y_{1m} par le vecteur qui minimise la distance $(\hat{Y}_1, \hat{Y}_2)$, c'est-à-dire, par celui qui maximise $RV(\hat{Y}_1, \hat{Y}_2)$.

Si l'indice *i* accolé à la matrice de variance-covariance S et au vecteur de la moyenne X indique que les *i* premiers individus ont participé à l'évaluation de cet estimateur, nous avons :

$$S_n = \frac{n-2}{n-1} S_{n-1} + \frac{1}{n}(\hat{X}_n - \hat{\bar{X}}_{n-1})(\hat{X}_n - \hat{\bar{X}}_{n-1})'.$$

Posons

$$(u \quad v \quad w) = \frac{1}{\sqrt{n}}(X_n - \bar{X}_{n-1}) \;\; et \;\; \frac{n-2}{n-1} S_{n-1} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix},$$

où *u,v et w* correspondent respectivement aux partitions des variables en Y_{1m}, Y_{1o}, et Y_2 et après des manipulations algébriques, nous obtenons :

$$RV(u) = \frac{2w^T A_{31} u + w^T w u^T u + \alpha}{\sqrt{\gamma}\sqrt{(u^T u)^2 + 2u^T (A_{11} + I(v^T v))u + 4v^T A_{21} u + \beta}}$$

avec $\alpha = tr\{A_{13} A_{31}\} + tr\{(A_{23} + vw^T)(A_{32} + wv^T)\}$

$$\beta = tr\{A_{11}^2 + 2A_{12} A_{21}\} + tr\{A_{22}^2\} + 2v^T A_{22} v + (v^T v)^2$$

$$\gamma = tr\{(A_{33} + ww^T)^2\}$$

Crettaz de Roten F. (1993) dans la section 3.2.3 de sa thèse montre que la fonction RV(u) possède un maximum car il faut maximiser cette fonction dans cette méthode. Cette méthode traite globalement l'individu en imputant simultanément des données manquantes. Une autre caractéristique de la méthode RV est son invariance par rapport à des translations, des transformations orthogonales et des homothéties de rapport constant.

Voici un exemple traitant des données artificielles permettant de visualiser ce que fait la méthode RV pour deux blocs de variables Y_1 et Y_2. Les données comprennent 5 individus et 4 variables :

$$X = \begin{array}{c} \begin{matrix} y_1 & y_2 & y_3 & y_4 \end{matrix} \\ \left(\begin{array}{cc|cc} 1 & 3 & 1 & 5 \\ 4 & 4 & 4 & 6 \\ 6 & 3 & 8 & 5 \\ 5 & 1 & 5 & 3 \\ ? & -1 & 7 & 1 \end{array}\right) \end{array}$$

où la première variable manque pour le dernier individu.

Dans ce cas :

$$RV(u) = \frac{4.062u^2 + 11.739u + 41.563}{9.076\sqrt{u^4 + 12.625u^2 + 3.354u + 28.752}}$$

$$avec\ u = \frac{y_1 - 4}{\sqrt{5}}$$

La figure 2.32 permet de trouver la solution de la méthode RV.

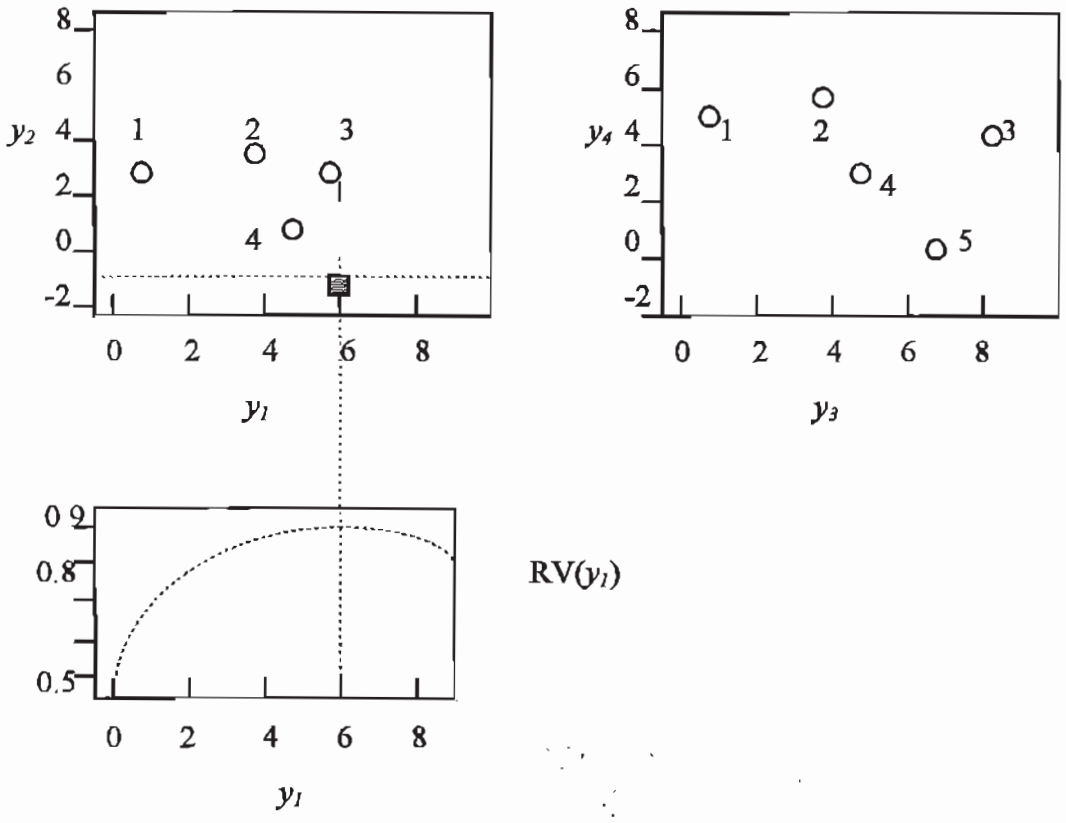


Figure 2.32 Illustration de la méthode RV

La valeur maximale de RV, RV_{max}=0.935, est obtenue pour une valeur de y_{1max}=6.07 (elle correspond à u_{max}=0.9257). La situation des différents points montre bien que cette imputation maximise la similitude entre les deux nuages de points

Aspects pratiques : la délimitation des variables en deux groupes est assez fréquente dans la pratique ; il peut s'agir de la répétition d'un certain nombre de tests ou de l'opposition entre variables endogènes et exogènes. La méthode RV peut s'appliquer même si les données n'ont pas naturellement deux groupes de variables, il suffit pour cela que la délimitation des deux groupes de variables induise un coefficient RV suffisamment élevé.

Si r vecteurs individus ont des données manquantes, nous traitons séparément et séquentiellement chacun des r vecteurs incomplets avec $(n-r)$ vecteurs complets.
Si des données manquantes sont localisées dans les deux groupes de variables (alors (u_1, u_2) représentent les parties inconnues), la formule $RV(u_1, u_2)$ se complique mais la procedure d'imputation s'applique de la même façon.

Condition d'application :

- Si le coefficient RV sur les données complètes est trop faible, il vaut mieux ne pas utiliser la méthode. Il est difficile de chiffrer a priori la structure de corrélation nécessaire : un test de nullité de ρv basé sur la distribution asymptotique permettra de définir une valeur critique de RV (Cléroux & Ducharme, 1989). Cette remarque s'applique aussi à la méthode de Buck et ses variantes.

- Si les données présentes pour l'individu ayant la ou les données manquantes sont mauvaises (données aberrantes), la méthode RV ainsi que les autres méthodes ne donneront pas de bons résultats. Cette remarque s'applique aussi au cas où les données complètes contiennent des données aberrantes. En effet, il est possible de modifier la méthode RV pour atténuer l'influence des données aberrantes sur l'imputation. L'idée est simple, elle consiste à utiliser une technique robuste pour calculer $\hat{\bar{X}}$ *et* $\hat{\Sigma}$, l'estimateur du vecteur d'espérance et de la matrice de variance-covariance basée sur les $n-1$ premiers individus (sans données manquantes) : $\hat{\bar{X}}^r_{n-1}$ *et* $\hat{\Sigma}^r_{n-1}$. On définit $S^*_n = \frac{n-2}{n-1}\hat{\Sigma}^r_{n-1} + \frac{1}{n}(X_n - \hat{\bar{X}}^r_{n-1})(X_n - \hat{\bar{X}}^r_{n-1})^T$. Les deux méthodes sont identiques sauf pour l'évaluation de $\hat{\bar{X}}$ *et* $\hat{\Sigma}$.

2.2.3. Imputation multiple

Lorsque nous faisons l'imputation simple, nous faisons comme si la valeur imputée était certaine, ce qui ne peut pas refléter la variabilité des échantillons sous un modèle pour des données manquantes.

L'imputation multiple proposée par Rubin (1978) consiste à remplacer chaque donnée manquante par m (>= *2*) valeurs d'imputations tirées de un ou plusieurs modèles d'estimation.

Si l'on se base sur un modèle d'estimation, cette méthode permet d'effectuer m imputations de meilleures probabilités, selon la probabilité de distribution sous le même modèle. Les m inférences obtenues sur les données ainsi complétées peuvent être combinées pour fournir une inférence qui reflètent l'incertitude due aux non-réponses. Si on choisit plusieurs modèles d'estimations, les différentes combinaisons des m inférences obtenues sur les données ainsi complétées reflètent l'incertitude au niveau du modèle. Cette variation des valeurs montre la sensibilité des données au modèle d'estimation de données manquantes.

L'imputation multiple (sous un ou plusieurs modèles) simule la distribution a postériori des données manquantes sous ce(s) modèle(s) et Rubin indique que cela nous permet d'avoir des variances correctes.

Les inconvénients consistent dans les points suivants :

- la complexité de calculs multiples à faire selon un ou plusieurs modèle(s) et le temps de calcul est considérable,

- les nombreuses données à gérer, car la taille d'échantillon est augmentée et éventuellement un redressement est à envisager.

2.3. ANALYSE HOMOGENE POUR DES DONNEES CATEGORIELLES

Considérons le cas où les observations sont décrites par p variables catégorielles. L'analyse homogène est une présentation de l'analyse des correspondances multiples qui se prête bien à une extension pour des données manquantes. Elle revient à chercher des quantifications simultanées des catégories et des individus (représentation sur des axes factoriels) telles que les individus soient proches des catégories qu'ils prennent et que les différents individus et les catégories des variables aient des valeurs aussi distinctes que possible. Nous présentons ce que sont les formules de l'analyse homogène pour données complètes avant de l'étendre à des données incomplètes.

2.3.1. Transformation optimale de données complètes et analyse homogène

Lorsque les variables mesurent plus ou moins la même propriété, il est possible de remplacer les observations sur les différentes variables d'une unité par une valeur d'une variable synthétique sans perdre trop d'informations. La petitesse des pertes est en fonction de l'homogénéité des variables. Pour évaluer le succès de cette substitution, nous allons définir un critère d'homogénéité et une fonction de perte. Le processus de quantification des variables en maximisant l'homogénéité des variables s'appelle l'analyse homogène.

Transformation non linéaire pour des variables quantitatives complètes :

Supposons que nous avons n unités et m variables $h_j = \begin{pmatrix} h_{j1} \\ \vdots \\ h_{jn} \end{pmatrix}$, $j=1,2,...m$.

La variable transformée par une fonction non linéaire φ_j est : $\phi_j(h_j) = \begin{pmatrix} \phi_j(h_{j1}) \\ \vdots \\ \phi_j(h_{jn}) \end{pmatrix}$ et

le score individuel est défini par $x = \frac{1}{m}\sum_{j=1}^{m} \phi_j(h_j) = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, égale à la moyenne des variables transformées.

Le but est de chercher une série de fonctions non linéaires $\varphi=(\varphi_1,\varphi_2,\cdots,\varphi_m)$ telles que les transformations $\varphi(h_j), j = 1,2,\ldots,m$, soient les plus proches possibles de x.

La variation totale des données transformées vaut :

$$\sum_{j=1}^{m}\left\|\phi_j(h_j)\right\|^2 = m\|x\|^2 + \sum_{j=1}^{m}\left\|x-\phi_j(h_j)\right\|^2$$

$$\text{T} \quad = \text{B} \quad + \quad \text{W}$$

$$= \text{Variation(interclasse)} + \text{Variation(intraclasse)}$$

- B la variation interclasse fait référence à une discrimination entre les différentes unités,
- W la variation intraclasse fait référence à un manque d'homogénéité des variables transformées

$$\phi_1(h_{1i}) \neq \phi_2(h_{2i}) \neq \cdots \neq \phi_m(h_{mi}),\ i=1,2,\ldots m.$$

En maximisant $\eta = \dfrac{B}{T}$, un critère de discrimination entre unités, nous cherchons à transformer les variables. La fonction de perte d'homogénéité est donc définie comme suit :

$$\sigma(x,\phi) = \frac{1}{m}\sum_{j=1}^{m}\left\|x-\phi_j(h_j)\right\|^2$$

Il s'agit ici d'une ACP non linéaire ou plus exactement semi-linéaire (El-Faouzi N.E., 1992). En pratique, on se limite à des données de transformation appartenant à des espaces de dimension finie comme les transformations spline (Van Rijckevorsel J.L.A , 1982).

Transformation non linéaire pour les variables qualitatives complètes :

Dans ce cas, nous voudrions quantifier les catégories des variables par une technique qui conduise à une solution optimale

Soit $H_{nxm} = (h_1, h_2, \ldots, h_m)$ le tableau brut, G_j le tableau des indicatrices de la variables h_j, $j=1,2,\ldots m$, $G_j = \left(g_{il}\right)_{n\times j_k}$,

$$g_{il} = \begin{cases} 1 \text{ si ième observation appartient à la caté gorie} l \text{ de la variable hj} \\ 0 \text{ sinon} \end{cases}$$

n - le nombre individuel, j_k- le nombre de catégories de la variable h_j . m- le nombre de variables.

$G = (G_1,G_2,...,G_m)$, est le tableau disjonctif complet, $Dj = Gj' Gj$ une matrice diagonale dont les éléments diagonaux correspondent aux marges des variables h_j,

$D=\begin{pmatrix} D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_m \end{pmatrix}$, $y_j = \begin{pmatrix} y_{j1} \\ \vdots \\ y_{jj_k} \end{pmatrix}$ la quantification de catégorie de la variable h_j,

$G_j y_j = \begin{pmatrix} q_{j1} \\ \vdots \\ q_{jn} \end{pmatrix}$, le score d'unité sur variable h_j ou bien les coordonnées des catégories de la variable h_j. On remplace la fonction de transformation non linéaire $\varphi(h_j)$ par $G_j y_j$, nous avons :

le score individuel x se définit $x = \frac{1}{m}\sum_{j=1}^{m} G_j y_j = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$,

Nous mesurons donc la fonction de perte d'homogénéité par

$$\sigma(x,y) = \frac{1}{m}\sum_{j=1}^{m} \left\| x - G_j y_j \right\|^2 \qquad (3)$$

Nous chercherons le minimum de $\sigma(x,y)$ par deux voies différentes : la maximisation de la discrimination entre individus et la maximisation de l'homogénéité des variables.

- Par la maximisation de la discrimination entre individus

Pour une valeur fixée du score individuel x, nous cherchons le minimum de $\sigma(x,y)$ en fonction des quantifications de catégorie y, noté $\sigma(x,*)$

$$\sigma(x,*) = \min_{y}\{\sigma(x,*)|y\}.$$

En mettant la dérivation partielle de l'équation ci-dessus égale à zéro, on trouve l'optimal y pour un x donné (Meulman J., 1982) :

$$y_j = (G_j{}'G_j)^{-1}G_j{}'x = D_j^{-1}G_j{}'x$$

ce qui montre que le y_j optimal pour un x donné est la moyenne des scores des individus appartenant aux catégories concernant la variable j.

On remplace y_j dans l'équation (3), on obtient :

$$\sigma(x,*) = \frac{1}{m}\sum_{j=1}^{m}\left\|x - G_j(G_j'G_j)^{-1}G_j'x\right\|^2 \qquad (4)$$

Posons $P_j = G_j(G_j'G_j)^{-1}G_j'$, P_j est un projecteur orthogonal de x sur un sous-espace engendré par les colonnes de G_j ;

$$\sigma(x,*) = \frac{1}{m}\sum_{j=1}^{m}(x - P_jx)'(x - P_jx) \qquad (5)$$

et posons $P_0 = \sum_{j=1}^{m} P_j$, nous obtenons :

$$\sigma(x,*) = \frac{1}{m}\sum_{j=1}^{m}(x'x - x'P_jx) = x'x - \frac{x'P_0x}{m} = x'x(1 - \frac{1}{m}\frac{x'P_0x}{x'x}) \qquad (6)$$

Pour éviter la solution «$x = U$» (U : un vecteur dont tous les éléments sont à 1), nous exigeons que $U'x = 0$. Nous évitons en plus la solution «$x = 0$» en demandant une minimisation de $\sigma(x,*)$ sur x satisfaisant $x'x = I$. Cela nous conduit à maximiser $x'P_0x$ sous la contrainte suivante : $U'x = 0$.

Posant $J_m = (I - \frac{UU'}{U'U})$, la contrainte sur x peut donc être traduite en une projection de J_m sur x, donc l'équation (6) avec la contrainte de normalisation de x conduit à :

$$\sigma(x,*) = 1 - \frac{1}{m}x'J_m'P_0J_mx \qquad (7)$$

Le minimum de $\sigma(x,*)$ sur x est en conséquence égal à

$1 - \frac{\lambda^2}{m}$ *où* λ^2 est la plus grande valeur propre de $J_m'P_0J_m$.

En effet, x est le vecteur propre de la matrice $GD^{-1}G'$ si on ignore le problème de normalisation.

Où $\frac{1}{m}\lambda^2$ est appelé **'le coefficient d'homogénéité'** de l'analyse.

- Par la maximisation de l'homogénéité entre variables

Nous minimisons la fonction de perte sur le score individuel x par rapport à un y donné.
La fonction de la **perte d'homogénéité** s'écrit

$$\sigma(*,y) = \min_{X} \{\sigma(*,y)\, x\}$$

Le minimum de $\sigma(*,y)$ par rapport à un y donné conduit à l'équation suivante (Meulman J., 1982) :

$$x = \frac{1}{m}\sum_{l=1}^{m} G_l y_l = \frac{1}{m} Gy \qquad (9)$$

Ici nous trouvons que le score individuel x optimal est la moyenne des catégories quantifiées correspondantes. On substitue (9) dans l'équation (3) :

$$\sigma(*,y) = \frac{1}{m}\sum_{j=1}^{m} (\frac{1}{m}Gy - G_j y_j)'(\frac{1}{m}Gy - G_j y_j) \qquad (10)$$

On remplace $G'G$ par C et $G_j'G_j$ par D_j, on obtient

$$\sigma(*,y) == \frac{1}{m} y'Dy(1 - \frac{y'Cy}{my'Dy}) \qquad (11)$$

La minimisation de $\sigma(*,y)$ sur y avec la contrainte $y'Dy = m$ conduit à la maximisation de $y'Cy$. On peut donc interpréter le problème comme suit : nous recherchons les quantifications des catégories telles que la somme des covariances des variables quantifiées $G_j y_j$ soit maximum, lorsque la somme des variances reste constante. Donc, on peut interpréter l'analyse homogène comme une analyse en composantes principales sur des variables catégorielles quantifiées

Pour éviter la solution triviale d'une constante pour y, en posant $J_D = (I - \frac{UU'D}{U'DU})$, donc on a $U'DJ_D y = 0$.

Nous introduisons les contraintes ci-dessus de normalisation dans la fonction de perte (11) :

$$\sigma(*, y) = 1 - \frac{1}{m^2} y' J_D C J_D y$$

Nous obtenons le minimum qui est égal à $1 - \frac{1}{m}\lambda^2$, λ^2 est valeur propre de $(J_D' G' G J_D)$ et y-vecteur propre de $(J_D' G' G J_D)$. Cette méthode est donc équivalente à **'L'analyse factorielle du tableau de Burt'.** (GJ_D) a le rang maximal $(\sum k_j - m)$, donc au plus $p_m = (\sum k_j - m)$ solutions non triviales. De cette façon, les vecteurs indésirables sont enlevés.

Cela nous montre que les deux façons de minimiser $\sigma(x,y)$ conduisent finalement au même résultat en terme de la fonction de perte. Il est évident que les valeurs pour x des deux solutions ne sont pas tout à fait les mêmes. Mais elles diffèrent seulement par un facteur multiplicatif à cause des contraintes de normalisation. Par contre, la matrice de corrélation des deux solutions est tout à fait la même.

Dans la minimisation de la perte discriminante, la catégorie quantifiée est au barycentre des scores individuels concernés. Dans la minimisation de la perte d'homogénéité, le score individuel est au barycentre des catégories quantifiées concernées. La solution stable et finale existe bien. Cela conduit à la méthode des moyennes réciproques : le score d'unité optimal par rapport à la quantification des catégories et la quantification des catégories optimales par rapport au score d'unité. Alors nous pouvons les utiliser pour définir un algorithme itératif qui approche d'une solution cible en remplaçant la procédure de la décomposition de la valeur propre. Nous allons faire un choix entre deux parcours . la normalisation de y ou la normalisation de x.

L'algorithme des moyennes réciproques avec la normalisation de x : HOMogeneity Analysis by Alternative Least Squares (HOMALS)

Le HOMALS prend la deuxième normalisation avec $x'x = n$, alors x devient un score standard. La raison du choix est que dans l'application, le nombre d'individus n est souvent beaucoup plus grand que le nombre des catégories $(\sum k_j)$; il est plus pratique sur la

présentation graphique que les points objets soient répartis de façon équilibrée dans toutes les directions et que les points catégorielles indiquent le centre de gravité des sous-groupes d'individus.

Ici, nous ne cherchons pas une solution complète de 'S V.D.', mais nous appliquons plutôt l'idée de la moyenne réciproque : la catégorie quantifiée est la moyenne des scores individuels appropriés (pour la première fois, c'est un choix arbitraire pour des scores individuels $x \neq 0$ et on rend x centré réduit). Une fois que nous avons des valeurs de x, on commence le calcul par l'étape 1. Le prochain calcul de score individuel est basé sur les catégories quantifiées obtenues et leur normalisation est telle que $x'x = I$ et $U'x = 0$, la procédure d'itération se traduit comme suit :

1. $\tilde{y} = D^{-1}G'x^t$
2. $\tilde{x} = \frac{1}{m}G\tilde{y}$
3. $x^{t+1} = \tilde{x}(\tilde{x}'\tilde{x})^{-\frac{1}{2}}$ *sous la condition* $U'x = 0$
4. *Test de convergence et retour à l'étape* 1.

Le calcul du score individuel et de la quantification s'arrête quand $(\tilde{x}'\tilde{x})$ s'approche d'un constante et qu'ils ne changent plus de valeurs. Il s'agit d'une méthode de moindres carrés alternatifs. Dans la pratique, cette itération peut devenir :

1. $\tilde{y} = D^{-1}G'x^t$,
2. $\tilde{x} = \frac{1}{m}G\tilde{y}$,
3. $x^{t+1} = \sqrt{n}\tilde{x}(\tilde{x}'\tilde{x})^{-\frac{1}{2}}$ *et centrer* x^{t+1},
4. *Test de convergence et retour à l'étape* 1.

l'étape 3 a changé et x est transformé tel que $x'x = n$. Donc la variance de x est égale à 1. lorsque x est multiplié par un facteur $\sqrt{n}$, y est automatiquement multiplié par le même facteur dans l'étape 1. Donc, cette procédure reste valable dans ce cas. Théoriquement il est plus commode de garder $x'x = 1$.

Dans l'analyse homogène unidimensionnelle, les variables sont quantifiées pour que les valeurs soient les plus homogènes possibles, donc il est possible que les variables soient remplacées par une seule variable synthétique de score individuel. Par ailleurs, dans certains cas, il est utile de

trouver plusieurs versions de la quantification en même temps. Par exemple, certaines variables contribuent peu à l'analyse sur la première dimension : mais il pourrait être intéressant de continuer à trouver si la deuxième meilleure solution donne plus d'importance à ces variables ce qui indique l'existence d'une forme d'homogénéité multiple. Pour des variables nominales, il n'est pas très logique de proposer de les quantifier en simples variables numériques, car cela est équivalent à les considérer comme des variables ordonnées.

2.3.2. Quantification de données qualitatives incomplètes

Meulman J. (1982) a introduit le processus de quantification de données catégorielles dans le cas de données incomplètes en maximisant le coefficient d'homogénéité.

La matrice indicatrice G_j possède une propriété spéciale : la somme d'une ligne est égale à 1 parce que chaque objet est dans une et une seule catégorie. Lorsqu'il existe des données manquantes, la somme d'une ligne du tableau disjonctif ne donne pas *1*. Donc la somme des lignes du tableau disjonctif n'est plus égale à *m*, le nombre des variables. Cela pourrait perturber certaines propriétés de l'analyse d'homogénéité Donc, on devait reformuler les critères, la minimisation de la fonction de perte en terme de matrice indicatrice incomplète.

Soit M_j la matrice diagonale *n x n* telle que

$M_j(i, i)$=1 si la i^{ème} observation est complète pour la variable *j*, 0 sinon.

M_j = I si il n'y a pas de données manquantes.

Réécrivons alors la fonction de perte $\sigma(x,y) = \frac{1}{m}\sum_{j=1}^{m}\left\|x - G_j y_j\right\|^2$ avec une matrice indicatrice incomplète G_j, on a :

$$\sigma(x,y) = \frac{1}{m}\sum_{j=1}^{m}(x - G_j y_j)' M_j (x - G_j y_j) \qquad (15)$$

Cette fonction de perte est adaptée à la fois au tableau disjonctif complet et incomplet.

En posant $M_0 = \frac{1}{m}\sum_{j=1}^{m} M_j = \frac{1}{m} M$,

$$\sigma(x,y) = \frac{1}{m}\sum_{j=1}^{m}(x' M_j x + y_j' G_j' M_j G_j y_j - 2x' M_j G_j y_j)$$

Comme $M_j G_j = G_j$, nous obtenons :

$$\sigma(x,y) = \frac{1}{m}(x' Mx + y' Dy - 2x' Gy) \qquad (16)$$

La minimisation de perte de cette fonction conduit alors aux équations simultanées :

$$\begin{cases} y = D^{-1} G' x \\ x = M^{-1} G y \end{cases} \qquad (17)$$

Le score individuel est au barycentre des catégories répondantes concernées et lorsqu'il n'existe pas de données manquantes, $M^{-1} = \frac{1}{m}\mathrm{I}$, c'est-à-dire que l'équation (17) s'adapte également aux cas des données complètes.

$$\begin{aligned} \sigma(x,*) &= \frac{1}{m}(x' Mx - x' G D^{-1} G x) \\ &= x' M_0 x - x' P_0 x \\ &= x' M_0 x (1 - \frac{x' P_0 x}{x' M_0 x}) \end{aligned}$$

$$\begin{aligned} \sigma(*,y) &= \frac{1}{m}(y' Dy - y' G' M^{-1} G y) \\ &= \frac{1}{m}\, y' Dy (1 - \frac{y' G' M^{-1} G y}{y' Dy}) \end{aligned}$$

Donc, la contrainte de normalisation sur y devient $\begin{cases} y' Dy = m \\ U' Dy = 0 \end{cases}$,

et la normalisation de x est $\begin{cases} x' M_0 x = 1 \\ U' M_0 x = 0 \end{cases}$.

En posant $J_M = (I - \frac{UU'M}{U'MU})$, $J_M x$ est M-centrée .

$$H = M^{-\frac{1}{2}} G D^{-\frac{1}{2}} - \frac{M^{\frac{1}{2}} \vec{U}\vec{U}' D^{\frac{1}{2}}}{\vec{U}M\vec{U}'} = M^{-\frac{1}{2}} J_M' G J_D D^{-\frac{1}{2}},$$

$$HH' = M^{-\frac{1}{2}} J_M' G D^{-1} G' J_M M^{-\frac{1}{2}} = m M^{-\frac{1}{2}} J_M' P_0 J_M M^{-\frac{1}{2}},$$

$$H'H = D^{-\frac{1}{2}} J_D' G' M^{-1} G J_D D^{-\frac{1}{2}}.$$

Si l'on fait le changement de variable suivant $\begin{cases} a = D^{\frac{1}{2}} y \\ z = M^{\frac{1}{2}} x \end{cases}$, on obtient :

$$\begin{aligned} \sigma(z,*) &= 1 - \frac{1}{m}(z' M^{-\frac{1}{2}} J_m' P_0 J_m M^{-\frac{1}{2}} z) \\ &= 1 - \frac{1}{m} z' HH' z \end{aligned} \qquad (18)$$

$$\begin{aligned} \sigma(*,a) &= 1 - \frac{1}{m}(a' D^{-\frac{1}{2}} J_D' G' M^{-1} G J_D D^{-\frac{1}{2}} a) \\ &= 1 - \frac{1}{m} a' H' H a \end{aligned} \qquad (19)$$

La minimisation des équations (18) et (19) est égale à $1 - \frac{1}{m}\lambda^2$ *où* λ^2 est la plus grande valeur propre de HH' pour (18) et $H'H$ pour (19) respectivement. Le résultat est similaire à celui des données complètes.

En ignorant pour l'instant la solution triviale, on a $H'H = D^{-\frac{1}{2}} G' M^{-1} G D^{-\frac{1}{2}}$ qui est composé d'un ***tableau de Burt modifié*** $G' M^{-1} G$. Pour donner une idée de la différence entre $G'G$ *et* $G' M^{-1} G$, nous allons comparer G' *et* $G' M^{-1}$ sur ce petit exemple à *4* individus et *3* variables :

$$\begin{pmatrix} ? & 1 & 2 \\ 2 & 2 & 1 \\ 1 & 3 & ? \\ 2 & 2 & 2 \end{pmatrix}$$

Soit $G_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$, $G_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$, $G_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}$, donc $M_0^{-1} = \begin{pmatrix} \frac{3}{2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{3}{2} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

$$G'M_0^{-1} = \left(\begin{array}{cccc} 0 & 0 & 1+\frac{1}{2} & 0 \\ 0 & 1 & 0 & 1 \\ \hdashline 1+\frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1+\frac{1}{2} & 0 \\ \hdashline 0 & 1 & 0 & 0 \\ 1+\frac{1}{2} & 0 & 0 & 1 \end{array}\right), \qquad G' = \left(\begin{array}{cccc} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ \hdashline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ \hdashline 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{array}\right) \leftarrow \text{catégorie}$$

individu

Posons $\tilde{G} = G'M_0^{-1}$, $\tilde{G}$ est quasiment identique à G' sauf sur les colonnes où des individus ont des données manquantes. Une constante égale à $\frac{k}{m-k}$ (k - nombre de données manquantes totales sur l'individu concerné et m - nombre des variables) s'ajoute sur les catégories répondantes. Les tableaux $G'G$ *et* $G'M^{-1}G$ sont :

$$Tableau\ G'G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 2 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 2 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 2 \end{pmatrix}$$

$$Tableau\ G'M_0^{-1}G = \begin{pmatrix} 1+\frac{1}{2} & 0 & 0 & 0 & 1+\frac{1}{2} & 0 & 0 \\ 0 & 2 & 0 & 2 & 0 & 1 & 1 \\ 0 & 0 & 1+\frac{1}{2} & 0 & 0 & 0 & 1+\frac{1}{2} \\ 0 & 2 & 0 & 2 & 0 & 1 & 1 \\ 1+\frac{1}{2} & 0 & 0 & 0 & 1+\frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1+\frac{1}{2} & 1 & 0 & 0 & 2+\frac{1}{2} \end{pmatrix}$$

En comparaison $G'G$ *et* $G'M^{-1}G$, on trouve que $G'M^{-1}G$ est égale à $G'G$ plus une constante, égale à $\frac{k}{m-k}$, sur les catégories répondantes de l'individu 1 : $\frac{1}{2}$ pour les 3ème et 7ème catégories et sur les catégories répondantes de l'individu 3 . $\frac{1}{2}$pour les 1ère et 5ème catégories $G'G$.

En terme général, nous avons tout à fait la même conclusion pour les données incomplètes que pour le cas des données complètes mis à part la matrice M. Mais avec les données manquantes, nous ne pouvons pas interpréter strictement l'analyse homogène comme une analyse en composantes principales sur données quantifiées, puisqu'on ne minimise pas la fonction de perte sur des catégories manquantes. L'équation (17) montre que le score individuel est la moyenne des catégories répondantes concernées et ensuite après avoir obtenu les scores individuels, une catégorie quantifiée est la moyenne des scores individuels des répondants concernés. Donc les données manquantes n'ont pas influencé les scores individuels. Ces catégories manquantes n'ont pas d'effet dans l'analyse puisque la fonction de perte n'est pas minimisée sur elles. Autrement dit, la quantification est basée sur les données disponibles en ignorant les données manquantes · c'est-à-dire les **"données manquantes passives"** dans l'analyse.

La fonction de perte s'écrit :

$$\sigma(X,Y)=\frac{1}{m}\sum_{j=1}^{m}(X-G_jY_j)'M_j(X-G_jY_j)$$

et la solution est

$$\begin{cases}Y_j = D_j^{-1}G_j'X\\ X = M^{-1}GY\end{cases}$$

En posant $M=\sum_j M_j$, la contrainte de normalisation devient donc

$$\begin{cases}U'MX=0 & (a)\\ X'\dfrac{M}{m}X=I & (b)\end{cases} \qquad (18)$$

La phase de normalisation (20) se décompose en la variante de l'orthogonalisation de *Gram-Schmidt* suivante :

$$\begin{aligned}&1.\ \tilde{Z}=(I-\frac{UU'M}{U'MU})X\\ &2.\ X=(\frac{M}{m})^{\frac{1}{2}}GRAM[(\frac{M}{m})^{\frac{1}{2}}\tilde{Z}]\end{aligned} \qquad (20)$$

L'étape 1 correspond à une procédure M-centrée : $U'M\tilde{Z}=0$. L'expression $(\frac{M}{m})^{\frac{1}{2}}\tilde{Z}$ rend $\tilde{Z}$ conforme à la contrainte (b) du (20) ci-dessus. $GRAM[(\frac{M}{m})^{\frac{1}{2}}\tilde{Z}]$ donne la décomposition de Gram-Schmidt noté S, $X=(\frac{M}{m})^{\frac{1}{2}}S$ transforme S tel que $S'\frac{M}{m}S=I$.

Cette décomposition est nommée la ***variante d'orthogonalisation de Gram-Schmidt***. Nous devons indiquer ici que la solution sortant de cet algorithme n'est pas exactement la solution des moindres carrés. Cette variante de *Gram-Schmidt* est une solution sous-optimale et la convergence de celle-ci a été montrée par A.Gifi (1990).

2.3.3. Quantification de données qualitatives incomplètes et imputation homogène par maximisation du coefficient d'homogénéité

Pour des données qualitatives complètes, $X = \frac{1}{m}GY$, Y-variable quantifiée, G-tableau disjonctif complet, la variation totale des données quantifiées :

$$\sum_{j=1}^{m}\left\|G_j Y_j\right\|^2 = m\|X\|^2 + \sum_{j=1}^{m}\left\|X - G_j Y_j\right\|^2 \qquad (3)$$

$$T = B + W$$

Le coefficient d'homogénéité : $\eta = \frac{B}{T}$ mesure à quel degré le score d'unité pourrait être considéré comme un représentant de chaque unité :

Plus la valeur η est grande, mieux $x_i - x_{i'}$ représente la différence entre unités *i et i'*. Nous considérons les données manquantes comme des paramètres qui vont être déterminés par l'ensemble des données disponibles. Les valeurs d'imputation pour données manquantes qui sont plus homogènes à l'ensemble des données disponibles sont dans la maximisation de η sur $X, y_1, y_2, \cdots, y_m$ et $g_1, g_2, \ldots, g_m$.

Le résultat de la maximisation de η est équivalent à celui de la minimisation de $\sigma(*)$ sous des contraintes.

La fonction de perte d'homogénéité est :

$$\sigma(X; y_1, \cdots, y_m, g_1^*, \cdots, g_m^*) = \sum_{j\in\Omega}\left\|X - g_j y_j\right\|^2 + \sum_{j\notin\Omega}\left\|X - g_j^* y_j\right\|^2$$

g_j^* - matrice indicatrice incomplète de la variable j.

Ω représente l'ensemble des variables ayant les réponses complètes.

Le problème étant de savoir où l'on impute "1" dans le vecteur manquant de g_j* Nous allons prendre le même principe que celui de la quantification : la minimisation de la fonction de perte d'homogénéité $\sigma(*)$. Buuren S.V. & Van Rijckevorsel J.L.A. en 1992 proposent un algorithme "*K-mean*" modifié pour résoudre ce problème.

L'algorithme "K-mean" modifié :

Si l'on suppose que nous commençons avec une certaine imputation initiale pour une donnée manquante, nous examinons chaque imputation l'une après l'autre et nous comparons le changement de catégorie courante *s* à une nouvelle catégorie *t* sur laquelle nous chercherions le minimum de perte d'homogénéité. Nous appliquons l'algorithme "*K-mean*" pour chaque variable et unité par unité, l'imputation s'exécute selon la fonction de perte. Supposons d_S, d_t, le nombre d'observations des catégories *s* et *t* de variable *j* et les quantifications y_s, y_t des catégories respectives. Si l'unité *i* a un score x_i et lorsque nous changeons l'imputation de la catégorie *s* à la catégorie *t*, alors Fisher (1958) montre que la nouvelle perte est égale à :

$$\sigma^*(\bullet) - \frac{d_s(x_i - y_s)}{d_s - 1} + \frac{d_t(x_i - y_t)}{d_t + 1}$$

Donc, la règle d'imputation :

$$si \ \frac{d_t(x_i - y_t)}{d_t + 1} < \frac{d_s(x_i - y_s)}{d_s - 1}$$

nous imputons la catégorie *t* à la place de celle de *s*. En même temps, les quantifications y_s, y_t des catégories respectives changent également,

$$\begin{cases} \hat{y}_s = y_s + \dfrac{x_i - y_s}{d_s - 1} \\ \hat{y}_t = y_t + \dfrac{x_i - y_t}{d_t - 1} \end{cases}$$, mais nous devrions nous assurer que *ds* >=1.

Algorithme principal de l'imputation homogène en maximisant le coefficient d'homogénéité

I. ***'Données manquantes passives'***, une phase de quantification dont nous ignorons les données manquantes dans l'analyse où nous calculons le score individuel et la quantification des variables selon le processus suivant (comme indiqué dans 2.3.2) :

X - valeur initiale d'un choix arbitraire (pour l'analyse multidimensionnelle, c'est une matrice de plein rang) ·

$$\begin{cases} y = D^{-1}G'x \\ x = M^{-1}Gy \end{cases}, \text{ avec la contrainte de normalisation } \begin{cases} U'MX = 0 \\ X'\dfrac{M}{m}X = I \end{cases}$$

On répète cette série de calculs et on s'arrête lorsque la valeur de perte ne s'améliore pas ou si le nombre d'itérations dépasse un certain nombre seuil.

II. Imputation initiale

Nous imputons une catégorie manquante en choisissant la catégorie dont la valeur de quantification est la plus proche du score individuel comme imputation initiale.

III. Imputation pour données manquantes et quantification des variables

Nous répétons le processus suivant avec une phase de contrôle indiqué dans la phase I :

- Calcul de X selon $X = \frac{1}{m}GY$, G étant pseudo-complet,

- Centrer, orthogonaliser X, X=*GRAM*(X) : par le processus indiqué dans (21)

- Utilisation de l'algorithme "*K-mean*" pour trouver l'imputation de la catégorie la plus homogène en prenant G pseudo-complet précédemment calculé comme valeurs initiales ;

- Pour chaque variable j faire :

  - Répéter le processus suivant jusqu'à ce qu'il n'y ait plus de changement d'imputation de catégorie ou (nombre de répétition >= n)

    - Utiliser l'algorithme "*K-mean*" pour trouver la catégorie qui rend minimale la perte sur $\sigma^{*}(\bullet)$,

      g_j^{*}<--l'imputation homogène

    calcul de $Y = D^{-1}G'X$.

  calcul $Y = D^{-1}G'X$.

Fin

2.3.4. Exemple d'analyse homogène pour l'imputation des données manquantes (Buuren S.V. & Van Rijckevorsel J.L.A., 1992)

Cette méthode se décompose en deux étapes :

1. définir une variable de score d'unité qui pourrait mesurer la différence entre les unités.
2. définir une règle d'imputation : des données manquantes sont reconstituées relatives à la valeur de score individuel et à la fonction de perte.

Ces deux étapes sont tout à fait interactives. Un changement de valeur du score individuel peut causer une modification d'imputation. Le phénomène inverse se produit également : un changement d'imputation a une action sur les variables de scores. Une homogénéité simultanée sur toutes les variables est exprimée par la variable de score. Cette variable est égale à la moyenne de toutes variables transformées (quantifiées). En comparant avec la méthode "hot-deck", cette variable ressemble à une partition d'unités en classes homogènes utilisée par la procédure "hot-deck" traditionnelle.

La différence avec la procédure "hot-deck" réside en ceci : toutes les variables interviennent en même temps dans la variable de score ; donc elles interviennent dans le choix de la valeur imputée. Mathématiquement, nous cherchons une imputation qui est la plus homogène pour toutes les variables. Nous faisons ici l'imputation des données manquantes en maximisant l'homogénéité de l'ensemble des données. Pour résoudre la distance entre deux catégories, nous transformons des variables qualitatives en variables quantitatives.

La difficulté avec les variables qualitatives est que la mesure de distance entre deux unités est difficile à évaluer. Nous résolvons ce problème par un processus de quantification des variables et de distribution à chaque unité d'un score La distance du score mesure une différence entre les unités.

Par exemple, pour des données suivantes :

Individu	**Revenu**	**Age**	**Car**
1	*x*	jeune	am
2	moyen	moyen	am
3	*y*	âgé	jap
4	bas	jeune	jap
5	moyen	jeune	am
6	haut	âgé	am
7	bas	jeune	jap
8	haut	moyen	am
9	haut	*z*	am
10	bas	jeune	am

x, y, z sont des données manquantes. Le problème est de trouver des valeurs de remplacement raisonnables pour *x, y et z*. Il existe 3 x 3 x 3 = 27 solutions possibles. Le choix de notre solution est une imputation qui rend le critère d'homogénéité η maximal.

Voici la liste des valeurs du coefficient η avec les 27 imputations possibles :

x	y	z	η	x	y	z	η	x	y	z	η
b	b	j	.70104	m	b	j	.63594	h	b	j	.61671
b	b	m	.77590	m	b	m	.72943	h	b	m	.66458
b	b	a	.76956	m	b	a	72636	h	b	a	.65907
b	m	j	.78043	m	m	j	.70106	h	m	j	.70106
b	m	m	.84394	m	m	m	.77839	h	m	m	.74342
b	m	a	.84394	m	m	a	84394	h	m	a	.74342
b	h	j	.78321	m	h	j	.73319	h	h	j	.68827
b	h	m	.84907	m	h	m	.80643	h	h	m	.74193
b	**h**	**a**	***.84964**	m	h	a	.80949	h	h	a	.74198

La recherche optimale est donc x=*b* ; y=*h* ; z=*a*.

Tableau des scores individuels :

	Score d'unité	
	Initial	**Final**
1 b* j a	- 1.43	- 1.33
2 m m a	0.79	0.66
3 h* a j	1.02	1.0
4 b j j	- 1.41	- 1.33
5 m j a	0.04	- 0.01
6 h a a	1.11	1.0
7 b j j	- 1.41	- 1.33
8 h m a	1.05	0.92
9 h a* a	1.02	1.0
10 b j a	- 0.58	- 0.59

Quantification des catégories :

Variable	**Quantification**	
	Initial	**Final**
Revenu : Bas	-1.13	1.15
Moyen	0.41	0.33
Haut	1.06	0.98
Age : Jeune	-0.96	-0.92
Moyen	0.92	0.79
Agé	1.07	1.00
Car : Jap	-1.41	-1.33
Am	0.63	0.57

Dans le tableau ci-dessus,

1. La quantification initiale ignore les données manquantes, elle correspond à la phase 'donnée manquante passive' citée dans la section 2.3.2. quantification des données incomplètes catégorielles.

2. L'imputation homogène initiale des données manquantes en choisissant la catégorie dont la valeur de quantification est la plus proche du score individuel : pour le premier individu, la catégorie "*b*" de la variable 'revenu' dont la quantification (*-1.13*) est plus proche de la valeur du score (*-1.43*), on choisit "*b*" comme l'imputation initiale Pour le troisième individu, la catégorie "h" de la variable 'revenu' dont la quantification (*1.06*) est plus proche de la valeur du score (*1.02*), on choisit "h" comme l'imputation initiale. Pour le neuvième individu, la catégorie "a" de la variable 'age' dont la quantification (*1.07*) est plus proche de la valeur du score (*1.02*), on choisit "a" comme l'imputation initiale

3. Nous recalculons les scores d'unité en utilisant les données pseudo-complètes à la suite de la deuxième étape. Nous poursuivons la troisième étape de calcul en utilisant le "K-mean" algorithme pour déterminer la valeur finale d'imputation. Nous poursuivons le même processus jusqu'à ce que le critère η ne s'améliore pas. Ici, la solution initiale est retenue comme celle finale.

Après un exemple illustrant le fonctionnement de la méthode, on travaille sur deux exemples issus de données réelles avec le programme "Mistress" (Buuren S.V., 1992). On cache aléatoirement quelques données et on les reconstitue ensuite. On mesure finalement les écarts entre des données réelles cachées et les données reconstituées.

Exemple 1 - considérons les données d'une étude sur des critères de classification de clou, vis et boulon. Elles ont été recueillies par Hartigan (1975, p.228), citées dans Gifi (1992) et se trouvent dans le tableau suivant 1. Elles contiennent 24 observations sur les 6 variables suivantes :

1. Thread	Yes=Y	No=N	
2. Head	Flat=F	Cup=C	Cone=O

3.Head indentation	None=N	Star=T	Slit=L
4. Bottom	Sharp=S	Flat=F	
5. Length	(in half inches)		
6. Brass	Yes=Y	No=N	

Les 24 observations sont les suivantes et le coefficient d'homogénéité de 0.64.

Observation																							
1	n	n	n	n	n	**n**	n	n	n	y	y	y	y	y	**y**	y	y	y	y	**y**	n	n	n	y
2	f	f	**f**	f	f	f	u	**u**	u	o	r	y	r	y	r	o	y	y	y	y	f	f	f	o
3	n	n	n	**n**	n	n	n	n	**n**	t	l	l	l	**l**	l	l	l	l	l	l	n	n	n	l
4	s	s	s	s	**s**	s	s	s	s	s	s	s	s	s	s	f	f	**f**	f	f	f	s	s	s	**s**
5	1	4	2	**2**	2	2	5	3	3	5	4	4	2	2	4	1	1	1	1	1	1	1	1	1	1
6	**n**	n	n	n	n	n	n	n	n	n	n	n	n	**n**	n	n	n	n	n	n	n	y	**y**	y	y

Les 14 valeurs 'en gras' ont été cachées. Ce qui équivaut à 10% de D.M., 14 valeurs manquantes.

Les valeurs reconstituées sont les suivants :

1	2	3	4	5	6
n	f	n	s	3*	y*
y		n	f		n
y		t*	f*		y

Parmi les 14 valeurs manquantes, 10 valeurs sont bien reconstituées, donc le taux de bien classées est de 71.4%.

Exemple 2 - les données concernent une hypothèse de Russett (1964) citée dans Gifi (1992) : les inégalités économiques entraînent l'instabilité politique. Les données quantitatives, que l'on recode en valeur ordinale, se trouvent dans l'Annexe B I, il s'agit de 47 pays avec les 9 variables suivantes :

- Il y a 3 variables pour mesurer la répartition inégale des terres et 2 variables pour mesurer l'économie globale du pays.

GINI - index de concentration

FARM- pourcentage de fermiers possédant au moins la moitié des terres

RENT- pourcentage de familles fermières qui louent entièrement leur terre

CNPR- production totale du pays

LABO- pourcentage du nombre de personnes travaillant dans le secteur agricole

- Il y a 4 variables pour évaluer la stabilité politique :

INST- nombre de changements chef d'état et le nombre d'années d'indépendance d'un Pays dans la période 1945-1961

ECKS- nombre de guerres internes apparues entre 1946-1961

DEAT- nombre de personnes tuées à cause de violences civiles

DEMO- degré de démocratie

Le coefficient d'homogénéité est de 0.54. Nous avons caché 21 données, soit environ 5%, et nous avons les reconstituées et obtenons 57.1% de bien classées et le coefficient de proximité (voir 2.4) est de 0.146. Les résultats se trouvent également dans l'Annexe B I.

Condition d'application :

Cette méthode est satisfaisante pour les données dont le niveau d'homogénéité η est assez élevé. C'est-à-dire le niveau d'homogénéité η, soit supérieur à *0.40*. Elle pourrait être utilisée dans l'analyse d'enquête avec un grand tableau de données où le nombre de questionnaires est beaucoup plus grand que le nombre de catégories de variables

Cette méthode a les avantages suivants :

- l'adaptation à tous types de données manquantes : *'D.M.C.A'*, *'D.M.A.'* et *'D.M.N.A'* et aux données mixtes : variables catégoriques et continues,
- la stabilité de la méthode et validation aisée selon le critère d'homogénéité η,
- sans contrainte sur le modèle de données.

L'inconvénient est la complexité et le coût de calcul qui sont relativement importants.

2.4. VALIDATION ET MESURE DE QUALITE DES DONNEES RECONSTITUEES

L'évaluation des résultats d'une méthode se fait en comparant des données vraies cachées avec des données reconstituées. Pour ce faire, on peut prendre un fichier complet et cacher aléatoirement des données (on fait semblant d'ignorer les valeurs cachées des variables). Puis on effectue des comparaisons entre données reconstituées et vraies données.

Le résultat est mesuré et évalué selon deux niveaux :

- ***Niveau global*** :

Les distributions des variables 'réelles' et reconstituées doivent être proches et présenter peu d'écarts significatifs Les relations entre variables 'réelles' doivent être les plus proches possibles après reconstitution. C'est-à-dire qu'on veut garder les mêmes marges et les croisements des variables.

- par comparaison de distribution des variables, on vérifie si les marges sont respectées :

  - pour les variables qualitatives, on peut se servir du chi-2 pour calculer une distance entre la marge reconstituée et la marge réelle,
  - pour les variables quantitatives, on compare de moyennes, variances, etc des deux distributions.

- par vérification de la préservation des relations entre variables, on vérifie si les croisements sont les mêmes :

  - pour calculer la distance entre deux tableaux croisés pour des variables qualitatives, on utilise la moyenne des statistiques chi-2 de Pearson :

$$x^2 = \sum_{ij} \frac{(m_{ij} - n_{ij})^2}{m_{ij}}$$

où m_{ij} - le nombre d'enregistrements qui appartient à la classe i de la variable x et la classe j de la variable y dans le fichier réel.

n_{ij} - celui de fichier reconstitué.

- on compare la corrélation des variables pour les variables quantitatives : la moyenne de l'écart des covariances $\overline{\text{cov}}$ (dans un grand nombre de simulation) :

$$\overline{\text{cov}} = \text{cov}(x, y) - \text{cov}(\hat{x}, \hat{y})$$

où $\hat{x}$ *et* $\hat{y}$ - variables reconstituées,

x et y - variables réelles.

- ***Niveau individuel*** :

Il s'agit de bonnes réponses sur les valeurs reconstituées. Puisque nous cherchons à reconstituer des données les plus proches possibles de ce qu'on aurait obtenu par l'interview des individus sur les questions manquantes. La validation au niveau individuel consiste à comparer individuellement les données vraies avec les données reconstituées.

Pour les variables quantitatives, on pourra utiliser :

soit l'écart absolu $\sum_i \frac{|y_i - \hat{y}_i|}{n}$, soit l'écart quadratique $\sqrt{\sum_i \frac{(y_i - \hat{y}_i)^2}{n}}$

où y_i est la vraie valeur supprimée pour le *ième* individu du fichier.

$\hat{y}_i$ est la valeur reconstituée correspondante.

Pour les variables qualitatives nominales : le taux de bonnes reconstitutions

Lejeune M. et Lebart L. (1995) proposent une validation par taux de bonnes reconstitutions dans le cadre de fusion des fichiers, nous pensons que ce critère peut être utilisé dans le cadre des données manquantes. Dès lors que nous souhaitons attribuer à un individu la réponse la plus probable, il est naturel d'utiliser comme critère, le taux de bien classées. Ce taux associé à la règle d'affectation se définit comme le pourcentage de données bien classées : il s'agit de données reconstituées par rapport aux données vraies. Si les probabilités correspondant à la variable J à k catégories sont $p_1, p_2, \cdots, p_k$, une règle d'affectation aléatoire respectant ces probabilités conduirait à un taux de bonnes affectations :

$$\tau = \sum_{i=1}^{k} p_i^2$$

Ce gain est alors défini comme le nombre de données de bien classées constatées, diminué du nombre d'identiques espéré ($\tau \cdot n_r$), sous l'hypothèse d'égalité des taux de bonnes affectations entre la règle utilisée et une affectation aléatoire.

***Pour les variables qualitatives ordinales** : le coefficient de proximité*

Saporta G. & Co V. (1996) proposent le coefficient de proximité suivant, qui pénalise les erreurs d'affectation, selon l'écart entre la modalité vraie et la modalité affectée. On utilise pour cela une matrice rendant compte des coûts d'erreur de classement, par exemple : pour une variable à cinq modalités, ce coût d'erreur $C=[c_{ij}]$ est normalisé de tel sort que le maximum $c_{ij}=1$.

	Valeur affectée				
Valeur vraie	*1*	*2*	*3*	*4*	*5*
1	*0*	*1/4*	*2/4*	*3/4*	*1*
2	*1/4*	*0*	*1/4*	*2/4*	*3/4*
3	*2/4*	*1/4*	*0*	*1/4*	*2/4*
4	*3/4*	*2/4*	*1/4*	*0*	*1/4*
5	*1*	*3/4*	*2/4*	*1/4*	*0*

Les coefficients varient entre *0* à *1*, ce qui facilite l'interprétation.

L'espérance du coût d'erreur global de classement, sous l'hypothèse d'affectation aléatoire respectant les proportion p_1 , p_2 , ... , p_k des catégories, vaut :

$$\tilde{d} = \sum_{i=1}^{k} \sum_{j=1}^{k} c_{ij}^{k} p_i p_j$$

Soit $\bar{d}$ la moyenne des coûts d'erreurs associées à une règle donnée sur l'ensemble des cases reconstituées. $\bar{d} - \tilde{d}$ mesure le gain par rapport à une affectation aléatoire.

CHAPITRE III : TRAITEMENT DES DONNEES MANQUANTES DANS LES METHODES D'ANALYSE DE DONNEES MULTIDIMENSIONNELLES

3.1. ANALYSE EN COMPOSANTES PRINCIPALES

Nous proposons, ici, une méthode d'analyse en composantes principales modifiée et adaptée aux données incomplètes dont le type des données manquantes est *'D.M.C.A.'* et le pourcentage de *D.M.* est faible (< 15 %). Notre but est de trouver un résultat d'analyse raisonnable et assez proche du résultat réel. Cette méthode consiste à trouver un depart raisonnable et à se rapprocher petit à petit, par le calcul, vers des valeurs finales stables Pour cela, nous introduisons un algorithme itératif : une étape d'imputation et une étape de calcul des nouvelles composantes principales. L'étude de comparaison avec des données cachées sur la base de simulations sera analysée.

3.1.1. Problèmes rencontrés sur l'A.C.P. des données incomplètes et méthodes choisies pour la comparaison :

Différentes possibilités ont été proposées pour l'A.C.P. des données incomplètes :

- Diagonalisation de la matrice de variance-covariance construite par la technique pairwise,
- Imputation des D.M. soit par la moyenne soit par l'estimation par régression etc.

La méthode en s'accommodant des données manquantes par pairwise cause des problèmes assez ennuyeux (Kim et Cury, 1978) : les valeurs propres peuvent être négatives ce qui conduit à des corrélations multiples supérieures à 1 ou négatives.

Le deuxième type de technique consiste à estimer les données manquantes par diverses méthodes, par exemple, l'imputation de la moyenne (employée par SPAD), l'imputation par l'estimateur de la régression, etc. et ceci est fait avant l'analyse ; c'est-à-dire qu'on effectue l'A.C.P après avoir complété les données.

Dans nos comparaisons, nous utiliserons :

- la méthode de l'imputation par la moyenne : on remplace les *D.M.* par la moyenne de variable calculée sur les données disponibles,

- la méthode de l'estimation des *D.M.* par *'MGV'* de la procédure *PRINQUAL* dans SAS (SAS/STAT User's Guide Vol.2, pp1266-1324, 1990).

C'est-à-dire que nous allons compléter des données par ces deux méthodes avant d'effectuer l'analyse et comparer les résultats de l'analyse avec celles de notre méthode.

La méthode de *MGV* du PRINQUAL (SAS) utilise un algorithme itératif de régression multiple afin de minimiser le déterminant de la matrice de covariance des variables transformées. Cette méthode transforme chaque variable (ayant D.M.) de façon la plus similaire possible à une combinaison linéaire des autres variables (au sens des moindres carrés). Sur chaque itération pour une variable, l'algorithme *MGV* alterne une régression multiple et une transformation optimale linéaire (option 'linear' est choisie). La régression multiple consiste à estimer (prédire) les D.M. d'une variable à partir des autres variables. C'est une méthode d'estimation des *D.M.* des données multivariées. La moyenne des colonnes est utilisée comme l'estimation initiale des *D.M.* dans cette procédure itérative. Chaque étape remplace les *D.M.* d'une variable par leur valeur estimée des autres variables et des estimations courantes des *D.M.*. En même temps, une transformation linéaire optimale est faite. L'itération est terminée lorsque les estimations des *D.M* et la transformation linéaire se stabilise ou lorsque le nombre d'itération maximum est atteint. Dans notre étude, les données originellement observées plus les estimations des D.M. par cette méthode sont retenues pour la comparaison.

3.1.2. Algorithme proposé

Nous proposons un algorithme qui ressemble à un processus *EM*, avec une phase initiale où les données sont initialement complétées. Une étape consiste à faire une *A.C.P.*, Une autre étape consiste à estimer de nouveau les *D.M.* et on itére ces deux étapes jusqu'à la stabilisation des paramètres : le résultat de l'*A.C.P.* et ici les *D.M.* sont considérées également comme les paramètres à estimer. Nous allons travailler sur l'*ACP* standardisée, c'est-à-dire sur des données centrées réduites. Les données manquantes sont estimées de façon optimale pour l' *A.C.P.*

I. Phase initiale :

Nous faisons une analyse en composantes principales sur une matrice de données complètes dont les données manquantes sont remplacées simultanement par la moyenne de chaque variable. Dans le cadre des *D.M.* complètement aléatoire, nous considérons que c'est un départ raisonnable.

II. Phase principale :

Elle se décompose en 3 étapes suivantes :

$$\begin{cases} a.\ \textit{imputation des D.M.} \\ b.\ \textit{effectuer une ACP} \\ c.\ \textit{test de stabilisation} \end{cases}$$

a. *Phase d'imputation* : nous employons la technique d'estimations des *D.M.* 'du plus proche voisin'. Elle consiste à sélectionner un plus proche voisin (selon une distance minimale) sur les composantes principales comme donneur pour un individu ayant des données manquantes et imputer des valeurs correspondantes pour des données manquantes de ce dernier.

La distance entre deux individus sur des composantes principales

Supposons qu'une unité x se présente comme un vecteur (x_1, x_2, ..., xp), dont les composantes principales sont (C_1,..., Cm),

$$\vec{C} = \vec{U}\vec{X}$$

$$C_i = \sum_{j=1}^{p} U_j X_j = (U_1, \cdots, U_p) \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}, \qquad i=1, 2, ..., l$$

Pour chaque unité incomplète x (x_1, ... , *,...,x_p), nous associons un projecteur Ax tel que

$$A_x x = (x_1, \cdots, 0, \cdots, x_p)$$

$$C^*_{ix} = U(A_x x), i=1, 2, ..., m$$

$$C^*_{iy} = U(A_x y),\ y \in \Omega$$ - l'ensemble des unités complètes

Ce projecteur A remplace les données manquantes par 0.

La distance entre x et y est donc définit comme suit :

$$d(x, y) = \sum_{i=1}^{l} (C^*_{ix} - C^*_{iy})^2$$

En effet, cette distance sur les composantes principales ignore la partie de distance concernant les *D.M.*. Lorsqu'il n'existe pas de *D.M.*, c'est une distance euclidienne normale.
y est le donneur de x si la distance $d(x,y)$ est la distance minimale entre x et y, $y \in \Omega$

b. *Phase A.C.P.* : on fait une analyse en composantes principales sur les données de nouveau complétées.

c. Phase contrôle : on répète les phases *a et b* jusqu'à ce que le critère d'arrêt soit satisfaisant :

Le critère d'arrêt se définit suivant la différence entre la valeur propre courante et l'ancienne valeur propre. Cette différence doit être inférieure à β (dans le programme β = *1E-5*) ou alors le nombre d'itération maximale m (m = *15*) est atteint.

Chaque étape d'imputation fait qu'une unité incomplète est reconstituée et approchée. La composante principale des pseudo-données complètes est plus proche lorsqu'on ignore complètement les données manquantes au départ. Chaque itération rapproche est ainsi du but : les données manquantes sont mieux estimées, le résultat de l'analyse se stabilise finalement. Dans la pratique, la convergence vers l'objectif est atteinte après environ 10 itérations. Cette méthode a été programmée dans le langage *IML de SAS* (*cf. Annexe A.I,* Programme *A.C.P. des données incomplètes*).

3.1.3. Critères de qualité

Ici, nous nous intéressons plutôt au résultat final de l'analyse en composantes principales qu'à la reconstitution des données individuelles. Donc des critères d'évaluation des résultats de l'analyse en composantes principales sont choisis. On voudra vérifier si le système orthogonal construit à partir des données disponibles en appliquant l'algorithme proposé est le même que celui construit à partir des données complètes cachées : c'est-à-dire la stabilité des axes. Cette question se traduit par deux indices : si les corrélations des axes principaux entre deux systèmes sont fortes et que les inerties expliquées par les premiers axes sont les mêmes. Ces deux critères sont calculés en se basant sur le résultat de *l'A.C.P.* des données simulées complètes et celui de la méthode concernée sur des données incomplètes. Ces deux critères sont les suivants :

I. L'inertie expliquée par les premiers axes par rapport à l'inertie réelle :

$$\sum_{i=1}^{p} (\lambda_i - \hat{\lambda}_i)^2$$

où λ_i *–vrai et* $\hat{\lambda}_i$ *–après imputation*

II. La corrélation (en valeur absolue) entre facteurs de même rang,

On montre également le résultat en ce qui concerne le cercle de corrélation à travers un exemple issu de données réelles.

3.1.4. Evaluation de résultats

Pour dégager les facteurs qui influencent la qualité de ces méthodes, les résultats seront comparés avec des données vraies cachées sur la base de simulation.

3.1.4.1. Jeu de données

La simulation des données de la loi normale multivariée se compose des étapes suivantes :

1. D'abord on génère un tableau de données $X=(x_1, x_2, \ldots, x_m)$ où x_j $(j=1, \ldots, m)$ suit la loi normale standard $N(0, 1)$ et $x_1, x_2, \ldots, x_m$ sont indépendants,

2. Ensuite on crée une matrice de corrélation R positive symétrique, la décomposition de Cholesky donne :

$$R = U'U \text{ , } U \text{ est une super matrice triangulaire}$$

Comme x_i, x_j sont indépendants, $X'X = I$, donc

$$(XU)'(XU) = U'X'XU = U'U = R$$

La transformation $Y=XU$ suit la loi normale multivariée avec la moyenne = **0** et la matrice de covariance ***R***.

En utilisant la technique ci-dessus, nous avons généré des données tirées d'une loi normale multivariée de moyenne **0** et dont la matrice de covariance a ses éléments hors diagonale égaux une constante r (r est comprise entre 0 et 1). La taille d'échantillon est de 120 sur 12 variables.

Nous avons créé des données manquantes du type *'D.M.C.A'* suivant trois pourcentages de données manquantes : *5%, 10% et 15%* sur l'ensemble de données. Pour un taux différent de *D.M.*, nous associons cinq jeux de données et nous évaluons le résultat sur la moyenne ces cinq jeux de données.

3.1.4.2. Résultats

Le résultat varie en fonction de la corrélation entre variables et du taux de *D.M.* Nous avons remarqué que presque dans tous les cas, la méthode proposée est supérieure au sens de récupération des variances de chaque dimension, c'est à dire sur le premier critère.

Taux de D.M. 5% :

Nous observons que les trois méthodes ont des avantages dépendants du niveau de corrélation R entre variables selon le critère *II* :

- lorsque *r* est inférieur à 0.45, la méthode de la moyenne est légèrement meilleure que la méthode proposée,
- lorsque *r* est entre 0.45 et 0.73, la méthode proposée est la meilleure,
- lorsque *r* est supérieur à 0.73, la méthode de transformation de *MGV* de *SAS* est la meilleure.

Les tableaux détaillés se trouvent dans Tab.1-Tab.4.

Tableau de corrélation entre facteurs de même rang
pour différentes méthodes et pour ***r =0.3***

Corrélation	Méthode proposée	Imputation par la moyenne	Méthode MGV
Axe1	0.96	0.97	0.17
Axe2	0.78	0.89	0.12
Axe3	0.56	0.71	0.4
Axe4	0.54	0.44	0.3
Axe5	0.44	0.43	0.38
Somme des carrés des résidus des inerties	0.003	0.04	6.3

Tab. 1

Tableau de corrélation entre facteurs de même rang

pour différentes méthodes et pour r =**0.5**

Corrélation	Méthode proposée	Imputation par la moyenne	Méthode MGV
Axe1	0.96	0.96	0.14
Axe2	0.93	0.89	0.11
Axe3	0.78	0 64	0.31
Axe4	0.52	0.54	0.53
Axe5	0.71	0 42	0.23
Somme des carrés des résidus des inerties	0.07	0.7	11

Tab. 2

Tableau de corrélation entre facteurs de même rang

pour différentes méthodes et pour r =***0.7***

Corrélation	Méthode proposée	Imputation par la moyenne	Méthode MGV
Axe1	0.96	0.92	0.97
Axe2	0.97	0.85	0.94
Axe3	0.77	0.34	0.60
Axe4	0.45	0.32	0.37
Axe5	0.31	0.14	0.14
Somme des carrés des résidus des inerties	0.06	0.15	0.02

Tab. 3

Tableau de corrélation entre facteurs de même rang
pour différentes méthodes et pour $r = 0.9$

Corrélation	Méthode proposée	Imputation par la moyenne	Méthode MGV
Axe1	0.98	0.70	0.98
Axe2	0.97	0.64	0.97
Axe3	0.56	0.20	0.82
Axe4	0.15	0.37	0.91
Axe5	0.41	0.05	0.72
Somme des carrés des résidus des inerties	0.00	0.25	0.00

Tab. 4

En conclusion, la méthode proposée est acceptable pour les différentes corrélations. La méthode *MGV* de *SAS* est satisfaisante seulement pour les données fortement corrélées. Par contre, la méthode de la moyenne n'est pas bonne pour les données très corrélées.

Taux de D.M. 10% :

C'est le même phénomène que nous observons pour le cas 5%. Nous remarquons que les trois méthodes ont des avantages dépendants du niveau de corrélation R entre variables selon le critère *II*:

- lorsque r est inférieur à 0.35, la méthode de la moyenne est légèrement meilleure que la méthode proposée,
- lorsque r est entre 0.35 et 0.75, la méthode proposée est meilleure,
- lorsque r est supérieur à 0.75, La méthode de transformation de *MGV* de *SAS* est la meilleure.

Voir tableaux Tab.5-Tab.8.

Tableau de corrélation entre facteurs de même rang

pour différentes méthodes et pour $r=0.3$

Corrélation	Méthode proposée	Imputation par la moyenne	Méthode MGV
Axe1	0.88	0.93	0.35
Axe2	0.78	0.85	0.22
Axe3	0.56	0.70	0.30
Axe4	0.55	0.30	0 19
Axe5	0 45	0.42	0.15
Somme des carrés des résidus des inerties	0.04	0.16	7.5

Tab. 5

Tableau de corrélation entre facteurs de même rang

pour différentes méthodes et pour $r=0.5$

Corrélation	Méthode proposée	Imputation par la moyenne	Méthode MGV
Axe1	0.92	0.88	0.35
Axe2	0.76	0.83	0.71
Axe3	0.81	0.64	0.51
Axe4	0.70	0.50	0.30
Axe5	0.9	0.36	0.12
Somme des carrés des résidus des inerties	0.01	0.35	18

Tab. 6

Tableau de corrélation entre facteurs de même rang
pour différentes méthodes et pour $r=0.7$

Corrélation	Méthode proposée	Imputation par la moyenne	Méthode MGV
Axe1	0 88	0.85	0.55
Axe2	0.89	0.73	0.76
Axe3	0.68	0.52	0.58
Axe4	0.50	0.54	0.44
Axe5	0.33	0.44	0.10
Somme des carrés des résidus des inerties	0.00	0.6	14

Tab. 7

Tableau de corrélation entre facteurs de même rang
pour différentes méthodes et pour $r=0.9$

Corrélation	Méthode proposée	Imputation par la moyenne	Méthode MGV
Axe1	0.93	0.31	0.98
Axe2	0.97	0.49	0.97
Axe3	0.36	0.59	0.65
Axe4	0.16	0.10	0.30
Axe5	0.84	0.05	0.35
Somme des carrés des résidus des inerties	0.00	1.05	0.09

Tab. 8

La conclusion reste en principe la même que celui du cas des *D.M.* de 5%. Seulement la méthode de moyenne est moins efficace lorsque le taux des *D.M.* augmente.

Taux de D.M. 15% :

Nous observons là encore que les trois méthodes ont des avantages différents selon le niveau de corrélation r entre variables selon le critère *II* :

- lorsque r est inférieur à 0.30, la méthode de la moyenne est légèrement meilleure que la méthode proposée,
- lorsque r est entre 0.30 et 0.85, la méthode proposée est meilleure,
- lorsque r est supérieur à 0.85, la méthode de transformation de MGV de *SAS* est la meilleure.

Voici les tableaux en détail selon les différentes corrélations.

Tableau de corrélation entre facteurs de même rang
pour différentes méthodes et pour ***r=0.3***

Corrélation	Méthode proposée	Imputation par la moyenne	Méthode MGV
Axe1	0.74	0.80	0 25
Axe2	0.53	0.67	0.26
Axe3	0.50	0.47	0.34
Axe4	0.57	0.36	0.55
Axe5	0.27	0.50	0.30
Somme des carrés des résidus des inerties	0.02	0.27	6.1

Tab. 9

Tableau de corrélation entre facteurs de même rang

pour différentes méthodes et pour $r=0.5$

Corrélation	Méthode proposée	Imputation par la moyenne	Méthode MGV
Axe1	0.78	0.72	0.45
Axe2	0.70	0.67	0.21
Axe3	0.30	0.40	0.35
Axe4	0.25	0.35	0.12
Axe5	0.1	0.25	0.2
Somme des carrés des résidus des inerties	0.05	0.57	15

Tab. 10

Tableau de corrélation entre facteurs de même rang

pour différentes méthodes et pour $r=0.7$

Corrélation	Méthode proposée	Imputation par la moyenne	Méthode MGV
Axe1	0.81	0.56	0.45
Axe2	0.83	0.62	0.35
Axe3	0.65	0.25	0.30
Axe4	0.45	0.35	0.23
Axe5	0.13	0.32	0.41
Somme des carrés des résidus des inerties	0.04	1.2	11

Tab. 11

Tableau de corrélation entre facteurs de même rang

pour différentes méthodes et pour $r=0.9$

Corrélation	Méthode proposée	Imputation par la moyenne	Méthode MGV
Axe1	0.78	0.34	0.94
Axe2	0 85	0.26	0.88
Axe3	0.89	0.44	0.73
Axe4	0.38	0.00	0.37
Axe5	0.09	0.01	0.64
Somme des carrés des résidus des inerties	0 00	1.9	0.02

Tab. 12

Par rapport au cas de *D.M.* à 5%, la méthode proposée résiste mieux aux nombreuses données manquantes : elle gagne plus de terrain par rapport aux deux autres méthodes. C'est-à-dire que la méthode proposée s'adapte à une plus grande variété de corrélation que donne le cas de *D.M.* à 5%.

La comparaison ci-dessus se base sur des données multinormales avec corrélations constantes. La question de l'influence de la normalité sur la qualité de méthodes se pose donc naturellement Nous tenterons d'y répondre sur la base d'un exemple issu de données réelles.

Exemple - Nous utilisons un extrait des données d'une enquête sur la "tolérance" réalisée par l'institut BVA portant sur 300 individus et 10 variables. Ces données se trouvent dans Annexe B II. Il s'agit des 10 variables suivantes :

eurv - Droit de vote des étrangers européens en France

nimp - Ne plus importer les produits des pays de bas salaires

mari - Contre les mariages entre gens de races différentes

toub - Toubon a raison de vouloir éliminer les mots étrangers

immi - Immigrés renforcent notre pays grâce à leur travail

homo - Homosexualité est un mode de vie que la société devrait accepter

iden (**d**) - Droit de la police d'exercer le contrôle d'identité

cult - Culture française s'enrichit de ses échanges avec les autres cultures

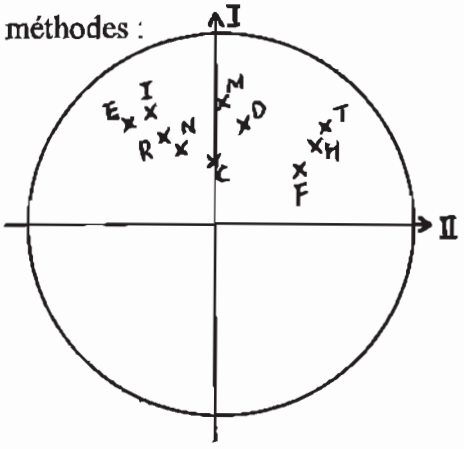
femm - Femmes devraient revenir leur rôle traditionnel dans la société

race - Rien en commun avec les autres races

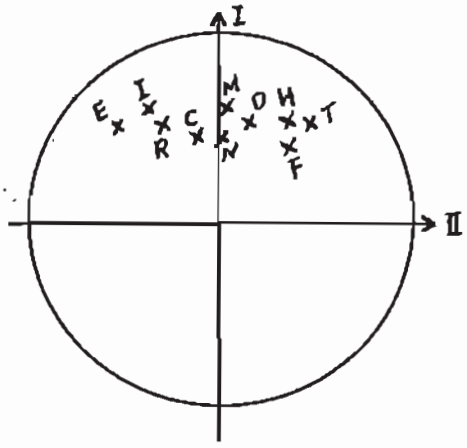
Le coefficient de corrélation moyen entre variables est de 0.22. On crée environ 8% de données manquantes aléatoirement dans le tableau Les corrélations en valeur absolue entre facteurs de même rang pour différentes méthodes est calculé :

Corrélation	Méthode proposée	Imputation par la moyenne	Méthode MGV
Axe1	0.90	0.92	0.54
Axe2	0.95	0.97	0.68
Axe3	0.93	0.83	0.38
Axe4	0.90	0.79	0.18
Axe5	0.82	0.91	0.13
Somme des carrés des résidus des inerties	0 00	0.05	3.74

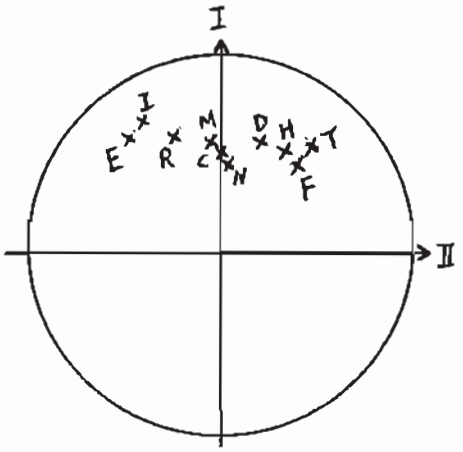
Nous voulons savoir également si les positions des variables sur les axes factoriels sont proches de celles calculées sur des données réelles. Voici les cercles de corrélation de différentes méthodes :



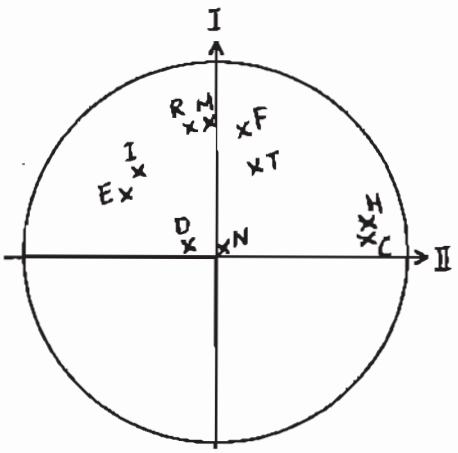
Données complètes réelles



Données reconstituées par la méthode proposée



Données reconstituées par
l'imputation de moyenne



Données reconstituées par
méthode MGV

Ces résultats confirment tout à fait les conclusions tirées sur des données simulées.

La comparaison de resultats montre que la méthode de l'imputation par régression ne s'avère pas très utile dans le cadre de l'utilisation ultérieure des données en ACP. Par rapport a la méthode proposée, l'imputation par la moyenne réduit la variation des données ; en effet la 'somme des carrés des résidus des inerties' est toujours plus élevée que par la methode proposée. En conclusion, le résultat de la méthode proposée semble satisfaisant dans les cas les plus courants.

3.2 L'ANALYSE FACTORIELLE DES CORRESPONDANCES

Proposée dans les années 60 par J.-P. Benzécri pour l'étude des tableaux de contingence (croisement de deux caractères nominaux), l'analyse des correspondances a été étendue par la suite au cas d'un nombre quelconque de caractères qui est équivalente à l'analyse homogène. Les statisticiens et notamment les utilisateurs de l'analyse des correspondances sont fréquemment confrontés au problème suivant : le tableau de leurs données est, pour une raison ou pour une autre, soit incomplet, soit entaché d'un certain nombre d'erreurs. Nous séparerons l'analyse factorielle des correspondances simples de l'analyse factorielle des correspondances multiples.

3.2.1 Analyse factorielle des correspondances simples aux données incomplètes

La méthode est présentée dans les années 70 par C. Nora pour l'étude des tableaux de contingence incomplets. Elle est étudiée ensuite par Benzécri et al. (Vol. 2, Chapitre III, 1980), par Greenacre (*pp* 236-244, 1984) et De Leeuw J. & Van der Heijden, P.G.M. (1988). Cette technique est spécialement conçue pour des tableaux avec des données manquantes auxquelles l'analyse factorielle des correspondances classique ne convient pas. La méthode de l'analyse factorielle des correspondances destinée au tableau de contingence incomplet est itérative et se présente comme suit : d'abord des données manquantes sont estimées, ensuite l'analyse factorielle des correspondances est effectuée, enfin le résultat est réutilisé pour l'estimation des données manquantes et ce processus est itéré jusqu'à la stabilisation des critères prédéfinits. L'algorithme se présente comme suit : on choisit d'abord la dimension h et une estimation initiale $X^{(0)}$ tel que $x_{ij}^{(0)} = p_{ij}$ pour l'élément (i, j) présent dans le tableau et un choix arbitraire pour l'élément (i, j) manquant dans le tableau S. La reconstitution d'ordre h est fait itérativement à l'aide de la formule de reconstitution :

$$x_{ij}^{(m+1)} = x_{i*}^{(m)} x_{*j}^{(m)} \left(1 + \sum_{\alpha=1}^{h} \frac{1}{\sqrt{\lambda_{\alpha}^{(m)}}} r_{i\alpha}^{(m)} c_{j\alpha}^{(m)}\right) \Big/ x_{**}^{(m)}$$

qui est appliqué pour toute paire (i, j) manquante dans le tableau S. Pour l'élément (i, j) présent dans le tableau, on l'affecte simplement selon $x_{ij}^{(m)} = p_{ij}$ pour tous m La solution, en général, dépend du choix de la dimension h. Benzécri semble plus favoriser la reconstitution itérative d'ordre zéro, c'est-à-dire pour tout élément (i, j) manquant dans le tableau S, on utilise la formule :

$$x_{ij}^{(m+1)} = x_{i*}^{(m)} x_{*j}^{(m)} \Big/ x_{**}^{(m)}$$

L'algorithme est convergent. En conclusion, cet algorithme permet d'appliquer les méthodes d'analyse factorielle à des tableaux de données incomplètes mais dans lesquels on dispose d'information suffisamment réparties pour chaque individu ou chaque caractère. Le taux de données manquantes, jusqu'à 23%, n'intervient que très peu dans les résultats obtenus Seul le temps de calcul en dépend très directement. Cet algorithme est d'autant plus sûr que les différentes reconstitutions d'ordre successif 0, 1, ..., 5 donnent des résultats très proches Ceci nous donne une valeur indicative permettant d'apprécier la validité de cet algorithme.

3.2.2 Analyse factorielle des correspondances multiples avec données incomplètes

Nous aimerions dire que la technique la plus simple utilisée dans l'ACM pour traiter des données manquantes est de créer une catégorie de non-réponse pour chaque variable. Mais lorsque des données manquantes ne correspondent pas à une attitude particulière, cette méthode n'est pas très pertinente. Escofier B. et Benali H. (1987) présentent une technique d'*A.C.M.* en cas de données manquantes et de modalités à faibles effectifs. Cette méthode destinée à des données incomplètes est une variante de l'*A.C.M.* classique. Les formules de calcul sont modifiées pour s'adapter au cas des données incomplètes. Elle résout simultanément le problème des *D.M.* et des modalités à faibles effectifs en minimisant leurs influences sur le résultat. Ici, nous nous intéressons au problème des données manquantes non significatives d'une attitude, c'est-à-dire des données manquantes complètement aléatoires. Dans cette méthode, on traite le tableau où les non-réponses sont codées zéro et les modalités rares sont supprimées. La marge sur l'ensemble des individus n'est plus constante mais est remplacée par une valeur constante sur l'ensemble des individus. Dans cette partie nous ne montrerons que la partie modifiée par rapport à l'*A.C.M.* classique. Rappelons qu'une autre méthode consiste à appliquer l'analyse homogène puisque cette dernière est équivalente à l'A.C.M. pour des données complètes.

Notations

Soient . - Q un ensemble de questions, - I un ensemble d'individus,

- J un ensemble de modalités de réponses à toutes les questions Q,

$card\, I=n$, $card\, J=card\, J1 + card\, J2 + ... + card\, JQ$,

$K_{IJ} = [K_{IJ1}, K_{IJ2}, ..., K_{IJQ}]$ le tableau disjonctif des variables indicatrices associées aux modalités de réponses.

$$\forall j \in J_q \subset J, \begin{cases} k_{ij} = 1 & \text{si l'individu i possède la modalité j} \\ k_{ij} = 0 \end{cases},$$

$$on\ note\ k_{i.} = \sum_{j \in J} k_{ij},\ k_{j.} = \sum_{i \in I} k_{ij},\ k = \sum_{i,j} k_{ij}.$$

Si un individu est représenté dans R_J par son profil ligne $\left\{ \frac{k_{ij}}{k_{i.}}, j \in J \right\}$, l'ensemble des profils lignes affectées des poids $\frac{k_{.j}}{k}$ représente le nuage des individus *N(I)*. Une modalité *j* est

représentée par son profil colonne $\left\{\frac{k_{ij}}{k_{.j}}, i \in I\right\}$, l'ensemble des profils colonnes affectés des poids $\frac{k_{i.}}{k}$ représente le nuage des modalités *N(J)*.

Tableau disjonctif incomplet $K'_{IJ'}$

Le tableau $K'_{IJ'}$ est obtenu à partir du tableau K_{IJ} en supprimant les colonnes correspondant aux modalités rares, les non-réponses à une question q étant codées 0 sur l'ensemble des modalités conservées de K_q. J' est donc le sous ensemble de J correspondant aux modalités conservées

Soit $K'_I = \left\{K'_{i.}, i \in I\right\}$ *et* $K'_J = \left\{K'_{.j}, j \in J\right\}$ les marges de $K'_{IJ'}$,
K' *le total des éléments de ce tableau*,
nous avons,

$$\forall i \in I, \forall j \in J', \begin{cases} k'_{ij} = K_{ij} \\ k'_{.j} = K_{.j} \end{cases} \qquad (1)$$

Par contre K' est différent de K, et en général $k'_{i.}$ est différent de $\frac{K'}{n}$ (K'_I n'est pas une marge constante sur I, comme ce serait le cas pour un tableau disjonctif complet).

On désignera par $k''_{IJ'}$ le tableau disjonctif complet issu de $K_{IJ'}$ en conservant les modalités rares, et en ajoutant des modalités de non-réponses pour les réponses manquantes.

Choix d'une distance pour l'étude du tableau disjonctif incomplet $K'_{IJ'}$

Distance du *khi-deux* :

L'A.F.C.M. classique utilise la distance du *khi-deux* entre deux profils lignes (resp. colonnes) qui s'écrit :

$$d^2(i,i') = \sum_{j \in J} \frac{K'}{K'_{.j}} \left(\frac{K_{ij}}{K'_{i.}} - \frac{K_{i'j}}{K'_{i'.}} \right)^2 ,$$

$$d^2(j,j') = \sum_{i \in I} \frac{K'}{K'_{i.}} \left(\frac{K_{ij}}{K_{.j}} - \frac{K_{ij'}}{K_{.j'}} \right)^2 .$$

Dans le tableau disjonctif incomplet $K'_{IJ'}$, le problème des modalités rares (contribuant fortement à la distance entre deux profils lignes) est résolu Par contre, si deux individus i et i' n'ont pas donné le même nombre de réponses ($K'_{i.} \neq K'_{i'.}$), une modalité j choisie simultanément par ces individus augmente leur distance car le terme $\frac{K_{ij}}{K'_{i.}} - \frac{K_{i'j}}{K'_{i'.}}$ n'est pas nul, ce qui est un inconvénient et pose un réel problème d'interprétation. Les auteurs pensent que cette métrique est donc inadaptée à l'étude de tableaux disjonctifs incomplets, ils proposent une variante de la distance du *khi-deux :*

On remplace la marge ($k'_{i.} = K'_{i.}, i \in I$) par la marge constante ($\frac{K'}{n}$,i ∈ I) partout où elle intervient : profil et poids des individus, métrique et origine des axes pour le nuage des profils colonne.

La distance entre les profils lignes est analogue à celles issues du tableau disjonctif complet $k''_{II''}$ obtenu en supprimant les termes provenant des non-réponses et des modalités rares, et les distances entre profils colonnes sont identiques à celle issues du tableau $k''_{II''}$.

Dualité et formules de transitions

Le facteur F_s du nuage *N(I)* se déduit des facteurs G_s du nuage *N(J)* par les formules de transitions suivantes où μ_s est la valeur propre d'ordre *s :*

$$F_s(i) = \left(\frac{n}{\sqrt{\mu_s} \sum_{j \in J'} k_{.j}} \right) \sum_{j \in J'} k_{ij} G_s(j) - \left(\frac{1}{\sqrt{\mu_s} \sum_{j \in J'} k_{.j}} \right) \sum_{j \in J'} k_{.j} G_s(j)$$

$$G_s(j) = \frac{1}{\sqrt{\mu_s}} \sum_{i \in I} \frac{k_{ij}}{k_{.j}} F_s(i)$$

Dans la première formule apparaît le terme $(\frac{1}{\sqrt{\mu_s \sum_{i \in I'} k_{.j}}}) \sum_{i \in I'} k_{ij} G_s(j)$ qui représente la coordonnée du centre de gravité G de nuage $N(J')$ sur l'axe F_s, et mesure le décalage du facteur quand l'origine des axes ne correspond pas à G. Ce terme, est presque nul en pratique, ce qui permet d'interpréter comme en *A.F.C.M* classique l'abscisse d'un individu comme le barycentre des modalités qu'il a choisies. La deuxième formule est exactement celle de l'*A.F.C.M.*

Elément supplémentaire et formule de reconstitution des données :

On peut mettre des modalités en éléments supplémentaires, particulièrement les modalités non-réponses et les modalités rares. L'abscisse $G_s(j^+)$ d'une modalité supplémentaire est définit par :

$$G_s(j^+) = \frac{1}{\sqrt{\mu_s}} \sum_{i \in I} \frac{k_{ij^+}}{k_{.j^+}} F_s(i).$$

A partir des formules de transition, on peut reconstituer le tableau initial en utilisant la formule :

$$k_{ij} = (\frac{k_{.j}}{n})(1 + \sum_{s} \frac{1}{\sqrt{\mu_s}} F_s(i) G_s(j)).$$

Nuage *N(I)* des profils des lignes

La distance entre deux profils lignes i *et* i' est définie par :

$$d^2(i,i') = \sum_{j \in J'} \frac{K'}{K_{.j}} (n\frac{K_{ij}}{K'} - n\frac{K_{i'j}}{K'})^2 = \frac{n}{K'} \sum_{j \in J'} \frac{1}{K_{.j}} (K_{ij} - K_{i'j})^2,$$

L'inconvénient rencontré dans la distance du khi-deux disparaît. L'analyse du nuage $N(I)$ est faite à partir de son centre de gravité qui sera pris comme origine des axes.

Nuage *N(J)* des profils des colonnes

La distance entre deux profils colonne j *et* j' est définie par :

$$d^2(j,j') = \sum_{i \in I'} K' \frac{n}{K'} (\frac{K_{ij}}{K_{.j}} - \frac{K_{ij'}}{K_{.j}})^2 = n \sum_{i \in I} (\frac{K_{ij}}{K_{.j}} - \frac{K_{ij'}}{K_{.j'}})^2,$$

Cette métrique possède la propriété intéressante d'équivalence distributionnelle.

L'analyse du nuage $N(J')$ est faite à partir du point de coordonnée $\frac{K'}{n}$ pris comme origine des axes. Les facteurs sur J' ne seront pas exactement centrés puisque le centre de gravité de $N(J')$ est $(\frac{K'_{i.}}{K'}, i \in I)$.

Considérons le tableau disjonctif complet $K''_{IJ''}$ obtenu, rappelons-le, en gardant les modalités rares, et en ajoutant des modalités de non-réponses pour les réponses manquantes. On a : $J'' \supset J' \supset J$, et la marge sur I de $K''_{IJ''}$ est constante Le nuage $N(J')$ des profils des colonnes de $K_{IJ'}$ est un sous nuage de $N(J'')$ associé à $K''_{IJ''}$, les profils sont identiques et la métrique de R_i définie par la marge constante. Cette méthode revient à analyser alors le sous nuage $N(J')$ du nuage complété $N(J'')$ en gardant la métrique et le centre de gravité associé à $N(J'')$.

Le résultat s'est montrée très efficace pour neutraliser l'effet des perturbations à cause des données manquantes et des modalités à faibles effectif, elle donne des résultats très stables Cette technique a été programmée et ajoutée dans la version de S.P.A.D (Système Portable d'Analyse des Données).

3.3. CLASSIFICATION AUTOMATIQUE SUR DONNEES QUANTITATIVES

Nous nous intéressons à la classification non hiérarchique, plus exactement celle des nuées dynamiques ou des centres mobiles. C'est une analyse en classes disjointes basée sur une distance euclidienne pour des variables quantitatives. Fèvre (1979) a étudié cette technique avec des données incomplètes. La méthode de Fèvre revient à ignorer les DM dans le calcul de distance en minimisant l'inertie intraclasse pour classer les unités.

3.3.1. Méthode de Fèvre

- **L'espace de représentation et la pseudo-distance** :

$E=\{ x_1, \ldots, x_n \}$ l'ensemble des unités $\in R^n$.

P_i , i = 1, 2, . ,n les poids respectifs.

$F=\{x^1, \ldots, x^p\}$, l'ensemble des variables $\in R^p$

$$x_i = \begin{bmatrix} x_i^1 \\ \vdots \\ x_i^j \\ * \\ \vdots \\ x_i^p \end{bmatrix}$$

où x_i^j = la valeur de la $j^{ème}$ variable sur l'individu i, les données manquantes sont notées par le signe '*'.

A chaque individu x_i, on associe à un projecteur A_i :

$$A_i : R^p -> R^p$$
$$x_i -> A_i(x_i)$$

$$A_i = \begin{bmatrix} 1 & & & & \\ & \ldots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & \ldots \\ & & & & & 1 \end{bmatrix}_{pxp} = \{a_{ij}\}_{pxp}$$

A_i est une matrice diagonale, les éléments non diagonaux sont égaux à 0 et les éléments diagonaux se définissent de la façon suivante:

$$a_{jj} = \begin{cases} 0 & \text{Si la valeur de la jème variable sur l'individu i est manquante} \\ 1 & \text{Sinon} \end{cases}$$

$$A_i(x_i) = \begin{bmatrix} x_i^1 \\ x_i^2 \\ 0 \\ x_i^l \\ 0 \end{bmatrix}$$

Approximation des distances

$$\forall (u,v) \in R^p x R^p, d^2(u,v) = (u-v)' M (u-v),$$

M - la métrique est une matrice définie positive et symétrique.

Distance d'un individu ayant des réponses complètes à un individu ayant des réponses incomplètes

$$X_l \in R^p -> A_i x_i \in R^p$$

La distance $d^2(u,x_l)$ d'un vecteur complet à un vecteur incomplet est l'approximation de la pseudo-distance $\hat{d}^2(u,x_l)$,

$$\hat{d}^2(u,x_i) = d^2(u,A_l x_l) = (u - A_l x_l)' M (u - A_l x_l) = d^2(A_l u - A_l x_l) = (u - x_l)' A_l M A_i (u - x_l)$$

Comparabilité des pseudo-distances .

Dans le calcul, on utilise des distances, soit

$$\forall (u,v) \in R^p x R^p,\ \hat{d}^2(u,x_i) = (u - x_i)' A_i M A_i (u - x_i)$$

$$\hat{d}^2(v,x_i) = (v - x_i)' A_i M A_i (v - x_i),$$

la distance $\hat{d}^2(u,x_i)$ est comparable à la distance $\hat{d}^2(v,x_i)$,

$$D(x,y) = \begin{cases} d^2(x,y) & x,y \text{ les vecteurs complets} \\ \hat{d}^2(x,y) & x \text{ ou } y \text{ est un vecteur incomplet} \end{cases}$$

donc la distance D est comparable dans l'espace E .

- **Le critère et l'approximation des inerties**

Le but est de trouver la partition en K classes qui sépare au mieux les classes. Les données sont supposées être réparties en K classes $C_1, C_2, \ldots, C_r$, $\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_k$ sont des pseudo-centres de gravité, alors la pseudo-inertie totale est :

$$T = \sum_{i=1}^{n} p_i D^2(x_i \hat{g}) = \sum_{l=1}^{k} \sum_{i \in c_l} p_i (x_i - \hat{g}_l)' A_i M A_i (x_i - \hat{g}_l) + \sum_{l=1}^{k} \sum_{i \in c_l} p_i (\hat{g} - \hat{g}_l)' A_i M A_i (\hat{g} - \hat{g}_l)$$

= inertie intra-classe $\hat{W}$ + inertie inter-classe $\hat{B}$.

Minimisation de la pseudo-inertie :

Soit un nuage de n individus $(x_1, x_2, \ldots, x_n)$, $n >= 1$, muni de poids $(p_1, p_2, \ldots, p_n)$, à chaque individu x_i est associé un projecteur A_i, la pseudo-inertie du nuage autour du g s'écrit:

$$\hat{I}(g) = \sum_{i=1}^{n} p_i \hat{d}^2(x_i, g) = \sum_{i=1}^{n} p_i (x_i - g)' A_i M A_i (x_i - g) \quad (3.22)$$

$$\text{quand } g = \left[\sum_{i=1}^{n} p_i A_i M A_i \right]^{-1} \left[\sum_{i=1}^{n} p_i A_i M A_i x_i \right]$$

$\hat{I}$ (g) est bien le minimum (Fèvre P., *1979, pp.135)*. Comme les pseudo-distances $\hat{d}^2(u, x_l)$, $\hat{d}^2(v, x_l)$ sont comparables, les différentes $\hat{W}$ sont comparables pour différentes partitions de classe.

Nous cherchons une partition minimisant $\hat{W}$, $\hat{T} = \hat{W} + \hat{B}$ où $\hat{T}$ est la pseudo-inertie totale du nuage, c'est une constante indépendante de la partition, il est donc équivalent de minimiser $\hat{W}$ ou de maximiser $\hat{B}$. La valeur de $\hat{W}$ peut servir à juger de la qualité de la partition

- **La fonction d'affectation**

$\forall x_i \in E$, $x_i \in C_l$ tel que

$$\forall m \in \{1, \ldots, k\} \begin{cases} m <= l \implies D(x_i, \hat{g}_l) < D(x_i, \hat{g}_m) \\ m >= l \implies D(x_i, \hat{g}_l) >= D(x_i, \hat{g}_m) \end{cases}$$, ce qui définit C_l de façon unique.

La méthode de Fèvre se déroule de la même façon que la méthode classique mais lorsqu'il n'existe pas d'individus complets dans une classe, l'algorithme ne fonctionne pas. Cet inconvénient pourrait devenir grave, lorsque les *D.M.* sont très dispersées, c'est à dire lorsque le nombre d'individus incomplets augmente.

3.3.2. Méthode proposée

Nous proposons une méthode grâce à l'imputation des *D.M.* qui évitera le défaut cité ci-dessus. Nous nous servons de l'algorithme de Fèvre comme l'étape initiale. Ensuite nous imputons les données manquantes de chaque individu par celle d'un individu le plus proche dans la classe. Nous recalculons les centres de gravité en utilisant les pseudo-données complètes et l'itération du processus se poursuit jusqu'à ce que l'inertie interclasse ne s'améliore pas ou le nombre maximum d'itération soit atteint. Nous introduisons cette méthode dans le contexte où le taux de données manquantes est inférieur à 10% et pour les données manquantes du type *'D.M.C.A.'*. Le programme en SAS IML se trouve dans *l'Annexe A.II.*

- <u>L'imputation de données manquantes d'une classe</u> C_l.

Nous proposons d'imputer les données manquantes par les valeurs correspondantes d'un individu de la même classe. En ignorant les valeurs des données manquantes de l'individu concerné, nous comparons les valeurs des autres variables à partir desquelles nous trouverons un donneur des données pseudo-complètes pour l'individu ayant des données manquantes La pseudo-distance servira à mesurer cette ressemblance.

Soit x un individu ayant des réponses incomplètes, u_j, $j = 1, \ldots, m$ les individus pseudo-complets de la classe C_l

si $\hat{d}(x,u_d) <= \hat{d}(x,u_j)$, $j=1, \ldots, m$. Cette distance se calcule en ignorant les *D.M.* de l'individu concerné x. Alors u_d est le donneur de l'unité x. Nous remplaçons les données manquantes de x, par les valeurs de celles de u_d.

Nous utilisons ce principe pour chaque donnée manquante de chaque classe C_l. Ainsi nous obtenons les pseudo-données complètes. Le centre de gravité utilisant les pseudo-données complètes se définit comme suit :

$$g_l = \left[\sum_{i \in c_l} p_i\right]^{-1} \left[\sum_{i \in c_l} p_i x_i\right] \quad \forall l \in \{1,2, \ldots ,k\} \qquad (3\ 41)$$

- **L'algorithme principal**

Phase préalable pour diviser les données en K classes, nous nous assurons qu'il existe au moins K individus complets en faisant une suppression des variables ayant très peu de réponses.

Phase I (initiale) : on n'estime pas les données manquantes

Nous tirons les K individus aléatoirement dans les sous-ensembles de données complètes pour faire K centres de gravité g_l $l=1, \ldots ,k$.

- L'affectation des données, y compris des données incomplètes, dans les différentes classes en utilisant la procédure indiquée dans 3.3.1 avec la pseudo-distance.

- Nous calculons les centres de gravité en utilisant la formule de (3.22).

- Itérer 1. et 2. et s'arrêter si la pseudo-inertie intraclasse ne s'améliore pas ou si le nombre d'itération maximal est atteint .

Phase II (principale) : la classification avec l'imputation des données manquantes.

- l'imputation de données manquantes comme indiquée au debut de la section 3.3.2.

- nous calculons les centres de gravité en utilisant les pseudo-données complètes servant la formule standard.

- la partition des données en K classes, la même processus que la méthode classique.

Nous répétons les étapes 1, 2, 3 et nous s'arrêtons dans les mêmes conditions que dans la phase I.

3.3.3. Méthode choisie pour la comparaison

La méthode de classification *FASTCLUS* de SAS du même type que notre méthode est choisie pour la comparaison. La procédure *FASTCLUS* combine une méthode efficace pour trouver les centres de gravité initiaux avec un algorithme standard d'itération pour minimiser la somme totale des inerties intraclasses. Elle utilise une méthode d'Anderberg (SAS/STAT Users's Guide, 1994) nommé 'la classification selon le plus proche centre'. La procédure *FASTCLUS* diffère des autres méthodes par la sélection des classes initiales. Cette procédure garantit le fait que toutes les distances entre les observations dans les mêmes classes sont inférieures aux distances entre les observations des différentes classes. Cette technique d'initialisation a pour objet de minimiser le nombre d'itérations nécessaire lors de l'affectation des individus aux classes. Elle se compose des étapes suivantes : on choisit la première observation complète comme le premier noyau. La prochaine observation complète qui sépare de premier noyau par une distance supérieure à celle spécifiée dans l'option *RADIUS* $=t$ sera sélectionnée comme deuxième noyau. Les autres observations complètes seront sélectionnées si elles sont séparées des autres noyaux d'au moins la distance t. Jusqu'à présent, si une observation n'est pas sélectionnée comme un noyau, deux tests suivants seront appliqués pour savoir si celle-ci peut remplacer un ancien noyau et ainsi devenir un nouveau noyau.

- test 1 : si la distance d'une nouvelle observation par rapport au centre le plus proche est supérieure à la distance minimale entre noyaux, un ancien noyau parmi les deux plus proches l'un à l'autre sera remplacé Celui qui aura la plus petite distance avec les noyaux restants, y compris le nouveau noyau, va être remplacé.

- test 2 : à la suite du test 1, si une observation n'est pas choisie comme un noyau, on remplace l'observation par le noyau le plus proche si la plus petite distance de l'observation à tous les centres, sauf celui le plus proche, est supérieure à la plus petite distance du plus proche noyau aux autres.

Pour traiter des *D.M.*, la procédure *FASTCLUS* utilise une distance modifiée en servant seulement les données disponibles. La distance entre un individu ayant des *D.M.* et le centre de gravité est basée sur les valeurs non manquantes. Cette distance est multipliée par le ratio du nombre de variables au nombre de valeurs non manquantes :

$$d = \sqrt{\frac{m}{p}\sum_{i=1}^{p}(x_i - g_{ki})^2}$$

où m - le nombre de variables,

p - le nombre dẽ réponses.

3.3.4. Critère d'évaluation et résultats

Le critère de qualité des méthodes :

Le pourcentage de bon classement est choisi pour évaluer les résultats des classifications Ce critère est calculé par rapport au résultat des données réelles complètes de la méthode SAS standard. On crée des données manquantes sur les données originelles pour comparer les résultats. Les jeux de données sont simulés de la même façon que dans la méthode de l'ACP. des données sont issues des lois multinormales standards de centre différent. C'est-à-dire que des données sont issues de deux lois normales multivariée soit N(**0**,**R**) soit N(**d**,**R**) de dimension 10 : *d* prend successivement la valeur 0.3, 0.5 et 1 et *r* est égal à 0.5 et r est égal à 0.5 ce qui équivaut à des distances de Mahalanobis D^2 entre les deux groupes de 1.636, 4.545 et 18.18. La taille d'échantillons est de 300 individus. Pour chaque cas différent, on associe 5 jeux de données pour calculer le résultat.

Taux de D.M. 5% :

A cause des différences de choix sur les centres de gravités initiaux entre notre méthode et *FASTCLUS*, nous comparons le résultat de notre méthode aux données incomplètes avec celui de SAS aux données complètes si et seulement si le résultat de notre méthode sur les données incomplètes est différent de celui de notre méthode sur les données complètes. Ainsi le mal classé est sûr d'être dû aux données manquantes et non aux choix sur des centres de gravités initiaux.

- Notre méthode est légèrement meilleure que FASTCLUS lorsque la distance de Mahalanobis D^2 est égale à 1.636 ou à 4.545.

- lorsque la distance de Mahalanobis D^2 entre deux groupes est égale à 18.18, FASTCLUS est légèrement supérieure à la notre méthode. Nous pensons que c'est peut-être dû au meilleur choix sur des centres de gravités initiaux puisque dans ce cas les deux classes sont très distinctes.

Tab. 1 Pourcentage de bien classés

D^2 / Méthode	1.636	4.545	18.18
Notre méthode	94.8%	90.2%	97%
Méthode SAS	94.2%	89.6%	97.2%

Taux de D.M. 10% :

Nous avons tout à fait la même conclusion que pour le taux de données manquantes à 5%.

Tab. 2 Pourcentage de bien classés

D^2 / Méthode	1.636	4.545	18.18
Notre méthode	90.2%	88%	91.8%
Méthode SAS	89.4%	86.2%	92.2%

En conclusion, notre méthode du traitement des *D.M.* dans le cadre de la classification est adéquate. La méthode ignorant les *D.M.* comme la technique utilisée dans *SAS* est acceptable. Dans la pratique, on peut améliorer notre méthode avec la technique de la sélection des centres de gravités initiaux, comme celle utilisée dans SAS FASTCLUS pour avoir une qualité de classification plus stable. Pour les taux de données manquantes supérieurs à 10%, nous suggérons de recourir aux méthodes de reconstitution des données manquantes spécifiques avant d'effectuer une analyse de classification

CONCLUSION

Cette partie de la thèse avait pour objet de décrire un ensemble de méthodes de traitement des données manquantes. En particulier nous avons étudié la méthode de l'analyse homogène développée par l'équipe Néerlandaise qui a une grande valeur dans la pratique : car elle permet de résoudre le problème des données manquantes de type non complètement aléatoire sur des variables nominales Nous avons d'ailleurs introduit cette méthode dans la fusion des fichiers.

Nous avons proposé pour l'analyse en composantes principales une technique appropriée aux cas des données incomplètes et la conclusion tirée à partir des comparaisons sur des données simulées et réelles est satisfaisante.

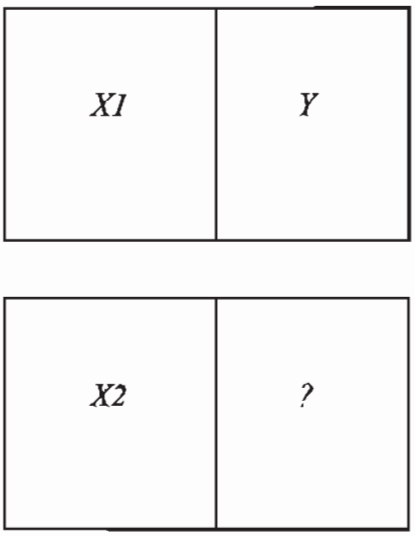
Un algorithme sur la classification automatique du type '*Nuées Dynamiques*' avec des données incomplètes est également proposé et évalué.

PARTIE II

LA FUSION DE FICHIERS

INTRODUCTION

La différence fondamentale entre le problème du traitement des données manquantes et celui de fusion de fichiers est que cette dernière traite des données manquantes apparaissant en bloc et non des données manquantes dispersées dans tout le tableau. Autrement dit, les données manquantes sont du genre "variables manquantes" dans un fichier de données. Ces variables manquantes sont présentées dans un autre fichier de données.



On rencontre souvent des cas de variables manquantes dans la pratique, par exemple dans les enquêtes : il peut s'agir de questions non posées parce que le questionnaire est trop long, ou bien de non-réponses aux questionnaires. Dans ce cas, la reconstitution des variables manquantes s'avère nécessaire à la qualité des résultats Cette technique est fréquemment appliquée par les praticiens dans le monde du marketing. L'autre type d'utilisation de cette technique est le suivant : les informations existent dans différentes sources et on voudrait créer une source unique et complète rassemblant celles des différentes sources. Dans ce cas, elles concernent une optimisation d'utilisation des informations.

Le problème posé à la fusion de fichiers est celui de la validation des données reconstituées et les utilisations ultérieures de ces données. Ces données simulées ressemblent - elles vraiment aux données réelles et avec quelle certitude ? Dans quel cas peut-on les utiliser sans causer trop de distorsion aux résultats d'analyse et sans conduire à une mauvaise

interprétation ? Est-ce que ces données permettent être utilisées pour faire de la statistique inférentielle au niveau individuel ou seulement au niveau collectif ? La plupart des méthodes existantes sont en principe des méthodes pour prévoir les comportements globaux et des données reconstituées que l'on considère comme satisfaisantes si elles permettent de produire des tableaux croisés corrects. La validation consistant à comparer les données reconstituées avec les données réelles n'est en générale pas appliquée, et de façon générale il y a peu d'étude de validité. Les méthodes ayant une bonne qualité individuelle sont encore rares.

Environ 90 % de données provenant d'enquêtes sont qualitatives ; La fusion de fichiers concerne souvent des données provenant d'enquêtes, une attention particulière sera donc portée au cas de données qualitatives.

Dans cette partie, nous présentons une méthode qui, non seulement, a une bonne qualité individuelle mais aussi respecte la structure des relations entre variables. Cette méthode (Saporta G. & Co V., 1996) est adaptée au problème général, où l'on n'exige pas que les données suivent un modèle statistique. Elle accepte des différences de structure et de taille des données entre le fichier donneur et le fichier receveur. La méthode que nous proposons a une approche tout à fait différente des méthodes classiques : il s'agit de simuler pour les variables manquantes des valeurs homogènes à l'ensemble des données existantes. Nous avons également proposé une technique d'évaluation des variables reconstituées au niveau individuel.

CHAPITRE IV : GENERALITES SUR LA FUSION DE FICHIERS

4.1. HISTORIQUE DU DEVELOPPEMENT DE LA FUSION DE FICHIERS

L'histoire de la fusion de fichiers n'est pas très ancienne : cette technique n'existe que depuis 20 ans et il y a peu de littérature sur ce sujet.

Wendt F. (1976, 1984) et Boucharenc L. (1981) sont parmi les premières personnes à présenter la technique de fusion. Cette méthode consiste à apparier un receveur dans le fichier Receveur et un donneur dans le fichier Donneur et à transférer ensuite les valeurs des variables correspondantes aux receveurs.

Santini G. (1984) développe et enrichit cette méthode en introduisant un type d'appariement entre receveur et donneur qui limite les copies d'un même donneur et utilise ainsi le plus grand nombre possible de donneurs.

Wendt F. (1984) propose une nouvelle méthode de fusion. Elle consiste à effectuer une typologie avec les variables communes de l'ensemble des deux fichiers. C'est au sein de chaque groupe que l'on effectue la fusion. Cette technique est utilisée essentiellement en Allemagne.

Rubin D.B. (1986) a proposé une méthode de régression paramétrique dans le cas où l'on a trois fichiers : un fichier receveur *A* sur *(x, y)*, un fichier donneur *B* sur *(y, z)* et un fichier supplémentaire *C* sur *(y, z) ou (x, y, z)*. Dans cette méthode de régression, d'abord, on déduit une valeur intermédiaire z_{int}, une prédiction par la régression de la variable à transférer *z* par rapport aux variables communes *(x, y)*. Puis, on détermine une valeur authentique *z* issu du fichier Donneur *B* qui a la valeur la plus proche de la valeur intermédiaire z_{int}.

Paass G. (1986) a présenté une méthode de régression non paramétrique dans le même contexte que le cas précédent. Il s'agit d'un double emploi de méthode d'imputation "*hot-deck*". Cette méthode consiste essentiellement à déterminer tout d'abord une valeur intermédiaire z_{int} à l'aide du fichier *C* par rapport aux variables communes (*x*, *y*) du receveur

A. Puis on détermine une valeur *z* authentique à partir du fichier donneur *B* en utilisant de nouveau le principe de "*hot-deck*" sur une distance euclidienne de *(X, Z)*.

Singh A.C., Mantel H.J., Kinck M.D. et Rowe G. (1993) présentent une autre technique de fusion traitant aussi le problème du cas précédent. Ils proposent une solution de remplacement à l'hypothèse d'indépendance conditionnelle en utilisant une information supplémentaire inclue dans le fichier C et améliorent les deux méthodes cités ci-dessus.

Hendrickson et Acott (1994) introduisent une technique de réseau neuronal dans la fusion à condition que les variables communes soient prédictives de celles à transférer. Cette méthode utilise les informations du fichier donneur comme modèle d'apprentissage. Les variables communes du fichier receveur sont les données d'entrées et les variables à transférer sont les informations à récupérer à la sortie dans le modèle du réseau de neurones construit par le fichier donneur.

L'AIMC (1994), Centre d'Information sur les Média à Bruxelles pratique des méthodes d'imputation multiple dans la fusion de fichiers et utilise un modèle probabiliste pour transférer isolément chacune des valeurs des variables.

STATIRO (1994) utilise la technique de fusion dans les enquêtes du type 'auto-administré', plus précisément une technique d'injection, par la recherche de sosies en se basant sur des résultats d'analyse des correspondances.

L'INSEE (1994) pratique une technique de fusion par imputation aléatoire dans différentes cellules. Cette méthode consiste en un choix d'un critère de partition de cellule, la partition des individus des deux fichiers en cellules et l'affectation à un individu d'un donneur pris au hasard avec (ou sans) remise dans la même cellule.

Jusqu'alors le problème de la validation des résultats de la fusion n'est pas standardisé. Lebart L. et Lejeune M. (1995) proposent des techniques de validation : la validation croisée et la validation par une approche bootstrap. La première est une méthode de comparaison de données réelles cachées, avec celles reconstituées par la fusion La deuxième est une procédure

de ré-échantillonage afin d'évaluer les statistiques, par exemple la variance, sur des variables fusionnées.

4.2. ENQUETES ET FUSION DE FICHIERS

Dans le cadre où l'environnement est suffisamment large au sens de la représentativité de la population, on peut admettre qu'un individu particulier d'une enquête peut se retrouver dans une autre enquête ; la similarité est mesurée par une distance dans un espace multidimensionnel. C'est-à-dire que les éléments d'un vecteur ne sont pas forcément identiques mais proches sous certains aspects. Donc, on peut marier deux individus issus de deux enquêtes différentes et imputer les vecteurs manquants de l'un par les valeurs renseignées de l'autre qui est à l'origine de la fusion de fichiers.

Quelques usages des fusions de fichiers :

Le champ d'application de la fusion de fichiers est très vaste. Voici quelques exemples :

1) En marketing, surtout pour les études de média-marché : les habitudes de la consommation médiatique par rapport aux comportements de consommation des produits

2) La fourniture d'estimations locales à partir d'une enquête nationale. Le panel de ménages décrit les consommations et le profil de consommation, y compris les patrimoines et équipements et joue un rôle de fichier donneur. Le fichier des communes, riche de la structure socio-économique de la population et des données d'équipement, intervient comme le fichier receveur pour constituer un fichier concernant la population locale contenant des informations complètes.

3) Il est difficile d'interroger une même personne sur sa consommation télévisuelle, l'audience de la radio et sa lecture de la presse. En fragmentant les médias entre plusieurs enquêtes indépendantes, on se demande alors :

- une émission et un quotidien attirent - ils le même type de clientèle ?

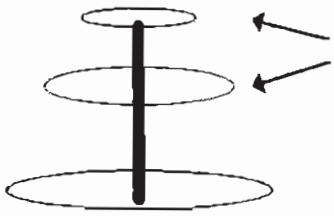
- la télévision détourne-t-elle de la presse sa clientèle potentielle ?

Un autre exemple simple est que deux échantillons issus d'un même fichier de clients sont respectivement soumis aux questionnaires *A et B*. Qu'auraient répondu aux questionnaires *B* les gens interrogés à propos du questionnaire *A* ?

Par ailleurs, pour les systèmes multicapteurs, on utilise des données provenant de plusieurs capteurs A ceux-ci qui peuvent être de nature très variée, signal, son, image... Il convient généralement d'adjoindre des données issues d'autres sources informatiques.

Conception d'enquêtes pour l'usage de fusion de fichiers :

Dans la pratique, on crée des questionnaires d'une enquête afin que celle-ci puisse servir de fichier donneur pour un autre fichier de données. L'enquête est établie de façon telle que les variables de base correspondent aux besoins pour la fusion. On crée une base de données contenant des variables à caractéristiques socio-démographiques, etc. C'est-à-dire qu'à chaque enquête, il y a des variables différentes et des variables communes. Pour cela, une enquête peut servir de fichier donneur et de fichier receveur d'enquêtes différentes. Ce système de fichiers multicouche est illustré dans la figure ci-dessous :



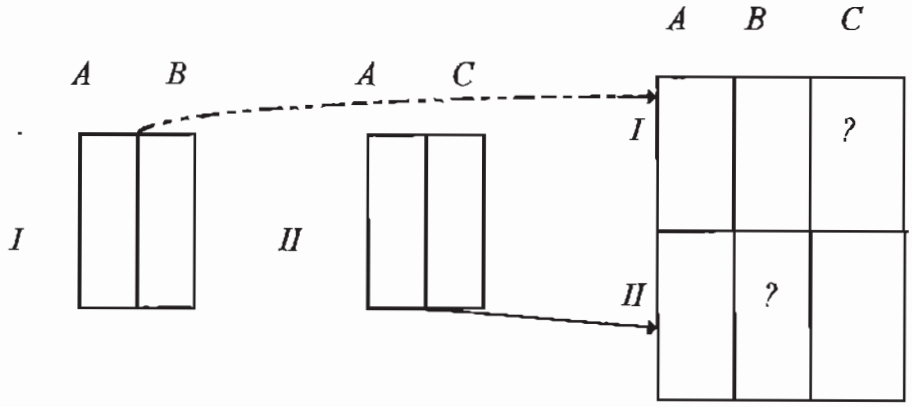
Différentes enquêtes considérées comme fichier donneur ou receveur particulier des autres.

Structure de base de la population : caractéristiques dérivées à partir des statistiques officielles, des enquêtes ou des bilans annuels.

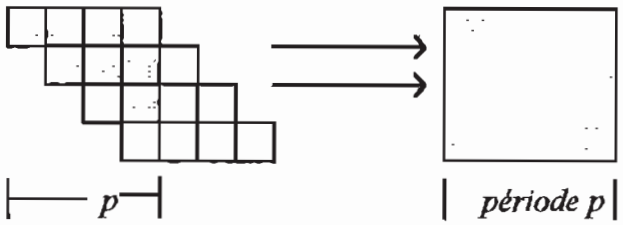
Ce système de base de données sert à la réalisation de la fusion peut s'utiliser dans des études chronologiques. Par exemple, pour la première année, on fait une enquête complète sur un sujet. Puis pour les années suivantes, on effectue seulement des petites enquêtes supplémentaires qui servent à indiquer l'évolution de ce sujet. Ces différents 'repères' nous signalent quand il faut renouveler les enquêtes entières.

On rencontre quatre cas différents de fusion dans la pratique (Santini G., 1986) :

1. *Enquête parallèle* : deux échantillons appareillés au sens de la représentativité de population sont en même temps interviewés sur les variables A, B et A, C et on voudrait avoir des informations des échantillons I et II sur des variables A, B et C :

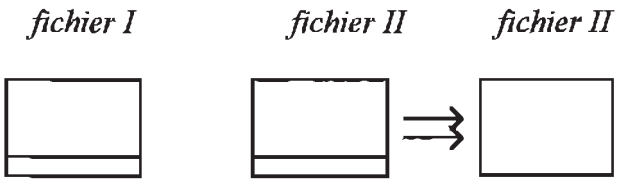


2. *Panel glissant* : l'échantillon est renouvelé par quart chaque semaine. Pour une période p, les données ne sont pas disponibles sur l'ensemble d'échantillon et on voudrait avoir des informations sur tous les sous-échantillons de la période p.



La caractéristique des 2 cas ci-dessus est : la ressemblance des échantillons (ou sous-échantillons) donneurs et receveurs, on appelle ce genre de fusion *fusion quasi-appareillée*.

3. *Ré-interviews* : le fichier II a les mêmes individus du fichier I, certaines personnes ne répondent pas au second questionnaire et on cherche avoir des informations sur la population entière du fichier II.



Dans ce cas, il peut y avoir une raison significative correspondante aux non-répondants. C'est aussi un problème d'unité manquante que nous avons cité dans la première partie.

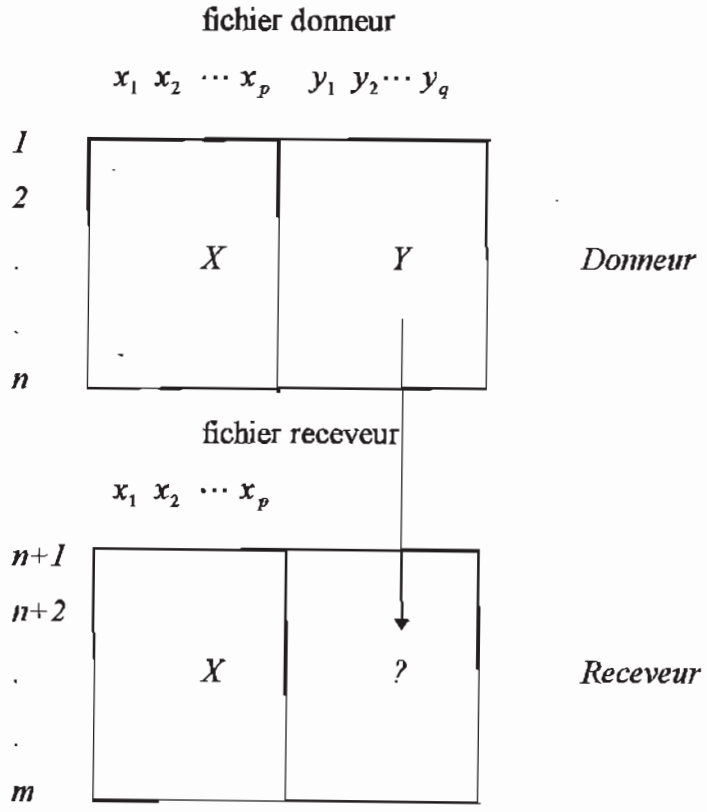
4. *Mise à jour* : les deux échantillons sont représentatifs de la population à étudier mais à des dates différentes. Le second échantillon destiné à une mise à jour n'est pas interviewé sur l'ensemble des questions. Dans ce cas, on parle de la *fusion évolutive* : supposons qu'il existe des échantillons différents mais représentatifs de la population relative aux deux époques d'interviews et postulons que l'évolution des phénomènes décrits par les variables à transférer ne change pas en intensité. En d'autres termes, la seule évolution admise des phénomènes est celle de la structure de la population.

4.3 PRINCIPALES NOTIONS SUR LA FUSION DE FICHIERS

Le but de la fusion de fichiers consiste à utiliser au maximum les informations existantes pour reconstituer les informations manquantes qui nous intéressent et qui n'étaient pas à l'origine renseignées. Il s'agit de simulations des informations manquantes dont nous avons besoin. Son principe est qu'à partir d'un bloc de variables renseignées, assez corrélées avec le bloc de variables à reconstituer, on estime les valeurs des variables non-renseignées. La différence avec le problème de données manquantes cité dans la première partie est que dans la fusion de fichiers, les données manquantes concernent des groupes de variables non renseignées, en fait, c'est sont des variables manquantes. Ici, il s'agit de rapprocher des fichiers issus de sources différentes.

Précisons ce que l'on entend par fichier donneur et fichier receveur : soit deux sources de fichiers qui contiennent des renseignements (ou variables) différents (dont une partie commune) sur des individus différents. L'une de ces sources sert de *fichier receveur* (ou fichier

cible), dans lequel des données sont reconstituées pour chaque variable manquante à partir des informations de l'autre source définie comme *fichier donneur*.



- variables critiques : une partie des variables communes dans les deux fichiers, receveur et donneur, servent principalement à reconstituer les valeurs des variables manquantes. Ces variables sont prédictives par rapport aux variables à reconstituer. Dans les méthodes classiques, ces variables critiques servent à déterminer pour chaque individu du fichier receveur ses donneurs éligibles.

- variable de rapprochement : une partie des variables communes, par un calcul de distance, permet de choisir pour chaque receveur le donneur le plus proche dans les méthodes classiques

Dans la plupart des méthodes existantes, on considère que la fusion de fichiers est un cas particulier de l'imputation par bloc. Elle consiste à donner à un receveur l'ensemble des valeurs d'un donneur pour les variables non renseignées. En fait, la fusion de fichiers peut être considérée comme l'application des méthodes d'imputation à grande échelle en utilisant les variables communes des deux fichiers. On complète les enregistrements du fichier receveur en imputant des valeurs authentiques (observées) de *Y* du fichier donneur en utilisant les relations entre *Y et X*. Un autre type de fusion consiste à estimer les valeurs de chacune des variables et

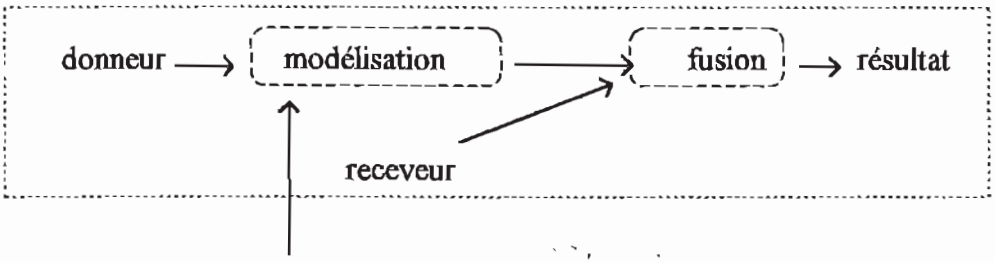
à ne pas imputer par bloc entier d'individu. La méthode que nous allons proposer appartient à ce dernier cas.

On peut distinguer quatre types de fusions (M Lejeune et L. Lebart, 1994) :

- l'injection : les fichiers donneur et receveur sont deux sous-échantillons d'une même enquête,
- la fusion unilatérale : fichiers donneur et receveur issus de deux enquêtes distinctes,
- la fusion réciproque : même type que la fusion unilatérale mais fusion dans les deux sens,
- la fusion en appariement : il existe trois fichiers *A(X, Y), B(X, Y) et C(X, Y, Z)*. Les fichiers *A* et *B* servent de fichiers receveurs Le fichier *C* sert de fichier donneur.

Les variables à transférer dans le fichier receveur sont considérées comme une simulation d'un modèle de distribution des variables du fichier donneur. Deux cas sont à envisager suivant que le modèle est explicite ou implicite avant la fusion (G. Santini, 1986)

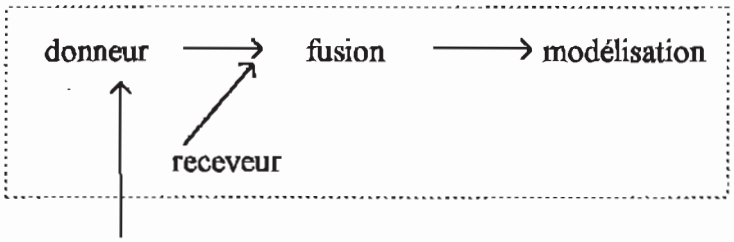
- *modélisation / fusion* : si une modélisation est construite à partir du fichier donneur et que l'information est transférée aux receveurs selon cette loi de probabilités, on dira qu'on est en présence d'un schéma de modélisation / fusion.



fichier contenant des caractéristiques et des probabilités individuelles.

Par exemple, dans une fusion, on effectue d'abord un redressement sur le fichier de donneur tel que la population soit conforme à celle du fichier receveur. Ensuite, on établit un modèle de régression avec les variables à reconstituer comme variables à expliquer, les variables prédictives variables explicatives. En utilisant ce modèle de régression, on fait la fusion. Dans ce cas, il s'agit de la *modélisation / fusion*.

- *fusion / modélisation* : par contre, si on fait une fusion sans avoir un modèle explicite et qu'on effectue après coup l'optimisation du fichier résultat, on parlera d'un schéma de fusion / modélisation :



fichier constituant une simulation des lois de probabilité latentes.

Par exemple, la fusion de type d'imputation 'hot-deck' avec une procédure d'optimisation globale est plutôt une fusion / modélisation.

Le schéma modélisation / fusion a une forme plus "organisée" et évite les problèmes de tailles d'échantillons et d'aléas relatifs aux simulations.

Les conditions nécessaires concernant des données pour faire la fusion :

1. Du point de vue de la population : le fichier donneur doit avoir suffisamment d'unités statistiques représentatives dans chaque sous-groupe de population contenu dans le fichier receveur. Si nécessaire, on pourra faire un redressement sur le fichier donneur.

2. Du point de vue des variables : il doit exister une assez forte liaison entre les variables à reconstituer et les variables prédictives (communes). Ceci est nécessaire pour obtenir une qualité satisfaisante des données reconstituées.

4.4. EVALUATION DES RESULTATS DE LA FUSION

L'évaluation des résultats de la fusion se fait en général de la même façon que pour les données manquantes partielles (section 2.4).

Dans le cas où l'objectif de fusion est de recréer une banque de données aussi proche que possible de celle qu'on aurait obtenue en interrogeant tous les individus sur toutes les questions, on doit envisager de modéliser les comportements au niveau individuel et rendre compte des écarts entre le comportement réel et le comportement simulé. Lejeune M. & Lebart L. (1995) montrent que les structures du fichier donneur et du fichier receveur peuvent coïncider parfaitement, mais cela ne dit rien sur la qualité prédictive des variables communes pour les variables transférées ; et cela ne nous dit rien sur les bonnes réponses individuelles reconstituées. La validation individuelle consiste à évaluer des bonnes réponses individuelles reconstituées. La validation individuelle et la validation globale sont bien deux critères distincts mais un bon résultat de la première entraîne celui de la deuxième. Mais l'inverse n'est pas toujours vrai : c'est-à-dire une bonne validation globale ne signifie pas forcement la validation individuelle. Pour avoir des résultats valides au niveau individuel un heureux concours de circonstances est indispensable : une bonne méthode de fusion et un bon pouvoir prédictif des variables communes. Pour un problème concret où l'on dispose d'un certain nombre des variables communes, c'est-à-dire des informations limitées, la validation se traduit par l'optimisation de l'utilisation des informations existantes : on profite au maximum des informations du fichier donneur pour la fusion. Autrement dit, en examinant la relation des deux blocs de variables dans le fichier donneur et selon le degré de liaison on peut déterminer la qualité espérée de la fusion. Dans tous les cas, les données fusionnées sont des données simulées, elles ne sont pas des observations réelles, donc des écarts entre les deux sont inévitables. Lejeune M. et Lebart L. (1995) conseillent d'accompagner chaque tableau d'un certain nombre d'individus observés afin d'avoir des résultats plus solides et proches de la situation réelle.

Lejeune M. et Lebart L. (1994) proposent la validation croisée comme suit :

- divisons au hasard un fichier complet en s parties comme s fichiers receveurs,
- fusionnons s fois pour chaque fichier receveur,
- calculons le taux d'erreur.

Le résultat est mesuré et évalué selon deux niveaux : niveau global et niveau individuel.

- ***Niveau global : (voir section 2.4).***

- ***Niveau individuel : (voir section 2.4).***

le coefficient de proximité (voir section2.4) .

le coefficient de proximité associé à l'affectation aléatoire respectant la proportion p_1 , p_2 , ... , p_k des catégories, vaut :

$$\tilde{d} = \sum_{i=1}^{k} \sum_{j=1}^{k} c_{ij}^{k} p_i p_j$$

Cette distance rend compte des structures de données ; en effet, elle varie en fonction de ces dernières. Par exemple, pour le tirage aléatoire d'une variable à six catégories :

si P={ 2, 0.6, 0.5, 0.5, 0.6, 1.8}/6,	$\tilde{d}$ *=0.467*	
si P={ 1, 1, 1, 1, 1, 1 }/6,	$\tilde{d}$ *=0.389*	
si P={0.5, 1.5, 0.5, 1, 1.5, 1 }/6,	$\tilde{d}$ *=0.37*	
si P={0.1, 0.1, 0.1, 0.1, 3.1, 2.5}/6,	$\tilde{d}$ *=0.161*	*(1)*
si P={0.1, 0.1, 0.1, 5.5, 0.1, 0.1}/6,	$\tilde{d}$ *=0.058*	*(2)*

Soit $\bar{d}$ le coefficient de proximité associé à une règle donnée sur l'ensemble des cases reconstituées $\bar{d}$ -$\tilde{d}$ mesure le gain par rapport à une affectation aléatoire.

Pour comparer les différentes méthodes, on évalue selon deux critères de validation et rend compte également de son adaptation aux diverses situations : par exemple, différentes proportions de taille ou différentes structures des deux fichiers.

CHAPITRE V : LES DIFFERENTES METHODES DE LA FUSION DE FICHIERS

5.1. LA FUSION PAR IMPUTATION ALEATOIRE

On appelle une **cellule** un groupe d'individus ayant des mêmes valeurs sur un ensemble de critères.

L'appariement (imputation) aléatoire comprend essentiellement les trois étapes suivantes :

- le choix d'un critère de partition par cellule,
- la partition de l'ensemble des individus de fichiers donneur et receveur en groupes (cellules) selon le critère de partition,
- l'affectation à chaque individu du fichier receveur d'un individu du fichier donneur, pris au hasard, avec ou sans remise dans la même cellule.

Ce genre de technique s'utilise sous une condition assez précise : il exige que les deux échantillons aient à peu près la même distribution des variables communes et soient à peu près de la même taille

On parle de poids *hétérogène* dans le cadre de l'appariement aléatoire par cellule lorsqu'un fichier est redressé par certaines variables qui ne sont pas inclues dans le critère de partition. Dans ce cas, si l'on applique sans précaution la méthode de l'appariement aléatoire par cellules, des biais de l'estimation (la moyenne ou la variance) sur les variables à transférer peuvent apparaître (*INSEE*, 1994) :

- Lorsqu'un redressement est effectué sur des échantillons du fichier donneur et qu'il existe une corrélation négative (positive) des variables à transférer avec le poids hétérogène à l'intérieur d'une cellule, alors on surestime (sous-estime) la moyenne des variables à transférer

Le remède est simple : chaque individu dans le fichier donneur sera démultiplié, c'est-à-dire, dupliqué proportionnellement à son poids (à l'intérieur d'une cellule).

- Lorsque des échantillons du fichier receveur sont redressés par un poids hétérogène, alors on doit démultiplier chaque individu du fichier receveur proportionnellement à son poids.

La démultiplication des receveurs a pour effet positif de réduire le bruit dû au caractère aléatoire de la méthode : la démultiplication du fichier receveur augmente la taille du fichier receveur et donc le nombre d'opérations d'affectation aléatoire (similaire à une imputation multiple). Nous concluons sur le redressement par les deux points suivants :

1. Il est préférable de démultiplier le fichier donneur, lorsque il existe une corrélation des variables à transférer avec le poids hétérogène à l'intérieur de cellule.

2. Démultiplier le fichier receveur n'a pas d'effet sensible, mais il a le mérite de réduire le caractère aléatoire de cette méthode de fusion.

Par exemple, (*INSEE, 1994*) applique l'imputation aléatoire sur une enquête de revenu. Une cellule regroupe des ménages de même *TAP*, où :

T - composition de famille : il existe six types de familles,
A - tranche d'âge : six tranches âge,
P - *CSP* : sept catégories

Il y a donc au total *252* cellules. La réduction du nombre de cellules (ex, *TAP->T ou A ou P seul*) et la proportion de la taille de fichier receveur par rapport au fichier donneur (*50%, 10%, 90%*) ont des résultats significativement différents. Seul le critère de partition TAP a des résultats satisfaisants. Le cas où la proportion du nombre de donneurs par rapport aux receveurs est de 10%, entraîne une faible variance des variables à transférer. Donc, le choix du nombre des critères de cellule et le rapport de la taille entre deux fichiers sont cruciaux, dans cette méthode.

L'avantage de cette méthode est le respect des relations entre les variables transférées. L'inconvénient est que cette méthode tend à trop homogénéiser les comportements simulés pour les individus du fichier receveur.

5.2. FUSION BASEE SUR UN "REFERENTIEL FACTORIEL"

La fusion basée sur "référentiel factoriel" obéit au principe suivant :

- *Recherche du référentiel factoriel* : on effectue une analyse des correspondances multiples sur le tableau des variables critiques (communes) à l'ensemble des données disponibles (donneur + receveur) et l'on conserve les k premiers axes de l'analyse, ce qui permet de positionner les observations dans un espace R^k et de calculer la distance entre les points sur les coordonnées factorielles. La similarité des individus est mesurée par cette distance.

- *Recherche de voisinage* : les deux fusions présentes ci-dessous utilisent la notion de voisinage d'un point dans un espace de k dimensions. Tous les points voisins sont compris dans une sphère dont le centre est 'a'. Dans le cas de faible densité, c'est le moyen efficace de trouver le voisin le plus proche. Dans le cas de haute densité, on diminue le rayon du cercle. Pour chaque receveur, on sélectionne un ensemble de donneurs dans un voisinage du receveur.

On présente ici deux techniques basées sur ce principe.

5.2.1. La fusion par "mariage" (Santini G., 1984)

1. Réalisation des mariages entre donneur et receveur :

L'algorithme rend difficile des liaisons multiples, un peu de la même manière que la saturation des valences d'une molécule interdit de nouvelles liaisons. Les individus sont mariés (donneur-receveur) en fonction de leur proximité d calculée sur coordonnées factorielles Si un individu donneur est déjà marié à δ autres individus, cette distance d est pénalisée par la formule suivante :

$$d^* = 1-(1-d)^{\delta}$$

δ - nombre d'individus déjà mariés à ce donneur

Il existe quatre types de 'mariages' :

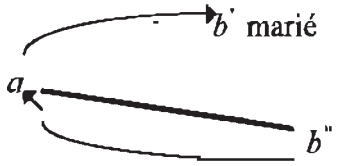
a - receveur ; b - donneur ;

- "coup de foudre" :

a est le plus proche voisin de b et réciproquement c'est la même chose et b n'est jamais copié.

--> On unit a *et* b.

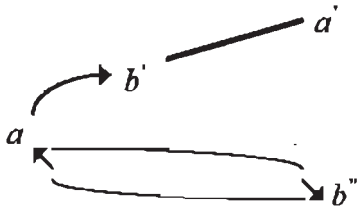
- "ami d'enfance" :



- b' est le plus proche voisin de a mais b' est marié. $b"$ n'est pas encore marié qui est le plus proche de a à part b'.

--> On unit a et $b"$.

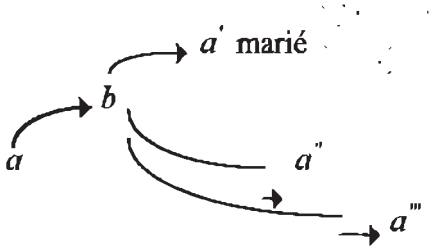
- "liaison adultère" :



a est trop loin de $b"$. la distance entre a et b', avec la pénalité attribué à b' (à cause de son premier mariage), est plus petite.

--> b' (marié à a') est réuni à a également.

- "assiduité" :



- b est le plus proche de a. Mais les plus proches voisins de b est a',$a"$ et $a"'$ qui sont tout mariés.

--> On unit a et b.

En dehors des cas cités ci-dessus, il existe bien d'autres cas plus complexes qui sont résolus ultérieurement, dans une phase d'optimisation au niveau global.

2. Choix final d'un donneur : parmi les donneurs potentiels, on choisit celui qui ressemble le plus au receveur sur un ensemble de variables signalétiques comme : l'âge, le sexe, la *C.S.P.* .

Certaines contraintes peuvent être imposées au processus du 'mariage' : par exemple, l'interdiction du mariage si deux individus n'appartiennent pas au même groupe socio-démographique. Ou encore, un critère général peut être imposé par la spécificité du problème concret, par exemple, une zone de transmission de radio, ou une date d'interview...

5.2.2. Fusion par recherche de sosies élaborée par Statiro (Sousselier J., 1994)

La recherche du plus proche voisin comprend :

- pour chaque unité de la population des receveurs et par le calcul d'une distance sur les coordonnées factorielles, m unités de la population des donneurs sont retenues si la distance avec ce receveur satisfait $D(r, d)<S$,

où r - receveur, d - donneur et S - un seuil de distance.

- une contrainte pour choisir son donneur : on cherche, par exemple, un donneur dans la même région géographique.

- une comparaison de la ressemblance du signalétique entre receveur et donneur

Un poids est utilisé pour différencier l'importance des variables signalétiques, une note globale est calculée sur toutes les variables signalétiques et les unités les plus ressemblantes sont retenues. Ce choix du poids s'effectue selon le degré de liaison entre les variables signalétiques et les variables à transférer.

Enfin, en cas d'égalité de la note globale signalétique, le donneur le moins utilisé est finalement retenu. Avec toutes ces opérations s'il n'existe pas de donneur pour un receveur, on recommence par la première étape en élargissant le rayon de voisinage S. C'est une méthode

hiérarchique et emboîtée, à chaque étape on garde les donneurs les plus crédibles et on sélectionne le donneur le plus ressemblant.

Cette méthode est utilisée dans les enquêtes "média-marché". Il y a deux fichiers de données à partir d'une même population : le premier est renseigné pour toute la population sur des informations signalétiques et des habitudes de lecture, le deuxième contient des questionnaires de type 'auto-administrés', pour les questions sur les styles de vie, les comportements et consommations, avec environ 60% de répondants. Le but de la fusion est de reconstituer le deuxième fichier complet.

La population des donneurs, les 60% de répondants complets, a des variables :

sexe, âge, ... , habitude de lecture, produits ...,
X Y

La population des receveurs, les 40% de répondants incomplets a des variables :

sexe, âge, ... , habitude de lecture
X

X - variables communes, *y* compris quelques variables prédictives de Y.

Y - variables à transférer, par exemple, équipements de la maison, voiture, loisirs ...

X *et* Y sont de type qualitatif.

- l'habitude de lecture porte sur plusieurs titres de presse :
  - la lecture des 12 derniers mois (*o, n*),
  - l'habitude (fréquence) de lecture,
  - la date de dernière lecture.

  Ce groupe des variables sert de *variables critiques.*

- les variables signalétiques : sexe, âge, C.S.P. , niveau d'étude... servent de *variables de rapprochement.*

Les résultats montrent que, au niveau global, les distributions des variables sur les populations des répondants et des reconstitués, sont très proches. En effet, il existe peu d'écarts significatifs sur des tableaux croisés entre les fichiers receveur et donneur.

Au niveau individuel, les signalétiques des donneurs et des receveurs sont très proches, car c'est bien l'unité la plus ressemblante qui est choisie comme donneur.

5.3. L'UTILISATION D'INFORMATION SUPPLEMENTAIRE COMME SOLUTION DE REMPLACEMENT A L'HYPOTHESE D'INDEPENDANCE CONDITIONNELLE DANS LA FUSION (Singh A.C., Mantel H.J., Kinck M.D. & Rowe G , 1993)

Le contexte du problème est le suivant : il existe deux fichiers *A et B*. Le fichier *A* contient une variable *X*, variable commune, et *Y*, une variable absente du deuxième fichier *B*. Le deuxième fichier *B* contient la variable commune *X* et *Z*, la variable à transférer. Ici, le problème diffère des autres en ce qu'il existe une variable *Y* dans le premier fichier *A*.

fichier receveur *A* :	*X*	*Y*	*?*
fichier donneur *B* :	*X*		*Z*

On veut transférer *Z* sur *A*, si *Y* était connu sur *B* on utiliserait l'information donnée par *X* et Y pour estimer *Z*. Comme *Y* est inconnu sur *B*, on n'utilise que l'information fournie par *X* ce qui peut causer un biais en régression (biais de modèle mal spécifié, non prise en compte de variables explicatives). Il faut donc supposer pour qu'il n'y ait pas de biais que *Y* et *Z* sont indépendant conditionnellement à *X*. Cette hypothèse (HIC) n'est guère vérifiable. Une façon de résoudre le problème consiste alors à disposer d'un troisième fichier *C* où toutes les variables ou l'ensemble réduit de variables *(Y, Z)* sont renseignées et cela nous permettra utiliser indirectement la relation entre *Y* et *Z*. Les informations sur l'ensemble complet de variables *(X, Y, Z)* ou l'ensemble réduit *(Y, Z)* contenue dans le fichier *C* peuvent être des informations antérieures, des informations de substitution (variables différentes mais ayant une haute corrélation) ou des informations disposées en tableaux de fréquence. On utilise ces informations en remplacement de l'HIC. Nous voulons compléter les enregistrements du fichier *A* en allant chercher les valeurs de Z dans le fichier *B* sur la base des informations contenues dans les fichiers *A, B et C* en utilisant les relations conjointes entre *(X, Y* et *Z)* ou *(Y, Z)*.

fichier receveur A :	X	Y	?
fichier donneur B :	X		Z
fichier donneur C :	X	Y	Z

- Si l'information C prend la forme de coefficients de corrélation, la méthode de régression est la suivante :

A partir de la relation *X, Y et Z de C,* on établit un modèle de régression linéaire de Z par rapport à X, Y : -

$$E(Z / X,Y) = \beta_0 + \beta_1 X + \beta_1 Y \qquad (3.1)$$

1. pour chaque couple *(x, y) de A*, on détermine une valeur z_{int} à l'aide de (3.1).
2. on remplace chaque triplet (x, y, z_{int}) par (x, y, z_{match}) où z_{match} est la valeur la plus proche de z_{int} dans B en utilisant la distance de (X, Z).

la méthode du type 'hot-deck' :

1 pour chaque couple *(x, y) de A,* on cherche une valeur z_{int} tirée de C qui rend la distance minimale sur *(X, Y)* si le fichier C *contient (X, Y et Z)* ou sur Y si C contient seulement *(Y, Z).*

2. on remplace (x, y, z_{int}) par (x, y, z_{match}), z_{match} est tirée du fichier B à l'aide de la distance minimale sur (X, Z).

- Si l'information prend la forme de proportion par case, on modifie les méthodes de fusion de fichiers par l'introduction de contraintes nominales :

On appelle ***Contraintes nominales*** un tableau de fréquences cible à respecter dans le fichier receveur.

On se sert des informations supplémentaires pour définir des contraintes nominales. Ces contraintes ont pour objet de conserver des liens catégoriques (théoriquement estimés par des modèles log-linéaires) suivant une partition de *(X, Y et Z)* acceptables dans le cas du fichier enrichi. On détermine ces liens en combinant des informations des fichiers *A, B et C.* Posons x^*, y^*, z^* comme les variables nominales correspondantes (après transformation si ce sont

des variables quantitatives), on peut alors construire la distribution des proportions par case pour les tableaux (X^*, Y^*, Z^*) au moyen du modèle log-linéaire suivant :

$$\log P_{ijk} = U + U_{1i} + U_{2j} + U_{3k} + U_{12ij} + U_{13ik} + U_{23jk} + U_{123ijk} \qquad (4.1)$$

Où P_{ijk} - la proportion relative à la case *(i, j, k)*.

Les numéros d'indicatrices *1, 2 et 3* désignent respectivement X^*, Y^*, Z^* . Evidemment, les fichiers *A et B* ne contiennent pas d'information sur les effets à deux facteurs U_{23} et les effets a trois facteurs U_{123}. Si ces effets sont posés égaux à zéro, cela revient à poser l'hypothèse de l'indépendance conditionnelle au sens qualitatif, $Y^* \perp Z^* / X^*$. Cependant, si nous disposons d'informations supplémentaires dans le fichier *C*, nous pouvons nous passer de cette hypothese parce qu'il est possible d'estimer les paramètres U_{23} et U_{123} à l'aide du fichier *C*.
L'information supplémentaire contenue dans le fichier *C*, qu'elle porte sur *(Y, Z)* ou *(X, Y, Z)* sert tout d'abord à définir des contraintes nominales sous la forme d'une distribution (X^*, Y^*, Z^*). Cette distribution pourra être obtenue par la technique du "balayage".

Supposons que l'information supplémentaire du fichier *C* prenne la forme d'une distribution de variables nominales, comme la distribution (X^*, Y^*, Z^*). La méthode du "balayage" correspond à faire un redressement sur l'effectif

Le principe est la suivante : par un exemple, nous avons besoin d'un tableau (X^*, Z^*) de *B* de manière à ce que ses fréquences marginales X^* concordent avec celles du fichier *A* (X^*, Y^*).

fichier *A*

	$y < 0$	$y >= 0$
$x < 0$	3	3
$x >= 0$	4	5

Tab.1

fichier *B*

	$z < 0$	$z >= 0$
$x < 0$	3	1
$x >= 0$	4	4

Tab.2

1. On va transformer les Tab.1, Tab.2, en tableau A' et B' respectivement en créant une variable "effectif" :

A'

X		Y		Effectif
$x < 0$	$x >= 0$	$y < 0$	$y >= 0$	
1		*1*		*3*
1			*1*	*3*
	1	*1*		*4*
	1		*1*	*5*

B'

X		Z		Effectif
$x < 0$	$x >= 0$	$z < 0$	$z >= 0$	
1		*1*		*3*
1			*1*	*1*
	1	*1*		*4*
	1		*1*	*4*

2. On fera un redressement sur la variable "effectif" *de* B' en prenant les fréquences marginales x^* de A' comme critère du redressement.

Le tableau de fréquence du fichier B' redressé est le tableau demandé : ses fréquences marginales X^* concordent avec celles du fichier A' (X^*, Y^*).

Pour le tableau de fréquences de trois dimensions C (x^*, y^*, z^*), le fichier correspondant prend la forme suivante :

C'

X		Y		Z		Effectif
$x<0$	$x>=0$	$y<0$	$y>=0$	$z<0$	$z>=0$	$n_{x_1y_1z_1}$
1		1		1		$n_{x_1y_1z_2}$
1		1			1	$n_{x_1y_2z_1}$
1			1	1		$n_{x_1y_2z_2}$
1			1		1	$n_{x_2y_1z_1}$
	1	1		1		$n_{x_2y_1z_1}$
	1	1			1	$n_{x_2y_1z_2}$
	1		1	1		$n_{x_2y_2z_1}$
	1		1		1	$n_{x_2y_2z_2}$

$n_{x_iy_jz_k}$ - fréquence relative à la case *(i, j, k)* du fichier C.

- si fichier C contient l'information sur (X Y, Z), les contraintes nominales peuvent être obtenues par les étapes suivantes :

1. $B(X^*,Z^*)$ est balayé de telle sorte que la fréquence sur x^* concorde avec celle du tableau (X^*,Y^*) de A'
2. $C(x^*,y^*,z^*)$ est balayé de telle sorte que le croisement (X^*,Y^*) concorde avec celui de $A'(X^*,Y^*)$ et celui de (X^*,Z^*) avec $B'(X^*,Z^*)$.

- si le fichier C contient l'information sur *(Y, Z)*, les contraintes nominales peuvent être obtenues par les étapes suivantes :

1. $B(X^*,Z^*)$ est balayé de telle sorte que la fréquence sur x^* concorde avec celle du tableau (X^*,Y^*) de A'
2. $C(x^*,y^*,z^*)$ est balayé de telle sorte que la fréquence y^* concorde avec celle du tableau de $A(X^*,Y^*)$ et X avec celui de $B'(X^*,Z^*)$,

3. prendre un tableau de trois dimensions (x^*, y^*, z^*) de fichier D avec n_{ijk}=1,

$D\,(x^*, y^*, z^*)$ est balayé selon trois critères $\begin{cases} (x^*, y^*) \text{ concord avec A } (x^*, y^*) \\ (x^*, z^*) \text{ concord avec B } (x^*, z^*) \\ (y^*, z^*) \text{ concord avec C } (y^*, z^*) \end{cases}$.

La méthode de régression avec les contraintes nominales :

La méthode est la même que la régression, mais des contraintes nominales sont appliquées. Ici, un ordre d'imputation est une nécessité lorsqu'on tire des valeurs authentiques de *Z* du fichier *B*. Une imputation est déterminée par sa distance minimale en *(X, Z)*, et par le fait que le nombre d'appariements réalisés dans cette classe ne dépasse pas le chiffre présenté par les contraintes nominales : la concordance est ainsi reconnue. Autrement, on rejette cette imputation et on examine le cas d'appariement qui vient au deuxième rang concernant la distance minimale. Le processus se poursuit jusqu'à ce que le fichier *A* soit complet. Les méthodes du type "hot-deck" avec les contraintes nominales ont des traitements d'imputations similaires à la méthode par la régression.

En conclusion, les imputations ont les avantages et inconvénients suivants :

- l'imputation par méthode de régression est plus satisfaisante au niveau individuel,
- les méthodes du type "hot-deck" sont plus satisfaisantes au niveau global,
- s'il n'existe pas d'informations supplémentaires, la méthode "hot-deck" est recommandée,
- s'il existe des informations supplémentaires :
    - la méthode régression est plus intéressante au niveau individuel,
    - la méthode "hot-deck" avec les contraintes nominales est plus intéressante au niveau global.
- s'il existe des informations substitutionnelles, alors la méthode "hot-deck" avec les contraintes nominales est recommandée.

CHAPITRE VI. : PROPOSITIONS NOUVELLES

6.1. L'ANALYSE HOMOGENE EN FUSION DE FICHIERS

On peut considérer le problème de la fusion de fichiers comme un cas particulier des données manquantes, lorsqu'on joint le fichier receveur à celui de donneur. C'est un tableau de données où il existe des données manquantes pour certaines variables : il s'agit de données manquantes sur des *variables à transférer*. Pour estimer des données manquantes, on n'est pas obligé d'imputer le bloc tout entier des données d'un individu. Comme ce qui a été étudié dans la première partie, l'analyse homogène permet d'estimer et d'imputer des données manquantes telles que l'on peut obtenir un ensemble de données les plus homogènes possibles. Pour chaque individu, on calcule un score qui est une mesure de ressemblance par rapport aux autres et pour chaque catégorie de variables, on les quantifie de la même façon. Cette méthode ne copie pas systématiquement tout le bloc entier des données d'un individu.

La méthode que nous essayons d'introduire a une approche tout à fait différente des méthodes classiques : elle se base sur un critère d'optimisation (maximiser l'homogénéité des données). Une homogénéité simultanée sur toutes les variables est exprimée par la variable de score. Cette variable est égale à la moyenne de toutes les variables transformées (quantifiées). Mathématiquement, nous cherchons l'imputation la plus homogène pour toutes les variables. Elle est mathématiquement robuste, adaptée à des données très générales sans contrainte de modèle. Elle peut accepter des différences de structure de population et de taille de fichiers donneur et receveur. De plus, le fait que cette méthode quantifie de façon optimale des catégories de variables nous permet de mesurer plus facilement le niveau homogène entre des variables et donc la prédictivité des variables communes aux *variables à transférer*. Cette prédictivité est une condition nécessaire à la fusion. Puisque le résultat est très stable par rapport au critère de l'homogénéité des données, la qualité du résultat peut être mesurée facilement par le niveau d'homogénéité des données. La validation des données reconstituées est ainsi simplifiée.

Nous introduisons la fusion de fichiers basée sur l'analyse homogène à la seule condition où le pouvoir prédictif des variables communes par rapport aux *variables à transférer* est assez fort : le niveau d'intercorrélation des données après quantification doit être assez élevé. C'est-à-dire que nous possédons des variables suffisamment prédictives par rapport aux *variables à transférer*. Par ailleurs, c'est une condition nécessaire à la fusion. Il est préférable de procéder à une sélection des variables critiques (prédictives) avant d'appliquer la fusion Une expérience de Santini G. (1986. 2) montre que, dans le cadre de fusion par imputation, la qualité du résultat s'accorde à un nombre compromis, un nombre ni trop grand et ni trop petit, de-variables critiques.

Puisque nous avons détaillé la méthode de l'analyse homogène dans la section 2.3.3 dans la première partie "traitement des données manquantes" et que nous n'avons pas change le principe de la méthode, nous ne soulignons ici que quelques termes différents :

la fonction de perte d'homogénéité dans la section 2.3.3.2. peut s'exprimer comme suit :

$$\sigma(x; y_1, \cdots, y_m, g_1^*, \cdots, g_m^*) = \sum_{j \in \Omega} (X - g_j y_j)^2 + \sum_{j \notin \Omega} (X - g_j^* y_j)^2 \qquad (1)$$

Ω représente l'ensemble des variables ayant les réponses complètes : les variables communes prédictives.

g_j^* : matrice indicatrice incomplète de la *$j^{ème}$ variable à transférer* .

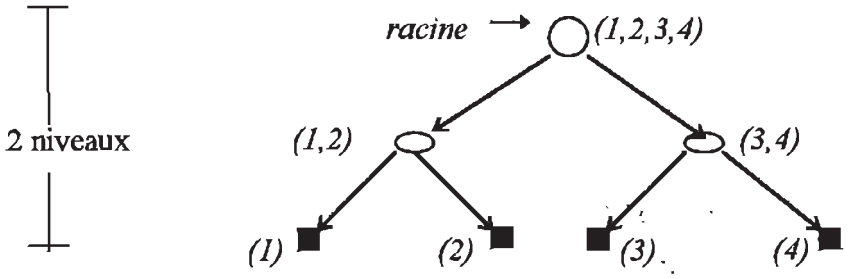
Les valeurs d'imputation les plus homogènes correspondent à la minimisation de σ sur X, y_1 , y_2 , ..., y_m *et* $g_1^*, g_2^*, \cdots, g_m^*$ *de (1).*

Dans le traitement des données manquantes, nous mesurons l'homogénéité sur l'ensemble des données. Dans la fusion, c'est plutôt la relation entre deux blocs de variables, le bloc des variables à transférer et le bloc des variables prédictives, qui est prise en compte dans la simulation des fichiers des données pour évaluer le résultat de la fusion.

6.1.1. Utilisation de la segmentation dichotomique pour les variables ayant un grand nombre de modalités

Pour des variables ayant un grand nombre de catégories, on propose une technique de 'segmentation dichotomique' qui regroupe les modalités des variables en deux groupes. On transforme la fusion de variables ayant un grand nombre de modalités en une série de fusions de variables à deux modalités.

Le principe de la transformation hiérarchique des variables à modalités nombreuses en variables à deux modalités est similaire à une procédure de segmentation dichotomique d'une variable : cette procédure s'effectue comme si on partait de la racine d'un arbre où nous aurions toutes les modalités d'une variable. On engendre chaque fois deux branches quand on descend dans l'arbre, c'est-à-dire que l'on sépare les modalités courantes en deux grandes parties homogènes dans lesquelles les modalités (quantifications) sont proches l'une de l'autre. Ceci se termine lorsqu'on n'aura plus de modalités à séparer et on aboutit finalement aux feuilles de l'arbre. Lorsqu'on sépare les modalités d'une variable en deux parties homogènes, on sépare en même temps l'ensemble des individus en deux sous-ensembles. Pour une variable à m modalités, nous aurons un arbre à n niveaux, si $n\text{-}1<\log_2^m<n$. Par exemple, pour des variables ayant 4 modalités, on pourra faire une segmentation dichotomique de la façon suivante si les quantifications des modalités 1 et 2 (3 et 4 respectivement) sont les plus proches :



Chaque noeud et racine d'arbre nécessitent une fusion sur l'ensemble des individus concernés. Le niveau d'arbre correspond au nombre de fusions consécutives. L'arbre ci-dessus a une racine et deux noeuds différents, nous avons donc au total trois fusions à effectuer sur des variables à transférer et à deux modalités.

Exemple - Considérons les données d'une enquête sur l'activité professionnelle. Elles représentent 315 individus : les 260 premiers individus constituent le fichier donneur et les 55 individus restants le fichier receveur. Les cinq variables suivantes sont prises dans la fusion.

Les trois variables communes sont :

Q1 - le sexe de l'enquêté(e),

Q2 - la situation actuelle de la personne interrogée (7 catégories),

Q3 - des conflits travail - vie personnelle (3 catégories).

Les deux variables à transférer sont :

Q4 - être au chômage ces douze derniers mois (3 catégories),

Q5 - exercer en ce moment une activité professionnelle (4 catégories).

L'application directe de la fusion donne un taux de *64%* de données bien classées. Les quantifications de catégories 1, 2, 3 et 4 de la variable Q5 par l'analyse homogène sur le fichier donneur sont : -0.92, -0.81, 1.08 et 1.16. On remarque que les catégories (1,2) et (3,4) de la variables Q5 sont relativement proches. A l'issue de la première fusion, on sépare l'ensemble des receveurs en deux groupes : les individus ayant les valeurs Q5 (1,2) dans le premier groupe et le reste des individus dans l'autre groupe. On sépare l'ensemble des donneurs en deux groupes : les individus ayant répondu Q5 (1,2) dans le premier groupe et le reste des individus dans l'autre groupe. On fait deux fusions respectivement sur deux groupes des individus en utilisant les variables communes Q1-Q4 A l'issue du résultat de la deuxième fusion, nous obtenons un taux de *74.5%* de données bien classées. Au niveau global, les marges simples des variables après la deuxième fusion (dichotomique) sont également mieux respectées, les marges simples des variables Q5 sont les suivantes :

Q5	*1*	*2*	*3*	*4*
Réelle	*26*	*3*	*21*	*5*
Fusion directe	*17*	*12*	*5*	*21*
Fusion dichotomique	*19*	*10*	*16*	*10*

Malgré l'intérêt que présente la technique, cependant on aimerait souligner que lorsque le nombre de variables à transférer et le nombre de modalités de variables augmentent, la méthode devient assez pénible à cause des nombreuses fusions à exécuter.

6.1.2. Exemples d'analyse homogène appliquée à la fusion de fichiers

Exemple 1 - le fichier donneur contient 10 individus sur 3 variables (INCOME, AGE, CAR) renseignées (même jeu de données au 2.3.4.) ; le fichier receveur contient *6* individus ayant seulement les variables INCOME et AGE renseignées. Dans le fichier receveur, la population de l'âge 'middle' est sur-représentée ; c'est à dire les structures de population de deux fichiers ne sont pas la même. Nous nous servons des deux variables du fichier receveur et des trois variables du fichier donneur pour reconstituer les valeurs de la variable CAR du fichier receveur.

OBS	INCOME	AGE	CAR	
1	middle	middle	am	fichier donneur
2	high	old	am	
3	low	young	jpn	
4	middle	young	am	
5	high	old	am	
6	low	young	jpn	
7	high	middle	am	
8	high	old	am	
9	low	young	am	
10	low	young	jpn	
11	high	old	am	partie reconstituée
12	middle	middle	am	
13	low	young	jpn	fichier receveur
14	high	middle	am	
15	middle	middle	am	
16	high	old	am	

Le coefficient d'homogénéité est de 0.86. En comparant les résultats obtenus avec les données du fichier donneur, nous constatons que les résultats sont raisonnables.

Exemple 2 : Considérons les données d'une étude sur le rendement de vendeurs par rapport à leurs capacités intellectuelles. Elles ont été recueillies par Johnson et Wichern en 1982, elles ont des valeurs quantitatives que l'on recode en valeur ordinale, se trouvent dans *l'Annexe B III.* Elle contient *50* observations sur les *7* variables suivantes :

$$X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} croissance\ des\ ventes \\ profit\ \ des\ ventes \\ ventes\ relatives\ aux\ nouveaux\ clients \end{pmatrix},$$

$$Y = \begin{pmatrix} x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \begin{pmatrix} note\ d'esprit\ de\ créativité \\ note\ d'habileté\ mécanique \\ note\ d'esprit\ d'abstraction \\ note\ d'esprit\ mathématique \end{pmatrix}.$$

Considérons le résultat des ventes X comme les variables à transférer et le profil du vendeur Y comme les variables prédictives. Suivant le principe de la validation croisée, nous faisons quatre fois la fusion pour 4 fichiers receveurs tirés au hasard du fichier complet. Nous obtenons un taux moyen de 71% de données bien classées avec un écart-type égal à 8.9 et un coefficient moyen de proximité de 0.16 avec un écart-type égal à 0.039. Le taux des données bien classées du tirage aléatoire est théoriquement de 34.6% et le coefficient moyen de proximité égal à 0.426. Les assez bons résultats individuels de la fusion par l'analyse homogène confirment l'optimalité de la méthode pour une valeur relativement élevée du coefficient d'homogénéité, qui est égale à *0.625.* Par exemple, nous prenons les douze dernières observations comme receveurs. Les valeurs enlevées sur les variables X sont les suivantes :

observations	*x1*	*x2*	*x3*
39	*3*	*3*	*3*
40	*3*	*3*	*3*
41	*2*	*2*	*2*
42	*1*	*2*	*1*
43	*2*	*3*	*3*
44	*1*	*1*	*1*
45	*1*	*2*	*1*
46	*2*	*3*	*3*
47	*1*	*1*	*1*
48	*1*	*1*	*1*
49	*2*	*2*	*3*
50	*3*	*3*	*2*

Les estimations obtenues par la fusion sont les suivantes :

observations	x1	x2	x3
39	3	3	3
40	3	3	3
41	2	2	2
42	2	2	2
43	3	3	3
44	1	1	1
45	2	2	2
46	3	3	3
47	1	1	1
48	1	1	1
49	3	3	3
50	3	3	3

Le tableau de fréquence correspondant pour les valeurs réelles cachées est :

	x1	x2	x3
1	5	3	5
2	4	4	2
3	3	5	5

Le tableau de fréquence correspondant après la fusion est :

	x1	x2	x3
1	3	3	3
2	3	3	3
3	6	6	6

Ici, le coefficient de proximité est de *0.1528* et le taux des données bien classées est egal a 75%.

Exemple 3 - Enfin, nous travaillons sur un grand fichier de données réelles "enquête 1000" (SPAD). Il contient 992 individus sur 7 variables suivantes :

Les 4 variables communes X de deux fichiers sont :

Q1 - l'âge de l'enquêté(e) en 5 tranches,

Q2 - la taille d'agglomération (en nombre d'habitants) en 5 modalités,

Q3 - l'heure de coucher en 7 tranches,

Q4 - l'âge de fin d'étude en 5 tranches.

Les 3 variables à reconstituer Y sont :

Q5 - la famille est le seul endroit où l'on se sente bien ? (O,N),

Q6 - le diplôme d'enseignement général le plus élevé obtenu (en 7 tranches),

Q7 - regardez-vous la télévision ? (en 4 modalités de fréquence).

192 individus tirés au hasard dans le fichier complet constituent le fichier receveur et les 800 individus restants le fichier donneur. Nous fusionnons cinq fois pour cinq fichiers receveurs. Le coefficient d'homogénéité est de 0 432. Nous obtenons un taux moyen de 48.6% de données bien classées avec un écart-type égal à 3.68. Le taux des données bien classées du tirage aléatoire est théoriquement de 33%.

Prenons les 800 premiers individus comme le fichier donneur et les 192 individus restants comme fichier receveur dont les valeurs réelles des variables à transférer sont cachées Nous faisons une comparaison sur deux méthodes : la fusion par l'analyse homogène et la méthode de fusion STATIRO dont le principe est étudié dans 5.2.2. Les résultats (*cf. Annexe B IV*) montrent bien les caractéristiques de deux familles de méthodes :

- Niveau individuel : le résultat obtenu par l'analyse homogène a un bon niveau individuel, le taux de bien classé est de 54%, celui de la méthode STATIRO est de 47%

- Niveau global : le résultat obtenu par la méthode STATIRO a un bon niveau global, selon les tableaux de fréquences croisées, qui est meilleur que la fusion par l'analyse homogène. La méthode STATIRO copie le bloc entier des données d'un donneur, cela permet de respecter le lien naturel entre variables à transférer et éviter ainsi l'incohérence des réponses entre les différentes variables La fusion par l'analyse homogène estime les valeurs manquantes à la manière d'un modèle de régression : c'est-à-dire que pour un X donné, on a une valeur Y par prédiction, la valeur plus probable (homogène) par le modèle. La prédiction individuelle est donc bonne et utile. Par contre, la méthode STATIRO impute pour différents individus ayant la même valeur de X différentes valeurs de Y (par exemple pour X=3421, les valeurs Y sont : 232,121,123,122,114,212). Au niveau individuel, on ne sait pas laquelle est la plus probable mais au niveau global cela permet de refléter la variation de réponses de l'échantillon.

Notons que pour la fusion par l'analyse homogène, nous utilisons la technique de segmentation dichotomique pour les variables Q6 et Q7. Les marges réelles et reconstituées selon les deux méthodes sur des variables Q5, Q6 et Q7 s'illustrent dans les tableaux suivants :

Q5	*Marges réelles*	*Analyse homogène*	*STATIRO*
1	*136*	*136*	*125*
2	*56*	*56*	*67*

Q6	*Marges réelles*	*Analyse homogène*	*STATIRO*
1	36	6	49
2	70	114	65
3	35	16	27
4	29	23	33
5	4	33	1
6	18	33	15
7	0	0	2

Q7	*Marges réelles*	*Analyse homogène*	*STATIRO*
1	100	118	100
2	36	18	43
3	37	29	31
4	19	27	18

Notre conclusion sur cette comparaison est donc la suivante : selon l'objectif de la fusion on choisit la méthode adéquate : la fusion par l'analyse homogène ou celle par la recherche de voisinage pour les différents avantages.

Exemple 4 - Pour mesurer l'efficacité au niveau individuel de la méthode, on va travailler sur des données simulées. Des données ordinales sont simulées selon les deux critères suivants :

1. Moyenne des coefficients de corrélation entre variables : quatre niveaux entre 0 et 1.
2. Structure de population entre les fichiers donneur et receveur (identique ou différente).

La combinaison des facteurs nous donne 8 cas différents. Pour chaque combinaison, on simule 50 jeux de données de 130 individus avec 8 variables à 5 modalités pour le fichier originel, 100 individus pour le fichier donneur, 30 individus pour le fichier receveur, 5 variables communes prédictives et 3 variables à transférer.

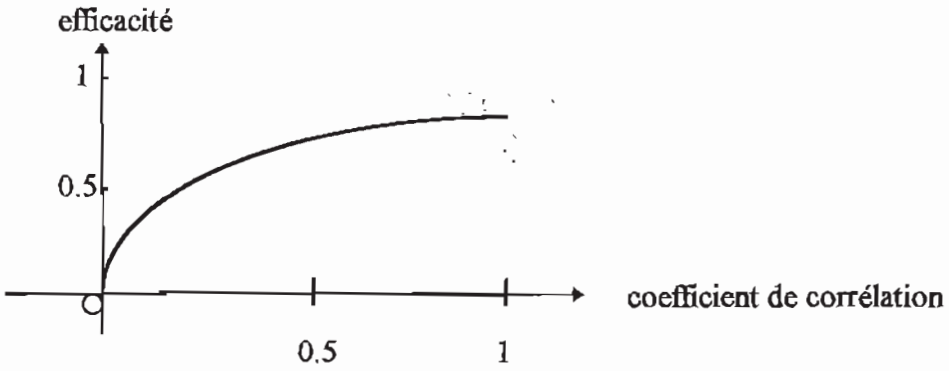
Nous utilisons le critère d'évaluation individuelle, c'est-à-dire le coefficient de proximité, pour mesurer le résultat de la méthode en comparant avec celui du tirage aléatoire présenté au chapitre 4.4. pour évaluer les résultats de la méthode. Le résultat obtenu ne varie pas sur le deuxième critère. En effet, nous avons obtenu le même résultat, que les deux fichiers aient ou non la même structure.

Valeur moyenne du coefficient de proximité sur les variables transférées

Coefficient de corrélation	0.001	0.269	0.555	0.840
Analyse homogène	0.379	0.327	0.243	0.124
Tirage aléatoire	0 275	0.370	0.290	0.357

Si nous possédons des données satisfaisant aux conditions nécessaires pour faire la fusion, mentionnées dans la section 4.3, la méthode basée sur l'analyse homogène est faite pour s'adapter aux différences de taille de fichiers. Car, dans le programme on ajoute chaque observation du fichier receveur au fichier donneur ; c'est sur cet ensemble de données qu'on reconstitue les valeurs manquantes de chaque receveur. On obtient donc des résultats de fusion indépendants de la taille des deux fichiers.

La qualité du résultat varie en fonction de la relation entre les deux blocs ; la relation entre la qualité du résultat et le coefficient de corrélation moyen peut être décrite comme suit :



Plus le coefficient de corrélation entre deux blocs est grand (c'est-à-dire que le coefficient d'homogénéité est grand aussi), meilleur est le résultat. Nous pouvons ainsi déterminer si les données reconstituées provenant de la fusion sont valables au niveau individuel. En comparant l'analyse homogène avec le tirage aléatoire, nous observons que lorsque la corrélation moyenne entre deux blocs est quasi nulle (proche à la valeur critique d'indépendance), le tirage aléatoire est meilleur. En général lorsque la corrélation moyenne entre deux blocs est supérieure à 0.3, la méthode "analyse homogène" est supérieure Notre méthode fournit un résultat remarquable au niveau individuel : au cas où elle ne trouve pas la bonne réponse, elle trouve souvent une réponse très proche de la réalité.

Les conditions d'application :

Dans la pratique, on peut valider automatiquement des résultats de fusion par le coefficient d'homogénéité. Lorsque le coefficient d'homogénéité est supérieur à 0.40, les données fusionnées atteignent un bon niveau individuel. Donc, l'usage de la fusion pour le niveau individuel est recommandable. Van Buuren S. & Van Rijckevorsel J L.A. (1992) montrent, au moyen d'un exemple, que l'imputation multiple par analyse homogène est possible et que le résultat obtenu est très intéressant au niveau global.

Dans le cadre d'un grand fichier où les individus ne sont pas très homogènes, nous proposons un processus de **classification** pour les répartir en plusieurs groupes homogènes avant d'appliquer notre méthode de fusion. Ceci afin d'augmenter la performance du résultat de la fusion. Ce processus se décompose en trois étapes suivantes :

1. On fait une classification sur les coordonnées factorielles basées sur des *variables à transférer* y de l'ensemble d'individus du fichier donneur. Chaque individu est donc associé à une classe n.

2. L'étape 2 peut se faire de deux manières différentes :
   - soit on utilise la technique de segmentation sur (x, n) du fichier donneur pour trouver les variables prédictives x (parmi des variables communes x) de n qui peuvent être utilisées comme un critère de classification a dans l'étape 3,

- soit on applique une analyse factorielle discriminante (x et n) à l'ensemble des individus du fichier donneur :
    - recherche du critère de classification noté a, c'est-à-dire les variables prédictives pour le numéro de classe n,
    - projeter les individus (sur x) du fichier receveur pour obtenir une classification.

3. Enfin, on utilise le critère a pour classifier les individus (sur x) du fichier receveur.

Ce processus nous permet d'avoir des classes assez homogènes dans lesquelles on effectue la fusion.

D'ailleurs, Saporta G. a suggéré d'ajouter des contraintes de cohérence pour vérifier la cohérence d'estimation des valeurs entre variables à transférer dans la méthode de fusion (du type non imputation par bloc). Par exemple, on rejette l'estimation 'cancer du sein pour un homme', etc. C'est évident que ce genre du problème ne se pose pas dans la fusion par l'imputation par bloc.

En conclusion, la méthode de fusion basée sur l'analyse homogène a les avantages suivants :

- une bonne qualité individuelle. En même temps, elle tient compte du lien entre variables,
- la stabilité du résultat,
- elle s'adapte à différentes structures de populations des fichiers,
- elle s'adapte à des différences de taille des fichiers.

Grâce aux travaux de Van Buuren S. & Van Rijckevorsel J.L.A. le programme de fusion par l'analyse homogène a pu être conçu. Il est fait pour l'usage de fichiers de données dont les tailles sont quelconques et sans limites. Il est programmé en langages SAS IML où les données sont présentées en mémoire sous forme matricielle Lorsque les tailles des fichiers sont grandes, c'est-à-dire que le nombre des variables et celui des observations est important, une grande place en terme de mémoire doit être prévue pour faciliter les calculs. Les temps de calculs dépendent de la grandeur des fichiers des données. Par exemple, sur un micro-ordinateur compatible PC 486-66, avec un fichier donneur contenant 800 individus sur 7 variables à 5 modalités et le fichier receveur 200 individus sur 5 variables à 5 modalités, le temps de calcul est d'environ 10 minutes. Un document concernant le programme se trouve dans l'annexe A III.

6.2 OPTIMISATION GLOBALE PAR APPLICATION DE CONTRAINTES NOMINALES DANS LA FUSION DE FICHIERS PAR IMPUTATION

Nous proposons une optimisation globale au cas où on se trouve dans la situation suivante :

- les tailles des fichiers donneur et receveur ne sont pas les mêmes, il y a parfois un grand écart entre les deux.
- les structures de données des fichiers donneur et receveur ne sont pas similaires.

Par exemple, si l'on se sert des méthodes d'imputation du type "hot-deck", d'imputation par cellule ou par la recherche de sosie et que l'on se retrouve dans la situation ci-dessus, on se demande combien de fois en moyenne un individu du fichier donneur va être copié et quels individus vont être copiés fréquemment. Il est évident que cela dépend du rapport des tailles des deux fichiers et des structures des deux fichiers. On cherche une référence raisonnable. Dans le fichier receveur, on doit avoir normalement les mêmes relations conjointes de *x, y* que dans le fichier donneur. C'est avec ces relations que l'on va effectuer la procédure de fusion. Lorsque les fichiers donneur et receveur sont de taille et de structure très différentes, la procédure de l'imputation nécessite un contrôle global afin de respecter la distribution de la population du fichier et la relation entre variables Pour ce faire le fichier donneur doit contenir obligatoirement des individus statistiquement représentatifs de chaque catégorie de population et, bien sûr, des informations correctes.

"Balayage" pour obtenir des contraintes nominales :

Nous voulons compléter les enregistrements du fichier receveur R en allant chercher les valeurs de Y dans le fichier donneur D sur la base d'informations contenues dans le fichier R *et* sur la base de relation conjointe entre X *et* Y dans le fichier donneur D. On se sert des informations de D pour définir des contraintes nominales. Ces contraintes ont pour objet de conserver des liens catégoriels (théoriquement estimés par des modèles log-linéaires) suivant une partition de *(X, Y)* acceptables dans le cas du fichier enrichi.

Posons (X^*, Y^*) comme les variables nominales correspondantes (ou après transformation), on peut alors présenter des proportions par case pour le tableau (X^*, Y^*) au moyen d'un modèle log-linéaire :

$$\log p_{ij} = U + U_{1i} + U_{2j} + U_{12ij}$$

p_{ij} – *la proportion relative à la case* (i, j)

S'il ne s'agissait pas de variables nominales dès le départ, on les transformerait et on peut obtenir le tableau de fréquences nécessaires, les contraintes nominales, en servant de la méthode "balayage". C'est-à-dire faire coïncider la fréquence X^* *du* (X^*,Y^*) obtenu à partir du fichier donneur avec celle du fichier receveur.

Dans la pratique, les variables X *et* Y sont en général des variables multidimensionnelles. Ici on présente un algorithme afin de passer du multidimensionnel à l'unidimensionnel et de pouvoir appliquer les contraintes nominales. Puisqu'il existe des partitions des individus en classes homogènes dans l'ensemble de données, alors une classe peut être choisie pour représenter un ensemble de variables d'un individu dans cette classe. Donc on peut prendre le numéro de sa classe pour remplacer $X_1(x_1,\cdots,x_p)$.

On voudrait classifier les variables X_1, variable prédictives, et Y, variables à transférer, de fichier donneur et remplacer par leur numéro de classe.

- l'étape 1:
  - faire une classification sur X_1 de fichier D :
    X_1^* - numéro de sa classe d'appartenance.
  - faire également une classification sur Y de fichier D :
    Y^* - numéro de sa classe d'appartenance.
    - nous avons un tableau de contingence (X^*,Y^*) de fichier D.
- l'étape 2
  - faire une classification sur X de fichier R selon le critère de classification de X_1 pour le fichier D, donc on obtient le tableau de fréquence X_2^* du fichier R.
  - on balaye (X^*,Y^*) de fichier D de la même façon que celle à la section 5.3 pour que X_1^* concorde avec X^*, celui de fichier R, et on obtient le tableau de fréquence $(X^*,Y^*)'$ noté C.

Les contraintes nominales sont une référence à suivre pour l'imputation : lorsque l'on tire des valeurs authentiques de Y du fichier D. On note la classe (X^*,Y^*) de l'enregistrement à compléter et si le nombre d'effectifs dans cette classe ne dépasse pas le chiffre présenté par des

contraintes nominales C, la concordance est retenue. Autrement, on rejette cette imputation et on examine le cas d'une imputation qui vient en deuxième rang en ce qui concerne la distance minimale. Le processus se poursuit jusqu'à ce que le fichier R soit complet.

En conclusion, l'application de contraintes nominales C nous fixe un seuil de nombres de copies raisonnable pour un groupe (ou un individu) du fichier donneur. Il est utile dans le cas où les fichiers donneur et receveur sont de tailles et de structures très différentes, que nous ayons une référence à suivre. Ces contraintes concernent deux blocs de variables *(X, Y)*, donc nous garderons la relation entre deux blocs de variables *X et Y* au sens de lien catégoriel.

CONCLUSION

Cette partie de la thèse avait pour objet de mettre au point une nouvelle méthode de fusion des fichiers. Les travaux qu'elle contient, ont permis de justifier l'utilisation de l'analyse homogène dans le contexte de la fusion. On y montre que la méthode a de bonnes qualités au niveau individuel et moyennes au niveau global.

Une proposition sur l'utilisation des contraintes globales dans le cadre de fusion par imputation en bloc est faite : elle a pour objet de fournir une contrainte nominale à suivre sur le nombre de copies à effectuer par un groupe de donneurs. Nous avons également présenté une évaluation individuelle pour des variables ordinales.

Le sujet sur la fusion de fichiers est riche car il peut se poursuivre dans plusieurs directions :

- des comparaisons avec d'autres méthodes de fusion restent à faire pour trouver les caractéristiques de chacune,
- les conditions nécessaires concernant les données pour effectuer la fusion restent à préciser et à standardiser.

ANNEXE AI : PROGRAMME DE ACP POUR DES DONNEES INCOMPLETES

/* tacpdm.sas */

```
%macro init1(data=_last_,out=fout,m=15,alfa=1E-5),

proc iml;
                                    /* ACP */
start corr;
  sum=hstd[+,];                     /* somme de colonne          */
  xpx=t(hstd)*hstd-t(sum)*sum/n;    /* n - nbr d'individus       */
  s=diag(1/sqrt(vecdiag(xpx)));
  corr=s*xpx*s;                     /* matrice de correlation    */
finish corr;


start std;                          /* n - nbr.d'individus       */
  mean=h[+,]/n;                     /* moyenne de colonne        */
  hstd=h-repeat(mean,n,1);          /* centrer h                 */
  ss=hstd[##,];                     /* somme de carre / colonne  */
  std=sqrt(ss/(n-1));               /* écart-type                */
  hstd=hstd*diag(1/std);            /* données centrées reduites */
finish std;

start impute;
   run std;
   run corr;
   call eigen(eval,evect,corr);
   freq=j(n,1,1),                   /* fréquence de repétition   */
   do i=1 to n;
    fi=0;
    f1=hstd[i,];
    if (dmid[i]=0) then do;
      do  i1=1 to n;
          f2=hstd[i1,];
          if all(f1=f2) then fi=fi+1;
      end;
      freq[i,1]=fi,
    end;
   end;
   valp=0;j=0;
   do while (valp/p <0.75);
    j=j+1;
    valp=valp + eval[j,1];
   end;
   nbv=j;                           /* nb.d'axes retenus                  */
  indm=0;
  indm=j(n,1,1);                    /* indice d'unité à imputer pour d.m. */
  do i=1 to n ;
    if (dmid[i]^=0) then do;        /* imputation des D.M                 */
        dmin=-1;                    /* initialisation                     */
        ci=j(nbv,1,1);
        ahstd=choose(a=0,0,hstd); /* ignore les d.m.                      */
        do j=1 to nbv ;
           ci[j,]=ahstd[i,]*evect[,j]; /* coordonnee sur les axes         */
```

```
        end;
        d=0;
        do i1=1 to n;
           if (dmid[i1]=0 ) then do;
                 aci=j(nbv,1,1);
                 pahstd=ahstd;
                 do j=1 to p;
                       if (ahstd[i,j]=0) then pahstd[i1,j]=0;
                 end;
                 do j=1 to nbv ;
                       aci[j,]=pahstd[i1,]*evect[,j];
                 end;
                 cci=aci-ci;
                 do l=1 to nbv;
                       cci[l,1]=cci[l,1]*eval[l,1]; /* poids d'axes: eval                    */
                 end;
                 d=cci[##,];                        /* distance entre 2 unités en ignorant d.m  */
                 if dmin=-1 then do,
                       dmin=d,
                       indm[i,1]=i1;
                 end;
                 if (d=dmin & dmin ^=-1) then do,   /* choix selon le prob.                  */
                       l=indm[i,1];
                       if freq[i1,1]>freq[l,1] then  indm[i,1]=i1,
                 end;
                 if d<dmin then do,
                       dmin=d;
                       indm[i,1]=i1,                /* indice la + pte distance  */
                 end;
           end;
        end;
        do j=1 to p ;
          if a[i,j]=0 then do;
             di=indm[i,1];
             h[i,j]=h[di,j];
          end;                                      /* imputation              */
        end;
     end,
   end;
finish impute;

start do_it;
   h=0; dmm=0;  p=0;  n=0;
   dmid=0; a=0; mvect=0; corr=0; hstd=0;
   valp=0; nbi=0;                   /* initialisation                                  */
   use &data;
   free h dmm dmid a  corr hstd ;
   read all var _num_ into h[colname=vars];
   doni=h;
                                    /* sas data ---> en matrice                        */
   dmm=choose(h=.,-1,0);            /* matrice DmM:0--presence  -1--DM                 */
   dmid=dmm[,+];
   dmid=choose(dmid=0,0,-1);        /* dmid(1:n),0-présence,-1--d.M  (unite)           */
   a=choose(dmm=-1,0,1);            /* a : mat.de projection. 0--d.m, 1--présence      */
   p=ncol(h) ;  n=nrow(h);
   do j=1 to p ;
      do i=1 to n ;
        if dmm[i,j]=-1 then do;
           h[i,j]=h[,j][:] ;        /* impution par la moyenne de la variable */
        end;
```

```
    end,
  end;
  run impute;
  valp=eval;
  ba=&alfa+1;
  do while(abs(ba)>&alfa & nbi<&m ) ;    /* phase de controle d'arrêt */
     run impute ;
     nbi=nbi+1;
     bam=t(eval-valp)*(eval-valp);
     ba=sum(bam),                        /* l'ecart de valeur propre  */
     valp=eval;
  end;
  create &out from h[colname=vars];
  append from h;
  close &out;
finish do_it;
                                         /*  matrice ---> sas data    */

run do_it;
quit;
%mend init1;
```

ANNEXE AII : PROGRAMME DE CLASSIFICATION AUTOMATIQUE POUR DES DONNEES INCOMPLETES

/* tfcla.sas */

```
%macro initc(data=_last_,out=fout, class=no ,k=3, mi1=5, mi2=8, alfa=1E-3);
 proc iml;
                                  /* prog. pour comparer avec sas    */
                                  /* classification automatique  avec taux de D.M <10% */
 start std;
  mean=h[+,]/n;
  hstd=h-repeat(mean,n,1);
  ss=hstd[##,];
  std=sqrt(ss/(n-1));
  hstd=h*diag(1/std);
 finish std;                      /* variables reduites              */

  start inig;                     /* centres de gravité initiales : k premiers unités */
   i=1;  j=1;  g=j(&k,p,1);
   do while (i<=&k);
     if(dmid[j]=0) then do;       /* l'unité complete                */
          g[i,]=hstd[j,];
          i=i+1;
     end;
     j=j+1;
   end;
 finish inig;

 start aff;                       /* affectation de classe           */
  ing=j(n,&k,0);                  /* indice des membres de classe    */
  ning=j(1,&k,0);
  do i=1 to n;
    indm=0;
    dmin=-1;
    ci=j(p,1,1);
     do i1=1 to &k;
       ci=t(hstd[i,]-g[i1,]);     /* projection à en ignorant d.m    */
       do j=1 to p;
          if (a[i,j]=.) then  ci[j,1]=0,
       end;
       d=ci[##,];
       if dmin=-1 then do;
            dmin=d;
            indm=i1;
        end;
       if d<dmin then do;
             dmin=d;
             indm=i1;             /* indice la + pte distance        */
       end;
     end;
    ing[i,indm]=1 ;               /*ing:1---présence,0---sinon      */
  end;
  ning=ing[+,];
 finish aff;

 start cent;                      /* calcul du centre de gravité     */

    do j=1 to &k;                 /* pour chaque classe              */
```

```
        ax=j(p,1,0);
        aa=j(p,p,0);
       do i=1 to n ;
        ai=i(p),
        if (ing[i,j]=1) then do;
             do l=1 to  p ;
              if (a[ı,l]=. ) then ai[l,l]=0;
             end;
             ax=ax+ai*t(hstd[i,]);
             aa=aa+ai,
         end;
       end;
      do l=1 to p,
         if (aa[l,l]=0)  then  goto fin;
       end;
       g[j,]=t(inv(aa)*ax);
     end;
fin:  i=i+1     ;
finish cent;

start inert;                          /*    calcul de l'inertie totale        */
   dt=0;
   do i=1 to &k;                      /* pour chaque centre de gravité    */
     dc=0;  ci=j(1,p,1);
     do i1=1 to n;
       di1=0;
       if(ing[i1,i]=1) then do;
            ci=hstd[i1,]-g[i,];
            do j=1 to p;
              if (a[ı,j]=.) then  ci[1,j]=0;
            end;
            di1=ci[,##];
       end;
       dc=dc+di1;
      end;
      dt=dt+dc;
   end,
finish inert;

start phase1;
 run std;
 run inig;
 nbi=0;  ba=&alfa+1;  dta=0;   dt=0;
      run aff;
 do  while (nbi<&mi1 &  ba>&alfa & all(ning >=1) ),

      run cent;
      run ınert;
      nbi=nbi+1;
      if nbi=1  then ba=&alfa+1;
      else   ba= abs(dta-dt)/dt,
      dta=dt;
      run aff;
 end;

finish phase1;

 start impute;
  run std;
  indm=0;   ci=j(p,1,1);   aci=j(p,1,1);
```

```
 indm=j(n,1,0);                    /* indice d'unité à imputer pour d.m.*/
do i=1 to n ;
  if (dmid[i]^=0) then do;         /* les d.m                          */
        dmin=-1;                   /* initialisation                   */
        ahstd=choose(a=0,0,hstd); /* ignore les d.m.                   */
        ci[,1]=t(ahstd[i,]);
     d=0;     pahstd=hstd;
     do i1=1 to n;
          aci[,1]=t(hstd[i1,]);
          aci=choose(ci=0,0,aci);
          cci=aci-ci,
          d=cci[##,];                    /* distance entre 2 unités en ignorant d.m */
        if dmin=-1  then do;
           dmin=d;
           indm[i,1]=i1;
         end;
         if d<dmin then do;
            dmin=d;
            indm[i,1]=i1;               /* indice de la + pte distance      */
         end;
      end;
      do j=1 to p ;
         if a[i,j]=0 then do;
            di=indm[i,1];
            h[i,j]=h[di,j];
         end;                          /* imputation                       */
      end;
  end;
 end;
finish impute;

start aff1;
  ing=j(n,&k,0);                       /* indice des membres de classe   */
  ning=j(1,&k,0);
  do i=1 to n,
    ndm=0;
    dmin=-1;
    ci=j(p,1,1);
     do i1=1 to &k;
       ci=t(hstd[i,]-g[i1,]);

       d=ci[##,];
       if dmin=-1 then do;
            dmin=d;
            ndm=i1;
        end;
       if d<dmin then do;
             dmin=d;
             ndm=i1;                   /* indice la + pte distance                */
       end;
ing[i,ndm]=1 ;                         /*ing:1---présence,0---sinon               */
  end;
  ning=ing[+,];
finish aff1;

 start cent1;                          /* calcul du centre de gravité              */
    do j=1 to &k;                      /* pour chaque classe                       */
       ax=j(p,1,0);
       do i=1 to n ;
        if (ing[i,j]=1) then do;
```

```
            ax=ax+t(hstd[i,]);
        end;
      end;
      g[j,]=t(ax/ning[1,j]);
    end;
finish cent1;

start inert1;                          /* calcul l'inertie totale                    */
  dt=0;
  do i=1 to &k;                        /* pour chaque centre de gravité              */
    dc=0;  ci=j(1,p,1);
    do i1=1 to n;
      di1=0;
      if(ing[i1,i]=1) then do;
          ci=hstd[i1,]-g[i,];
          di1=ci[,##];
      end;
      dc=dc+di1;
    end;
    dt=dt+dc;
  end;
finish inert1;


start phase2;
 nbi=0;  ba=&alfa+1,  dt=0;  dta=0;
 r=j(p,1,.);                           /* poids pour les variables          */

  run impute;
  run aff1;
 do  while (nbi<&mi2 &  ba>&alfa & all(ning >=1) ) ;
      run cent1;
      run inert1;
      nbi=nbi+1;
      if nbi=1 then ba=&alfa+1;
      else  ba=abs(dta-dt)/dt;
      dta=dt;
      run impute;
      run aff1;
 end;
/*   print,,"RESULTAT:";
    print,,"l'indice de l'unité imputée:",,t(indm);

  print,,"l'effetive de classe de notre methode:",,ning;   */
 finish phase2;

 start do_itc;
   h=0; dmm=0;  p=0;  n=0;
   dmid=0; a=0; mvect=0; hstd=0;   ing=0; cla=0;                                    */
   valp=0; nbi=0; ning=0; r=0;       /* initialisation,
  use  &data;
   free h dmm dmid a   hstd ning r ing cla ;
   read all var _num_ into h[colname=vars];
/*   print,,"données initiales:",,h ;          */
    doni=h;                          /* sas.data ---> en matrice                  */
    dmm=choose(h=.,-1,0);            /* matrice sur DM: 0--presence, -1--DM       */
    dmid=dmm[,+];
    dmid=choose(dmid=0,0,-1);        /* dmid(1:n),0-présence,-1--d.M  (unite)     */
    a=choose(dmm=-1,.,1);            /* a : mat.de projection. 0--d.m,1--présence  */
   p=ncol(h) ;  n=nrow(h);
```

```
                                    /* centres de gravité initiales : k premiers unités */
  do j=1 to p ;
    do i=1 to n ;
      if dmm[i,j]=-1 then do;
        h[i,j]=h[,j][:] ;              /* init1: imputée par la moyenne de la variable */
      end;
    end;
  end;
 run phase1;
 if(&m12>0) then run phase2;
 cla=j(n,1,0);
 do i=1 to n;
     do j=1 to &k;
          if ing[i,j]=1 then cla[i,1]=j;
     end;
 end;
 create &out from h[colname=vars];
 append from h;
 close &out;       vars='class';
 create &class from cla[colname=vars];
 append from cla;
 close &class,
 finish do_itc;
                                    /*  matrice ---> sas.data        */
run do_itc;
quit;
%mend initc;
```

FUSION V.1 DOCUMENTATION

DESCRIPTION

Ce document décrit le programme FUSION V.1. La Macro SAS Fusion réalise la technique de fusion par l'analyse homogène pour variables qualitatives. Cette technique maximise la somme des p plus grandes valeurs propres de la matrice de corrélation des données imputées quantifiées. Cette méthode est adaptée pour des variables avec fort lien ($\eta > 0.3$).

Le programme FUSION fait appel à MISTRESS "Missing data imputation by maximizing internal consistency" créé par Van Buuren S. & Van Rijckevorsel J.L A.(1992).

ORGANIGRAMME

Phase 1 : quantification des variables qualitatives du fichier Donneur en utilisant itérativement les formules des moyennes réciproques.

$$Y = D^{-1}G'X$$
$$X = \frac{1}{m}GY$$

Phase 2 : pour chaque individu r du fichier Receveur + l'ensemble du fichier Donneur, on recherche les valeurs à transférer les plus homogènes au fichier Donneur.

l'étape 1 - imputation initiale : nous imputons une catégorie manquante en choisissant la catégorie dont la valeur de quantification est la plus proche du score individuel.

L'étape 2 - reimputation telle que la somme des ndim plus grandes valeurs propres de la matrice de corrélation des variables quantifiées optimales soit maximale.

I. L'utilisation de l'algorithme "*K-mean*" pour trouver l'imputation de la catégorie la plus homogène en prenant G pseudo-complet précédemment calculé comme valeurs initiales.

II. Un changement de valeur du score individuel peut causer une modification d'imputation. Le phénomène inverse se produit également, c'est-à-dire un changement d'imputation a une action sur les variables de score. Le recalcul des quantifications est nécessaire.

III. Test : la valeur de la fonction de perte d'homogénéité se stabilise ou le nombre d'itération sont atteint. Réitération des procédures I et II jusqu'à la satisfaction du critère choisi.

MATERIEL REQUIS :

Fusion est développée et testée sur un PC 486DX2-66 avec SAS/IML 6.10.

TYPE DE VALEUR DE DONNEES :

Le Macro Fusion est appelée par %fusion(< paramètre1 = valeur1 > <, paramètre2 = valeur2 > ...). Tous les paramètres sont optionnels. Il est nécessaire de préciser les paramètres quand ses valeurs différent des valeurs par défaut, et ces paramètres peuvent être spécifiés dans n'importe quel ordre dans la Macro appelée. Les valeurs par défaut sont précédées par un signe "=" dans la liste suivante.

Le Macro suppose que les paramètres sont dans les catégories suivantes.

LISTE DE PARAMETRES :

datad = _LAST_ Nom de fichier donneur de format SAS. Si aucun nom n'est donné, le dernier SAS data set créé est utilisé.

datar = _LAST_ Nom de fichier receveur de format SAS. Si aucun nom n'est donné, le dernier SAS data set créé est utilisé.

out = IMPUTED Nom de fichier de format SAS contenant fichier donneur et les données complètes (après la fusion) du fichier receveur. Si aucun nom n'est précisé, un SAS data set IMPUTED sera créé dans la librairie courante. Le nom du fichier résultat doit être différent de ceux des fichiers d'entrée. Si out = _NONE_ aucun fichier de sortie n'est généré.

var = _NUM_ La liste de variables d'entrée inclus dans l'analyse. Les noms des variables doivent être individuellement entre accolade, ex. var = { revenu voiture sexe }. La Macro n'accepte pas les règles conventionnelles comme 'v1 - v5'. Les variables numériques et caractères peuvent être mélangées. Alternativement, à la place des variables en liste, trois options (mot-clé) peuvent être utilisées:

- var = _NUM_ : toutes les variables numériques.
- var = _CHAR_ : toutes les variables de type caractère
- var = _ALL_ : toutes les variables.

Par défaut, toutes les variables numériques sont lues et chaque valeur unique de variable définit une catégorie.

vgtvar = _NONE_ Nom d'une variable qui contient des valeurs entières non négatives comme poids d'unité ligne. Cette option est spécialement utilisée pour lire les réponses de profil fréquence ou des données de tableau croisé. Poids de valeur égale à 0 élimine la ligne de l'analyse. Noter que les données sont répliquées par le poids d'unité ligne en un nouveau data set avant de commencer l'analyse, ce qui permet de fabriquer facilement une

matrice disjonctive. La variable poids doit être listée parmi les variables d'entrée, mais elle est éliminée avant que l'analyse commence. Après redressement, ce programme imprime les observations après expansion et les variables. Par défaut les données ne sont pas redressées.

nid = 1 — Dimensionnée de la solution. Le nombre de dimension ne doit pas dépasser le nombre total des catégories moins le nombre de variables. Par défaut, la fusion par analyse homogène choisit nid = 1.

prit = 1 — Contrôle le nombre d'affichage.
prt = 0 : affichage les remarques d'avertissement et message d'erreurs.
prt = 1 : idem, plus l'affichage les remarques d'avertissement des sommaires de différentes étapes d'imputation
prt = 2 : idem, plus de l'affichage historique.

maxit1 = 100 — Nombre maximum d'itérations de la phase quantification. Cette phase est utilisée pour obtenir une quantification raisonnable avant que la phase d'estimation optimale commence maxit1 = 0 est interdit. Le changement de valeur par défaut n'est pas conseillé dans l'analyse actuelle et permet seulement de vérifier le comportement de cet algorithme.

crit1 = 1E - 6 — Critère de convergence de la phase quantification. Le changement de cette valeur n'est pas conseillé.

maxit2 = 100 — Nombre maximum d'itération de la phase de fusion. maxit2 = 0 est interdit, dans ce cas la fusion n'est pas effectuée.

crit2 = 1E - 7 — Critère de convergence de la phase de fusion. Le changement de cette valeur n'est pas conseillé.

file = print — Nom de fichier pour imprimer les messages et le résultat. Deux noms de fichier sont spécialement intéressants : «file = print» imprimer sur output windows (par défaut), et «file = log» imprime sur le log windows. Macro peut générer des messages d'avertissement et d'erreur. Des messages d'avertissement sont imprimés dans le fichier output et l'exécution continue Si une erreur est apparue, le programme s'arrête. La liste d'avertissement et d'erreur accompagné d'une courte explication est donnée ensuite.

AVERTISSEMENT :

1. Attention : variable x a 32 catégories

Si une variable contient plus de 25 catégories, c.a.d. elle prend plus de 25 valeurs différentes. Ce message est surtout pour attirer attention sur une variable quantitative dans la base de donnée, l'utilisateur peut recoder ce genre de variable en moins de catégories.

2. Attention : 8 lignes vides éliminées.

Les observations contenant uniquement des données manquantes sont éliminées de l'analyse.

3. Attention : 2 colonnes vides éliminés.

Les variables contenant seulement des données manquantes sont éliminées de l'analyse.

4. Attention : nombre de dimension remise à 1.

Si ndim est spécifié de façon inapproprié, le programme trouve une valeur voisine valide, on peut demander une dimension maximale par lui donner une grande valeur, ex., ndim = 999

5. Attention : fréquence trop petite pour X^2-test.

Ce message est imprimé si un des cas suivants est survenu :

- S'il n'y a que deux catégories et que la fréquence calculée (espérance) est inférieure à 5
- Si la fréquence calculée (espérance) est inférieure à 1 pour n'importe quelle catégorie.
- Si le pourcentage des fréquences ayant moins de 5 dépasse 20%, voir Siegel & Castellan (1988. p49), dans ce cas, la p-valeur du test ne peut être utilisée.

6. Attention : variable x n'est pas trouvée.

Une variable dont son nom dans "var = option" n'est pas trouvée dans le SAS data set datar = data ou datad = data. Cette variable ne participe pas dans l'analyse. Vérifier dans " var = option ", "datar = data" ou "datad = data".

ERREUR :

1. Erreur : variable x ne contient pas de colonnes valides.

La variable contient seulement des données manquantes et ne peut pas être analysée. Eliminer cette variable de l'analyse.

2. Erreur : dépendance linéaire dans la routine wgram.

Un message d'erreur est apparu si le rang des données est plus petit que le nombre de dimension demandée. Essayer diminuer ndim.

3. Erreur : matrice de données ne contient aucune information;

Les données ne contiennent que des d.m.. Une cause possible est une mauvaise affectation du data set d'entrée ou une erreur de poids dans la variable de poids. Vérifier dans " datar = option ", "datad = option" ou "wgtvar = option " ;

4. Erreur : la variable n'a pas de catégorie;

Aucune catégorie valide n'est trouvée pour la variable. Vérifier le data set d'entrée.

5. Erreur : trop peu de catégories dans variable x .

Cette erreur est apparue si la variable contient seulement une catégorie valide, ce qui peut se produire dans une variable binaire. Vérifiez le fichier de donnée d'entrée.

6. Erreur : catégorie vide dans variable x .

Variable contient catégorie qui n'a jamais été observée. Cette erreur ne doit jamais se produire.

7. Erreur : variable poids non trouvée.

La variable spécifique dans « wgtvar = option » n'est pas dans le fichier d'entrée de SAS. Vérifier si cette variable est dans la liste de «var = option » et dans «wgtvar = option».

8. Erreur : variable caractère ne peut être variable de poids.

Une variable de type caractère est désignée comme une variable de poids. Choisissez une variable de poids du type numérique.

9. Erreur : variable dans le fichier receveur a un nombre de catégories différent que celui dans le fichier donneur.

CONTRAINTE :

Chaque variable commune dans les fichiers receveur et donneur doit avoir absolument le même nombre de catégories,sinon engendrer les messages d'erreur et l'exécution s'arrête.

ANNEXE B I. DONNEES DE L'EXEMPLE DANS 2.3.4

Données reélles **Données avec l'estimation**

OBS	GINI	FARM	RENT	GNPR	LABO	INST	ECKS	DEAT	DEMO
554425321	1	5	5	4	4	2	5	3	2	1
659714111	2	6	5	9	7	1	4	1	1	1
452525212	3	4	5	2	5	2	5	2	1	2
235616221	4	2	3	5	6	1	6	2	2	1
653146353	5	6	5	3	1	4	6	3	5	3
552446322	6	5	5	2	4	4	6	3	2	2
122714211	7	1	2	2	7	1	4	2	1	1
653325222	8	6	5	3	3	2	5	2	2	2
553435342	9	5	5	3	4	3	6	3	4	2
552435232	10	5	5	2	4	3	5	2	3	2
455435353	11	4	5	5	4	4	5	3	5	3
112625111	12	1	1	2	6	2	5	1	1	1
453334233	13	4	5	3	3	3	4	2	3	3
553336323	14	5	5	3	3	3	6	3	2	3
453246323	15	4	5	3	2	4	6	3	2	3
553346223	16	5	5	3	3	4	6	2	2	3
232636212	17	2	3	2	6	3	6	2	1	2
234626322	18	2	4	4	6	2	6	3	2	2
553345333	19	5	5	3	3	4	5	3	3	3
453336222	20	4	5	3	2	3	6	2	2	2
453245343	21	4	5	3	2	4	5	3	4	3
235142321	22	2	3	5	1	4	2	3	2	1
555356243	23	5	5	5	3	5	6	2	4	3
232525211	24	2	3	2	5	2	5	2	1	1
554526322	25	5	5	4	2	2	6	3	2	2
122326222	26	1	2	2	3	2	6	2	2	2
342145213	27	3	4	2	1	4	5	2	1	3
333725111	28	3	3	3	7	2	5	1	1	1
335615211	29	3	3	5	6	1	5	2	1	1
454715111	30	4	5	4	7	1	5	1	1	1
459445223	31	4	5	9	4	4	5	2	2	3
332625211	32	3	3	2	6	2	5	1	1	1
443436233	33	4	4	3	4	3	6	2	3	3
559235233	34	5	5	9	2	3	5	2	3	3
234355241	35	2	3	4	3	5	5	2	4	1
111533223	36	1	1	1	5	3	3	2	2	3
343244353	37	5	4	3	2	4	4	3	5	3
455431223	38	4	5	5	4	3	1	2	2	3
233713111	39	2	3	3	7	1	3	1	1	1
123713111	40	1	2	3	7	1	3	1	1	1
344231213	41	3	4	4	2	3	1	2	1	3
445615211	42	4	4	5	6	1	5	2	1	1
453815212	43	2	5	3	8	1	5	2	1	2
554525221	44	5	5	4	5	2	5	2	2	1
653635243	45	6	5	3	6	3	5	2	4	3
342612212	46	3	4	2	6	1	2	2	1	1
111441213	47	1	1	1	4	4	1	2	1	3

ANNEXE B II. DONNEES DE L'EXEMPLE II DANS 3.1

OBS	EURV	NIMP	MARI	TOUB	IMMI	HOMO	IDEN(D)	CULT	FEMME	RACE
1	4	4	2	2	4	4	4	2	1	2
2	5	2	1	1	1	1	4	1	1	1
3	5	4	1	4	4	2	2	4	2	5
4	1	2	1	2	2	2	2	2	1	1
5	2	5	4	4	4	2	1	1	2	5
6	5	5	2	1	4	1	5	2	1	5
7	2	4	1	2	4	2	2	2	2	2
8	5	5	5	2	4	4	2	1	2	1
9	1	5	1	1	4	1	5	1	4	1
10	5	4	1	5	2	2	1	2	1	1
11	1	2	1	1	1	2	1	1	1	2
12	4	4	1	2	2	1	4	1	1	1
13	5	5	1	5	5	2	5	1	5	1
14	5	5	1	3	2	4	5	5	2	2
15	2	5	5	5	5	1	5	5	5	1
16	5	5	1	5	1	5	5	5	1	1
17	2	1	1	1	2	2	4	1	1	1
18	5	4	2	2	5	3	4	2	1	2
19	5	5	2	2	4	2	4	2	2	2
20	5	1	1	5	5	2	4	1	1	1
21	4	4	5	2	2	4	5	2	2	2
22	5	3	2	5	4	5	5	4	1	5
23	2	2	2	1	2	1	5	1	1	2
24	1	5	5	2	4	4	2	1	1	2
25	5	5	4	4	5	5	5	2	5	1
26	2	2	1	1	4	2	5	5	1	1
27	2	1	1	2	2	1	4	1	2	2
28	5	4	4	5	4	2	5	4	4	4
29	4	5	2	5	5	4	5	1	5	2
30	5	5	1	1	5	5	5	1	5	1
31	5	5	1	1	5	1	2	1	1	1
32	2	2	2	2	2	4	2	2	1	1
33	5	5	5	5	5	5	4	5	5	2
34	4	4	5	4	4	4	2	4	5	2
35	2	4	1	2	2	2	5	2	1	1
36	2	2	1	4	1	2	5	1	5	1
37	4	5	2	5	3	5	5	2	1	2
38	5	5	5	1	5	1	2	5	1	1
39	2	5	1	1	2	2	5	2	1	4
40	5	5	4	1	4	2	4	2	1	4
41	2	1	1	4	5	2	5	1	1	2
42	4	4	1	2	4	5	2	2	5	2
43	5	4	4	5	5	5	5	5	5	3
44	2	1	1	1	1	2	1	2	4	1
45	2	4	1	1	2	2	2	2	1	1
46	5	4	1	5	5	1	5	1	4	1
47	5	5	1	4	4	5	4	2	2	5
48	1	5	1	1	4	2	5	2	1	2
49	2	5	5	1	5	1	5	5	5	2
50	5	5	1	1	5	1	1	1	1	2
51	5	5	2	2	4	1	5	5	4	2
52	5	4	1	4	5	4	2	1	1	3
53	2	4	2	5	2	4	2	1	1	1
54	4	5	1	1	1	1	5	1	1	1
55	2	5	2	4	2	2	1	2	1	2
56	5	1	1	1	5	5	1	5	2	1
57	1	5	5	5	4	2	1	1	4	5
58	5	5	5	5	5	5	5	5	5	5
59	4	4	2	2	4	2	4	4	2	4
60	5	4	4	4	4	4	4	1	2	1
61	4	4	4	2	5	1	1	2	4	4
62	1	4	2	2	5	1	2	1	1	2
63	1	5	1	5	2	5	5	1	1	1
64	1	4	1	2	1	1	2	1	1	1
65	5	5	5	5	5	1	4	2	4	4
66	1	2	1	2	1	1	1	1	1	1
67	5	5	5	5	5	4	4	5	1	4
68	2	2	2	4	2	2	2	1	2	1
69	2	4	1	2	4	2	2	2	2	1
70	2	4	2	2	1	2	5	2	1	2
71	2	2	1	1	2	2	1	2	1	1
72	2	5	1	2	2	1	4	1	1	2
73	4	3	3	3	4	2	3	3	4	3
74	5	1	1	4	4	2	4	2	4	5
75	4	4	4	4	5	4	4	2	1	3
76	4	2	1	2	4	2	2	2	1	1
77	5	5	1	1	5	2	4	2	1	1
78	5	5	1	4	4	1	4	2	1	1
79	1	2	1	1	2	1	1	1	2	1
80	5	5	1	5	1	4	5	1	5	1
81	2	4	2	1	5	1	5	1	1	1
82	4	5	1	4	1	2	5	2	1	5
83	2	2	5	1	5	5	2	5	1	1
84	2	4	1	2	4	2	1	1	5	2
85	5	4	1	1	5	2	1	2	4	4
86	2	4	1	2	4	1	2	1	1	2
87	1	2	5	5	5	5	5	2	4	2
88	2	2	1	2	5	1	2	1	1	1
89	2	2	1	1	2	1	2	1	1	2
90	1	5	1	5	4	1	5	1	5	1
91	2	4	1	4	1	2	4	2	5	2
92	1	5	2	5	2	4	5	1	2	4
93	4	4	1	4	5	1	2	4	2	1
94	2	2	1	1	5	1	5	1	4	2
95	1	2	1	2	4	2	4	2	2	2
96	2	4	4	4	5	2	4	2	4	2
97	4	4	4	2	4	4	4	5	1	2
98	2	5	1	2	1	5	4	2	1	1
99	4	4	4	5	2	1	5	2	5	4
100	4	5	1	1	2	2	5	1	1	2
101	2	3	2	5	4	2	3	2	4	4
102	5	3	2	4	4	2	4	2	1	2
103	5	2	1	2	5	1	2	5	5	4
104	2	2	1	2	4	2	1	1	1	1
105	2	4	1	3	3	5	5	5	5	1
106	2	5	2	1	5	5	4	1	1	4
107	4	4	5	5	4	2	4	2	4	4
108	1	4	4	4	4	5	5	1	1	1
109	1	5	1	1	2	1	1	1	1	2
110	4	5	1	5	5	1	1	1	5	5
111	5	5	4	5	5	3	5	2	2	5
112	5	1	4	5	4	4	5	1	4	1
113	5	2	2	5	2	1	5	4	4	4
114	5	4	1	1	4	2	5	2	1	1
115	2	4	2	4	2	1	4	2	1	1
116	2	4	1	2	1	2	4	1	2	1
117	4	4	1	4	1	2	1	1	1	1
118	2	5	1	1	1	1	5	1	2	1
119	3	3	3	3	3	2	4	3	3	3
120	2	4	1	2	1	1	2	1	1	1
121	2	4	1	1	1	1	1	1	4	1
122	4	1	1	1	2	1	1	1	1	1
123	2	5	4	2	4	4	4	2	4	4
124	5	5	5	5	5	5	5	5	4	5
125	2	2	1	2	3	3	2	2	2	2
126	1	5	4	3	3	4	3	5	5	4
127	2	5	1	2	4	2	4	4	2	2
128	1	5	4	4	5	5	5	5	5	1
129	3	4	2	2	2	3	4	2	4	3
130	2	5	2	4	4	4	4	2	2	2
131	4	4	1	2	2	2	1	2	1	1
132	4	4	1	1	2	2	4	2	4	2
133	2	2	5	4	4	2	4	2	1	4
134	5	5	2	2	5	2	2	1	2	2
135	1	5	5	5	5	2	5	1	2	1
136	4	1	5	1	5	1	2	1	1	5
137	2	2	2	2	4	1	5	4	2	1
138	4	5	5	5	4	5	5	5	4	4
139	1	5	1	3	1	1	2	4	5	2
140	5	5	2	1	1	1	5	5	1	5
141	4	5	4	4	4	1	5	1	1	4
142	4	2	1	2	4	1	5	4	1	2
143	4	5	1	1	5	1	5	2	1	1
144	1	2	2	3	2	2	1	1	1	1
145	4	4	2	1	2	1	2	1	1	1
146	5	4	5	5	5	5	1	1	1	5

147	2	5	2	5	5	4	5	4	2	4
149	5	5	1	4	5	1	5	1	1	1
151	4	5	2	2	4	2	1	2	3	2
153	2	1	1	4	1	1	2	1	1	1
155	4	4	1	2	1	2	1	2	1	2
157	2	2	1	4	2	1	1	1	3	1
159	2	3	1	3	3	3	3	2	3	1
161	5	2	4	1	4	2	5	2	1	4
163	5	5	4	5	5	1	5	2	1	1
165	4	5	1	2	4	2	2	2	2	1
167	4	5	1	2	4	4	5	1	1	4
169	4	2	1	2	4	2	2	2	1	2
171	1	5	1	1	2	1	2	1	1	1
173	1	5	5	5	5	5	5	1	1	1
175	5	5	1	1	5	4	5	5	2	5
177	5	4	5	2	5	2	4	1	5	5
179	2	4	2	2	4	2	2	2	1	2
181	3	5	5	5	1	3	5	2	5	3
183	2	4	1	1	4	5	5	1	2	2
185	5	4	1	4	4	2	4	4	1	2
187	1	5	1	1	2	2	1	1	5	1
189	1	5	1	4	2	1	4	2	1	1
191	2	4	1	1	1	1	1	1	1	1
193	2	4	1	1	2	2	5	2	4	1
195	4	5	5	2	4	4	5	1	4	4
197	2	2	1	1	1	1	1	2	1	1
199	5	5	1	2	5	1	5	1	1	4
201	2	5	1	4	4	2	4	2	4	2
203	2	4	2	2	4	2	2	2	1	2
205	2	2	2	4	2	2	1	2	2	2
207	1	4	1	4	2	1	1	1	1	1
209	4	4	1	2	5	1	5	2	4	1
211	2	2	1	2	2	1	2	1	1	1
213	5	5	2	1	4	1	4	1	1	2
215	2	1	4	2	4	2	4	4	4	2
217	1	5	4	1	5	1	5	2	4	1
219	1	4	1	2	5	2	5	1	5	1
221	5	4	1	4	2	1	5	1	1	1
223	4	2	1	2	5	2	4	2	2	4
225	2	5	2	5	4	1	2	2	2	2
227	1	5	1	5	1	4	5	4	2	1
229	1	5	2	1	1	1	1	2	1	4
231	4	4	2	4	4	1	4	2	2	2
233	2	5	1	1	5	2	5	2	1	5
235	5	5	2	1	5	1	5	2	4	1
237	2	4	1	2	2	1	2	1	1	2
239	2	2	1	2	2	1	1	1	1	1
241	2	4	1	2	2	2	2	2	1	1
243	2	2	2	1	2	2	2	2	2	2
245	1	2	1	1	2	1	1	1	1	1
247	1	5	1	5	1	2	5	1	5	1
249	1	2	1	4	3	5	2	1	1	1
251	5	5	5	5	5	1	5	2	1	1
253	5	2	1	1	4	1	1	1	1	1
255	2	2	4	4	2	1	2	1	1	1
257	2	2	1	4	1	1	4	1	1	1
259	2	2	1	1	4	2	5	1	1	1
261	2	4	2	2	4	2	4	4	1	2
263	1	2	1	1	4	5	2	2	2	1
265	5	1	1	2	2	2	5	2	1	1
267	5	5	5	1	5	1	5	1	2	5
269	5	5	5	2	5	1	5	1	1	4
271	1	2	1	1	4	1	2	1	1	1
273	1	1	1	2	1	1	1	1	1	1
275	2	4	2	5	5	2	4	3	4	4
277	1	1	1	1	1	1	1	1	1	1
279	2	5	1	2	1	2	5	5	2	1
281	1	5	4	2	4	1	5	1	2	2
283	4	4	3	2	4	2	2	2	1	2
285	2	1	1	1	1	2	2	2	1	1
287	4	4	5	3	4	4	4	4	5	4
289	4	4	1	1	1	2	4	1	1	1
291	1	5	1	1	5	2	5	1	1	4
293	4	4	4	5	5	4	4	2	4	2
295	1	5	1	1	1	1	1	1	1	1
297	1	5	1	1	2	5	5	1	1	1
299	1	2	1	1	1	1	2	2	1	1

148	2	2	1	2	2	2	2	5	1	1
150	1	4	2	2	2	1	5	2	2	2
152	5	4	5	2	1	2	5	4	1	1
154	1	5	4	4	4	1	5	2	4	4
156	1	5	1	2	4	2	1	1	1	1
158	4	4	2	4	2	3	4	2	5	1
160	5	5	4	1	4	2	4	1	1	2
162	4	5	2	3	4	3	5	4	1	1
164	5	5	1	4	2	1	1	5	1	3
166	1	5	1	2	1	1	4	1	4	1
168	5	1	4	5	5	2	5	2	4	3
170	3	3	2	5	1	5	5	1	1	1
172	5	1	1	4	1	4	5	1	1	1
174	2	5	1	1	4	2	2	1	2	4
176	2	4	2	2	4	4	5	2	1	2
178	1	1	1	4	1	2	4	1	4	1
180	2	5	1	2	4	1	5	2	2	4
182	1	5	1	2	5	1	1	1	5	2
184	1	5	1	5	1	1	5	1	4	1
186	2	4	1	1	5	1	1	2	5	2
188	5	2	1	1	2	4	2	2	2	1
190	2	4	1	2	2	2	4	2	1	1
192	2	2	1	2	4	1	1	1	1	1
194	5	5	5	5	4	5	5	1	5	5
196	1	2	1	2	1	2	1	2	2	1
198	2	5	2	2	5	3	2	2	2	2
200	1	4	1	2	2	1	1	1	1	1
202	2	2	1	2	4	5	4	2	2	4
204	1	4	1	1	2	1	2	1	1	1
206	1	5	1	1	1	1	5	1	5	1
208	1	1	1	1	5	2	2	1	1	1
210	4	4	1	2	4	2	4	2	1	4
212	5	5	4	2	3	1	4	2	5	1
214	4	4	2	4	4	2	4	2	1	2
216	2	5	1	1	5	1	2	2	2	2
218	1	5	2	5	5	2	4	2	1	1
220	2	4	4	5	2	5	1	2	1	1
222	5	5	1	5	5	2	3	5	4	2
224	5	5	3	1	5	1	1	5	1	1
226	1	1	1	2	1	2	3	1	1	1
228	5	5	4	4	5	2	5	2	1	5
230	5	5	1	4	5	5	5	4	4	5
232	5	2	5	1	2	4	2	2	4	3
234	2	4	1	4	2	2	4	2	1	1
236	5	5	5	4	5	2	5	2	1	5
238	4	5	2	5	4	1	5	4	5	2
240	4	1	1	2	2	2	2	2	2	1
242	4	5	1	1	5	1	5	4	1	2
244	5	4	5	5	2	2	2	1	1	5
246	1	2	2	2	1	1	5	1	5	2
248	5	5	4	1	4	2	4	2	4	1
250	2	4	1	2	1	2	2	1	1	1
252	3	5	4	2	3	4	5	3	4	2
254	5	5	4	5	2	1	5	2	3	1
256	1	1	1	1	2	1	1	1	1	1
258	1	1	4	4	5	5	4	2	5	3
260	4	1	2	3	3	2	4	3	5	4
262	1	2	1	2	1	1	2	1	1	2
264	2	5	4	2	4	3	2	4	2	4
266	4	4	4	2	2	2	4	1	2	2
268	1	3	3	2	1	5	4	1	1	1
270	5	5	1	1	5	1	5	2	1	2
272	5	5	1	2	4	4	5	2	2	2
274	5	4	3	5	5	2	4	4	5	1
276	2	4	1	1	5	5	5	5	3	1
278	3	1	1	4	3	3	5	1	1	1
280	1	2	1	1	2	1	4	1	1	1
282	2	4	4	5	4	5	4	2	1	3
284	2	4	4	2	2	3	2	3	4	2
286	5	1	1	1	2	1	1	1	1	1
288	2	1	1	1	3	2	2	2	1	1
290	5	5	2	2	4	1	2	2	1	2
292	5	5	2	5	4	5	5	4	2	2
294	5	5	5	5	4	4	5	1	5	5
296	5	5	5	5	5	1	5	1	1	1
298	2	2	1	1	2	1	1	1	2	1
300	1	5	1	3	3	5	2	2	1	3

Les coordonnees sur le circle de correlation (axi1 et axe2)

	c1c	c2c,	c1r	c2r (notre méthode),	c1m	c2m (moyenne),	c1s	c2s(MGV)
EURV	0.5373542	-0.469326	0.5484255	-0.514544	0.53441	-0.482659	0.3074321	-0.522547
NIMP	0 4887731	-0.177962	0.458691	0.012831	0.4440107	0.007041	0.0003285	-0.00881
MARI	0.6707031	0.0316918	0.6532597	0.0788423	0.6382979	-0 003564	0.8168585	-0.051051
TOUB	0.5538336	0.474074	0.5702694	0.4946332	0.5275937	0.4562611	0.4650359	0.1818249
IMMI	0.6028451	-0.418684	0.5977839	-0.430193	0.5748441	-0.476199	0 4760615	-0.51218
HOMO	0 4778985	0.462561	0.5400597	0.3670638	0.519621	0.3507406	0.2551788	0.8241637
IDEN(D)	0.5546984	0.1520429	0 5784711	0.1400867	0.5444853	0.237106	0 0461682	-0.129094
CULT	0.4913078	0.0061521	0.4865603	-0.195837	0.4793813	-0 000112	0.2113716	0.8178605
FEMM	0.4748758	0.383596	0.5056134	0.3702044	0.4526891	0.4255859	0.739109	0.1108689
RACE	0.5999775	-0.3047	0 5776057	-0.296532	0.5590807	-0.382504	0.8033303	-0.121419

ANNEXE B.III : DONNEES SUR "VENDEUR"

x1-x3 : résultat de vente
x4-x7 : profil du vendeur

x1	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*
93.0	96 0	97.8	9.0	12.0	9.0	20.0
88.8	91.8	96.8	7.0	10 0	10.0	15.0
95.0	100.3	99 0	8.0	12.0	9.0	26.0
101.3	103.8	106.8	13.0	14 0	12.0	29.0
102.0	107.8	103.0	10.0	15.0	12 0	32.0
95.8	97.5	99.3	10 0	14.0	11.0	21 0
95.5	99.5	99.0	9.0	12.0	9.0	25.0
110.8	122.0	115.3	18.0	20.0	15.0	51.0
102.8	108.3	103.8	10.0	17.3	13.0	31.0
106.8	120.5	102.0	14.0	18.0	11.0	39.0
103.3	109.8	104.0	12.0	17.0	12.0	32.0
99.5	111.8	100.3	10.0	18.0	8.0	31.0
103.5	112.5	107.0	16.0	17 0	11.0	34.0
99.5	105.5	102.3	8.0	10.0	11.0	34.0
100.0	107.0	102.8	13.0	10.0	8.0	34.0
81.5	93.5	95.0	7.0	9.0	5.0	16.0
101.3	105.3	102.8	11.0	12.0	11.0	32.0
103.3	110.8	103.5	11.0	14 0	11.0	35.0
95.3	104 3	103.0	5.0	14.0	13.0	30.0
99.5	105.3	106.3	17.0	17.0	11.0	27.0
88.5	95.3	95.8	10.0	12.0	7.0	15.0
99.3	115.0	104.3	5.0	11.0	11.0	42.0
87.5	92.5	95.8	9.0	9 0	7.0	16.0
105.3	114.0	105 3	12.0	15.0	12.0	37.0
107.0	121.0	109.0	16.0	19.0	12.0	39.0
93.3	102.0	97.8	10.0	15.0	7.0	23.0
106.8	118.0	107.3	14.0	16.0	12.0	39.0
106.8	120.0	104.8	10 0	16.0	11.0	49.0
92.3	90.8	99.8	8.0	10.0	13.0	17.0
106.3	121.0	104.5	9.0	17.0	11.0	44 4
106.0	119.5	110.5	18.0	15.0	10.0	43.0
88.3	92.8	96.8	13.0	11.0	8.0	10.0
96.0	103.3	100.5	7.0	15.0	11.0	27.0
94.3	94.5	99.0	10.0	12.0	11.0	19.0
106.5	121.5	110.5	18.0	17.0	10.0	42.0
106.5	115.5	107.0	8.0	13.0	14.0	47.0
92.0	99.5	103.5	18.0	16.0	8.0	18.0
102.0	99.8	103.3	13.0	12.0	14.0	28.0
108.3	122.3	108.5	15.0	19.0	12 0	41.0
106.8	119.0	106.8	14.0	20.0	12.0	37.0
102.5	109.3	103.8	9.0	17.0	13.0	32.0
92 5	102.5	99.3	13.0	15.0	6.0	23.0
102.8	113.8	106.8	17.0	20.0	10.0	32.0
83.3	87.3	96.3	1.0	5.0	9 0	15.0
94.8	101.8	99.8	7.0	16 0	11.0	24.0
103 5	112.0	110.8	18.0	13.0	12.0	37.0
89.5	96.0	97.3	7.0	15.0	11.0	14 0
84.3	89.8	94.3	8.0	8.0	8.0	9.0
104.3	109.5	106.5	14.0	12.0	12.0	36.0
106.0	118.5	105.0	12.0	16.0	11.0	39.0

Matrice de corrélation pour variables : ***x1- x7***

	x1	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*
x1	*1*						
x2	*.9261*	*1*					
x3	*.884*	*.8425*	*1*				
x4	*.572*	*.5415*	*.7004*	*1*			
x5	*.7079*	*.745*	*.6368*	*.5892*	*1*		
x6	*.6744*	*.4654*	*.6411*	*.1469*	*.3874*	*1*	
x7	*.9273*	*.9443*	*.8526*	*.4126*	*.5739*	*.5664*	*1*

Données sur 'Vendeur' en valeur ordinale recodifiée

x1	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*
1	1	1	1	1	1	2
1	1	1	1	1	1	1
1	2	1	1	1	1	2
2	2	3	2	2	2	2
2	2	2	1	2	2	2
2	1	1	1	2	2	2
2	1	1	1	1	1	2
3	3	3	3	3	3	3
2	2	2	1	3	2	2
3	3	2	2	3	2	3
2	2	2	2	3	2	2
2	3	2	1	3	1	2
2	3	3	3	3	2	2
2	2	2	1	1	2	2
2	2	2	2	1	1	2
1	1	1	1	1	1	1
2	2	2	2	1	2	2
2	3	2	2	2	2	2
2	2	2	1	2	2	2
2	2	3	3	3	2	2
1	1	1	1	1	1	1
2	3	2	1	1	2	3
1	1	1	1	1	1	1
3	3	3	2	2	2	3
3	3	3	3	3	2	3
1	2	1	1	3	1	2
3	3	3	2	3	2	3
3	3	2	1	3	2	3
1	1	1	1	1	2	1
3	3	1	1	3	2	3
3	3	3	3	2	1	3
1	1	1	2	1	1	1
2	2	2	1	2	2	2
1	1	1	1	1	2	1
3	3	3	3	3	1	3
3	3	3	1	1	3	3
1	1	2	3	3	1	1
2	1	2	2	1	3	2
3	3	3	3	3	2	3
3	3	3	2	3	2	3
2	2	2	1	3	2	2
1	2	1	2	2	1	2
2	3	3	3	3	1	2
1	1	1	1	1	1	1
1	2	1	1	3	2	2
2	3	3	3	1	2	3
1	1	1	1	2	2	1
1	1	1	1	1	1	1
2	2	3	2	1	2	3
3	3	2	2	3	2	3

ANNEXE IV : L'EXEMPLE DE LA FUSION DES FICHIERS SUR L'ENQUETE 1000

Q1 - âge de l'enquêté(e) en classes :
1 moins de 20 ans
2 20 - 29 ans
3 30 - 49 ans
4 50 - 64 ans
5 65 ans et plus

Q2 - taille d'agglomération (en nombre d'habitants) .
1 moins de 2.000
2 2.000 - 20.000
3 20.000 - 100.000
4 plus de 100.000
5 Paris

Q3 - heure de coucher :
1 21h. ou avant
2 entre 21h. et 22h.
3 entre 22h. et 23h.
4 entre 23h. et 24h.
5 après minuit
6 variable
7 non-réponse

Q4 - âge de fin d'étude :
1 moins de 18 ans
2 18 - 19 ans
3 20 - 24 ans
4 25 - 26 ans
5 26 ans et plus

Q5 : la famille est le seul endroit où l'on se sente bien ?
1 oui
2 non

Q6 : diplôme d'enseignement général le plus élevé obtenu :
1 aucun
2 CEP ou fin études
3 BEPC-BE-BEPS
4 baccalauréat (1/2)
5 brevet sup.
6 université, gde. école
7 autre

Q7 - regardez-vous la télévision ?
1 tous les jours
2 assez souvent
3 pas très souvent
4 jamais

Fichier Receveur

Données Réelles Calcul AH Calcul Statiro

OBS	Q5	Q6	Q7	AH5	AH6	AH7	ST5	ST6	ST7
1	1	5	2	1	2	1	1	4	1
2	1	3	2	1	1	2	1	3	3
3	1	2	1	1	2	1	1	2	3
4	2	6	3	2	6	4	2	4	4
5	1	1	1	1	2	1	1	2	1
6	1	2	1	1	2	1	2	3	2
7	2	4	1	2	5	3	1	3	2
8	1	2	1	1	2	1	1	3	1
9	2	2	1	1	2	1	1	1	2
10	1	2	3	1	2	1	1	1	1
11	2	4	2	2	5	3	2	4	2
12	1	2	1	1	2	1	1	2	1
13	1	1	3	1	2	1	1	2	1
14	1	1	1	1	2	1	1	1	3
15	1	3	2	1	2	1	1	2	2
16	2	2	1	1	2	1	1	1	1
17	1	2	3	1	2	1	1	2	3
18	1	3	3	1	3	2	2	1	1
19	2	1	3	2	5	3	1	1	1
20	1	1	4	1	2	1	2	2	1
21	1	4	3	2	6	4	2	6	2
22	1	2	3	1	2	1	1	2	4
23	1	3	2	1	2	1	1	3	1
24	1	2	1	1	2	1	1	1	4
25	1	2	3	1	2	1	1	1	2
26	2	4	4	2	6	4	1	3	4
27	2	6	3	2	6	4	2	4	3
28	1	4	1	1	1	2	1	4	1
29	1	4	3	1	2	1	1	2	1
30	1	3	1	2	5	3	2	3	2
31	1	2	1	1	2	1	1	2	1
32	1	1	1	1	2	1	1	1	2
33	1	4	3	1	1	2	1	4	1
34	2	2	1	1	2	1	1	1	1
35	1	2	1	1	2	1	1	2	1
36	1	1	1	1	2	1	2	2	1
37	2	4	1	2	6	4	2	4	2
38	2	3	3	2	6	4	2	6	2
39	1	4	2	2	6	4	2	6	4
40	1	1	3	1	2	1	1	1	2
41	1	2	2	1	2	1	1	1	1
42	2	6	4	2	6	4	2	2	3
43	1	4	3	2	5	3	1	3	3
44	1	3	2	1	2	1	2	3	4
45	1	2	2	1	2	1	1	2	2
46	1	1	1	1	2	1	1	1	4
47	2	2	1	1	2	1	2	2	1
48	1	4	3	2	6	4	2	4	4
49	1	2	1	1	2	1	1	1	1
50	1	3	1	2	5	3	1	4	1
51	1	1	1	1	2	1	1	4	1
52	1	3	1	1	2	1	1	2	2
53	1	2	1	1	2	1	1	1	1
54	1	2	1	1	2	1	1	1	1
55	1	3	1	1	2	1	1	1	1
56	1	5	3	2	6	4	2	5	1
57	1	2	4	1	2	1	1	2	4
58	1	2	1	1	2	1	2	3	1
59	1	2	2	1	2	1	1	2	1
60	1	3	1	2	5	3	2	2	2
61	2	6	3	2	5	3	2	4	3
62	2	2	4	1	2	1	1	2	2
63	1	4	2	1	2	1	1	2	1
64	1	1	1	1	2	1	1	2	1
65	2	4	4	2	6	3	1	4	2
66	1	2	1	2	5	3	1	3	2
67	1	1	1	1	2	1	1	1	4
68	1	1	1	1	2	1	1	2	1
69	2	4	4	2	6	3	2	2	3
70	1	3	2	1	2	1	2	2	1
71	2	4	3	2	5	3	2	4	3
72	1	1	1	1	2	1	1	2	1
73	1	2	1	1	2	1	1	1	1
74	1	1	2	1	2	1	1	1	4
75	2	3	4	1	2	1	1	2	1
76	1	1	1	1	2	1	1	2	1
77	1	2	1	1	2	1	1	2	1
78	2	3	1	1	3	1	1	4	2
79	1	2	1	1	2	1	1	1	2
80	1	3	1	1	2	1	2	1	1
81	2	6	3	2	6	3	1	3	2
82	2	3	4	1	2	1	2	2	1
83	1	6	1	2	6	4	2	4	2
84	1	3	1	2	6	4	2	4	4
85	1	4	4	1	2	1	1	1	1
86	1	2	1	1	2	1	1	2	1
87	1	1	1	1	2	1	1	2	3
88	2	3	1	2	5	3	1	4	3
89	2	3	3	2	6	3	2	4	2
90	1	2	1	1	2	1	2	2	1
91	1	1	2	1	3	1	1	1	1
92	1	2	1	1	2	1	2	1	1
93	2	3	1	1	2	1	2	2	3
94	1	1	1	2	5	3	2	1	1
95	2	3	2	1	2	1	1	1	1
96	1	4	1	1	3	2	1	4	1
97	1	6	1	2	6	4	2	6	4
98	1	4	2	1	3	2	2	3	2
99	1	1	1	1	2	1	1	1	1
100	2	4	1	2	6	4	2	4	3
101	1	1	2	2	6	4	2	6	3
102	2	4	3	2	5	3	2	3	2
103	1	3	3	1	2	1	2	3	1
104	1	2	1	1	2	1	1	1	4
105	1	1	2	1	2	1	2	2	1
106	1	1	1	1	2	1	1	3	1
107	2	4	1	1	3	2	1	1	1
108	1	3	2	2	5	3	2	6	1
109	1	2	1	1	2	1	1	1	1
110	1	1	1	1	2	1	1	1	1
111	1	2	1	1	2	1	1	2	1
112	1	1	1	1	2	1	1	1	2
113	1	2	1	1	2	1	1	1	1
114	1	2	1	1	2	1	2	2	1
115	1	1	2	1	2	1	1	1	1
116	1	3	4	1	2	1	2	1	1
117	1	3	3	2	6	3	1	1	1
118	1	3	2	1	2	1	1	2	1
119	1	6	2	2	5	3	1	4	3
120	2	2	3	1	3	2	2	4	3
121	2	3	1	1	2	1	1	2	1
122	2	3	1	2	6	4	1	7	2
123	1	2	1	1	2	1	2	1	3
124	1	2	2	2	5	3	2	1	1
125	2	2	2	1	2	1	1	2	1
126	1	2	1	1	2	1	1	1	1
127	1	1	1	1	3	2	1	4	3
128	1	2	3	2	5	3	1	3	1
129	1	4	4	1	2	1	1	3	3
130	2	2	1	1	2	1	1	1	1
131	1	4	2	1	3	2	2	4	2
132	1	2	1	1	3	2	1	2	1
133	1	3	2	1	1	2	1	3	3
134	1	6	1	1	2	1	1	6	1
135	1	3	4	1	2	1	1	2	1
136	2	1	1	1	2	1	1	1	1
137	2	6	3	2	6	4	2	6	3
138	1	2	2	1	2	1	1	2	1
139	1	3	1	1	3	2	2	3	4
140	1	1	1	1	3	2	1	3	2

141	1	4	1	2	5	3	1	3	2
143	1	2	1	1	2	1	1	3	1
145	1	1	1	1	2	1	1	2	1
147	1	2	1	1	2	1	1	2	1
149	1	6	4	2	6	4	2	6	3
151	2	1	1	1	2	1	1	1	1
153	2	4	2	1	3	2	1	2	1
155	2	2	1	1	2	1	1	2	1
157	1	2	1	1	2	1	2	2	1
159	1	1	1	1	2	1	1	2	1
161	2	2	1	2	5	3	2	4	2
163	1	2	1	1	2	1	1	6	3
165	1	2	2	1	2	1	1	2	1
167	1	2	3	2	5	3	1	2	1
169	2	6	4	2	6	4	2	4	3
171	1	2	1	1	2	1	1	2	1
173	1	1	1	2	6	3	2	3	3
175	2	2	2	1	2	1	1	1	1
177	2	2	3	1	2	1	2	1	2
179	2	4	3	2	5	3	2	4	4
181	1	2	1	1	2	1	2	2	2
183	1	4	3	2	6	4	2	6	3
185	1	6	2	2	6	4	2	6	1
187	1	5	3	1	2	1	1	2	1
189	2	2	1	1	1	1	1	2	1
191	1	6	4	2	6	4	2	4	2

142	1	1	3	1	2	1	1	2	2
144	1	2	1	1	2	1	1	2	1
146	1	3	1	1	3	2	1	3	1
148	2	5	2	1	3	2	1	6	2
150	2	1	4	1	2	1	1	2	1
152	1	2	1	1	2	1	1	2	1
154	2	2	2	1	2	1	1	2	1
156	1	2	1	1	2	1	1	2	2
158	2	2	1	1	2	1	2	4	2
160	2	3	2	1	3	2	2	4	3
162	1	2	1	1	1	1	1	4	2
164	1	2	3	1	2	1	1	7	1
166	2	6	1	2	6	4	2	6	3
168	1	4	3	2	5	3	2	1	3
170	2	6	4	2	6	4	2	3	1
172	1	3	1	2	6	4	2	4	2
174	2	2	1	1	2	1	1	2	1
176	1	2	1	1	2	1	1	1	2
178	2	2	1	2	5	3	2	2	2
180	2	4	2	2	6	4	1	3	2
182	1	6	3	2	6	4	2	6	4
184	2	3	1	1	2	1	1	1	4
186	1	1	4	1	2	1	1	2	1
188	1	6	1	1	2	1	1	4	1
190	1	2	3	1	2	1	1	1	3
192	1	2	2	1	2	1	1	2	1

Statistiques Marges

Q1 / Q5	1	2	3	4	5	Total
1	3	24	44	29	36	136
2	5	23	17	8	3	56
Total	8	47	61	37	39	192

Q1 / AH5	1	2	3	4	5	Total
1	4	22	38	35	37	136
2	4	25	23	2	2	56
Total	8	47	61	37	39	192

Q1 / ST5	1	2	3	4	5	Total
1	3	24	29	35	34	125
2	5	23	32	2	5	67
Total	8	47	61	37	39	192

Q2 / Q5	1	2	3	4	Total
1	22	10	40	64	136
2	3	5	19	29	56
Total	25	15	59	93	192

Q2 / AH5	1	2	3	4	Total
1	24	10	46	56	136
2	1	5	13	37	56
Total	25	15	59	93	192

Q2 / ST5	1	2	3	4	Total
1	24	10	42	49	125
2	1	5	17	44	67
Total	25	15	59	93	192

Q3 \ Q5	1	2	3	4	5	6	Total
1	21	38	63	4	9	1	136
2	4	8	32	7	3	2	56
Total	25	46	95	11	12	3	192

Q3 \ AH5	1	2	3	4	5	6	Total
1	24	37	69	1	3	2	136
2	1	9	26	10	9	1	56
Total	25	46	95	11	12	3	192

Q3 \ ST5	1	2	3	4	5	6	Total
1	22	30	60	5	6	2	125
2	3	16	35	6	6	1	67
Total	25	46	95	11	12	3	192

Q4 \ Q5	1	2	3	4	5	Total
1	89	25	19	1	2	136
2	24	16	11	3	2	56
Total	113	41	30	4	4	192

Q4 \ AH5	1	2	3	4	5	Total
1	109	22	4	0	1	136
2	4	19	26	4	3	56
Total	113	41	30	4	4	192

Q4 \ ST5	1	2	3	4	5	Total
1	90	23	10	1	1	125
2	23	18	20	3	3	67
Total	113	41	30	4	4	192

Q6 \ Q5	1	2	3	4	5	6	Total
1	32	53	23	16	3	9	136
2	4	17	12	13	1	9	56
Total	36	70	35	29	4	18	192

AH6 \ AH5	1	2	3	4	5	6	Total
1	6	114	16	0	0	0	136
2	0	0	0	23	33	33	56
Total	6	114	16	23	33	33	192

ST6 \ ST5	1	2	3	4	5	6	7	Total
1	40	50	17	13	0	3	2	125
	21	26	9	7	0	2	1	65
2	9	15	10	20	1	12	0	67
	5	8	5	10	1	6	0	35
Total	49	65	27	33	1	15	2	192
	26	34	14	17	1	8	1	100

Q7 \ Q5	1	2	3	4	Total
1	76	27	24	9	136
2	24	9	13	10	56
Total	100	36	37	19	192

AH7 \ AH5	1	2	3	4	Total
1	118	18	0	0	136
2	0	0	29	27	56
Total	118	18	29	27	192

ST7 \ ST5	1	2	3	4	Total
1	79 41	24 13	13 7	9 5	125 65
2	21 11	19 10	18 9	9 5	67 35
Total	100 52	43 22	31 16	18 9	192 100

Q1 \ Q6	1	2	3	4	5	Total
1	2	3	6	12	13	36
2	0	17	22	15	16	70
3	2	12	13	7	1	35
4	4	9	11	3	2	29
5	0	0	0	0	4	4
6	0	6	9	0	3	18
Total	8	47	61	37	39	192

Q1 \ AH6	1	2	3	4	5	Total
1	0	0	2	2	2	6
2	0	20	29	31	34	114
3	4	2	7	2	1	16
4	0	0	0	0	0	0
5	4	8	9	2	0	2
6	0	17	14	0	2	33
Total	8	47	61	37	39	192

Q1 \ ST6	1	2	3	4	5	Total
1	1	5	16	12	15	49
2	0	17	16	19	13	65
3	2	11	9	4	1	27
4	5	8	13	2	5	33
5	0	0	0	0	1	1
6	0	5	7	0	3	15
7	0	1	0	0	1	2
Total	8	47	61	37	39	192

Q2 \ Q6	1	2	3	4	Total
1	5	3	13	15	36
2	11	5	22	32	70
3	5	4	10	16	35
4	3	3	8	15	29
5	1	0	1	2	4
6	0	0	5	13	18
Total	25	15	59	93	192

Q2 \ AH6	1	2	3	4	Total
1	2	0	1	3	6
2	19	9	35	51	114
3	3	1	10	2	16
4	0	0	0	0	0
5	1	2	10	10	23
6	0	3	3	27	33
Total	25	15	59	93	192

Q2 \ ST6	1	2	3	4	Total
1	6	4	13	26	49
2	13	8	18	26	65
3	4	0	11	12	27
4	2	3	10	18	33
5	0	0	0	1	1
6	0	0	5	10	15
7	0	0	2	0	2
Total	25	15	59	93	192

Q3 \ Q6	1	2	3	4	5	6	Total
1	8	8	15	2	3	0	36
2	8	21	35	2	3	1	70
3	6	10	16	1	1	1	35
4	1	5	17	4	1	1	29
5	1	0	2	0	1	0	4
6	1	2	10	2	3	0	18
Total	25	46	95	11	12	3	192

Q3 \ AH6	1	2	3	4	5	6	Total
1	0	1	5	0	0	0	6
2	23	33	52	1	3	2	114
3	1	3	12	0	0	0	16
4	0	0	0	0	0	0	0
5	0	4	10	5	4	0	23
6	1	5	16	5	5	1	33
Total	25	46	95	11	12	3	192

Q3 \ ST6	1	2	3	4	5	6	Total
1	15	11	16	2	3	2	49
2	9	21	30	2	2	1	65
3	0	5	17	2	3	0	27
4	1	4	24	2	2	0	33
5	0	0	0	0	1	0	1
6	0	5	7	2	1	0	15
7	0	0	1	1	0	0	2
Total	25	46	95	11	12	3	192

Q4 \ Q6	1	2	3	4	5	Total
1	30	4	1	0	1	36
2	64	6	0	0	0	70
3	16	14	5	0	0	35
4	3	15	10	1	0	29
5	0	1	3	0	0	4
6	0	1	11	3	3	18
Total	113	41	30	4	4	192

Q4 \ AH6	1	2	3	4	5	Total
1	1	5	0	0	0	6
2	107	3	3	0	1	114
3	1	14	1	0	0	16
4	00	0	0	0	0	0
5	4	14	4	1	0	23
6	0	5	22	3	3	33
Total	113	41	30	4	4	192

Q4 \ ST6	1	2	3	4	5	Total
1	42	6	1	0	0	49
2	59	5	1	0	0	65
3	9	12	4	2	0	27
4	1	18	11	1	2	33
5	0	0	1	0	0	1
6	1	0	11	1	2	15
7	1	0	1	0	0	2
Total	113	41	30	4	4	192

Q7 \ Q6	1	2	3	4	Total
1	24	5	4	3	36
2	48	10	10	2	70
3	16	10	5	4	35
4	7	7	10	5	29
5	0	2	2	0	4
6	5	2	6	5	18
Total	100	36	37	19	192

AH7 \ AH6	1	2	3	4	Total
1	2	4	0	0	6
2	114	0	0	0	114
3	2	14	0	0	16
4	0	0	0	0	0
5	0	0	23	0	23
6	0	0	6	27	33
Total	118	18	29	27	192

ST7 \ ST6	1	2	3	4	Total
1	31	8	4	6	49
2	48	9	6	2	65
3	9	10	5	3	27
4	7	12	10	4	33
5	1	0	0	0	1
6	3	3	6	3	15
7	1	1	0	0	2
Total	100	43	31	18	192

Q1 / Q7	1	2	3	4	5	Total
1	2	19	28	27	24	100
2	3	10	13	3	7	36
3	3	12	12	5	5	37
4	0	6	8	2	3	19
Total	8	47	61	37	39	192

Q1 / AH7	1	2	3	4	5	Total
1	1	20	29	34	34	118
2	3	2	9	1	3	18
3	4	12	11	2	0	29
4	0	13	12	0	2	27
Total	8	47	61	37	39	192

Q1 / ST7	1	2	3	4	5	Total
1	2	17	26	23	32	100
2	2	15	13	10	3	43
3	3	10	14	2	2	31
4	1	5	8	2	2	18
Total	8	47	61	37	39	192

Q2 / Q7	1	2	3	4	Total
1	14	7	34	45	100
2	9	2	11	14	36
3	2	3	7	25	37
4	0	3	7	9	19
Total	25	15	59	93	192

Q2 / AH7	1	2	3	4	Total
1	19	10	36	53	118
2	5	0	10	3	18
3	1	5	11	12	29
4	0	0	2	25	27
Total	25	15	59	93	192

Q2 / ST7	1	2	3	4	Total
1	21	6	33	40	100
2	1	5	15	22	43
3	3	3	9	16	31
4	0	1	2	15	18
Total	25	15	59	93	192

Q3 \ Q7	1	2	3	4	5	6	Total
1	12	28	51	2	6	1	100
2	2	6	24	1	2	1	36
3	5	10	12	6	4	0	37
4	6	2	8	2	0		19
Total	25	46	95	11	12	3	192

Q3 \ AH7	1	2	3	4	5	6	Total
1	23	34	55	1	3	2	118
2	1	3	14	0	0	0	18
3	0	4	13	5	6	1	29
4	1	5	13	5	3	0	27
Total	25	46	95	11	12	3	192

Q3 \ ST7	1	2	3	4	5	6	Total
1	10	25	53	3	7	2	100
2	6	11	19	3	4	0	43
3	5	5	17	2	1	1	31
4	4	5	6	3	0	0	18
Total	25	46	95	11	12	3	192

Q4 \ Q7	1	2	3	4	5	Total
1	70	19	9	0	2	100
2	21	7	8	0	0	36
3	13	12	8	3	1	37
4	9	3	5	1	1	19
Total	113	41	30	4	4	192

Q4 \ AH7	1	2	3	4	5	Total
1	109	5	3	0	1	118
2	0	17	1	0	0	18
3	4	19	5	1	0	29
4	0	0	21	3	3	27
Total	113	41	30	4	4	192

Q4 \ ST7	1	2	3	4	5	Total
1	77	15	6	1	1	100
2	18	14	11	0	0	43
3	9	10	7	3	2	31
4	9	2	6	0	1	18
Total	113	41	30	4	4	192

ANNEXE C

DECOMPOSITION EN VALEUR SINGULIAIRE D'UNE MATRICE :

- **Vecteur singulier** : soit une matrice A_{nxm}, avec n>m et le rang k<m, vecteurs x_i et y_i satisfassent l'équation suivante sont dit vecteur singulier de A :

$$\begin{aligned} Ay_i &= x_i\psi_i \\ A'x_i &= y_i\psi_i \end{aligned} \qquad (1)$$

L'équation (1) implique :

$$\begin{aligned} A'Ay_i &= y_i\psi_i^2 \\ AA'x_i &= x_i\psi_i^2 \end{aligned}$$

Sous forme de matrice, $X = \{x_1, \cdots, x_k\}$, $Y = \{y_1, \cdots, y_k\}$, l'équation (1) s'écrit .

$$\begin{aligned} AY &= X\Psi \\ A'X &= Y\Psi \end{aligned} \quad avec\ X'X = I,\ Y'Y = I \qquad (2)$$

- **Vecteur singulier** : Ψ est une matrice diagonale dont l'élément diagonal non négatif ψ_i est appelé valeur singulière de A.

L'équation (2) implique :

$X'AY = \Psi$ où Ψ est appelé la forme canonique de A.

$A = X\Psi Y'$ est appelé la Décomposition de Valeur Singulière (noté SVD) de A.

Pour une matrice A_{nxm} le rang k avec n>m et m>=k, le SVD de A :

$$A = X\Psi Y',$$

où X_{nxk} satisfait $X'X = I$, Y_{mxk} satisfait $Y'Y = I$, Ψ_{kxk} est une matrice diagonale avec ψ_i>0.

L'équation (2) multipliée respectivement par A' et A, on obtient :

$$\begin{aligned} A'AY &= A'X\Psi = Y\Psi^2 \\ AA'Y &= AY\Psi = X\Psi^2 \end{aligned}$$

Les vecteurs colonnes de Y sont appelés **vecteur propre** de matrice $(A'A)$, les éléments diagonals de Ψ^2 sont appelés **valeurs propres** correspondantes. Si on garde la convention de normalisation des 2I, X peut être calculé par $X = AY\Psi^{-1}$ où Ψ a p valeurs propres non nulle, Y a p vecteurs propres. Si A est une matrice de valeur réelle, ψ_i est une valeur réelle également. Donc la valeur propre ψ_i^2 de $(A'A)$ n'est jamais négative.

BIBLIOGRAPHIE

I. SUR LE TRAITEMENT DES DONNEES MANQUANTES

Anderson E.B., Hand D., Heiser J.W., Langeheine R., Meuleman J J , Saporta G. & Whittaker J. (1993), *Analysis of categorical data ; Theory and applications*, The 4th Course in the ECAS Program, Leiden, The Netherlands.

Beale E.M.L. & Little R.J.A. (1975), Missing values in multivariate analysis. *J. Roy. Stat. Ass., Ser. B*, **37**, *pp* 129-145.

Benali H. & Escofier B. (1987), Stabilité de l'analyse factorielle des correspondances multiples en cas de données manquantes et de modalités à faibles effectifs, *Rev. Statistique Appliquée*, XXXV(1), *pp* 41-52.

Benali H. (1988), Données manquantes et modalité à faible effectif en analyse des correspondances multiples et conditionnelle, in Data analysis and informatics (E.Diday ed.), V. Elsevier Science Publishers B.V., North-Holland.

Benzécri J.P. et al. (1980), *Pratique de l'analyse des données*, volume 3, Dunod, Paris.

Berger Y. (1995), Estimation de la variance dans le cas d'un plan de sondage à probabilités inégales sans remise : L'algorithme de CHAO, 27èmes Journées de Statistique, Jouy en Josas.

Bouroche J.M. & Saporta G. (1980), *L'Analyse des Données. Collection "Que sais-je ?"* , 5ème édition. Presses Universitaires de France, Paris.

Buck S.F. (1960), A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J. Roy. Stat. Soc., Ser. B, 22, pp* 302-306.

Buuren S.V. & Van Rijckevorsel J.L.A. (1991), Fast least squares imputation of missing data, *Psychometrics and Research Methodology*, Department of Psychology, Leiden University, The Netherlands.

Buuren S.V. & Van Rijckevorsel J.L.A. (1992), Imputation of missing categorical data by maximizing internal consistency, *Psychometrika, vol.57, n°.4, pp.*567-580.

Buuren S.V. & Van Rijckevorsel J.L.A. (1992), Data augmentation and optimal scaling, Statistiek en Informatica, *Statistiekreeks 03*.

Buuren S.V. (1992), Mistress V1.17 Documentation, Statistiek en Informatica, *Statistiekreeks 07*.

Caron N. (1996), Les principales techniques de correction de la non-réponse et les modèles associés, Document de travail n° ***9604***., INSEE.

Castellano R. (1993), Missing data imputation : The case study of an italian income survey, Proceedings 49th Session of ISI, book 1 *pp.*217-218.

Celeux G. (1988), Le traitement des données manquantes dans le logiciel SICLA, Rapport de Recherche INRIA, N°102.

Celeux G. & Diebolt J. (1988), A random imputation principle : the stochastic EM algorithm, Rapport de Recherche l'INRIA, N° 901.

Christoffersson A. (1965), *A method for component analysis when the data are incomplete.* Seminar communication. University Institute of statistics, Uppsala.

Cleroux R. & Ducharme G.R. (1989), Vector correlation for elliptical distributions, Comm. Statist. *Theory Meth. 18, pp* 1441-1454.

Crettaz de Roten F. (1993), Données manquantes en statistique multivariée: Une nouvelle méthode basée sur le coéfficient RV, *Thèse de l'Ecole Polytechnique Fédérale de Lausanne*, Lausanne.

Crettaz de Roten F. (1993), Imputation de données manquantes à l'aide du coefficient RV, Proceedings 49th Session of ISI, book 1, *pp*.217-218.

Crettaz de Roten F. (1995), Une méthode robuste d'imputation des données manquantes basée sur le coefficient RV, 27èmes Journées de Statistique, Jouy en Josas.

Crettaz de Roten F. & Helbling J.M. (1996), Données manquantes et aberrantes · le quotidien du statisticien analyste de données, *Rev. Statistique Appliquée, XLIV (2) pp*.105-115, Paris.

Morineau A. (1996), Reconstitution des données manquantes, Rapport de Recherche, C.I.S.I.A., Paris.

Daniel H. & Freeman J.R. (1987), *Applied categorical data analysis*, Marcel Dekker Inc., New York.

Daniel K., Gres J.D., Graham K. & Singh M.P. (1990) : *Panel surveys*, Wiley.

Dear R.E. (1959), A principal component missing data method for multiple regression models. System Development Corporation, Technical Report SP-86.

De Leeuw J. (1973), *Canonical analysis of categorical data*. DSWO, Press.

De Leeuw J. & Van der Heijden, P.G.M. (1987), Correspondence analysis of incomplete contingency tables. Communication personnelle.

Dempster A.P. Laird N.M. & Rubin D.B. (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Stat. Ass., Ser. B. 39, pp* 1-38.

Dempster A.P. & Rubin D.B. (1983), Overview, in Incomplete data in sample surveys, ***vol.2***. *Theory and annotated bibliography*, New York : Academic Press.

Deville J.C. & Särndal C.E. (1994), Variance estimation for the regression imputed Horvitz-Thompson estimator, *Journal of Official Statistics, vol. 10. n°4, pp.*381-394, Statistics Sweden.

Deville J.C. (1995), Non réponses et données manquantes, Compte rendu de "Méthodes d'Enquêtes de l'I.N.E.D.", Paris.

Droesbeke F., Fichet B & Tassi P. (1987), *Les sondages*, Paris, Economica.

Droesbeke J.J. & Lavallée P. (1995), La non-réponse dans les enquêtes probabilistes, 27èmes Journées de Statistique, Jouy en Josas.

El-Faouzi N E. (1992), Extensions non-linéaires de l'analyse en composantes principales, *Thèse de docteur* Université Montpellier II.

Escofier B. (1981), Traitement de questionnaires avec non-réponses et analyse des correspondances avec marges modifiées et analyse multicanonique avec contrainte. Rapport de Recherche I.R.I.S.A. n° 146.

Escoufier Y (1973), Le traitement des variables vectorielles, *Biometrics* **29**, *pp.*751-760.

Federspiel C.F., Monroe R.J. & Greenberg B.G. (1959), An investigation of some multiple regression methods for incomplete samples. *Mimeo serie n°.**236***. Institute of statistics, North Carolina.

Fevre P.(1979), Optimisation en classification automatique, Rapport de Recherche INRIA, Tome 1, Paris.

Fisher W.D. (1958), On grouping for maximum homogeneity *JASA, **53**, pp* 789-798.

Ford B.L. (1983), An overview of hot deck procedures, In *Incomplete data in sample surveys, vol.2.* Theory and annotated bibliography (Madow W.G , Olkin I. & Rubin D.B, Eds), New York : Academic Press.

Frane J.W. (1976), Some simple procedures for handing missing data in multivariate analysis, *Psychometrika* 41, *pp.* 409-415.

Fuchs C. (1982), Maximum likelihood estimation and model selection in contingency tables with missing data, *JASA. 77, pp* 270-278.

Gifi A. (1990), *Nonlinear multivariate analysis*, Wiley Chichester.

Gleason T.L. & Staelin R. (1975), A proposal for handling missing data, *Psychometrika 40. pp* 229-252.

Goldstein M. & Dillon R.W. (1978), *Discrete discriminant analysis*, J.Wiley, NewYork

Gouno E. (1995), Algorithmes stochastiques appliqués à l'estimation du taux de defaillance dans un contexte de données manquantes, 27èmes Journées de Statistique, Jouy en Josas

Gourieroux Ch. (1984), *Econométrie des variables qualitatives*, Economica, Paris.

Grangé D. & Lebart L. (1993), *Traitement statistique des enquêtes*, Dunod, Paris.

Greenacre M.J. (1984), *Theory and applications of correspondence analysis*, Academic Press, Paris.

Grosbras J M. (1987), *Méthodes statistiques des sondages*, Economica, Paris.

Haitovsky Y. (1968), Missing data in regression analysis. *J. Roy. Stat. Soc., Ser. B, 30, pp.*67-82.

Hartley H.O & Hocking R.R. (1971), The analysis of incomplete data (with discussion). *Biometrics, 27, pp* 783-823.

Herzog T.N. & Rubin D.B. (1983), Using multiple imputations to handle non-response in sample surveys, vol **2**, *Theory and annotated bibliography*, New York.

Hoffman G. (1993), Méthodologies de détection et de traitement des valeurs aberrantes dans un panel de pharmacies, Mémoire ISUP, Université Paris VI.

Jaupi L. (1992), Méthodes robustes en analyse en composantes pincipales, *Thèse de doctorat* C.N.A.M., Paris.

Johnson R.A. & Wichern D.W. (1982), *Applied multivariate statistical analysis*, Englewood Cliffs : Prentice-Hall.

Kim J. & Curry J. (1978), The treatment of missing data in multivariate analysis. In *Survey design and analysis (*D.F. Alwin Ed.), Sage Pubns., Beverly Hills, *pp* 91-116.

Lacourly N. (1974), Problèmes statistiques posés par le dépouillement d'enquêtes alimentaires, *Thèse de 3eme cycle* Faculté des Sciences de Paris

Little R.J.A. (1982), Models for nonresponse in sample survey, *JASA 77, pp* 237-250.

Little R.J.A. & Rubin D.B. (1987), *Statistical analysis with missing data*, New York, Wiley.

Little R.J.A. & Schluchter M.D. (1985), Maximum likelihood estimation for mixed continuous and categorical data with missing values, *Biometrika 72, pp* 497-512.

Little R.J.A. (1988), A test of missing compltely at random for multivariate data with missing values, *JASA **83**, pp* 1198-1202.

Meng X.L. & Rubin D.B. (1991), Using EM to obtain asymptotic variance-covariance matrices : *the SEM algorithm, JASA **86**, pp* 899-909.

Meulman J. (1982), *Homogeneity analysis of incomplete data*, Pswo Press.

Michael A., Hidiroglou, Douglasdrew J.& Gerald B.Gray (1993), Cadre pour l'évaluation et la réduction de la non-réponse dans les enquêtes, *Techniques d'enquête*, vol ***19***, n°1 *pp.91-105.*

Milidi M.A. (1993), A methode of estimation of missing values based on the redundancy index, Proceedings 49th Session of ISI, book 2 *pp.*178-179.

Miller A J (1993), Missing values in multiple regression, Proceedings 49th Session of ISI, book 2, *pp.*181-182.

Nadif M. & Govaert (1993), Binary clustering with Missing data, *Applied stochastic models and data analysis, vol.9, pp.*59-71.

Nora C. (1975), Reconstitution de données manquantes, *Thèse de doctorat,* Université de Paris VI.

Orchard T. & Woodbury M.A. (1972), A missing information principle : theory and applications. Proc. 6th. Berkeley Symp. Math. Stat. and Prob., **1**, *pp* 697-715, University of California Press, Berkeley.

Pages J. (1995), Eléments de comparaison entre l'analyse factorielle multiple et la méthode statistique, 27èmes Journées de Statistique, Jouy en Josas

Press S.J. & Scott A.J. (1976), Missing variables in bayesian regression, JASA 71, *pp.*366-369.

Roderick J.A. Little & Donald B.Rubin (1986), *Statistical analysis with missing data*, Wiley, New York.

Round J.I. (1993), Incorporating fragmented and incomplete data in the social accounts : A sam approach, Proceedings 49th Session of ISI, book 2. pp.355-356.

Rubin D.B. (1976), Inference and missing data, *Biometrika* ***63***, *pp* 581-592.

Rubin D.B. (1978), Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse, *Imputation and Editing of Faulty or missing survey data*, U.S. Departement of Commerce, *pp* 1-23.

Rubin D.B. (1981), The Bayesian boostrap, *Ann. Statist.* ***9***, *pp* 130-134.

Rubin D.B. (1987), *Mutiple imputation for nonresponse in surveys*, Wiley, New York.

Russo A. & Demetrio F.P. (1993), Weighting adjustement for unit non response in two stage stratified sampling, Proceedings 49th Session of ISI, book 2. pp.367-368

Saporta G. (1990), *Pobabilités, analyse des données et statistique*, Editions Technip, Paris.

Seber G.A F (1984), *Multivariate observations*, New York, Wiley.

Simon G.A. & Simonoff J.S. (1986), Diagnostic plots for missing data in least squares regression, *JASA* ***81***, *pp* 501-509

Timm N.H. (1970), The estimation of variance-covariance and correlation matrices from incomplete deta. *Psychometrika,* ***35***, *pp* 417-437.

Van Rijckevorsel J.L.A. & Bijleveld C.C.J.H (1992), A reader on Applying statistics in Public Health and Prevention, TNO Institute of Preventive Health Care., Leiden.

Van Rijckevorsel J.L.A. (1982), Canonical Analysis with B-spline, in Compstat 82, Proceeding in Comp.Stat.(H. Caussinus & Al, Eds), Physika-Verlag, *pp*. 393-398.

Vardi Y., Shepp L.A. & Kaufman L. (1985), A statistical model for positron emission tomography, *JASA* ***80***, *pp* 8-37.

Widmaiex U. (1994), Non response and weighting possibilities, Newsletter of the entreprise panel projet-the Panelist, *Eurostat.*

Wilks S.S (1932), Moments and distributions of estimates of population parameters from fragmentary samples. *Ann. Math. Stat. 3, pp* 163-195.

Wold H. (1966), *Nonlinear estimation by iterative least squares procedures.* In F.N. David (Ed), Festschrift Jerzy Neyman. Wiley, New York.

Yates F. (1933), The analysis of replicated experiments when field results are incomplete. Emp. *J. Exp. Agric., 1, pp* 129-142.

II. SUR LA FUSION DE FICHIERS

Abbruzzese L. (1991), Indagini sui media - single source o fusione ? AISM, Gennaio-Marzo.

Acott R. (1994), Even cleaner data using neural data ascription. SGSA Conférence.

Antoine J (1985), Fusion d'enquetes, CESP, Paris.

Antoine J. & Santini G. (1987), Fusion techniques : Alternative to single-source methods. European Research, August.

Arbeitsgemeinschaft media-analyse E.V.(AG.MA) und Media-micro-census Gmbh (1993). Der datentransfer im partnerschaftsmodell der Arbeitsgemeinschaft Media-analyse.

Baker K., Harris P., O'Brien J. (1989), Data fusion : An appraisal and experimental evaluation *Journal of the Market Research Society*, Vol **31**, No.2, 1989, *pp*.153-212.

Bermingham J. (1981), Similating readership data surveys in Great Britain-A review of current practices, First International Readership Symposium.

Boucharenc L. (1981), Les techniques de fusion Appliques aux Etudes CESP, Memorandum

Carpenter R. & Wilcox S. (1995), Data fusion in the British National Readership survey-An experiment, Mirror Group Newspapers & RSMB Television Researche Ltd.

Centre d'information sur les media ASBL (1994), Test de méthodes de fusion.

Dansk Media Komite. (1985), Fusion-An overview by an outside observer, Salzburg 3 - Readership Researche, Symposium.

Den Haute V. Vondermeesch I. & Locker R. (1994), Test de méthodes de fusion, Rapport de Recherche CIM, Bruxelles.

Francoz D. (1995), Estimation des cessations d'entreprises: Méthode et résultats, 5ème Journée de Méthodologie Statistique, INSEE.

Gonzalez P.L. & Rioux B. (1990), Selecting the best subset of variables in principal component analysis, Compstat, Physica-verlag Heidelberg for IASC.

Hendriekson K & Acott R. (1994), Beyond fusion-using neural networks to improve media targeting, Séminaire on "From door to door to satellite : Media reseach for more effective planning", Athens.

INSEE (1994), Appariements aléatoires de deux fichiers . Budgets de famille et revenus fiscaux, *Conseil Economique et social, CES/AC, **70/6***, Genève.

IPSOS (1996), Les méthodes de fusion, *Systems & technology*, n°**16**, Paris.

Laniel N. (1995), La refonte de l'échantillon de l'enquête sur la population active, Rapport de Recherche, Octobre, INSEE.

Lebart L. & Lejeune M. (1995), Assessment of data fusions and injections, Encuentro International AIMC sobre Investigacion de Medios, Madrid.

Lejeune M. & Lebart L. (1994), On the assessment of data fusions and injections, CESP, Paris.

Lejeune M. (1995), De l'usage des fusions de données dans les études de marché, Proceedings 50th Session of ISI, Tome LVI. pp.923-935, Beijing.

Lokker R. (1994), Les techniques de fusion : Comment les évaluer ? CIM News, Bruxelles

Paass G. (1986), Statistical match : Evaluation of existing procedures and improvements by using additional information. in *Microanalytic Simulation Models to Support Social and Financial Policy* (Eds G.H Orcutt, J.Merz et H. Quinke), pp.*401-422*, Amsterdam : Elsevier Science.

Parent M.C (1995), Une base d'analyse longitudinale de données d'entreprises : Problèmes et résultats, Rapport de Recherche INSEE.

Riandey B. (1993), Enquêtes de référence, greffe d'enquêtes et fusion entre fichiers d'enquêtes, Séminaires de Méthodes d'Enquêtes de l'I.N.E.D., Paris.

Rubin D.B. (1986), Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics, 4, pp* 87-94.

Santini G. (1983), Clustering techniques, Readership Research, Montreal.

Santini G. (1984), Fusion d'enquêtes ou rapprochement de données, IREP, Paris.

Santini G. (1986), An experiment to validate fusionned files obtained by the referential factorial method, ESOMAR, Helsinki.

Santini G. (1986), Fusion processes : A conceptual and pratical approache, SAMRA, Johannesburg.

Santini G. (1986), Méthodes de fusion Nouvelles réflexions, nouvelles expériences, nouveaux enseignements, IREP, Paris.

Santini G (1988), Validation of data fusion techniques · What can statistical theory do for us ? Readership Symposium, Barcelone.

Santini G. (1989), Fusion in perspective, Television Research INT. Symposium, Tarritown.

Saporta G. & Co V. (1996), Data fusion : A new method based on homogeneity analysis, Sino-French Workshop on Advanced Data Analysis Methods in Industry and Management, Beijing.

Singh A.C., Mantel H.J., Kinck M.D. & Rowe G. (1993), Appariement statistique : L'utilisation d'information supplémentaire comme solution de remplacement à l'hypothèse d'indépendance conditionnelle, *Techniques d'Enquête, pp.67-89.*

Sousselier J. (1995), Un programme de fusion de fichiers, STATIRO, Paris.

Verger D. & Contencin D. (1995), Les imputations économetriques dans l'enquête sur les Revenus Fiscaux, 5ème Journée de Méthodologie Statistique, INSEE.

Wendt F. (1976), 'Beschreibung Eine Fusion' Schriftenreihe Band 21, Gruner and Jahr A.G. and Co.May.

Wendt F. (1984), The AG.MA model, in Proceedings of the 2nd International Symposium Media Research, Montreal 83, Ed. H. Henry North-Holland, *pp* 393-403.

Wiegand J. (1986), The combining of two separately derived data-set into an integrated intermedia planning system . The German 'Model of Partnership', New Developments in Media Research., ESOMAR, Helsinki, Finland.