Imputation multiple en présence de variables-flags

Modou DIA

CEPS-INSTEAD à Differdange, Luxembourg

Dans la conduite d'une enquête, l'existence de données manquantes est un phénomène courant. Elle peut avoir lieu au niveau d'une grappe, d'une unité d'une grappe ou d'un item d'une unité ou d'une grappe répondante. Elle est susceptible d'engendrer des biais dans l'analyse des résultats. Différentes attitudes sont adoptées selon le contexte : ne rien faire, exploiter uniquement les observations totalement renseignées, pondérer ou repondérer et enfin imputer les données des items manquants ou unités manquantes. Cette dernière attitude recoupe la problématique qui motive la rédaction de ce document, à savoir une application de l'imputation multiple avec des flags à l'aide du logiciel « Imputation-Variance-Estimation » (IVEware)[1]. La compréhension du mécanisme de la non-réponse est un passage obligé à cet effet. L'approche méthodologique est fortement dépendante du type de non-réponse. L'utilisation du programme IVEware n'est envisageable que si la non-réponse est de type *Missing Completely At Random* (MCAR), ou de type *Missing At Random* (MAR) ou *Ignorable* à la différence du type de réponse *Not Missing At Random* (NMAR) ou *Non-Ignorable*. Le choix d'une méthode d'imputation multiple s'explique par la nécessité de contourner les limites inhérentes aux méthodes dites simples, qu'elles soient déterministes ou stochastiques.

La pratique courante qui consiste à imputer une valeur unique à la donnée manquante présente de graves lacunes dans la mesure où elle ne restitue pas toute l'incertitude de la distribution des données imputées. La conséquence est la sous-estimation de la variance de la variable imputée par l'assimilation des données imputées à des données observées.

L'imputation multiple, en générant plusieurs valeurs issues d'une distribution adéquate, permet d'éviter cet écueil. Mais cela ne suffit pas dans la mesure où la présence de variables-flags dans sa mise en œuvre peut influer sur les résultats. Ce cas de figure se présente tout particulièrement lorsque la variable à imputer comporte des observations dites « non-concernées » comme par exemple dans le cas de retraités dans une variable relative au salaire. Le fait de leur attribuer une valeur nulle ou un code de donnée manquante diminue ou augmente les moments d'ordre 1 et d'ordre 2 qui servent à initialiser l'algorithme Expectation-Maximization (EM) qui est le noyau du programme IVEware. On pourrait imaginer de renoncer aux flags en restreignant l'imputation aux seuls salariés, mais cela reviendrait à renoncer à la structure de corrélation des revenus, une des forces du programme IVEware en cohérence avec le plan de sondage. En outre, il n'y a pas une étanchéité au niveau de la perception des salaires et des retraites, les salaires et les pensions de retraite peuvent être cumulés par des individus tels que les militaires et les préretraités.

Avec le programme IVEware, nous nous assignons comme objectif de réaliser et de comparer deux cas d'imputation multiple des salaires bruts : un premier avec flags et un autre sans flags. Le

[1] Par RAGHUNATHAN T.E. , LEPKOWSKI J.M., VAN HOEWYK J., and SOLENBERGER P. (1999), http://www.isr.umich.edu/src/smp/ive/

critère de jugement sera l'ampleur de l'erreur par rapport aux sources administratives agrégées disponibles.

Auparavant, nous aborderons les caractéristiques générales des méthodes d'imputation multiple ainsi que les présupposés et les traits saillants des applications à réaliser.

1. Principales caractéristiques des différentes méthodes

La Librairie des imputations multiples est principalement composée de quatre familles : les méthodes d'échantillonnage et de ré-échantillonnage, les méthodes non-paramétriques ou semi-paramétriques, les modèles classiques basés sur l'Estimateur de Maximum de Vraisemblance (EMV) et les modèles d'inspiration bayésienne ou de « Data Augmentation » (DA) pour ne citer que celles qui obéissent aux mécanismes MCAR et MAR.

La méthode d' « échantillonnage et de ré-échantillonage » est une variante du calcul de la précision des estimateurs en générant plusieurs échantillons sur lesquels est appliquée la même procédure d'imputation. Les méthodes non paramétriques et paramétriques telles que le Hot Deck et la régression stochastique utilisent une initialisation différente du générateur des nombres aléatoires à chaque exécution.

Les modèles de l'EMV recouvrent une variété de composantes traitant des données catégorielles, continues et discrètes selon des lois appropriées avec des valeurs des paramètres fixées d'une manière déterministe ou d'une manière aléatoire. L'algorithme EM est le plus répandu dans cette catégorie.

Les modèles d'inférence bayésienne (DA) tirent partie d'une loi a priori informative ou non-informative pour construire d'abord la distribution a posteriori des paramètres, ensuite pour suivre un processus presque similaire à l'algorithme EM. Le modèle DA est un type de Markov Chain Monte Carlo (MCMC). Le MCMC est un ensemble de méthodes pour générer des pseudo-tirages aléatoires à partir des chaînes de Markov. Une chaîne de Markov est une séquence de variables aléatoires dont la probabilité de chaque élément dépend du précédent :

$$\{X_t : 1, 2, ...\} \text{ où } P(X_t / X_0, ..., X_{t-1}) = P(X_t \mid X_{t-1})$$

Une chaîne de Markov est entièrement définie par sa valeur initiale $P(X_0)$ et la règle de transition $P(X_t \mid X_{t-1})$.

Soit θ les paramètres du modèle, en général une matrice de variance-covariance associée au vecteur des moyennes calculées à partir les données observées. C'est le même θ qui sera repris pour toute la suite du document.

Soit Y la variable à imputer
Soit $Y = (Y_m, Y_{obs})$, Y_m est la partition des données manquantes de Y et Y_{obs} est la partition des données observées de Y, il s'ensuit ce processus pour l'algorithme du DA :

-Etape initiale : $Y_m^{t+1} \sim P(Y_m \mid Y_{obs}, \theta^{(t)})$

- Processus itératif : $\theta^{(t+1)} \sim P(\theta \mid Y_{obs}, \theta^{(t+1)})$

Ainsi est créée une chaîne de Markov :
$Y_m^1, \theta^{(1)}, Y_m^2, \theta^{(2)}, ...$ qui converge en distribution.

Pour la suite, nous nous concentrons uniquement sur les modèles de types EMV classiques et DA.

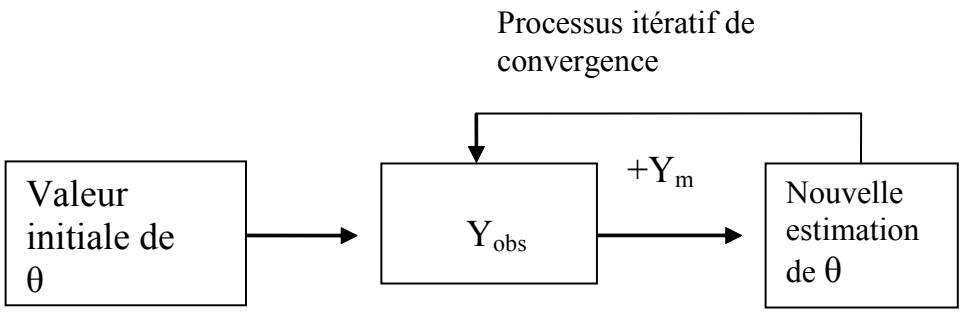
1.1. Les différentes approches au niveau algorithmique

Les EMV classiques et les DA sont les seuls modèles dont la création est uniquement motivée par une approche singulièrement plurielle de l'imputation. Elles sont aussi les plus répandues. Les algorithmes EM et MCMC sont respectivement qualifiés pour exhiber leur philosophie et leur structure. Ainsi se limiterons-nous à ces deux derniers algorithmes pour exposer les grandes lignes des modèles associés. Il est utile de rappeler que ces modèles sont envisagés sous la seule optique des mécanismes de réponse MCAR et MAR.

L'EM est une application de l'EMV en présence de données manquantes. A partir des paramètres du modèle θ, les coefficients du modèle d'estimation des valeurs manquantes sont calculés. A l'aide de ces derniers, les valeurs manquantes sont estimées. On recommence le processus en calculant cette fois-ci le paramètre avec les données observées et les données estimées. L'algorithme converge après un certain nombre d'itérations initialement fixé ou dès que la différence des θ entre deux itérations successives est inférieure à un seuil initialement fixé.

Soient Y la variable à imputer, Y_{obs} les observations renseignées, Y_m les observations manquantes et θ les paramètres du modèle

L'algorithme peut-être ainsi résumé :



Critère de convergence $i = 1, 2, ...K$ fixé ou $\left|\theta^i - \theta^{i-1}\right| \leq \left|\sum\right|$, un seuil initialement fixé.

Quant au MCMC, c'est une séquence de variables aléatoires dans laquelle la distribution d'un élément dépend des valeurs fournies par la séquence précédente. Les données peuvent présenter une configuration quelconque, par contre elles sont supposées suivre une distribution multivariée normale. Son algorithme est un peu plus complexe que l'EM et comprend 4 étapes :

1) Initialisation des valeurs des paramètres θ de départ avec un aléa ou par le recours à l'algorithme EM.

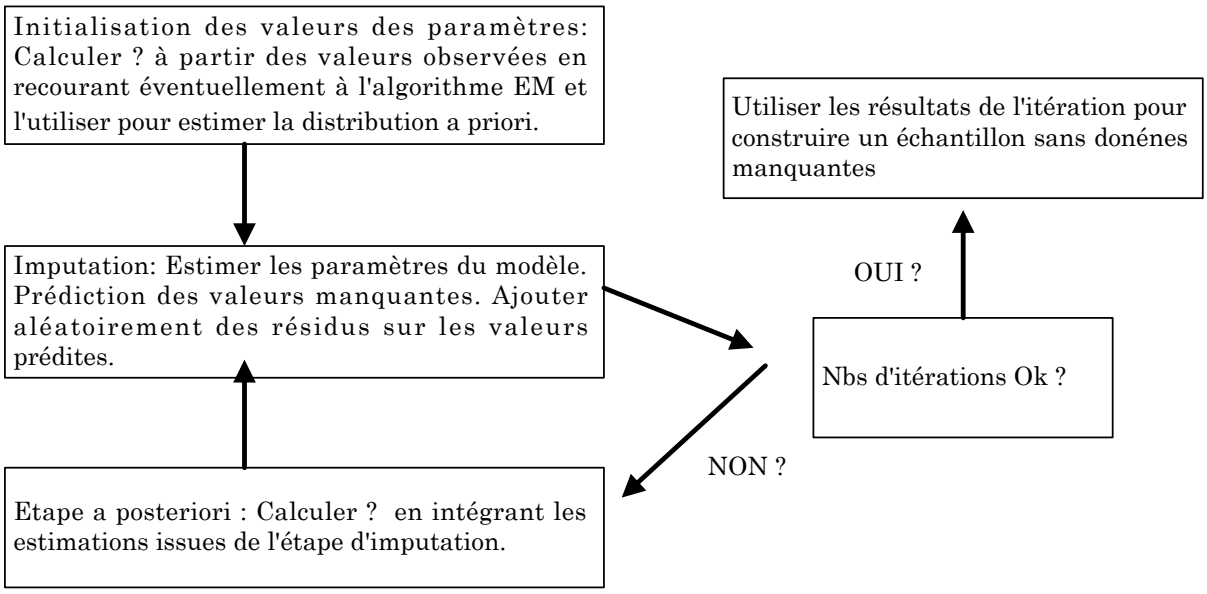
2)Utilisation des valeurs courantes des moyennes et de la matrice de variance-covariance (θ) pour estimer tous les coefficients de la régression des variables avec données manquantes sur les variables sans données manquantes. Les valeurs prédites de toutes les données manquantes sont générées à partir des résultats de la régression en y ajoutant aléatoirement des résidus de la distribution normale du modèle.

3) Estimation des valeurs du paramètre θ à partir des données complètes. Avec ce nouveau θ, on opère un tirage aléatoire sur la distribution a posteriori des θ.

4) Retourner à l'étape 2 et ainsi de suite jusqu'à la convergence de l'algorithme.

Contrairement à l'algorithme EM qui converge vers un optimum local, l'algorithme MCMC converge vers une distribution.

Algorithme du MCMC



1.2. Critères de convergence

Dans l'imputation multiple, on distingue deux niveaux d'itérations : le premier aboutit à un échantillon de données complètes à la fin de chaque cycle ; le second est relatif au nombre d'échantillons de données dont on veut disposer. La combinaison des résultats permet de calculer les variances et par voie de conséquence des intervalles de confiance « fiables » en prenant en compte les erreurs d'échantillonnage ET les erreurs liées aux imputations.

Le nombre d'itérations et le nombre d'imputations sont évidemment fonction du taux de données manquantes et de la puissance prédictive des variables explicatives : plus le taux de non-réponses partielles est élevé, plus le nombre d'imputations et le nombre d'itérations devront être importants, comme le montre la relation ci-après :

Efficience de l'estimateur $= (1+\gamma/m)^{-1}$ (1 +γ/m) $^{-1}$, γ étant le taux de données manquantes et m étant le nombre d'imputations.

D'où le tableau suivant :

Tableau 1 : Efficience de l'imputation multiple en %
Taux de données manquantes = 0.1, 0.2, …., 0.9
Nombre de pseudo-échantillons imputés =3, 5, 10, 20, …50

	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
3	97	94	91	88	86	83	81	79	77
5	98	96	94	93	91	89	88	86	85
10	99	98	97	96	95	94	93	93	92
20	100	99	99	98	98	97	97	96	96
30	100	99	99	99	98	98	98	97	97
40	100	100	99	99	99	99	98	98	98
50	100	100	99	99	99	99	99	98	98

La question de l'arbitrage sur la priorité à accorder au nombre d'itérations ou au nombre d'imputations demeure. Elle traduit le souci légitime de procéder à des améliorations et à des raffinements croissants selon la vision rétrospective d'une histoire récente de l'imputation qui s'est progressivement sophistiquée en gravitant successivement les paliers déterministes, ensuite aléatoires des méthodes simples d'imputation, puis des méthodes multiples d'imputation dont certaines sont enrichies d'un zeste aléatoire dans le choix des paramètres □. Cependant une évidence s'impose sur le fait que le nombre d'itérations doit être suffisamment important pour atteindre une solution stationnaire.

Par ailleurs, pour les méthodes de type DA (Data Augmentation), il existe une alternative entre une variante séquentielle et une variante parallèle. La première consiste, après une période de chauffage d'un nombre d'itérations donné W de produire tous les $W + m * L$ itérations un échantillon de résultats, $m = 1, 2, 3, ..., M$ (nombre de pseudo-échantillons) et L un nombre d'itérations jugé suffisant. La seconde profite de la puissance d'un système informatique distribué pour lancer plusieurs M exécutions avec des valeurs initiales affectées aux paramètres θ différentes pour chacune d'entre elles.

Le défaut attribué à la variante séquentielle est le risque de non-indépendance des M échantillons tandis que la convergence de la distribution a posteriori est plus fiable, surtout pour les derniers échantillons de la séquence. Les défauts et les avantages opposés sont attribués à la variante parallèle : les vertus de l'indépendance et les lacunes au niveau de la convergence. Il ne faudrait pas oublier le gain en temps d'exécution de la version parallèle sur un système informatique parallèle, surtout s'il s'agit d'un gros fichier.

1.3. Calcul de la variance

L'imputation multiple génère, pour chaque variable Y, M valeurs imputées qui vont contribuer à permettre de calculer la variance sur M pseudo-échantillons de données.

1.3.1. Cas scalaire

Soient : $\hat{\theta}_i, \overline{V}_i, 1, 2i =, ...M$, les M estimateurs respectivement de θ et de sa variance V calculés à partir des M pseudo-échantillons de données complétées à partir d'un modèle d'imputation.

Soient : - L'estimateur global du vecteur des moyennes de θ : $\overline{\theta} = \frac{1}{M}\sum_{i=1}^{M}\hat{\theta}$

- La variance intra-imputation : $\overline{V} = \frac{1}{M}\sum_{i=1}^{M} V_i$

- La variance inter-imputation s : $B = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\theta}_i - \overline{\theta}\right)^2$

- La variablilité totale pour $\overline{\theta}$: $T = \overline{V} + \left(1 + \frac{1}{M}\right)B$

- La variance totale : $VT = \overline{V} + \left[(M+1)/M\right] * B$

- Les degrés de liberté : $DDL = (M-1)\left[1 + (1/\overline{r}_M)\right]$

$$\overline{r}_M = \left[(M+1)/M\right] * (B/\overline{V})$$

Ces résultats sont dus à Rubin et Raghunathan (1991).

1.3.2. Cas vectoriel

Le paramètre estimé $\hat{\theta}_i$ devient un vecteur. $\hat{V}_i$ est sa matrice de variance-covariance estimée. Les expressions développées ci-dessus de $\overline{\theta}$, $\overline{V}$ et VT sont identiques tandis B devient :

$$B = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\theta}_i - \overline{\theta}\right)\left(\hat{\theta}_i - \overline{\theta}\right)^T$$

2. Applications sur le logiciel IVEware de Michigan

Deux procédures d'imputation seront appliquées aux données individuelles du *Panel Socio-Economique Liewen zu Lëtzebuerg* (PSELL3) avec le programme IVEware:

- La première impute les salaires en attribuant des valeurs nulles aux données manquantes des individus « non-concernés »

- La seconde réalise l'imputation de ces dernières, c'est-à-dire impute des valeurs pour les non-salariés, quitte à remplacer leurs résultats par des valeurs nulles après la procédure d'imputation grâce aux « variables-flags ».

Les résultats de ces modèles seront ensuite confrontés à une distribution des salaires issue d'une source administrative.

2.1. Mise en œuvre de la méthode

La méthode d'imputation proposée découle de l'algorithme EM.

2.1.1. Généralités

Sous sa forme simple c'est-à-dire en une seule passe, elle peut être considérée comme un modèle séquentiel de régression. Son algorithme peut être décrit de la façon ci-après :

Soient :

X un ensemble de variables explicatives sans données manquantes ;

$Y_1, Y_2, \ldots, Y_k$, un ensemble de variables dépendantes ordonnées selon le taux croissant de données manquantes.

La séquence des imputations est déterminée par la factorisation que voici :

$$\left[Y_1 / X\right], \left[Y_2 / X, Y_1\right], \left[Y_k / X, Y_1, \ldots, Y_{k-1}\right]$$

où $\left[Y_i / X\right]$ est la distribution conditionnelle jointe de Y sachant X . Autrement dit après chaque itération i , la variable Y_i qui vient d'être imputée s'ajoute à l'ensemble des variables explicatives.

Selon la nature de la variable Y_i à imputer, le modèle de régression peut revêtir cinq formes :

a) une régression linéaire généralisée si la variable Y_i est continue ;

b) une régression logistique si la variable Y_i est binaire ;

c) une régression polytomique si Y_i est une variable catégorielle ;

d) une régression log-linéaire (loi de Poisson) si Y_i est une variable discrète finie ;

e) une régression pour une variable mixte si Y_i est une variable mixte telle qu'une quantité de stocks positive ou nulle, s'il y a lieu).

La diversité des types de variables qui peuvent être traités ainsi que la prise en compte de leur structure corrélative constituent un point fort des méthodes de régressions généralisées séquentielles.

2.1.2. Enjeux des variables-flags

Soient une variable $Y = (y_1, y_2, \ldots, y_i, \ldots, y_n)$ et F_i, le flag associé à chaque observation de y_i telles que :

- $F_i = 1$, si y_i est observé ;
- $F_i = 2$, si y_i est manquant ;
- $F_i = 3$, si y_i est « non-concerné » par exemple dans le cas d'un salaire pour un chômeur ou des allocations de maternité à un homme.

On présume que pour $F_i = 1$ ou 2, l'individu est « concerné ».

Soient :

- r, le nombre d'observations à valeurs renseignées ;
- m, le nombre d'observations à valeurs manquantes ;
- nc, le nombre d'observations non-concernées ;

tels que : $n = r + m + nc$

Dans l'implémentation classique du programme d'« IVEware » par leurs auteurs, une valeur nulle est affectée à toutes les observations correspondant aux flags dont la valeur est égale à 3. Le symbole « manquant » est attribué aux observations, dont le flag est égal à 2, qui sont ainsi exclues du calcul des valeurs initiales des paramètres de θ. Ce choix peut considérablement affecter le mode de calcul et par ricochet les valeurs des paramètres θ, surtout si on se trouve comme dans la plupart des cas dans le cadre d'un modèle multivarié où on ne peut pas se restreindre pour chaque variable Y_i aux observations correspondant aux flags 1 et 2. La conséquence est la sous-estimation des paramètres θ. Il en résultera vraisemblablement des biais de sous-estimations des variables imputées.

Illustrons cela par un calcul simple des moyennes des y_i selon le statut accordé aux observations « non-concernées ».

Si les observations « non-concernées » sont supposées égales à zéro, alors pour le calcul initial de θ, on a :

$$\overline{Y} = 1/(r + nc) * \sum_{Fi=1}^{n} y_i$$

et pour les valeurs suivantes de θ :

$$\overline{Y} = (1/(r + m + nc)) * (\sum_{Fi=1}^{n} y_i + \sum_{Fi=3}^{n} y_i)$$

Dans le cas contraire, les valeurs de $\overline{Y}$ pour le calcul des valeurs initiales ou non-initiales de θ seront respectivement :

$$\overline{Y} = 1/r \sum_{F_i=1}^{n} y_i \text{ et } \overline{Y} = (1/(r + m + nc)) * (\sum_{F_i=1}^{n} y_i + \sum_{F_i=2}^{n} y_i + \sum_{F_i=3}^{n} y_i)$$

car les observations « non-concernées » sont imputées, même si elles seront remises à zéro à la fin du processus grâce aux valeurs de leurs flags.

Cependant, la correction probable des biais de sous-estimation risque de provoquer des effets pervers en déformant la structure des corrélations entre les variables. En effet, la structure de corrélation peut être malmenée par le fait que dans le processus séquentiel, la covariance calculée prend en compte des observations « non-concernées » : des salaires imputés à des chômeurs ; des pensions de retraite attribuées à des salariés.

Par ailleurs, les distorsions éventuelles de la structure de corrélation constituent un moindre mal si l'utilisation des flags « non-concernés » sont limités aux variables continues de sorte que les corrélations induites par les variables catégorielles, dont les croisements produisent des strates relativement homogènes, sont préservées. Toutes choses égales par ailleurs, il est logique de supposer que les salaires et les retraites sont corrélés, à Catégorie Socio-Professionnelle (CSP), à niveau d'éducation et à expérience professionnelle équivalents.

2.1.3. Implémentation du modèle d'estimation

Les principes élémentaires de construction d'un modèle prédictif recommandent de choisir les variables explicatives parmi les variables structurantes de l'analyse de l'enquête, les variables les plus corrélées à la variable à expliquer, les variables les plus pertinentes par rapport au phénomène de la non-réponse et les variables les plus significatives du point de vue du plan de sondage.

Pour imputer le salaire brut dans le cadre de l'enquête *Panel Socio-Economique Liewen zu Lëtzebuerg* (PSELL3), les variables candidates sont les suivantes :

a) Variables continues : le temps de travail, le ratio salaire net-salaire brut, la pension de retraite ;
b) Variables catégorielles : le secteur public ou privé, le niveau d'éducation, la CSP, le sexe, la classe d'imposition, la nature de la carte d'impôt (principale ou additionnelle) ;
c) Variables discrètes : l'âge, l'expérience professionnelle, le nombre d'enfants donnant droit à la modération d'impôt.

Cependant, pour éviter une discordance entre le salaire brut et le salaire net, c'est le ratio salaire brut- salaire net qui sera introduit dans le modèle au lieu du salaire net. En tenant compte du taux d'imposition marginal maximal et les cotisations sociales, il est contraint d'évoluer entre les bornes 0.55 et 1. Le salaire brut sera soumis à des contraintes relatives aux minima légaux et aux queues de distributions des salaires observés.

Enfin, ce modèle est exécuté avec ou sans flags avec pour tous les types de revenus dans le modèle.

2.2. Interprétation des résultats des imputations des salaires bruts

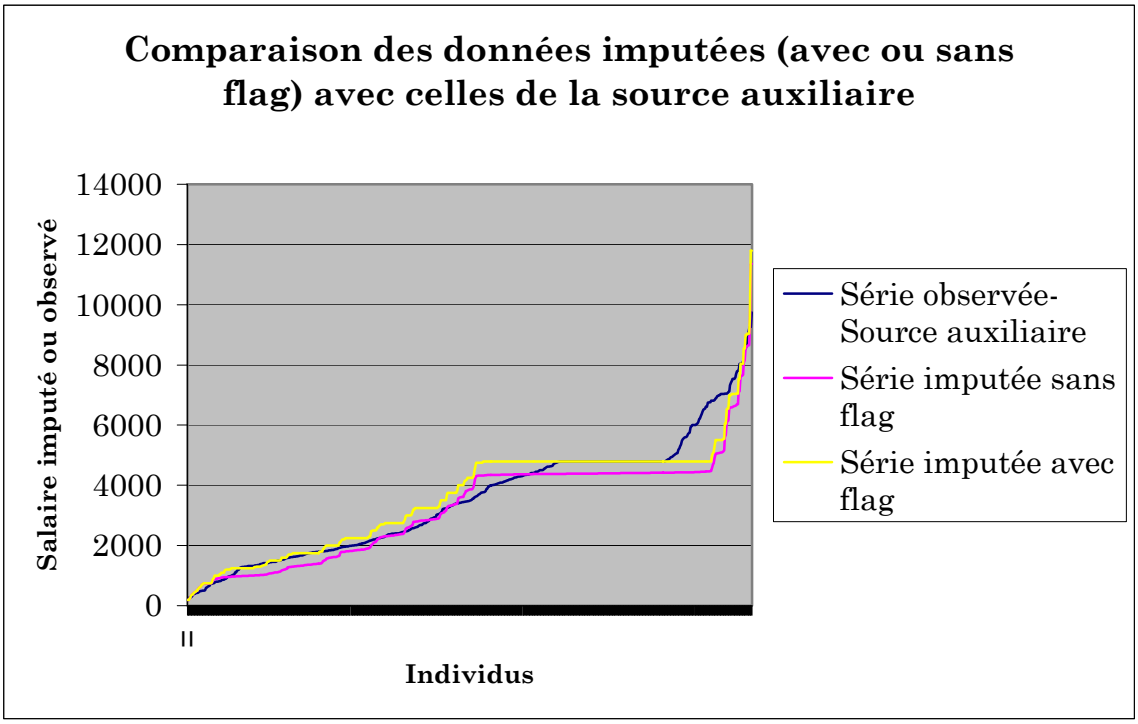
Deux exécutions, dont une avec flag et l'autre sans flag, produisant chacune 50 pseudo-échantillons de résultats ont été lancées. Chaque pseudo-échantillon possède 7675 observations dont 3997 sont dites « non-concernées » et 3678 dites « concernées ». Ces dernières sont composées de 2724 valeurs observées et 954 valeurs initialement manquantes qui sont imputées. Le tableau 2 fournit des statistiques les concernant. On constate que la moyenne pour les imputations avec flag est plus élevée que la moyenne des valeurs observées alors que c'est le phénomène inverse pour les imputations sans flag. Un plus grand crédit doit être accordé aux imputations avec flag qui surestiment la moyenne dans la mesure où les individus à fort potentiel salarial, avec plus de qualification et d'expérience professionnelle, sont plus nombreux parmi les observations à données manquantes.

Tableau 2 : Moyenne et écartype des valeurs collectées durant l'enquête et des valeurs imputées

FLAG \ type de données		Observées	Imputées
Sans flag	Moyenne	3259,69	3201,02
Avec flag	Moyenne	3259,69	3557,32
Sans flag	Ecar-type	2114,16	1765,41
Avec flag	Ecar-type	2114,16	1821,09

Parmi les 954 observations à données manquantes, 656 d'entre elles sont renseignées au niveau d'une source de données administrative. Une comparaison a été faite entre leurs valeurs et celles provenant des imputations avec flag et des imputations sans flag. Pour des raisons de confidentialité relative à l'accès aux micro-données détenues par un organisme public, un fichier des données imputées avec des identifiants anonymisés leur a été transmis pour réaliser le graphique 1 et le tableau 3.

Graphique 1



L'allure des 3 courbes ci-dessus confirme les résultats antérieurs, c'est-à-dire par rapport aux valeurs observées dans la source auxiliaire une sous-estimation des salaires imputés sans flag et une surestimation des salaires imputés avec flag.

Tableau 3 : Erreur entre les valeurs imputées et les valeurs observées dans la source externe

Flag	Nb_imputations	Erreur_imputation	Nb_observations
NON	50	2087,45	656
OUI	50	1871,20	656

Dans le tableau 3, le calcul, en terme de distance euclidienne de l'erreur entre les valeurs imputées et les valeurs observées dans la base, donne un avantage aux valeurs imputées avec flag. Il en résulte en fin de compte une confirmation de notre hypothèse.

3. Conclusion

En définitive, si on peut constater une relative unanimité quant à la supériorité des méthodes d'imputation multiple sur celles dites simples, sa concrétisation dépend des conditions de leur mise en œuvre : l'utilisation du programme IVEware, avec flag ou sans flag, en est une illustration.

Bibliographie

[1] ALLISON Paul D. (2000), Multiple imputation for Missing Data : A cautionary Tale , Sociological Methods and Research, 28, 301-309 Sage publications.

[2] ALLISON Paul D. (2001), Missing Data, Sage publications.

[3] CARON Nathalie, Les principales techniques de correction de la non-réponse et les modèles associés, Série des Documents de Travail « Méthodologie Statistique » N° 9604.

[3] CARTWRIGHT M.(2003), Data Imputation Techniques for Software Engineering : Case for Support.

[4] DEMPSTER A.P., LAIRD N.M. and RUBIN D.B., Maximum likehood estimation from incomplete data via EM algorithm. Journal of Royal Statistical Society, Series B, 39, 1-38.

[5] DIA M. (2003), Réponses aux non-réponses, Document de recherché, MS N°2003-1, ID :08-03-0016I, CEPS-INSTEAD, Grand Duché du Luxembourg.

[6] DONZE Laurent, Imputation to correct the item nonresponse for the KOF/ETHZ-Innovation Survey 1999, Zurich, September 2000.

[7] Eurostat, Règles d'imputation transversales et application aux micro-données, Doc PAN 47/1995

[8] HAZIZA David, Inférence en présence d'imputation : un survol, Journées de Méthodologie Statistique, 16-17 décembre 2002 à Paris

[9] HORTON N., LIPSITZ S.R. (2001), Multiple Imputation, Comparison of Software Packages for Regression Models With Missing Dat, The American Statistical Association, August 2001, Vol. 55, N° 3.

[10] HOX J.J. (1999), A review of current software for handling Missing data, Kwantitatieve Methoden (1999), 62, 123-138.

[11] PARKER H. (2001), Multiple Imputation for Multivariate Continuous Data, Honour Thesis, University of Queensland.

[12] RUBIN D.B., LITTLE R.J.A. (1987), Statistical Analysis with missing data

[13] RAGHUNATHAN T.E. , LEPKOWSKI J.M., VAN HOEWYK J., and SOLENBERGER P. (1999), A multivariate technique for multiply imputing missing value using a sequence of regression models. Unpublished manuscript. Contact : teraghu@umich.edu

[14] SCHAFER J. L. (1997), Analysis of Incomplete Multivariate Data. London : Chapman & Hall.
Schulte E. Nordholt, Imputation: Methods, Simulation Experiments and Pratical Examples.

[15] WAYMAN J.C., Mutiple Imputation For Missing Data : What is IT and How Can I Use It ?, Johns Hopkins University, Annual Meeting of the American Educational Research Association, Chicago, IL.

Sites Web des logiciels d'imputation

SOLAS : http://www.statsol.ie/solas/solas.htm
PROC MI (Multiple Imputation) : http://www.sas.com/
NORM, CAT, MIX, PANEL : http://www.stat.psu.edu/~jls/
AMELIA : http://gking.harvard.edu/stats.shtml
SIRNORM : http://yates.coph.usf.edu/research/psmg/web.html
(??)
MICE(Multivariate Imputation by Chained Equations) : http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm
IVEware : http://www.isr.umich.edu/src/smp/ive/

Inventaire des logiciels ou macros disponibles

Macros ou Logiciels / Caractéristiques	AMELIA	SIRNORM	MI	IVE	SOLAS	MIX	NORM	CAT	MICE	PAN(EL)
Non ou semi-Paramétrique									*	
Echant-Ré-échantillonnage	*	*								
DA			*		*	*	*	*		*
EMV-Classique				*					*	
Variables continues		*	*	*	*	*	*		*	
Variables discrètes				*		*			*	
Variables Catégorielles				*		*		*		
Environnement	GAUSS	SAS	SAS	SAS		S-PLUS	S-PLUS	S-PLUS	S-PLUS	S-PLUS
Autonome	*				*	*	*	*		*
Windows	*	?	*	*	*	*	*	*	*	*
Unix		?	*	*		*	*	*	*	
Autres Syst. D'exploitation		?	*							*
Gratuit	*	*		*		*	*	*		*
Commercial			*		*					