Fusion de fichiers : une nouvelle méthode basée sur l'analyse homogène

Gilbert Saporta, Vila Co

▶ **To cite this version:**

Gilbert Saporta, Vila Co. Fusion de fichiers : une nouvelle méthode basée sur l'analyse homogène. Gildas Brossier; Anne-Marie Dussaix. Enquêtes et sondages, Dunod, pp.81-93, 1999, 9782100040230. hal-01124586

HAL Id: hal-01124586
https://hal.science/hal-01124586

Submitted on 23 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FUSION DE FICHIERS : UNE NOUVELLE METHODE BASEE SUR L'ANALYSE HOMOGENE

Gilbert SAPORTA, Vila CO

Conservatoire National des Arts et Métiers
292, Rue Saint-Martin - 75141 Paris cedex 03 - France

RESUME

La technique de fusion est largement utilisée en Europe mais peu étudiée sur le plan statistique. Le champ d'application de la fusion de fichiers est très vaste, par exemple, dans le domaine du marketing (Riandey B., 1993) et des enquêtes d'audience-médias (Carpenter R. & Wilcox S., 1995). La fusion statistique de fichiers est un outil indispensable pour rassembler des informations provenant de différentes sources. C'est en fait une technique d'estimation de sous-tableaux manquants qui s'applique au cas où des blocs entiers de variables n'ont pas été observés.

Pour des données qualitatives nous étudions une méthode de fusion basée sur l'analyse homogène. L'évaluation de la méthode est faite simultanément sur données réelles et simulées selon deux critères distincts. Une comparaison avec une méthode classique est effectuée, au moyen d'un exemple sur données réelles.

La méthode proposée (Saporta G. & Co V. 1996, Co V.1997) est adaptée au problème général, où l'on n'exige pas que les données suivent un modèle statistique. Elle a une bonne qualité individuelle. Elle accepte des différences de structure et de taille des données entre les fichiers donneur et receveur. La plupart des méthodes existantes sont en principe des méthodes pour prévoir les comportements globaux. Les méthodes ayant une bonne qualité individuelle sont encore rares. Comme la validation des données reconstituées est une phase primordiale de la fusion, une technique d'évaluation des variables (qualitatives ordinales) reconstituées au niveau individuel a été proposée.

I. INTRODUCTION

Parfois nous nous trouvons dans la situation suivante :

- Fichier de données abîmé ou introuvable.
- Questions non répondues dans une enquête ou questions non posées parce que le questionnaire est trop long.
- Informations provenant de différentes sources.

Le but de la fusion des fichiers consiste à utiliser au mieux les informations existantes pour reconstituer les informations non renseignées. La fusion de fichiers peut nous fournir les simulations de ces informations manquantes dont nous avons besoin. Son principe est, à partir d'un bloc de questions répondues, et suivant sa relation avec le bloc de variables à reconstituer, d'estimer les valeurs des variables non renseignées.

La technique de la fusion est souvent utilisée dans le monde du marketing, surtout dans le domaine des études de media-marché : l'habitude de l'usage médiatique vis à vis du comportement de consommation de produits.

Exemple 1 - Parfois il est difficile d'interroger une même personne sur sa consommation télévisuelle, radiophonique 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 un même type de clientèle ?
- La télévision détourne- t- elle de la presse sa clientèle potentielle ?

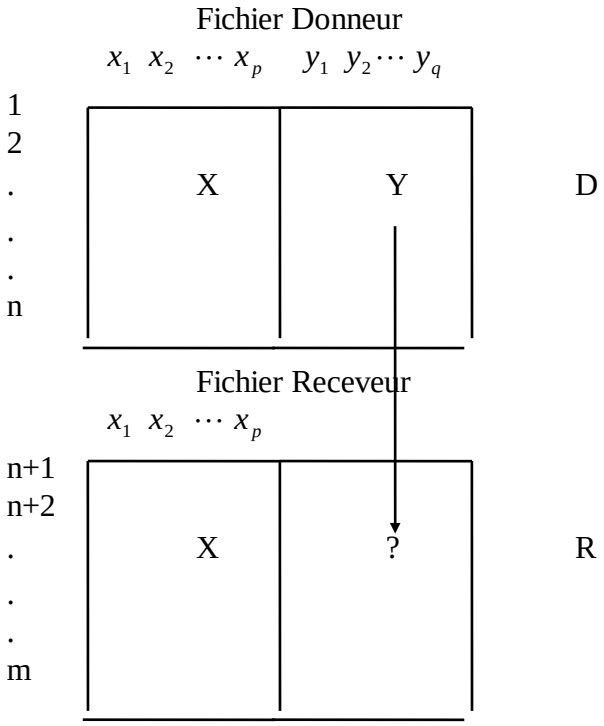
Exemple 2 - La fourniture d'estimation locale à partir d'une enquête nationale. Le panel de ménages décrivant des consommations et profils de consommation (y compris les patrimoines et équipements), joue le rôle de fichier donneur. Le fichier des communes, contenant la structure socio-économique de la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

II. GENERALITES SUR LA FUSION DE FICHIERS

La fusion de fichiers est une technique d'estimation de sous-tableaux manquants où des blocs entiers de variables n'ont pas été observés. Autrement dire, les données manquantes sont du genre variables manquantes dans un fichier, ces variables manquantes étant présentes dans un autre fichier de données.

Fichier donneur et fichier receveur, considérons deux sources de données qui contiennent des renseignements (ou variables) différents sur des individus différents. L'une de ces sources sert

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



Dans la plupart des méthodes existantes, on considère que la fusion de fichiers est un cas particulier de l'imputation par bloc qui consiste à donner à un receveur l'ensemble des valeurs d'un donneur pour les variables non renseignées. La fusion de fichier d'enquêtes peut être considérée comme une imputation à grande échelle en utilisant les variables communes des deux fichiers. On complète les enregistrements du fichier R en imputant des valeurs authentiques (observées) de Y à l'aide de l'information du fichier D en utilisant les relations entre Y et X.

Une technique de fusion fréquemment utilisée en France (par Statiro et IMS) est la Fusion basée sur "Référentiel Factoriel" (Santini, 1984) qui repose sur les points suivants :

- Variables critiques : une partie des variables communes dans les deux fichiers receveur et donneur sert 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 l'individu du fichier receveur ses donneurs éligibles.

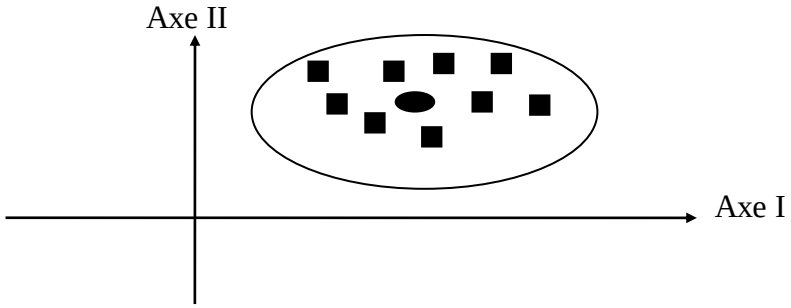
- Variables de rapprochement : une partie des variables communes, par un calcul de distance, permettant de choisir pour chaque receveur le donneur éligible le plus proche par rapport à ces variables dans les méthodes classiques.

1.Recherche du référentiel factoriel :

Ce processus commence de la façon suivante : On effectue une analyse des correspondances multiples sur le tableau des variables critiques communes à l'ensemble des données disponibles (donneur + receveur) et on conserve les k premiers axes de l'analyse, ce qui permet de positionner les observations dans un espace R^k et de calculer les distances entre points sur les coordonnées.

2.Recherche de voisinage :

Pour chaque receveur, on sélectionne un ensemble de donneurs dans un voisinage du receveur.



3.Choix final d'un donneur :

1).Parmi les donneurs potentiels, on choisit celui qui ressemble le plus au receveur sur les variables de rapprochement, qui sont des variables signalétiques comme: l'âge, le sexe, la C.S.P. etc.

2).On évite d'utiliser trop souvent le même donneur en utilisant une fonction de pénalité (Santini, 1986) et le donneur le moins copié dans la méthode 'Statiro' est finalement retenu.

III. ESTIMATION DES DONNEES PAR L'ANALYSE HOMOGÈNE

Un autre type de fusion consiste à estimer simultanément les valeurs de chacune des variables sans imputer un bloc entier de valeurs. La méthode que nous proposons appartient à cette dernière catégorie.

L'analyse homogène, développée par des chercheurs Néerlandais (De Leeuw, Gifi, Meulman, Van Buuren et Van Rijckevorsel) est une présentation de l'analyse des correspondances multiples qui se prête bien à une extension pour des données manquantes. L'analyse homogène repose sur la maximisation d'un critère de cohérence interne entre coordonnées des unités statistiques et des modalités des variables.

III.1 Transformation optimale de données et analyse homogène

Quand les variables mesurent plus ou moins la même propriété, il est possible de remplacer les observations sur les différentes variables par une valeur d'une variable synthétique sans perdre trop d'informations. La petitesse des pertes varie en fonction de l'homogénéité des variables. Pour évaluer le succès de cette substitution, on définit un critère d'homogénéité et une fonction de perte. Le processus de maximisation d'homogénéité des variables conduit à l'analyse homogène qui est similaire à de l'analyse des correspondances multiples.

- Pour des variables quantitatives complètes :

Si nous avons *n* unités et *m* variables y_j, j=1, 2, ... , m, $y_j = \begin{pmatrix} y_{j1} \\ \vdots \\ y_{jn} \end{pmatrix}$

La variable transformée $\phi_j(y_j) = \begin{pmatrix} \phi_j(y_{j1}) \\ \vdots \\ \phi_j(y_{jn}) \end{pmatrix}$,

Le score individuel $X = \frac{1}{m}\sum_{j=1}^{m} \phi_j(y_j) = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

Nous cherchons une transformation ϕ la plus proche possible de X.

La variation totale des données transformées est :

$$\sum_{j=1}^{m} (\phi_j(y_j))^2 = mx^2 + \sum_{j=1}^{m} (x - \phi_j(y_j))^2$$

T = B + W = Variation (interclasse) + Variation (intraclasse)

La variation interclasse B mesure la discrimination entre les différentes unités. La variation intraclasse W se rapporte à un manque d'homogénéité des variables transformées :

$$\phi_1(y_{1i}) \neq \phi_2(y_{2i}) \neq \cdots \neq \phi_m(y_{mi})$$, i=1, ... , n

Le coefficient d'homogénéité $\eta = \frac{B}{T}$, mesure à quel point le score individuel peut être considéré comme représentant de chaque unité. Plus la valeur η est grande, mieux $x_i - x_{i'}$

représente la différence entre unités i et i'. En maximisant $\eta = \frac{B}{T}$, qui est un critère de discrimination entre individus, nous cherchons à transformer les variables de façon optimale.

La fonction de perte d'homogénéité se définit comme suit:

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

- Pour des variables qualitatives complètes :

Soit G_j le tableau d'indicatrices de la variable j : $G_j = (g_{il})_{n \times j_k}$, j=1, 2, ... , m

$g_{il} = \begin{cases} 1 & \textbf{si la i ième observation tombe dans la catégorie l de la variable j} \\ 0 & \textbf{sinon} \end{cases}$

n - nombre d'unités, j_k - nombre de catégories de variable j.

Soit G = (G_1,G_2,...,G_m), le tableau disjoinctif complet,

alors, si $y_j = \begin{pmatrix} y_{j1} \\ \vdots \\ y_{jj_k} \end{pmatrix}$ est la quantification des catégories de la variable j ,

on a $G_j Y_j = \begin{pmatrix} g_{j1} \\ \vdots \\ g_{jn} \end{pmatrix}$ Score individuel sur la variable j, j=1, 2, ... , m

S'il y a homogénéité parfaite, alors :

$$X = G_1 y_1 = G_2 y_2 = \cdots = G_m y_m = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$ X- score individuel où $X = \frac{1}{m} \sum_{j=1}^{m} G_j Y_j$.

La fonction de perte d'homogénéité est definie par :

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

Le minimum $\sigma(x, \phi)$ sur X, ϕ vaut $\sigma(*,*) = 1 - \lambda_+^2$ (Gifi A., 1990), où $\lambda_+^2 = \frac{1}{m}$ fois de la plus grande valeur propre de matrice de corrélation des données transformées. Nous cherchons le score individuel optimal par rapport à la quantification des variables et la quantification des

variables optimale par rapport au score individuel. La minimisation $\sigma(X,Y)$ sur X et Y (J. Meulman, 1982) nous donne :

$$\begin{cases} \tilde{y}_j = D_j^{-1} G_j' X^t \\ \tilde{X} = \dfrac{1}{m} \sum_{l=1}^{m} G_l y_l \\ X^t = \tilde{X} (\tilde{X}' \tilde{X})^{-\frac{1}{2}} \quad \text{avec } U' X = 0 \text{ et } X' X = 1 \end{cases}$$

III.2 Quantification de données qualitatives incomplètes et imputation homogène en maximisant le coefficient d'homogénéité [Buuren S.V. & Van Rijckevorsel J.L.A., 1992].

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.

- La fonction de perte d'homogénéité se définit 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 des réponses complètes
g_j* : Matrice indicatrice incomplète de variable j (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 , ... , g_m de (1).

Le problème est de savoir où l'on impute "1" dans le vecteur manquant g_j*. Van Buuren S. & Van Rijckevorsel J.L.A. (1992) proposent un algorithme "*K-means*" modifié pour estimer les données manquantes. Nous allons utiliser l'algorithme 'K-means' en minimisant la fonction de perte d'homogénéité $\sigma(*)$.

Algorithme de "K-means" modifié :

Supposons 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 pour minimiser la perte d'homogénéité. Soit d_s, d_t, les nombres d'observations des catégories s et t de variable j ; y_s, y_t, les quantifications respectives des catégories. L'unité i a un score x_i, si nous changeons l'imputation de la catégorie s à t. Ficher(1958) montre que la nouvelle perte est égale à :

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

$si\ \frac{d_t(x_i - y_t)}{d_t + 1} < \frac{d_s(x_i - y_s)}{d_s - 1}$, nous imputons la catégorie t à la place de s. En même temps, y et y_s changent aussi,

$$\begin{cases} \hat{y}_s = y_s + \frac{x_i - y_s}{d_s - 1} \\ \hat{y}_t = y_t + \frac{x_i - y_t}{d_t - 1} \end{cases}$$, Mais nous devons nous assurer que ds >=1.

Exemple d'analyse homogène pour l'imputation des données manquantes

(Van Buuren S. & Van Rijckevorsel J.L.A., 1992)

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 <u>la liste des valeurs du coefficient</u> 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 solution optimale est donc *x=b (Revenu bas) ; y=h (Revenu haut); z=a(Car am).*

Cette méthode produit un très bon résultat à condition que le niveau d'homogénéité soit assez élevé.

III.3. L'analyse homogène 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 partie précedente, 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.

Nous introduisons la fusion de fichiers basée sur l'analyse homogène à la seule condition que le pouvoir prédictif des variables communes par rapport aux *variables à transférer* soit 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) montre que, dans le cadre de fusion par imputation, la qualité du résultat est optimale pour un nombre ni trop grand et ni trop petit, de variables critiques.

Puisque nous avons détaillé la méthode de l'analyse homogène dans la partie (III.2) et que nous n'avons pas changé le principe de la méthode, soulignons ici quelques aspects différents : 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.

IV. ÉTUDE COMPARATIVE

Pour valider le résultat, il faut disposer d'un fichier pour lequel on connaît toutes les valeurs. Pour cela, on va prendre un fichier de données et faire comme si on ignorait certaines variables que l'on reconstitue par fusion. Le résultat est mesuré et évalué selon deux critères distincts :

- *niveau global :* les distributions des variables 'réelles' et reconstituées doivent être proches ; les relations entre variables 'réelles' doivent être les plus proches possibles après reconstitution [V. CO, 1997].

- *niveau individuel :* on compare individuellement valeurs estimées et valeurs réelles.

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

- Divisons au hasard un fichier complet en s parties comme s fichiers de receveurs.
- Fusionnons s fois pour chaque fichier de receveur.
- Calculons le taux d'erreur standard.

Pour des variables qualitatives ordinales, Saporta G. & Co V. (1996) ont proposé un **coefficient de proximité** pour la validation au niveau individuel construit comme suit :

L'idée est de pénaliser 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 :

	Valeur affectée				
Vraie valeur	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 coûts sont normalisés de telle sorte que le maximum $c_{ij}=1$.

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

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

$\tilde{d}$ est un écart entre données réelles et données affectées pour une variable. 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 $\bar{d}$ la moyenne des coûts d'erreur associée à une règle donnée sur l'ensemble de cases reconstituées, $\bar{d} - \tilde{d}$ mesure le gain par rapport à l'affectation aléatoire.

Exemple 1 - 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. Moyennes des corrélations des variables : quatre niveaux entre 0 et 1.
2. Structure de population entre les fichiers D et R (identique ou différente).

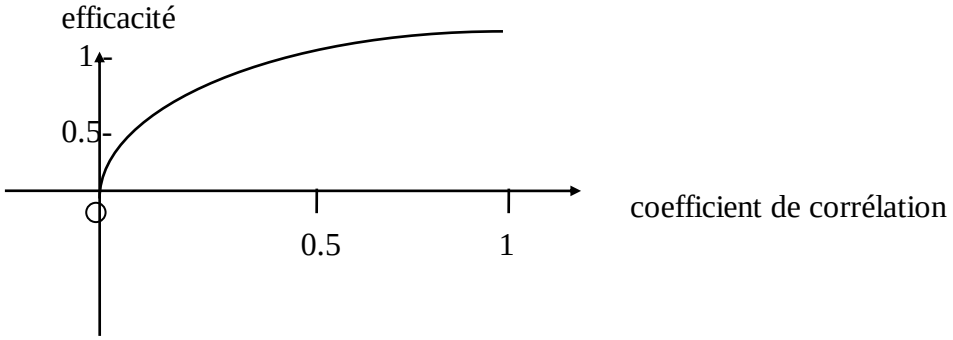
La combinaison des facteurs nous donne 16 cas différents. Pour chaque combinaison on simule 50 jeux de données de 130 individus avec 8 variables comme fichier originel, 100 individus comme fichier donneur et 30 individus comme fichier receveur, 5 variables communes prédictives, 3 variables à transférer.

Nous utilisons notre coefficient de proximité pour évaluer les résultats de notre méthode. Le résultat ne varie pas sur le deuxième critère, en fait nous avons obtenu les mêmes résultats que les deux fichiers soient (ou non) la même structure.

Tab.1. Le coefficient de proximité $\bar{d}$ sur les 3 variables transférées

	Corrélation entre deux blocs			
	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

La qualité du résultat varie en fonction de la relation entre les deux blocs :



La qualité du résultat varie en fonction de la relation entre les deux blocs. Plus le coefficient de corrélation entre deux blocs est grand, meilleur est le résultat. Nous pouvons ainsi déterminer si les données reconstituées provenant de la fusion sont significatives. Il est difficile de comparer la méthode de fusion avec celle du tirage aléatoire car le résultat de cette dernière dépend fortement de la structure de variables, est instable et les relations entre variables ne sont pas considérées. En comparant l'analyse homogène avec le tirage aléatoire, nous

observons que lorsque la corrélation moyenne entre deux blocs est inférieure à la valeur critique sous l'hypothèse d'indépendance, le tirage aléatoire est meilleur. Dans le cas contraire la méthode « analyse homogène » est supérieure, à l'exception de quelques cas extrêmes de structure de données (variables), avec une corrélation moyenne entre deux blocs inférieure à 0.3.

Exemple 2 - 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 des 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).

a. Comparaison au niveau individuel avec une méthode de tirage aléatoire

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 de données bien classées du tirage aléatoire est théoriquement de 33%.

b. Comparaison avec la méthode classique « Référentiel factoriel »

- 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. Les résultats montrent bien les caractéristiques de deux familles de méthodes:

Niveau individuel : le résultat obtenu par l'analyse homogène a un bon niveau individuel, selon le taux de données bien classées, qui est meilleur que la fusion Statiro :

Méthode	*Bien classé*
Analyse homogène	*54%*
Statiro	*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. 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

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.

Notre conclusion sur cette comparaison est donc la suivante : selon l'objectif de la fusion on choisira la méthode adéquate : la fusion par l'analyse homogène ou celle par la recherche de voisinage.

V. 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,

- elle s'adapte à différentes structures de populations des fichiers,

- elle s'adapte à des différences de taille des fichiers.

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.

BIBLIOGRAPHIE

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.

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***.

Carpenter R. & Wilcox S. (1995), Data fusion in the British National Readership survey-An experiment, Mirror Group Newspapers & RSMB Television Researche Ltd.

Co V. (1997), Méthodes statistiques et informatiques pour le traitement des données manquantes, *Thèse de doctorat*, C.N.A.M, Paris.

De Leeuw J. (1973), *Canonical analysis of categorical data*. DSWO, Press.

Gifi A. (1990), *Nonlinear multivariate analysis*, Wiley Chichester.

INSEE (1994), Appariements aléatoires de deux fichiers : Budgets de famille et revenus fiscaux, *Conseil Economique et social, CES/AC, **70/6***, Genève.

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.

Lokker R. (1994), Les techniques de fusion : Comment les évaluer ? CIM News, Bruxelles.

Meulman J. (1982), *Homogeneity analysis of incomplete data*, Dswo Press.

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.

Santini G. (1986), Fusion processes : A conceptual and pratical approache, SAMRA, Johannesburg.

Santini G. (1986), An experiment to validate fusionned files obtained by the referential factorial method, ESOMAR, Helsinki.

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.

Saporta G. (1990), *Probabilités, analyse des données et statistique,* Editions Technip, Paris.

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*.

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.