FUSION ET GREFFES DE DONNEES

Gilbert Saporta

Conservatoire National des Arts et Métiers
Chaire de Statistique Appliquée-CEDRIC
saporta@cnam.fr

Nicolas Fischer

EDF Division Recherche et Développement et CNAM-CEDRIC
Nicolas.Fischer@edf.fr

Cet article fait partie des Actes des 5èmes JOURNEES MODULAD qui ont eu lieu les 16 et 17 novembre 2000 à E.D.F. (Clamart).

Résumé :

La fusion statistique de fichiers a pour but de compléter un fichier « receveur » où certaines variables ne sont pas renseignées (questions non posées) à l'aide d'un ou plusieurs fichiers « donneurs » portant sur d'autres individus. Le fichier donneur comprend bien sûr des variables communes ainsi que les variables d'intérêt renseignées pour tous les individus. Les remplacements de données manquantes se font soit par des méthodes d'imputation basées sur des proches voisins (injection) soit à l'aide de méthodes explicites de type régression.

Les greffes d'enquêtes poursuivent des objectifs proches, en ce sens qu'il s'agit par exemple de positionner des résultats d'un sondage (une analyse factorielle) sur ceux d'un autre en utilisant des variables passerelles, mais sans nécessairement chercher à estimer les données manquantes. Cet exposé présentera la problématique, les principales techniques utilisées, les critères de validation, ainsi que les dangers potentiels ....

1.Introduction

Fusions de fichiers et greffes d'enquêtes sont liées à la combinaison de données provenant de sources différentes: ces techniques font donc partie, mais en amont, du processus de « data mining » ou d'extraction des connaissances. Le problème ici n'est pas d'extraire de l'information d'une base unique mais de fusionner différentes bases provenant d'enquêtes, de sources administratives, de fichiers clients, de données socio-économiques agrégées, etc. Chaque base peut être constituée d'unités statistiques différentes ou d'agrégation de ces unités à différents niveaux.

1.1 Fusion de fichiers

L'objectif d'une fusion de fichiers est d'obtenir une base unique réunissant tous les individus concernés, où toutes les variables sont renseignées. Le problème peut se formaliser en termes de deux fichiers: le premier contient les observations de $p+q$ variables mesurées sur n_0 unités, le second des observations sur un sous-ensemble de p variables pour n_1 unités. Souvent n_0 est faible par rapport à n_1. Si les variables communes sont désignées par des tableaux X, on a le schéma suivant:

X_0	Y_0
X_1	?

Il s'agit alors de remplir la partie vide de la table précédente: c'est un problème assez spécial d'estimation de données manquantes où un bloc entier manque car les variables n'ont pas été collectées.

La fusion de sources de données est un ensemble de techniques qui se sont développées à partir des années 80 dans le domaine des études de marché ([BAK89]), et plus spécialement dans les études média. D'une part dans ces types d'enquêtes il est souvent impossible de poser toutes les questions d'intérêt qui sont trop nombreuses, et afin de réduire la charge pesant sur les enquêtés on procède avec des échantillons indépendants où les questions sont partagées. D'autre part, les utilisateurs d'enquêtes d'audience des médias souhaitent pouvoir rapprocher des données provenant de fichiers différents afin de réduire les budgets consacrés aux enquêtes, de connaître la position des non répondants dans une enquête par rapport à celle des répondants.

On constate actuellement un intérêt croissant dans ces techniques d'enrichissement de fichiers avec la disponibilité de sources d'information: citons les études consommateurs où on dispose d'un coté de données exhaustives de transactions et de l'autre d'enquêtes de satisfaction. L'usage de sources multiples a attiré également l'attention des militaires pour des problèmes de détection voir par exemple le congrès Fusion2000[1] et le nouveau journal « Information Fusion ».

1.2. Greffe d'enquêtes

Dans cette méthodologie proche, on ne cherche pas à estimer les variables manquantes mais à "coller" ou projeter les résultats d'une enquête S1 sur l'espace de référence défini par une enquête S0.

Proche des méthodes procustéennes, la greffe d'enquête a été développée dans le contexte de l'ACP et de l'ACM. Son but est d'ajouter en points supplémentaires sur S0 les individus et les variables de S1 ([BON86]).

Le schéma est ici :

X_0	Y_0	
X_1		Y_1

X_0 et X_1 ont les mêmes variables tandis que Y_0 et Y_1 sont des tableaux de variables spécifiques à chaque enquête.

Pour des variables numériques, la procédure est la suivante:

- On effectue une ACP de $(X_0\ Y_0)$, on retient k composantes C_0 que l'on régresse sur les variables communes X_0, d'où la formule d'approximation $\hat{C}_0 = X_0 b_0$
- On positionne les unités de S1 dans le plan principal de S0 par $C_1 = X_1 b_0$
- On positionne les variables de Y_1 dans le cercle des corrélations de S0 en calculant les corrélations entre Y_1 et C_1

[1] http://www.onera.fr/fusion2000)

On utilise donc deux fois la méthode des points supplémentaires (les variables supplémentaires sont positionnées grâce aux individus supplémentaires) combinée avec une approximation des composantes principales. Pour obtenir de bons résultats il est nécessaire que X_0 et Y_0 soient bien corrélées , pour pouvoir reconstituer les composantes principales de S0, et que X_1 etY_1 soient aussi bien corrélées.

2. Les modèles et les différentes méthodes utilisées pour la fusion de données

2.1. Contexte

Le but de la fusion de fichiers est d'utiliser les informations existantes dans un fichier pour reconstituer les informations absentes d'un autre fichier. Le principe est d'estimer au mieux les valeurs de variables non-renseignées à partir d'un bloc de variables renseignées assez corrélées avec le bloc de variables à reconstituer.

La fusion de données est une technique particulière du traitement des données manquantes.

Pour traiter les valeurs manquantes ou inexistantes dans un fichier de données, deux approches sont envisageables.

La première méthodologie consiste à compléter la non-réponse par une valeur plausible. C'est ce que l'on appelle plus communément les méthodes d'imputation.

La seconde approche est de considérer uniquement la population des répondants et de leur affecter des pondérations nécessaires pour compenser les non-réponses, les interviewés absents par extrapolation. On parle alors de techniques de repondération ([LIT87]).

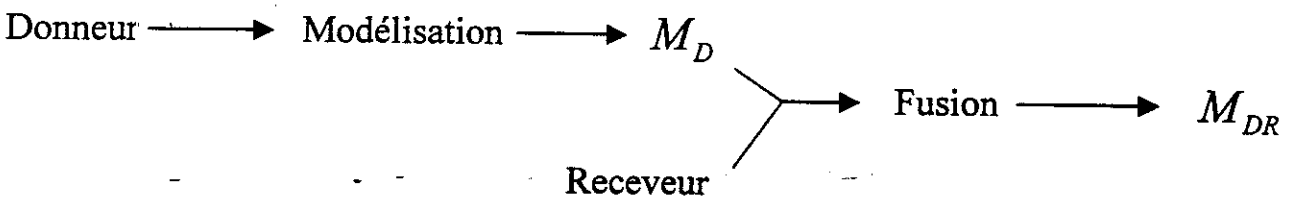
Ici nous traiterons plus particulièrement des techniques d'imputation.

Toute procédure de traitement des données manquantes doit, soit explicitement, soit implicitement modéliser le processus de création des données manquantes.

Deux schémas sont ainsi envisageables pour le transfert des variables spécifiques dans le fichier receveur selon que le modèle de distribution des variables du fichier donneur soit explicite ou implicite.

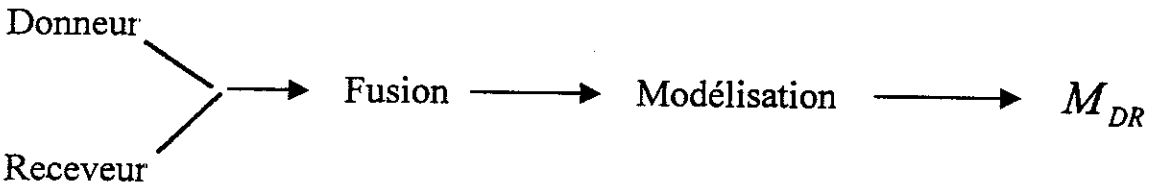
Modèle explicite : schéma modélisation / fusion :

Une modélisation est construite sur le fichier donneur et l'information est transférée aux individus receveurs suivant ce modèle.



Modèle implicite : schéma fusion / modélisation :

Une fusion est réalisée directement entre les deux fichiers sans avoir de modèle explicite et ensuite le fichier résultat est optimisé.



Par ailleurs, pour réaliser une fusion dans de bonnes conditions et obtenir une qualité satisfaisante sur les données reconstituées, il est nécessaire de vérifier préalablement que la taille de la population du fichier donneur est suffisamment importante par rapport au fichier receveur et que les variables communes (c'est à dire celles présentes dans les deux échantillons donneur et receveur à la fois) et les variables spécifiques (c'est à dire les variables de l'échantillon donneur qui sont à reconstituer dans l'échantillon receveur) possèdent des liaisons relativement fortes entre elles.

2.2. L'estimation basée sur des modèles explicites

On se placera dans un cadre paramétrique puisque les méthodes utilisées vont tenir compte des données existantes, c'est à dire recueillies au cours des diverses enquêtes. En effet, l'estimation de paramètres d'une loi d'une variable présentant des données manquantes devra se référer à toute l'information existante sur cette variable et également sur les autres variables du fichier.

Chaque valeur manquante peut être estimée grâce à des techniques classiques de régression (régression linéaire, régression logistique, modèle linéaire généralisé). Chaque variable de $\mathbf{Y_0}$ est modélisée à partir des coordonnées de $\mathbf{X_0}$ qui joue le rôle de variables explicatives. Le modèle généré est alors appliqué au fichier receveur. Le principe de ces méthodes est de prédire les valeurs manquantes en utilisant un modèle de régression adapté aux variables observées. Il y a plusieurs possibilités et nous allons en citer quelques unes :

- **une régression simple** en prenant la variable la plus corrélée.
- **une régression multiple** en prenant le meilleur sous-ensemble de variables explicatives, utilisant un modèle pas à pas, ou la méthode de Furnival et Wilson d'exploration optimisée de toutes les possibilités.
- **une analyse de variance** , cas particulier de la régression lorsque la variable explicative X est nominale et la variable à expliquer est quantitative.
- **une analyse discriminante,** lorsque la variable à expliquer est multiclasse ou **une régression logistique** lorsque la variable expliquée est dichotomique. On impute alors par la catégorie la plus probable.

Toutes ces techniques, bien que simples présentent au moins deux inconvénients majeurs: les variables sont estimées une par une et non conjointement: ainsi ces techniques d'estimation ne prennent pas en compte les corrélations éventuelles entre les variables, ce qui peut induire des résultats incohérents. Si aucune vérification des résultats trouvés n'a été prévue à la fin du processus d'estimation, des résultats incohérents peuvent se produire, comme par exemple un jeune homme de 20 ans qui serait retraité.....

On peut aussi appliquer **la méthode du maximum de vraisemblance**. Le principe de cette méthode, sous l'hypothèse que les données qualitatives proviennent d'un échantillon d'une variable aléatoire multinomiale, est le suivant :

Les paramètres de la loi multinomiale sont estimés par l'algorithme EM ([DEM77]) que l'on décrit ici dans son principe. Partant d'une estimation des paramètres de la loi, il s'agit d'un algorithme itératif utilisant alternativement deux étapes. L'étape E (E comme espérance) consiste à déterminer l'espérance conditionnelle de chaque donnée manquante sachant les données observées et l'estimation courante des paramètres. L'étape M (M comme

maximisation) consiste à calculer les estimateurs du maximum de vraisemblance des paramètres, les formules faisant usage des lois conditionnelles des données manquantes. De manière naturelle, à la convergence de l'algorithme EM, on attribue à chaque donnée manquante la valeur la plus probable pour l'estimation obtenue des paramètres de la loi multidimensionnelle. De cette façon, tous les individus qui présentent des données manquantes pour les mêmes variables sont complétés de manière identique. Mais la méthode du maximum de vraisemblance ne prévient pas non plus des estimations incohérentes.

Un autre inconvénient de ces techniques d'estimation est le suivant : deux unités ayant les mêmes valeurs (lignes identiques dans $\mathbf{X_1}$) auront le même estimateur pour leur variable Y, d'où une variabilité insuffisante dans $\mathbf{Y_1}$;

La technique d'imputation multiple ([RUB87]), consiste à imputer chaque donnée par m (≥ 2) valeurs obtenues par tirage dans un ou plusieurs modèles d'estimation. Puis on fait l'analyse des données sur chacun des m jeux de données ainsi complété.

L'estimateur final d'un paramètre quelconque sera la moyenne des m estimations ainsi réalisées. L'imputation multiple sous un ou plusieurs modèles permet de simuler la distribution a posteriori des données manquantes sous ce ou ces modèles et d'obtenir des variances correctes. Les inconvénients majeurs sont la complexité des calculs sous un ou plusieurs modèles, le temps de calcul considérable et la quantité à stocker et à gérer.

2.3 La fusion de données par maximisation de la cohérence interne ou de l'homogénéité

Proposée par [BUU92], Co ([COV97]) et Saporta ([SAP99]) ont étudié une technique pour données qualitatives basée sur la présentation « hollandaise » de l'analyse des correspondances multiples appelée « homogeneity analysis », cf. [DEL73], [GIF90]. Cette présentation permet de faire de l'ACM avec données manquantes ([MEU82]) mais notre propos ici est d'estimer ces données et non de les ignorer.

L'ACM d'un tableau disjonctif **G=(G1|G2|...|Gm)** revient à minimiser sur X et Y la fonction de perte suivante:

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

où $\mathbf{Y_j}$ est la matrice des coordonnées des catégories et $\mathbf{X}$ la matrice des coordonnées des unités le long des axes principaux.

L'idée est alors la suivante: lorsqu'il y a des données manquantes on estime les catégories de telle sorte que la fonction de perte soit minimale, en d'autres termes pour que l'on ait une ACM avec les plus grandes valeurs propres.

Formellement, on minimise sur $\mathbf{X}$, $\mathbf{Y_j}$ et $\mathbf{G_j^*}$:

$$\sigma(X; Y_1, \ldots, Y_m, G_1^*, \ldots, G_m^*) = \sum_{j \in \Omega} \left\| X - G_j Y_j \right\|^2 + \sum_{j \notin \Omega} \left\| X - G_j^* Y_j \right\|^2 \qquad (2)$$

où Ω est l'ensemble des variables complètes et $\mathbf{G_j^*}$ une matrice d'indicatrices d'une variable où les données manquantes ont été complétées.

L'exemple suivant ([BUU92]) montre comment imputer pour maximiser la première valeur propre dans une ACM avec 3 variables à 3 catégories:

Table 1 : ***Tableau avec données manquantes***

Unit	Income	Age	Car
1	*x*	young	am
2	medium	medium	am
3	*y*	old	jap
4	low	young	jap
5	medium	young	am
6	high	old	am
7	low	young	jap
8	high	medium	am
9	high	*z*	am
10	low	young	am

Il existe donc 27 façons d'imputer les trois valeurs manquantes x,y,z.

Table 2 : *Résultats des 27 ACM*

x	y	z	λ_1	x	y	z	λ_1	x	y	z	λ_1
l	l	j	.70104	m	l	y	.63594	h	l	y	.61671
l	l	m	.77590	m	l	m	.72943	h	l	m	.66458
l	l	o	.76956	m	l	o	.72636	h	l	o	.65907
l	m	j	.78043	m	m	y	.70106	h	m	y	.70106
l	m	m	.84394	m	m	m	.77839	h	m	m	.74342
l	m	o	.84394	m	m	o	.84394	h	m	o	.74342
l	h	j	.78321	m	h	y	.73319	h	h	y	.68827
l	h	m	.84907	m	h	m	.80643	h	h	m	.74193
l	**h**	**o**	***.84964**	m	h	o	.80949	h	h	o	.74198

La solution optimale en terme d'homogénéité maximale est x = « low », y = « high », z = « old ». C'est en quelque sorte la solution la plus plausible si les donnée sont unidimensionnelles.

Bien sur une telle recherche exhaustive est infaisable pour des données réelles de fusion de grands fichiers, et des heuristiques sont nécessaires ([COV97])

Cette méthode n'évite pas le défaut déjà signalé de réduire la variabilité, car deux unités ayant les mêmes données se verront imputer les mêmes valeurs des variables manquantes, si on ne procède pas à des imputations multiples.

Le choix du nombre d'axes de l'ACM doit aussi être précisé car une solution unidimensionnelle est en général trop réductrice.

Pour résumer, les techniques d'estimation basées sur des modèles explicites semblent plus adaptées pour estimer quelques données manquantes que d'estimer des blocs de centaines de données manquantes comme dans le cadre de la fusion de données.

2.3 Les méthodes implicites: fusion par appariements intra-cellulaires, imputation par Hot-Deck, la méthode des plus proches voisins etc....

Il existe des techniques plus simples que les précédentes qui reposent sur le principe de donner aux variables du fichier receveur toute l'information et les renseignements détenus par les variables du fichier donneur. Nous allons présenter ici quelques méthodes qui utilisent des modèles implicites.

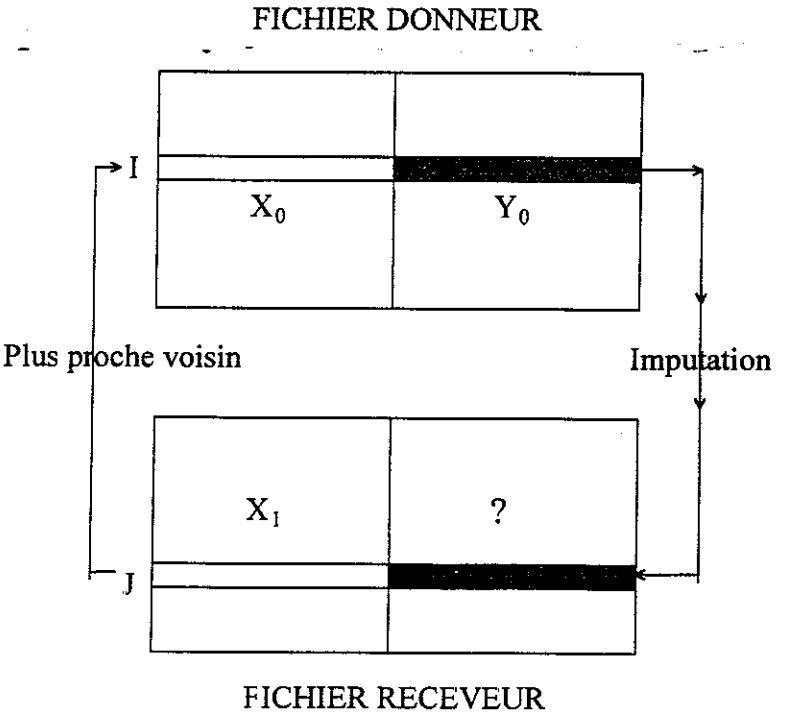


Figure 1 : ***Schéma de fusion par "copier-coller"***

2.3.1 La fusion par appariements intra-cellulaires

Souvent, on dispose de plusieurs fichiers de renseignements sur une population. Les domaines couverts sont spécifiques à chaque fichier mais relativement liés et une partie des informations est présente dans chaque fichier.

Un objectif naturel est alors d'apparier ces fichiers pour affiner et améliorer les liens entre ces domaines de renseignements.

La méthode statistique de l'appariement aléatoire par cellule, utilisée en particulier à l'INSEE, repose sur le partitionnement des échantillons de chaque enquête en cellules regroupant des individus semblables. On affecte ensuite à un individu du fichier receveur un individu du fichier donneur pris de façon aléatoire sans remise dans la même cellule.

Une condition nécessaire d'utilisation de cette méthode est que les deux échantillons aient une distribution similaire suivant les variables communes. De plus, le choix du nombre de critères de cellulage et le rapport des tailles entre les deux fichiers sont essentiels: il faut que les tailles des deux échantillons ne diffèrent pas trop l'une de l'autre.

Un avantage majeur de cette méthode est de respecter les relations entre les variables transférées. Malheureusement, elle a tendance à homogénéiser les comportements simulés pour les individus du fichier receveur.

2.3.2 Les méthodes d'imputation de type « hot-deck »

Le principe est le suivant : la valeur manquante est remplacée par la valeur observée chez un répondant « proche », le « donneur ».

Cette méthode se divise elle-même en plusieurs procédés :

- le *hot-deck d'ensemble* : le donneur est choisi de façon aléatoire parmi les répondants.
- le *hot-deck par classe* : le donneur est choisi de façon aléatoire dans la classe à laquelle appartient le receveur.
- le *hot-deck séquentiel* : le fichier à compléter défile. Si une unité est manquante, alors on lui impute la valeur renseignée par l'individu le plus « récent » du tableau de données et appartenant à la même classe. Cette procédure exige une valeur initiale.
- le *hot-deck hiérarchisé* : une suite de critères C1, C2,..., Ck est utilisée. On remplace l'unité défaillante par une unité ayant les mêmes valeurs pour C1, C2,..., Ck. S'il n'en existe pas alors on la remplace par une unité ayant les mêmes valeurs pour C1, C2,..., Ck-1; etc.

- le *hot-deck métrique* ou méthode du plus proche voisin : on construit une distance notée d(i,j) entre unités en fonction de variables clés bien renseignées qu'elles ont en commun. Si l'unité k est défaillante, on lui impute la valeur observée chez son plus proche voisin « donneur potentiel ». S'il y a ex-æquo, alors on impute une de ces valeurs possibles aléatoirement.

2.3.3 La fusion sur référentiel factoriel

Cette méthode fait partie des méthodes qui utilisent des modèles implicites et est fréquemment utilisée en France. Son principe défini par [SAN84] repose sur les deux points suivants:

- les 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.

- les 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.

D'autre part, cette méthode de fusion comprend la recherche de deux éléments essentiels :

- celle du référentiel factoriel. On commence par effectuer une analyse factorielle (par exemple une analyse des correspondances multiples) sur le tableau des variables critiques communes ou sur une sélection de variables communes les plus discriminantes des variables spécifiques, à l'ensemble des données disponibles c'est à dire les données relatives aux individus donneurs et receveurs. Ensuite on conserve les l premiers axes de l'analyse. Ainsi on positionne les observations dans un espace réel de dimension l, sur lequel on introduit une distance euclidienne calculée à partir des coordonnées factorielles qui mesure la similarité « donneur-receveur ». L'intérêt d'une analyse en axes principaux est qu'elle filtre les informations, élimine les effets marginaux de données perturbatrices.

- celle de voisinage d'un receveur. Pour chaque individu receveur, on sélectionne un ensemble de donneurs dans un voisinage du receveur. Ce voisinage peut-être défini soit à partir d'une sphère centrée sur chaque receveur de rayon r, soit à partir de la recherche des k plus proches donneurs.

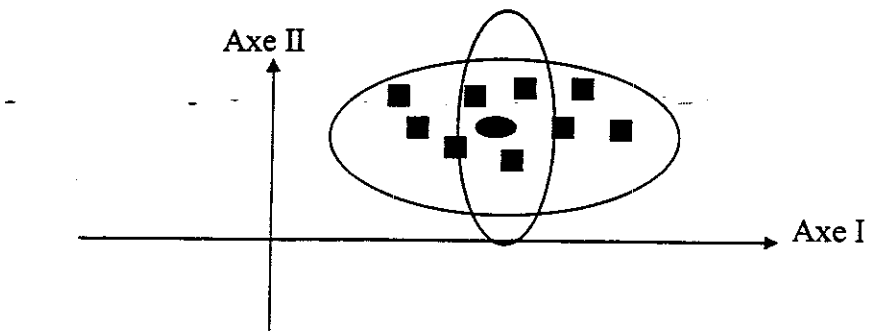


Figure 2 : *choix du donneur*

On choisit ensuite parmi les donneurs potentiels celui qui ressemble le plus au receveur sur les variables de rapprochement, qui sont des variables signalétiques comme l'age , le sexe, la classe socio-professionnelle, etc.

D'autre part, on peut éviter d'utiliser trop souvent le même donneur en utilisant une fonction de pénalité.

On peut distinguer deux méthodes de fusion utilisant un référentiel factoriel sont :

- la fusion par « mariage »
- la fusion par recherche de sosies

La fusion par « mariage »

Cette méthode de fusion s'appuie sur le calcul d'une distance dans le référentiel factoriel entre les individus donneurs et les individus receveurs.

Cette distance permet de définir ce que G. Santini appelle les « mariages » entre un individu receveur et un individu donneur. Les individus sont mariés en fonction de leur proximité d calculée sur les coordonnées factorielles. Une méthode élémentaire de fusion serait une minimisation de la distance moyenne calculée sur les réponses. Cependant G. Santini veut éviter qu'un individu donneur soit marié plusieurs fois. En fait, il faut éviter qu'un même individu donneur transmette son information plusieurs fois c'est à dire à plusieurs individus receveurs. Pour pallier ce problème de mariages multiples, on peut introduire un système de pénalités pour les individus donneurs dans le cas où ces derniers donneraient leur information à plusieurs individus receveurs. On pénalise alors la distance, c'est à dire que si un individu donneur est déjà marié à n individus receveurs, cette distance d est pénalisée par la formule suivante :

$$d' = 1 - (1 - d)^n$$

G. Santini a imaginé 6 types différents de relations de voisinage par « mariage » :

On note : A l'individu receveur, B l'individu donneur.

- le mariage par « coup de foudre » (voisins réciproques) : si A est le plus proche voisin de B et réciproquement B est le plus proche voisin de A et n'a jamais été marié, alors A et b sont immédiatement mariés.

- le mariage avec « l'ami d'enfance » : si B est le plus proche voisin de A, mais que B est déjà marié à A' , alors A sera marié à B' qui est le plus proche voisin de A après B.

- le mariage par « adultère » : une variation du cas précédent est nécessaire quand la distance entre B' et A est plus grande que la distance pénalisée entre A et B (puisque B est déjà marié a A'). Donc nous marions ensemble A et B.

- le mariage par « assiduité » : un autre cas apparaît quand on voudrait unir A et B, mais B est le plus proche voisin de A', A" et A"' avec lesquels il est déjà marié. Finalement nous marions A et B ensemble.

- le mariage de « raison » : après les cas simples illustrés ci-dessus, il existe d'autres mariages pour lesquels les décisions sont plus complexes. Ces mariages sont réalisés en utilisant des méthodes d'optimisation de distances à des niveaux globaux

- le mariage des « irréductibles » : ce cas traite des individus restant n'ayant pas trouvé de mari; de tels cas sont dus principalement aux faiblesses des règles d'optimisation utilisées. Nous cherchons alors d'autres règles d'optimisation permettant d'obtenir un mariage avec cet irréductible.

La fusion sur référentiel factoriel permet d'imposer des contraintes à la fusion par mariage comme interdire des mariages à l'intérieur de cellules disjointes définies sur des critères comme le sexe ou l'age ou encore la catégorie socio-professionnelle ou encore comme le nombre d'utilisation d'un même individu donneur.

Néanmoins la fusion par « mariage » n'est pas restreinte au seul cas de la fusion sur référentiel factoriel. En effet [RAS97] l'utilise dans un cadre plus théorique de fusion de fichiers de données continues ayant des distributions « classiques ».

La fusion par recherche de sosies

Son principe s'appuie sur la recherche du sosie d'un individu à l'aide d'une distance calculée dans le référentiel factoriel. Cette méthode et ses dérivées sont fréquemment utilisées dans le cas de fusion d'enquêtes de consommation.

Cette méthode se décompose en trois étapes :

1. La première étape est la recherche du référentiel factoriel qui est obtenu par une analyse factorielle des variables critiques. Puis on ensuite, comme indiqué précédemment, on introduit une distance euclidienne qui mesure la similarité des individus. Pour chaque individu receveur, on retient les m unités de la population des donneurs dont la distance D avec le receveur vérifie la condition suivante :

$$D(r, d)<S$$

où r est un individu du fichier receveur, d un individu du fichier donneur et S un seuil de distance.

2. La seconde étape impose des contraintes pour choisir le donneur parmi les m unités de la population des donneurs.

3. La dernière étape est une comparaison de la ressemblance signalétique entre l'individu donneur et l'individu receveur. On utilise pour cela l'agrégation multicritère. On calcule un poids, suivant le degré de liaison entre les variables de signalétiques et les variables spécifiques à transférer du donneur vers le receveur, pour différencier l'importance des variables signalétiques. On a alors une note globale sur toutes les variables. Les individus les plus ressemblants sont retenus.

En cas d'échec c'est à dire dans le cas où on n'a pas trouvé de donneur pour le receveur, on recommence ces trois étapes en élargissant le rayon du voisinage S de la première étape.

Au niveau global, les résultats montrent que les distributions des variables sur les populations des répondants et des individus reconstitués sont très proches. En fait, il existe peu d'écarts significatifs dans les tableaux croisant les fichiers donneur et receveur.

Au niveau individuel, la signalétique des individus donneurs et celle des individus receveurs sont sensiblement les mêmes du fait que l'individu le plus ressemblant a été choisi comme l'individu donneur.

3. Une nouvelle méthode fusion de tableaux de données existants par construction d'un échantillon de données virtuelles

3.1. Contexte

Les méthodes de fusion développées précédemment consistent essentiellement à transférer de l'information d'un (ou plusieurs) individu(s) à un autre individu. D'autres méthodes récentes ont été développées résonnant à l'inverse sur les variables. Il n'est alors plus question d'apparier des individus mais de modéliser les variables spécifiques à partir des variables communes.

Derquenne ([DER99]) présente une méthode pour construire un échantillon de données virtuelles à partir de plusieurs tableaux de données existants que nous résumons dans le cadre de la fusion par modélisation : comme les grandes entreprises ne disposent pas en général d'une unique base de données qui regrouperait toutes les informations disponibles pour un même client relevées au cours des différentes enquêtes ou sondages pour des raisons déjà évoquées. Pour renseigner leur base clientèle à partir d'enquêtes, un projet de simulateur a été

lancé pour générer un échantillon d'individus virtuels. La démarche se décompose en deux étapes qui sont :

- la construction du premier échantillon virtuel fondé sur l'échantillon primaire(on développera la notion d'échantillon primaire par la suite).Cette construction est à base d'analyse des correspondances multiples.
- une greffe statistique d'un échantillon secondaire sur le premier échantillon artificiel afin de construire le second échantillon virtuel. La seconde étape utilise un outil statistique également : les modèles linéaires généralisés. Cette étape se répétera autant de fois voulues pour obtenir l'échantillon virtuel final.

3.2. Méthode pour générer un échantillon d'individus virtuels

Deux types d'échantillons de données sont disponibles :

- le premier: l'*échantillon primaire* noté $\mathbf{X}$ contient des variables du plan d'échantillonnage notées X_{MP}. Ces dernières renseignent sur l'age de l'interviewé, la taille de l'agglomération ou il demeure, sa profession .... L'échantillon primaire contient également des variables mesurées notées X_M qui donnent des informations sur les caractéristiques du logement et du ménage....
- le second : *les échantillons secondaires* notés $\mathbf{Y}^1,....,\mathbf{Y}^K$ comprennent des variables du plan d'échantillonnage dont quelques unes sont communes à X_{MP} , on les notera Y_{MP} et des variables mesurées communes et non communes à X_M notées $Y_{MP}^{(k)}$.

La construction de l 'échantillon d'individus virtuels se fait en deux étapes principales :

1. la première étape est la construction du *premier échantillon virtuel* fondé sur l'échantillon primaire $\mathbf{X}$.
2. la seconde étape est la greffe statistique d'un échantillon secondaire $\mathbf{Y}^1$ (sélectionné) sur le premier échantillon virtuel afin d'obtenir le *deuxième échantillon virtuel.* Cette seconde étape est répétée autant de fois qu'il y a d'échantillons secondaires puisqu'on

greffe au fur et à mesure les échantillons secondaires $\mathbf{Y}^2,\ldots,\mathbf{Y}^K$ les uns à la suite des autres. On obtient à la fin de ce processus *l'échantillon virtuel final*.

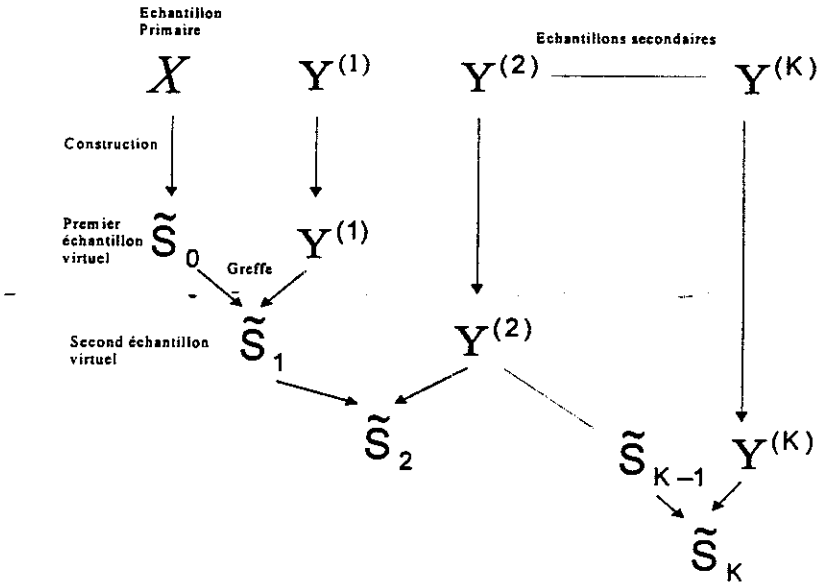


Figure 3 : ***schéma de la construction***

3.3. Construction du premier échantillon virtuel $\tilde{S}_0$ à partir de l'échantillon primaire

Cette construction utilise l'échantillon primaire défini préalablement. On considère un vecteur noté X_{MP} de taille Q constitué de Q coordonnées notées $\left(X_{MP(1)},\ldots,X_{MP(Q)}\right)$, qui sont les variables du plan d'échantillonnage. On introduit un second vecteur noté X_M de taille R constitué de R variables mesurées qui peuvent être nominales, ordinales ou discrétisées. En fait dans cette étape, on applique sur les variables du plan d'échantillonnage notées X_{MP} une Analyse des Correspondances Multiples (ACM) afin de constituer les composantes principales de l'ACM. Les variables X_{MP} jouent alors le rôle de variables actives. Les variables supplémentaires sont les variables mesurées contenues dans le vecteur X_M.

On sélectionne les composantes principales qui correspondent aux valeurs propres supérieures à 1/MP(Q), soit Z'.

A cette étape, on construit des groupes de variables les plus corrélées avec les composantes principales. Pour cela, on recherche les variables initiales du plan d'échantillonnage les plus liées à chacune des composantes principales au sens du rapport de corrélation η^2.

On procède ensuite au tirage des individus virtuels. Ce tirage se fait en deux étapes, après être passé d'un espace continu qui est celui des composantes principales à un espace discret afin de déterminer une distribution empirique. Ce passage se fait par une discrétisation de chaque composante principale, d'où une distribution empirique sur des pavés de cet espace.

Disons pour simplifier que l'on tire alors N individus virtuels de cette distribution observée, d'où le premier échantillon virtuel $\widetilde{S}_0$

3.4. Greffe statistique fondée sur les échantillons secondaires

Maintenant on va construire le *deuxième échantillon virtuel* noté $\widetilde{S}_1$ choisi parmi les échantillons secondaires $\mathbf{Y}^1,\ldots,\mathbf{Y}^K$. Ce choix est basé sur le nombre de variables d'échantillonnage communes avec l'échantillon primaire et les variables de l'échantillon secondaire jugées importantes dans le cadre de l'étude.

Soit G_1 le nombre de variables non communes à greffer sur le *premier échantillon virtuel* $\widetilde{S}_0$.
La construction se fait en deux étapes :

- la première étape est le redressement de l'échantillon secondaire par rapport au informations auxiliaires du plan d'échantillonnage de l'échantillon primaire, à l'aide d'une méthode de calage sur marge [DEM40].

- La seconde étape est la greffe une par une des G_1 variables de l'échantillon secondaire redressé sur le premier échantillon virtuel $\widetilde{S}_0$ à l'aide d'un modèle linéaire généralisé. On aboutit à l'estimation de la probabilité d'appartenance d'une des "réponses" possibles de la variable à greffer en fonction des strates issues des variables statistiquement "explicatives". Alors G_1 modèles sont construits. La première variable à greffer correspond à la variable qui a fourni la meilleure adéquation avec les données. N individus artificiels sont tirés de la distribution estimée de la première variable, sachant les caractéristiques communes des variables « explicatives » significatives dans le

premier échantillon virtuel $\widetilde{S}_0$. Une nouvelle variable est créée pour commencer à construire le deuxième échantillon virtuel $\widetilde{S}_1$. Après cela, il reste G_1-1 pour lesquelles G_1-1 modèles sont construits etc.

Le processus de sélection des variables et le tirage des N individus virtuels est à nouveau appliqué pour construire un troisième échantillon virtuel, et ainsi de suite jusqu'à l'échantillon virtuel final.

4.Validation

Comment mesurer la qualité des techniques de fusion? Comme il n'y a en général pas de modèle pour les données, on doit utiliser des procédures empiriques où on estime des données connues mais cachées que l'on compare ensuite aux vraies valeurs;: ce sont les méthodes de validation croisées, de bootstrap etc. cf [COM99].

4.1. Mesures et conditions de validité

Quels indicateurs prendre? La reconstitution des valeurs pour chaque individu est en général un critère trop sévère. Les utilisateurs sont d'ailleurs en général moins intéressés par des prévisions individuelles que par des prévisions au niveau d'un groupe. Mais retrouver les distributions marginales ne suffit pas puisqu'il suffirait alors de tirer aléatoirement les valeurs imputées! Le point essentiel est de conserver les structures de covariance, ou certains croisements pour des variables qualitatives.

Nous avons vu auparavant que les méthodes étaient de deux typas: fusion avec collage du vecteur entier du donneur, ou régression variable par variable. Le premier type est en général moins bon pour la reconstitution de données individuelles mais garde bien la structure de corrélation et évite les incohérences. L'inverse est vrai pour le deuxième type de méthodes. Dans tous les cas il est nécessaire pour avoir des résultats satisfaisants d'avoir:

- Un nombre suffisant de variables communes
- Des corrélations élevées entre le bloc des variables communes et celui des variables à impute.

- Une structure commune entre le fichier donneur et le fichier receveur, c'est à dire que les distributions des variables communes ou critiques doivent être à peu près les mêmes dans les fichiers donneur et receveur, sinon les résultats risquent d'être biaisés. Des redressements sont alors souvent nécessaires.

4.2 Un exemple ([SAP99])

Nous avons utilisé le jeu de données classique de SPAD: une étude sociologique portant sur 992 interviews, séparés aléatoirement en deux fichiers l'un de 800 unités pour le fichier donneur et l'autre de 192 pour le fichier receveur. 7 variables qualitatives ont été retenues:

4 variables communes:

Q1 - classe d'age(5 catégories),
Q2 - taille d'agglomération (5 catégories),
Q3 - heure de coucher (7 catégories),
Q4 - age de fin d'études (5 catégories).

3 variables d'opinion Y à imputer:

Q5 - La famille est le seul endroit où on se sent bien ? (oui, non)
Q6 - Plus haut diplôme obtenu (7 catégories),
Q7 - Taux d'écoute TV (4 catégories).

Les tableaux 3 et 4 présentent les résultats comparés de la fusion sur référentiel factoriel (FRF) de l'analyse homogène, et d'une affectation aléatoire. Les résultats illustrent bien ce qui était annoncé plus haut: la recherche de l'homogénéité maximale donne de meilleurs résultats individuels, mais est moins bonne en termes de distribution marginale et croisée (faute de place les croisements sont omis) que la fusion sur référentiel factoriel.

Table 3 : *performances individuelles*

Méthode	*Classifications correctes*
Aléatoire	*49%*
Homogénéité max.	*54%*
FRF	*47%*

Table 4 : *performances marginales*

Q5	*Vraies marges*	*Homogénéité max*	*FRF*
1	*136*	*136*	*125*
2	*56*	*56*	*67*
Q6			
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			
1	100	118	100
2	36	18	43
3	37	29	31
4	19	27	18

La fusion par homogénéité maximale estime les données manquantes comme un modèle de régression: l'estimation est la plus vraisemblable en termes d'homogénéité et est unique pour un vecteur donné des X alors que FRF peut donner plusieurs imputations différentes: par exemple avec X=3421, on a pour Y 6 estimations différentes (avec des fréquences différentes): 232,121,123,122,114,212 qui représentent mieux la variabilité des réponses possibles.

5. Conclusions

Elles sont de deux sortes: techniques et déontologiques.

D'un point de vue méthodologique, la fusion de données est un problème de données manquantes massives, et les statisticiens devraient être intéressés dans le développement et la validation de nouvelles méthodes. En plus des méthodes passées en revue dans cette communication, on peut suggérer d'autres voies de recherche du côté des algorithmes d'apprentissage non-linéaires (réseaux de neurones).Il est clair que les méthodes de fusion répondent à un besoin fréquemment exprimé par les praticiens et les gestionnaires de données qui souhaitent fournir à l'utilisateur final une base unique sans « trou ». La prudence s'impose

quand on utilise des « données » qui sont en réalité des estimations et non des valeurs observées: de selles données ne devraient jamais être utilisées à un niveau individuel, mais uniquement agrégé. Une conséquence perverse de ces méthodes peut être un moindre effort de collecte, puisque l'on peut reconstituer des données.

Un autre danger des méthodes de fusion concerne la confidentialité et la protection de la vie privée: de nombreux pays ont des lois du type « Informatique et Liberté » sur la protection des données personnelles, et la constitution de fichiers nominatifs. Avec de telles méthodes, on arrive à la situation suivante: des données qui n'ont pas été recueillies sont estimées et peuvent être ajoutées dans des fichiers à l'insu des individus concernés. Il y a là un paradoxe quand on passe aux efforts fournis parles Instituts Nationaux de Statistique pour préserver la confidentialité de leurs fichiers. Les greffes d'enquêtes sont moins sujettes à suspicion, puisqu'elles ne cherchent pas à estimer des données individuelles.

REFERENCES

[ALU97] Aluja-Banet T., Morineau A., Rius R. (1997) La greffe de fichiers et ses conditions d'application. Méthode et exemple. in *Enquêtes et sondages*, G.Brossier, A.M.Dussaix (Eds), Dunod, Paris, 94-102

[BAK89] Baker K., Harris P., O'Brien J. (1989), Data fusion: An appraisal and experimental evaluation, *Journal of the Market Research Society*, 31, 153-212

[BON86] Bonnefous S., Brenot J., Pagès J.P. (1986), Méthode de la greffe et communications entre enquêtes, in *Data Analysis and Informatics vol 4*, E.Diday (ed), North-Holland, Amsterdam, 603-617

[BUU92] Buuren S.V. & Van Rijckevorsel L.A. (1992) : Imputation of missing categorical data by maximizing internal consistency, *Psychometrika*,57, 567-580.

[COV97] Co V. (1997) *Méthodes statistiques et informatiques pour le traitement des données manquantes*. Ph.D., CNAM, Paris.

[COM99] Comyn M. (1999) *Modélisation et validation des rapprochements et fusions de fichiers d'enquêtes*. Ph.D., ENST, Paris.

[DEL73] De Leeuw J. (1973), *Canonical analysis of categorical data*. Dswo, Leiden.

[DEM40] Deming W.E., Stephan F.F. (1940) On a least square adjustment of sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

[DEM77] Dempster A. P., Laird N. M., Rubin D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* B39.

[DER99] Derquenne C. (1999) Une méthode pour construire un échantillon de données virtuelles à partir de plusieurs tableaux de données existants. *In Combiner des données de sources différentes, recueil du Symposium 99 de Statistique Canada*.

[GIF90] Gifi A. (1990), *Nonlinear multivariate analysis*, Wiley , New-York.

[LEJ95] Lejeune M. (1995) De l'usage des fusions de données dans les études de marché, *Proceedings 50th Session of ISI-Beijing*, Tome LVI, 923-935

[LIT87] Little R.J.A., Rubin D.B. (1987) *Statistical analysis with missing data*, Wiley, New-York.

[MEU82] Meulman J. (1982), *Homogeneity analysis of incomplete data*, Dswo, Leiden.

[RAS97] Rassler S., Fleischer K. (1997) Aspects concerning data fusion techniques. Discussion paper 16/1997, Nuremberg.

[RUB87] Rubin D.B. (1987), *Mutiple imputation for nonresponse in surveys*, Wiley, New-York.

[SAN84] Santini G. (1984) La méthode de fusion sur référentiel factoriel. *Séminaire IREP*, mars 1984.

[SAP99] Saporta G, Co V.(1999): Fusion de fichiers: une nouvelle méthode basée sur l'analyse homogène, in *Enquêtes et sondages*, G.Brossier, A.M.Dussaix (Eds), Dunod, Paris, 81-93.