REVUE DE STATISTIQUE APPLIQUÉE

J.-P. NAKACHE
A. GUEGUEN

Analyse multidimensionnelle de données incomplètes

Revue de statistique appliquée, tome 53, n° 3 (2005), p. 35-62
<http://www.numdam.org/item?id=RSA_2005__53_3_35_0>

Rev. Statistique Appliquée, 2005, **LIII** (3), 35-62

ANALYSE MULTIDIMENSIONNELLE DE DONNÉES INCOMPLÈTES

J.-P. NAKACHE*, A. GUEGUEN**

* *CNRS/INSERM U88-IFR 69*
** *INSERM U88-IFR 69*

RÉSUMÉ

Ce papier concerne principalement l'utilisation des imputations multiples pour analyser des données multidimensionnelles incomplètes. Cette méthodologie est appliquée à des données, déjà utilisées par les mêmes auteurs dans un récent article publié dans la RSA, pour étudier l'effet prédictif de l'état de santé perçu (ESP) sur la mortalité jusqu'à 6 ans avant la survenue du décès. Mais la présente étude concerne 14 956 hommes à risque au début de l'année 1991 et suivis dans la période 1991-2002 au cours de laquelle on observe 575 décès. De plus les valeurs de l'ESP manquantes depuis 1990 ont été remplacées dans le précédent article par la dernière valeur déclarée. Elles sont ici imputées en utilisant une méthode récente d'imputations multiples basée sur une procédure MCMC connue sous le nom de *data augmentation* et particulièrement utile dans les problèmes de données manquantes.

Le but d'un modèle d'imputation est de refléter correctement l'incertitude des données manquantes et de préserver les aspects importants de la distribution des données et des importantes relations entre les variables.

Les données manquantes de l'ESP ont été imputées à partir de différents sous-ensembles de données homogènes en incluant dans le modèle d'imputation l'âge et la pcs : sous-ensemble de sujets vivants en 2002 et les différents sous-ensembles de sujets décédés en 2002, 2001,…, 1991.

Mots-clés : *Imputations simples et multiples, modèle d'imputation, procédure MCMC, data augmentation, algorithme EM, données de survie groupées, covariables dépendant du temps, modèle discret de Cox, risque relatif.*

ABSTRACT

The main purpose of this paper concerns the use of multiple imputations in the analysis of incomplete multidimensional data. This methodoly is applied to data, already used in a recent paper by the same authors in RSA, to study the predictive effect of self-rated health (SHR) on mortality until 6 years before death occurrence. But the present study includes 14956 men of the cohort Gazel at risk at the beginning of 1991 and followed during the period 1991-2002. Moreover the missing values of SHR, present in the data since 1990, have been replaced by the last declared SHR value in the preceding paper. They have been here imputed using a recent and flexible multiple imputation method based on data augmentation MCMC procedure and especially useful in missing-data problems.

The goal of an imputation model is to properly reflect uncertainty and to preserve important aspects of the data distribution and important relationships between the variables.

The SHR missing data have been imputed from different homogeneous sub-samples introducing age and pcs in the imputation model : sub-sample of subjects alive in 2002 and the different sub-samples of subjects died in 2002, 2001, …, 1991.

Keywords : *Simple and multiple imputations, data augmentation, EM algorithm, grouped survival data, time-dependent covariates, discrete Cox model, relative risk.*

Introduction

Le problème des données manquantes se pose souvent dans le traitement statistique de données, surtout quand on analyse des données multidimensionnelles. Dans les études épidémiologiques par exemple, les données sont recueillies au moyen de questionnaires ou d'interviews ou peuvent être extraites de dossiers de malades de services hospitaliers. Certaines de ces données peuvent être manquantes si des sujets refusent de répondre à certaines parties du questionnaire, si certains sujets ne se souviennent pas de certains évènements ou si des données ou certains dossiers sont perdus pour des raisons techniques lors de la collecte des données.

Ce problème de données manquantes se pose également quand on analyse des données longitudinales pour lesquelles il arrive que certaines données ne soient pas disponibles pendant une ou plusieurs périodes de collecte des données.

Une pratique courante dans les analyses épidémiologiques est d'éliminer de l'analyse tous les sujets pour lesquels des observations sont manquantes et d'analyser les données complètes. On fait ainsi implicitement l'hypothèse que les données sont manquantes complètement au hasard.

Dans une analyse multidimensionnelle même si, pour chaque variable, le taux de données manquantes est bas, la proportion de sujets à éliminer peut être grande car il est probable que peu de sujets aient des données complètes. Ce sous-ensemble de données complètes n'est plus représentatif de l'échantillon total, ce qui rend ce type de méthode inefficace en ce sens que les résultats sont biaisés et l'analyse manque de puissance. Cette approche fait perdre beaucoup d'information sur les données, surtout quand plusieurs covariables sont concernées et quand une grande partie des sujets ont des données incomplètes pour au moins une des variables.

Une alternative à cette dernière approche et qui est utilisée communément dans des analyses épidémiologiques à partir de covariables catégorielles (variables qualitatives ordinales ou nominales) est d'ajouter aux différentes catégories une modalité « manquant », ce qui permet de garder tous les sujets dans l'analyse mais qui introduit certainement un biais qui peut être sévère même quand les données sont *manquantes complètement au hasard.*

Un rappel des méthodes classiques d'imputation simple et de l'hypothèse de données manquantes au hasard est présenté dans les paragraphes 1 et 2. La méthode d'imputations multiples, fondée sur la méthode MCMC connue sous le nom de *data augmentation*, fait l'objet du paragraphe 3 qui constitue la partie principale de l'article. Le paragraphe 4 est consacré à l'application de cette technique à des données de la cohorte Gazel [Goldberg et Leclerc, (1994)] déjà utilisées dans un récent article des mêmes auteurs [Nakache *et al*, (2004)], mais dans la période 1989-1999 et en

remplaçant toute valeur manquante pour l'ESP par la dernière valeur déclarée. Il s'agit dans le présent article de 14956 hommes de la cohorte Gazel vivants au début de 1991 et suivis jusqu'en 2002, avec le même but, à savoir l'étude de l'effet prédictif de l'état de santé perçu sur la mortalité, en prenant en compte dans le modèle, l'âge, la pcs et la période calendaire.

1. Imputations simples

Depuis assez longtemps et jusque dans les années 1990, les statisticiens ont eu recours à des procédures d'imputation simple consistant à remplacer une donnée manquante par une valeur estimée prédite ou simulée, ce qui permet de garder tous les sujets d'un échantillon et de pouvoir ainsi effectuer des analyses en utilisant les logiciels standards. Ces méthodes d'imputation simple sont faciles à mettre en œuvre mais elles présentent des inconvénients en ce sens qu'elles introduisent pratiquement toutes un biais dans les estimations des paramètres [Little et Rubin, 1987]. Greenland et Finkle (1995) et Vach et Blettner [1991] montrent de tels biais dans le contexte de la régression logistique, présentent des expériences de simulation qui révèlent les limites de ces méthodes et notent que même l'analyse sur données complètes peut conduire à de meilleurs résultats que ceux obtenus par ces méthodes d'imputation simple. Parmi ces méthodes d'imputations simples, les plus courantes sont les suivantes :

1.1. Imputation d'une valeur manquante par la moyenne

Une première technique d'imputation consiste à remplacer toutes les données manquantes par la moyenne de la variable. Ainsi si $x_1, x_2, \ldots, x_a$ sont les valeurs observées d'une variable et si $x_{a+1}, x_{a+2}, \ldots, x_n$ sont des valeurs manquantes pour cette même variable, chaque valeur manquante est remplacée par :

$$\overline{x}_{\text{obs}} = \frac{1}{a} \sum_{i=1}^{a} x_i$$

La figure 1, qui fournit les histogrammes de la variable avant et après imputation, montre que la distribution de la variable est déformée quand les valeurs manquantes sont remplacées par la moyenne.

La figure 2 représente le nuage de points obtenu avec en ordonnée une variable avec données manquantes et en abscisse une autre variable sans données manquantes. Les observations qui sont imputées sont représentées en gris et on observe que la corrélation entre les deux variables est déformée après imputation.

La substitution des valeurs manquantes, pour une variable, par la moyenne de cette variable entraîne donc une déformation de la distribution de la variable et affaiblit ses relations avec d'autres variables.

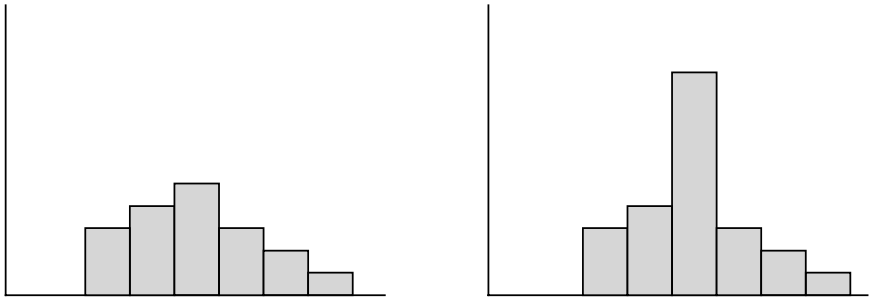


FIGURE 1
Remplacement des données manquantes par la moyenne

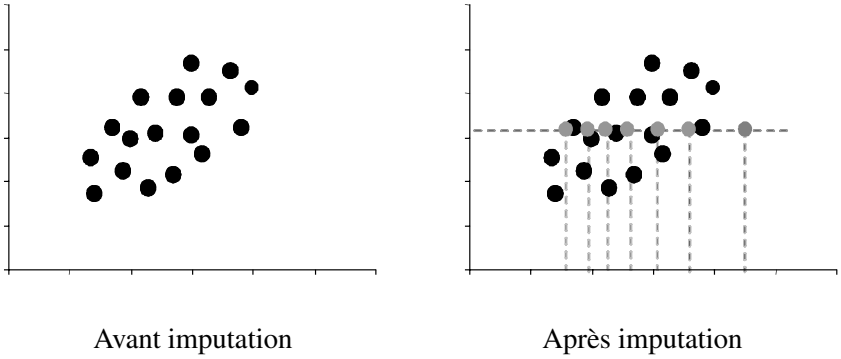


FIGURE 2
Remplacement des données manquantes par la moyenne

1.2. Imputation d'une valeur manquante par la valeur prédite par un modèle de régression

Une amélioration de la technique précédente est la suivante : on considère une variable X pour laquelle toutes les données sont observées $(x_1, x_2, \ldots, x_n)$ et une variable Y pour laquelle des données sont manquantes ($y_1, y_2, \ldots, y_a$ sont observées et $y_{a+1}, y_{a+2}, \ldots, y_n$ sont manquantes).

On effectue la régression de Y sur X, à partir des données complètes $\{(x_i, y_i)\,;\ i = 1, \ldots, a\}$, et on impute chaque donnée manquante sur la variable Y par la valeur prédite par la droite de régression. La distribution de Y n'est pas dans ce cas déformée, mais la corrélation entre X et Y est augmentée artificiellement, comme le montre la figure 3.

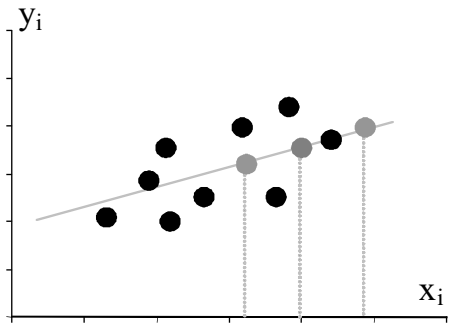


FIGURE 3
Remplacement des données manquantes par la valeur prédite par un modèle de régression

1.3. Imputation d'une valeur manquante par la valeur prédite plus un résidu aléatoire

Une imputation simple un peu plus correcte que la précédente consiste à remplacer une valeur manquante par une valeur prédite par le modèle à laquelle on ajoute un résidu aléatoire normal de moyenne nulle et de variance s^2 : $\widehat{y}_i + e_i$ avec $e_i \sim \mathcal{N}(0, s^2)$. Dans ce cas (figure 4), ni la distribution de Y, ni la corrélation avec X ne sont déformées, et on peut penser que la méthode d'imputation est correcte. On a cependant fait l'hypothèse suivante : pour chaque valeur de X fixée, la distribution des Y manquants est la même que celle des Y observés, puisqu'on a remplacé les valeurs de Y manquantes par des valeurs tirées au sort dans la distribution des Y observés. Cette hypothèse est connue sous le nom *d'hypothèse de données manquantes au hasard*.

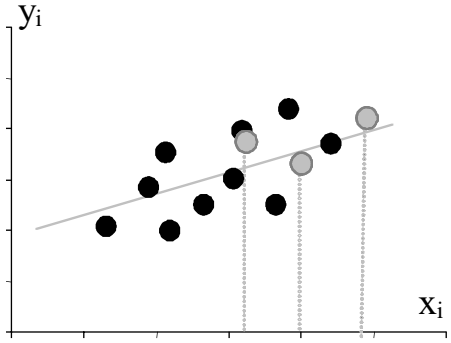


FIGURE 4
Remplacement des données manquantes par la valeur prédite + un résidu aléatoire

2. Hypothèse de données manquantes au hasard

Les probabilités de non-réponse peuvent dépendre des valeurs qui sont observées, mais pas de celles qui sont manquantes. On est, dans ce cas, dans l'hypothèse de données manquant au hasard qui permet mathématiquement d'éviter un modèle probabiliste explicite pour les données manquantes. Le fait qu'une donnée soit manquante au hasard pour une variable (par exemple l'ESP dans les données utilisées pour illustrer la méthode présentée) dépend uniquement de valeurs observées pour cette variable et aussi pour d'autres variables comme l'age, le sexe et la pcs du sujet.

Cette hypothèse est non testable : on ne peut pas tester si les données sont manquantes au hasard ou non, à moins de connaître les valeurs réelles des données manquantes. Cette hypothèse que les données sont manquantes au hasard ou non est relative : elle peut être vraie ou non selon les données observées que l'on prend en compte.

Ce qui nous amène au choix du modèle d'imputation : il doit inclure les variables qui interviennent dans l'analyse qu'on souhaite réaliser, les variables prédictives du fait que les données sont manquantes et toutes les variables susceptibles d'apporter de l'information si elles sont liées aux variables pour lesquelles on a des données manquantes. Plus le modèle d'imputation est riche, plus on a de chance que l'hypothèse que les données sont manquantes au hasard soit vraie.

3. Méthode des imputations multiples

Quelque soit la méthode d'imputation simple utilisée, les valeurs imputées ne sont que des prévisions des vraies valeurs inconnues. Même si les valeurs manquantes sont imputées correctement, c'est-à-dire telles que les distributions des variables et les relations entre elles soient parfaitement préservées, le tableau ainsi complété ne permet pas de tenir compte de l'incertitude des données imputées. Toute analyse qui ignore l'incertitude des valeurs manquantes conduit à des erreurs sur les degrés p de significativité et les intervalles de confiance des paramètres estimés, d'où l'intérêt d'utiliser des méthodes plus sophistiquées qui fournissent des imputations multiples qui sont plusieurs valeurs plausibles d'une donnée manquante.

La méthode des imputations multiples suppose que les données sont manquantes au hasard. Quand plusieurs variables présentent des données manquantes, la procédure est un peu plus compliquée et il s'agit d'une procédure itérative.

Ces méthodes d'imputations multiples qui sont restées longtemps inutilisées à cause sans doute de leur complexité, connaissent un développement important en raison de la disponibilité de logiciels permettant de les appliquer.

Le but des imputations multiples n'est pas de prédire les données manquantes avec la plus grande précision ni de décrire les données de la meilleure façon possible, mais de refléter correctement l'incertitude des valeurs manquantes et de préserver les aspects importants des distributions et les relations importantes entre variables.

Une méthode d'imputations multiples, proposée par Rubin (1978), est restée largement inconnue et non utilisée par des non experts à cause principalement d'un

manque d'outils de calculs pour créer des imputations multiples. L'application de cette méthodologie à la création d'imputations multiples a été en pratique possible bien plus tard grâce à la mise au point de méthodes de calcul connues sous le nom de Monte Carlo Markov Chain (MCMC) [Gilks, Richardson et Spiegelhalter,(1996)] parmi lesquelles les algorithmes *Gibbs sampling* et *Metropolis-Hastings* [Demster *et al*., (1977)] et l'algorithme *Data augmentation* [Tanner et Wong (1987)] qui ont entraîné des progrès considérables dans l'analyse statistique de données incomplètes. Ce sont des méthodes itératives de tirage pseudo-aléatoires dans des distributions compliquées, permettant de créer une *chaîne de Markov* qui converge vers la distribution prédictive *a posteriori* des données manquantes.

L'algorithme «*Data Augmentation*», qui a été adapté et implémenté par J.L. Schafer [Schafer et Olsen, (1998); Schafer, (2000)] pour la création d'imputations multiples, est l'algorithme utilisé dans la méthode des imputations multiples présentée dans la suite. Cette méthode des imputations multiples s'effectue en trois étapes : (1) remplacement de chaque valeur manquante par $m > 1$ valeurs simulées (en général $m = 5$) en utilisant l'algorithme «*data augmentation*» (DA), ce qui conduit à m tableaux de données complétées, (2) analyse de chacun de ces m tableaux et (3) utilisation des règles de Rubin pour combiner les m résultats de ces analyses et obtenir un ensemble global de paramètres estimés et de leurs écarts-type.

3.1. Remplacement de chaque valeur manquante par $m > 1$ valeurs simulées

Chaque valeur manquante est remplacée par un ensemble de m valeurs plausibles tirées de la distribution prédictive $P(Y_{\text{manq}}|Y_{\text{obs}}, \theta)$ [Dempster *et al*., (1977)]. Exceptés pour certains cas triviaux, les distributions de probabilité (distributions prédictives des données manquantes) à considérer pour l'obtention de bonnes imputations multiples sont très compliquées.

Algorithme «data augmentation» *(DA)*

L'algorithme DA repose sur une approche bayésienne dans laquelle les données et les paramètres sont considérés comme des variables aléatoires et il est donc nécessaire de spécifier la distribution *a priori* des paramètres. L'algorithme simule alternativement données manquantes et paramètres à l'aide d'une chaîne de Markov qui, au bout d'un certain temps, se stabilise et converge en probabilité; la distribution des paramètres se stabilise vers la distribution *a posteriori* des paramètres et la distribution des données manquantes se stabilise vers la distribution prédictive *a posteriori* des données qui est la distribution utilisée pour créer de bonnes imputations multiples. L'algorithme DA consiste en deux étapes :

Étape I : Imputation des valeurs manquantes

Tirage au hasard de $Y_{\text{manq}}^{(t+1)}$ dans la distribution de probabilité $P(Y_{\text{manq}}|Y_{\text{obs}}, \theta^{(t)})$

Étape P : Distribution *a posteriori* des paramètres θ

Tirage au hasard de $\theta^{(t+1)}$ dans la distribution de probabilité $P(\theta|Y_{\text{obs}}, Y_{\text{manq}}^{(t+1)})$, ce qui crée une chaîne de Markov $(Y_{\text{manq}}^{(1)}, \theta^{(1)}), (Y_{\text{manq}}^{(2)}, \theta^{(2)}), \ldots$ qui converge en probabilité vers $P(Y_{\text{manq}}, \theta|Y_{\text{obs}})$, distribution prédictive *a posteriori* des données

(utilisée pour les imputations). Cet algorithme est un schéma de Gibbs particulier [Roberts, (1996)].

Les annexes 1 et 2 présentent le déroulement des procédures EM et DA dans les cas respectifs de deux variables binaires et de deux variables continues.

Utilisation de l'algorithme DA pour générer des imputations multiples

L'algorithme DA, spécialement utile dans les problèmes de données manquantes, requiert des valeurs de départ pour les paramètres θ : les estimations des paramètres fournies par l'algorithme EM [Dempster *et al.*, (1977)] constituent un bon choix des valeurs de départ.

L'algorithme EM est utilisé pour maximiser la fonction de vraisemblance et donc le point de départ de l'algorithme est une approximation de l'estimation du maximum de vraisemblance.

On obtient itérativement des $Y_{\text{manq}}^{(t)}$ en faisant tourner la procédure DA

$$\theta^0 \longrightarrow (Y_{\text{manq}}^1) \longrightarrow \theta^1 \longrightarrow (Y_{\text{manq}}^2) \longrightarrow \theta^2 \longrightarrow (Y_{\text{manq}}^3) \longrightarrow \ldots \longrightarrow \theta^t \longrightarrow (Y_{\text{manq}}^{t+1}) \longrightarrow \ldots$$

jusqu'à la convergence de la chaîne de Markov précédente vers sa mesure invariante $P(Y_{\text{manq}}, \theta | Y_{\text{obs}})$. On obtient ainsi un tableau imputé T1 (tableau de données complétées). En pratique, on construit m tableaux imputés qu'on obtient de la manière suivante :

- on détermine le nombre d'itérations k correspondant à un temps de chauffe (*burn-in*) pour s'assurer que la chaîne de Markov a bien convergé; en général le nombre d'itérations nécessaire à la convergence de l'algorithme DA est du même ordre de grandeur que le nombre d'itérations nécessaire à la convergence de l'algorithme EM. De plus le nombre d'itérations à effectuer pour atteindre la convergence peut être évalué à l'aide de graphiques. Pour s'assurer une marge de sécurité, il est préférable de choisir une valeur de k supérieure à celle donnée par l'algorithme EM et celle suggérée par les validations graphiques;
- on construit ensuite une chaîne de Markov de longueur mk et on stocke les m valeurs de Y_{manq} obtenues après k itérations, $2k$ itérations et … mk itérations. On peut également construire m chaînes de Markov indépendantes et chacune de longueur k et stocker les m valeurs de Y_{manq} correspondant à la fin de chacun des cycles.

Convergence de l'algorithme DA

La chaîne est dite convergente ou a atteint la stationnarité après t itérations, si $\theta^{(t)}$ est indépendant de $\theta^{(0)}$, $\theta^{(2t)}$ est indépendant de $\theta^{(t)}$, etc. Cette convergence, qui est théoriquement compliquée à vérifier, peut être contrôlée en pratique en examinant les graphiques des séries de θ et les fonctions d'autocorrélations entre séries d'estimations des paramètres successives qui sont présentées dans l'annexe 3.

3.2. Analyse statistique identique de chacun des m fichiers de données complétées

Les m imputations fournissent m ensembles de données complétées sur lesquels on réalise la même analyse. On obtient ainsi des estimations des paramètres avec leurs intervalles de confiance, pour chacun des tableaux de données complétées.

3.3 Combinaison des résultats des m analyses : règles de Rubin

Les résultats issus de la même analyse, effectuée sur chacun de ces m tableaux (estimations des paramètres et écarts-type), sont ensuite combinés en utilisant des règles simples (règles dues à Rubin [1987]) qui donnent lieu à des estimations globales des paramètres et leurs intervalles de confiance, reflétant ainsi l'incertitude des données manquantes.

Une fois les imputations multiples créées, les fichiers peuvent être analysés par la plupart des méthodes appropriées à des données complètes. On peut par exemple soumettre les fichiers à un modèle de régression linéaire ou à un modèle logistique en utilisant n'importe quel logiciel statistique standard.

Méthode de Rubin

Dans l'application présentée, on utilise le logit $g(x)$ comme fonction de lien entre la probabilité de décès $\pi(x) = (PY = 1|x)$ et p variables explicatives $x_1, x_2, \ldots, x_p$ formant le vecteur x , dont l'expression est :

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Il s'agit alors d'estimer les paramètres $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ du modèle logistique. Pour une raison de simplicité d'écriture, on notera dans la suite $\widehat{\beta}$ un des paramètres estimés et $\widehat{\sigma}^2_{\hat{\beta}}$ sa variance estimée. Les m analyses sur tableaux imputés fournissent m estimations plausibles $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_m$ de β et les m variances correspondantes $\widehat{\sigma}^2_{\hat{\beta}_1}, \widehat{\sigma}^2_{\hat{\beta}_2}, \ldots, \widehat{\sigma}^2_{\hat{\beta}_m}$.

L'estimation par imputations multiples du paramètre β est : $\overline{\beta} = \frac{1}{m}\sum_{i=1}^{m}\widehat{\beta}_i$. La variance totale du paramètre estimé $\overline{\beta}$ a 2 composantes qui tiennent compte de la variabilité intra-tableau et inter-tableau et qui sont :

- la variance d'imputation intra : $V_{\text{intra}} = \frac{1}{m}\sum_{i=1}^{m}\widehat{\sigma}^2_{\hat{\beta}_i}$ qui est la moyenne des variances estimées des paramètres obtenues pour les m tableaux imputés
- la variance d'imputation inter : $V_{\text{inter}} = \frac{1}{m-1}\sum_{i=1}^{m}(\widehat{\beta}_i - \overline{\beta})^2$ qui est la variance de l'ensemble des estimations $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_m$.

La variance d'imputation totale est la somme des deux composantes avec un terme de correction additionnel qui tient compte de l'erreur de la simulation dans $\overline{\beta}$,

soit :

$$V_{\text{tot}} = V_{\text{intra}} + \left(1 + \frac{1}{m}\right) V_{\text{inter}}$$

La taille de V_{inter}, relativement à V_{intra}, est le reflet de l'information contenue dans la partie manquante des données relativement à la partie observée.

Remarque. – Si les données sont complètes (pas de données manquantes) alors $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_m$ sont identiques, V_{inter} est nulle et V_{tot} est tout simplement égale à V_{intra}.

Un intervalle de confiance (IC) à 95 % approximatif peut être obtenu : $\overline{\beta} \pm 1.96\sqrt{V_{\text{tot}}}$. En général, il est préférable de calculer les intervalles de confiance en utilisant l'approximation $\overline{\beta} \pm t_v\sqrt{V_{\text{tot}}}$, où t_v est le quantile de la distribution de Student dont le degré de liberté v est :

$$v = (m-1)\left(1 + \frac{mV_{\text{intra}}}{(m+1)V_{\text{inter}}}\right)^2$$

Ces règles de combinaison des résultats sont utilisées dans les procédures MI et MIANALYZE du logiciel SAS [SAS/STAT Software (2001)].

3.4. Fraction d'information manquante

Rubin (1987) montre qu'une estimation de la fraction (ou taux) d'information manquante à propos de la quantité β est : $\gamma = \dfrac{r + 2/(v+3)}{r+1}$ où $r = \dfrac{(1+m^{-1})V_{\text{inter}}}{V_{\text{intra}}}$ $= \dfrac{(V_{\text{tot}} - V_{\text{intra}}}{V_{\text{intra}}}$ est l'accroissement relatif de variance due aux données manquantes. Cette fraction γ est une notion différente du pourcentage connu de données manquantes car elle dépend des liaisons entre variables. D'autre part, plus γ est grand, plus lente est la convergence de l'algorithme. Ces quantités γ et r sont des outils diagnostics importants en pratique; elles donnent une idée de l'importance des données manquantes dans une application.

3.5 Efficacité d'une estimation basée sur m imputations

Rubin (1987) montre aussi que l'efficacité d'une estimation basée sur m imputations est approximativement égale à : $\left(1 + \frac{\gamma}{m}\right)^{-1}$ où γ est le taux d'information manquante. Le tableau 1, qui est extrait de [Schafer, (2000)], fournit l'efficacité (en %) en fonction de m et de γ. Il montre que le gain diminue rapidement après très peu d'imputations. Par exemple : pour $\gamma = 0.3$ (qui correspond à 30 % d'information manquante ce qui représente un taux modérément important dans beaucoup d'applications), avec $m = 5$ imputations on atteint déjà 94 % d'efficacité. Cette efficacité passe à 97 % quand $m = 10$: un gain plutôt léger en doublant l'effort de calcul. Dans la plupart des applications, l'avantage qu'on a à considérer plus de 5 imputations est minime.

TABLEAU 1
Pourcentages d'efficacité en fonction de m (nombre d'imputations) et γ (taux d'information manquante) [Schafer J.L., 2000]

	γ				
m	.1	.3	.5	.7	.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	95
20	100	99	98	97	96

Remarque. – Le lecteur pourra se reporter à l'ouvrage de Schafer (2000) pour tous les détails théoriques concernant la fraction d'information manquante et l'efficacité d'une estimation basée sur m imputations.

4. Application

Les données de cette application sont des données de la cohorte Gazel [Goldberg et Leclerc, (1994)] qui ont été déjà utilisées dans un récent article des mêmes auteurs [Nakache *et al*., (2004)], mais dans la période 1989-1999 et en remplaçant toute valeur manquante de l'ESP par la dernière valeur de l'ESP déclaré. Dans le présent article la population étudiée est constituée des 14 956 hommes de la cohorte vivants au début de 1991 et suivis dans la période 1991-2002 au cours de laquelle on observe 575 décès pour cette population.

On cherche, comme dans le précédent article, à étudier l'effet prédictif de l'état de santé perçu sur la mortalité. De nombreuses études ont montré que l'état de santé perçu est prédictif de la mortalité, mais en utilisant une seule mesure de l'état de santé perçu déclaré à l'entrée dans l'étude.

Les données de la cohorte Gazel [Goldberg et Leclerc, (1994)], du fait qu'on a tous les ans la déclaration de l'état de santé perçu, permettent de répondre à une question plus précise : l'effet prédictif de l'état de santé perçu sur la mortalité est-il le même un an après, deux ans après ou même six ans après? Pour analyser ces données, on a utilisé le modèle de Cox [1972] sur données groupées qui est équivalent à un modèle logistique binaire [Hosmer et Lemeshow, (1989); Nakache et Confais, (2003)] avec l'année comme intervalle de temps. L'effet prédictif de l'état de santé perçu a été estimé après ajustement sur la pcs (catégorie socio-professionnelle), l'âge et la période calendaire.

4.1. Modèle d'imputations multiples

La méthode des imputations multiples a été utilisée puisque des valeurs de l'ESP sont manquantes dans les données à partir de 1990. Les imputations multiples ($m = 5$) ont été créées séparément pour les sous-groupes suivants : les hommes vivants en 2002, les hommes décédés en 2002, les hommes décédés en 2001, les hommes décédés en 2000,…, les hommes décédés en 1992 et les hommes décédés en 1991, en incluant dans le modèle d'imputation l'ESP pour les années allant de 1989 à 2002, l'âge en 2002 et la pcs (en 3 classes binaires avec la classe des cadres et ingénieurs comme classe de référence). Pour chaque sous-groupe considéré, 5 tableaux de données complétées ont été obtenus à partir desquels ont été constitués les 5 fichiers contenant chacun d'eux, les hommes décédés et vivants pendant la période 1991-2002 avec l'information concernant l'ESP de 1990 à 2002, l'âge du sujet en 1989 et la pcs.

Pour estimer le risque relatif de décès quand le sujet a déclaré un mauvais état de santé un an avant, on procède de la manière suivante : on considère les sujets vivants au début de 1991 et on relie la probabilité de décéder durant l'année 1991 à la déclaration d'un bon ou d'un mauvais état de santé en 1990. On considère également les sujets vivants au début de 1992 et on relie la probabilité de décéder durant l'année 1992 à la déclaration d'un bon ou d'un mauvais état de santé en 1991 et ainsi de suite pour toutes les années. Toutes ces analyses sont réalisées conjointement en tenant compte du fait qu'un même sujet intervient plusieurs fois. Pour l'analyse deux ans avant, on procède de manière semblable : on considère l'ensemble des sujets vivants au début de 1992 et on relie la probabilité de décéder durant l'année 1992 à la déclaration d'un bon ou d'un mauvais état de santé en 1990, c'est-à-dire deux ans avant.

L'analyse de ces données de survie groupées a été effectuée, comme dans l'article précédent, en utilisant la procédure LOGISTIC de SAS [SAS/STAT Software, (2001)] à partir du fichier constitué comme suit : un sujet vivant en 2002 est représenté dans le fichier par 12 observations (ou lignes) correspondant aux années de 1991 à 2002 et un sujet décédé est représenté par autant d'observations que d'années de 1991 à l'année de son décès : ainsi un sujet qui décède en 1995 est représenté dans le fichier par 5 observations de 1991 à 1995. Chacune des observations du fichier contient la période calendaire, le statut pendant cette année (décédé ou vivant), la pcs (en 3 classes binaires avec la classe des cadres et ingénieurs comme classe de référence), l'âge pendant cette année en 3 classes binaires (46-50, 51-55 et 56-65 avec la classe 41-45 comme classe de référence), et l'ESP (bonne santé, mauvaise santé avec bonne santé comme classe de référence) 1an, 2 ans, jusqu'à 6 ans avant l'année en question.

Le tableau 2 fournit les résultats du modèle de l'effet prédictif de l'ESP sur le décès l'année suivante (risques de décès à t en fonction de l'ESP à $(t-1)$, ajusté sur l'âge, la PCS et la période calendaire). Le tableau 3 résume les résultats des 6 modèles correspondant à l'effet prédictif de l'état de santé perçu sur la mortalité un an après, deux ans, jusqu'à six ans après.

TABLEAU 2
Modèle de Cox sur données groupées par année (imputations multiples)
Risques relatifs RR de décés et IC 95 % au temps t en fonction de l'ESP déclaré à $(t-1)$ (t variant de 1991 à 2002)*

Variable	**RR***	IC 95%	
ESP déclaré à t-1 = bonne santé	1		
ESP déclaré à t-1 = mauvaise santé	3.65	2.92	4.63
Age 41-45 (1)	**1**		
Age 46-50 (2)	1.06	0.65	1.72
Age 51-55 (3)	1.48	0.90	2.45
Age 56-65 (4)	2.32	1.36	3.98
Ingénieurs. Cadres (3)	**1**		
Prof. intermédiaires (4)	1.53	1.23	1.91
Employés (5)	1.91	1.30	2.81
Ouvriers (6)	2.19	1.67	2.88
Période 2002	**1**		
Période 1991	0.90	0.55	1.47
Période 1992	1.04	0.65	1.67
Période 1993	0.99	0.61	1.61
Période 1994	1.02	0.63	1.64
Période 1995	0.72	0.44	1.20
Période 1996	0.95	0.58	1.54
Période 1997	0.99	0.61	1.61
Période 1998	0.90	0.55	1.48
Période 1999	1.08	0.67	1.76
Période 2000	0.93	0.57	1.55
Période 2001	1.01	0.61	**1.67**

* ajusté sur l'âge, la pcs et la période calendaire

4.2. Imputations simples : remplacement par le dernier état de santé déclaré

Les données de Gazel, du fait qu'on a tous les ans la déclaration de l'état de santé perçu, permettent d'utiliser la méthode d'imputation simple des données de l'ESP manquant qui consiste à compléter le tableau des données en remplaçant toute valeur manquante de l'ESP par la dernière valeur déclarée. On a donc appliqué sur ces données cette méthode d'imputation simple, utilisée dans le précédent article, dans le but de comparer les résultats obtenus par les deux méthodes. Le tableau 4 fournit les résultats des six modèles appliqués à ces données après imputation simple, et la figure 5 constitue une représentation des résultats obtenus par les deux méthodes. L'examen des différents risques de cette figure 5 obtenus en remplaçant toute valeur manquante de l'ESP par la dernière valeur déclarée montre que pour des individus vivants au début d'une année donnée, le risque de décéder durant l'année est environ trois fois plus important si l'individu s'est déclaré en mauvais état de santé l'année précédente et ce, après ajustement sur l'age, la pcs et la période calendaire. Le risque est environ deux fois plus important si on prend en compte la déclaration de l'état de santé deux ans

TABLEAU 3
Modèle de Cox sur données groupées par année (imputations multiples)
Risques relatifs RR de décés et IC 95 % au temps t en fonction de l'ESP déclaré à $(t-j)$ avec $j = 1, \ldots, 6$ (t variant de 1991 à 2002)*

à	ESP déclaré	RR*	IC95%	
t-1	Bonne santé	1		
	Mauvaise santé	**3.65**	**2.92**	**4.63**
t-2	Bonne santé	1		
	Mauvaise santé	**2.38**	**2.01**	**2.90**
t-3	Bonne santé	1		
	Mauvaise santé	**1.92**	**1.50**	**2.35**
t-4	Bonne santé	1		
	Mauvaise santé	**2.05**	**1.51**	**2.45**
t-5	Bonne santé	1		
	Mauvaise santé	**1.35**	**1.03**	**1.81**
t-6	Bonne santé	1		
	Mauvaise santé	**1.55**	**1.23**	**2.06**

* ajusté sur l'âge, la pcs et la période calendaire

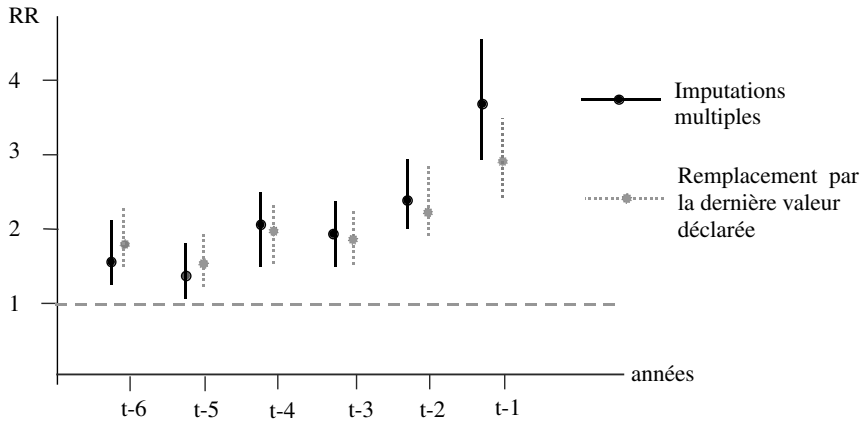


FIGURE 5
Risques relatifs de décés et IC 95 % au temps t en fonction de l'ESP déclaré à $(t-j)$ avec $j = 1, \ldots, 6$: ajusté sur l'âge, la pcs et la période calendaire

TABLEAU 4
Modèle de Cox sur données groupées par année (imputations simples)
Risques relatifs RR de décés et IC 95 % au temps t en fonction de l'ESP déclaré à $(t-j)$ avec $j = 1, \ldots, 6$ (t variant de 1991 à 2002)*

à	ESP déclaré	RR*	IC95%	
t-1	Bonne santé	1		
	Mauvaise santé	**2.90**	**2.42**	**3.50**
t-2	Bonne santé	1		
	Mauvaise santé	**2.21**	**1.90**	**2.83**
t-3	Bonne santé	1		
	Mauvaise santé	**1.85**	**1.52**	**2.25**
t-4	Bonne santé	1		
	Mauvaise santé	**1.95**	**1.50**	**2.32**
t-5	Bonne santé	1		
	Mauvaise santé	**1.50**	**1.20**	**1.95**
t-6	Bonne santé	1		
	Mauvaise santé	**1.78**	**1.47**	**2.25**

* ajusté sur l'âge, la pcs et la période calendaire

avant et on voit que le risque relatif de décès est toujours significativement différent de 1 si on prend en compte la déclaration de l'état de santé jusqu'à six ans avant.

C'est ce résultat qui intéresse les épidémiologistes : on voit très clairement que le risque n'est pas le même si on considère l'état de santé un an avant ou six ans avant; cependant, la plupart des études réalisées sur ce sujet, et qui n'ont qu'une mesure de l'état de santé à l'entrée dans l'étude font, en utilisant un modèle de survie classique [Cox, (1972)], implicitement l'hypothèse que les risques sont proportionnels c'est-à-dire que les risques relatifs sont constants dans le temps. Ils estiment un risque relatif moyen pour toute la période, sous estimé un an avant, et surtout surestimé six ans avant ce qui peut être grave en ce qui concerne l'interprétation de ce résultat.

La figure 5 montre aussi que les deux méthodes d'imputations fournissent les mêmes résultats quand on se place entre deux et six ans avant. Par contre, l'effet prédictif d'un mauvais état de santé sur la mortalité un an après est très différent selon qu'on considère la méthode des imputations multiples ou la méthode de remplacement de l'ESP par la dernière valeur déclarée. Ceci est du au fait que dans la méthode des imputations multiples, on s'appuie, pour les sujets qui sont décédés, sur la distribution observée de l'état de santé des sujets un an avant qu'ils ne décèdent alors que dans la méthode de remplacement, on ne tient pas compte de cette information.

5. Logiciels

Les procédures MI et MIANALYZE du logiciel SAS [SAS/STAT Software, (2001)] et également un logiciel extrêmement bien documenté dû à Schafer permettent de réaliser des imputations multiples. Ce dernier logiciel, implémenté sur PC avec Window(95/NT), est constitué de plusieurs procédures : NORM pour des données multidimensionnelles continues, CAT pour des données multidimensionnelles catégorielles, MIX pour des données mixtes : continues et catégorielles et PAN pour des données groupées, données multi-niveaux comme les données longitudinales. Ces procédures sont disponibles sous forme de package dans S+ et sont aussi téléchargeables gratuitement sur le net : http ://stat.psu.edu/ jls/misoftwa.html.

La méthode MCMC de la procédure MI du logiciel SAS est exactement la procédure NORM de Schafer qui suppose la multinormalité des données.

Les données réelles se conforment rarement à des modèles, comme le modèle multinormal. Mais la méthode d'imputations multiples est assez robuste quand on s'écarte du modèle d'imputation. En pratique, avant d'utiliser la procédure MCMC de MI, les variables qui ont des distributions très dissymétriques sont transformées (en prenant le logarithme, …) pour être plus proches de distributions normales. S'il s'agit de variables binaires ou ordinales, il est souvent acceptable d'imputer sous l'hypothèse de multinormalité et d'arrondir les valeurs continues imputées à la modalité la plus proche. Pour une variable binaire une valeur imputée inférieure ou égale à 0.5 (respectivement supérieure à 0.5) sera changée en 0 (respectivement en 1). Quand aux variables nominales, elles sont transformées en variables indicatrices en prenant une des modalités comme référence. Les valeurs imputées sont ensuite transformées dans leur échelle originale après imputation.

6. Conclusion

Effectuer des analyses sur données incomplètes en faisant l'hypothèse que les données sont manquantes complètement au hasard, peut poser des problèmes de puissance et surtout de biais. Il est toujours possible d'effectuer ces analyses et d'aborder le problème de ce biais potentiel au moment de la discussion des résultats. Mais il est actuellement possible de prendre en compte ce problème dès l'analyse en utilisant la méthode des imputations multiples qui est attractive pour plusieurs raisons : (1) elle permet d'utiliser les méthodes statistiques des logiciels standards, (2) un ensemble de m imputations peut être utilisé pour plusieurs analyses : il n'est souvent pas nécessaire de ré-imputer quand une nouvelle analyse est effectuée, (3) les inférences statistiques finales -p-values, intervalles de confiance, etc.– obtenues à partir des imputations multiples – sont généralement valables puisqu'elles incorporent l'incertitude des non réponses et (4) elle est très efficace même pour des petites valeurs de m.

Le choix d'une méthode appropriée de traitement des données manquantes doit être déterminé en tenant compte de la raison pour laquelle la donnée est manquante. Des études de validation sont nécessaires pour savoir si les données manquantes se produisent dans la population étudiée ou dans des sous-groupes spécifiques, si elles

dépendent ou non d'autres facteurs : caractéristiques démographiques comme l'âge, le sexe, … ou si elles sont associées à la variable réponse d'intérêt.

L'utilisation des logiciels disponibles actuellement pour créer des imputations multiples demande un certain investissement; en effet, la plupart du temps les procédures d'imputations sont itératives et il faut savoir utiliser les outils permettant de vérifier que l'algorithme a bien convergé. Et enfin, il y a un travail à mener préalablement pour choisir correctement le modèle d'imputation.

Annexe 1

A1.1 Algorithme EM dans le cas de 2 variables binaires

Soient Y_1 et Y_2 deux variables binaires prenant les valeurs 1 et 2 et pouvant présenter des valeurs manquantes. L'objectif est d'estimer les paramètres $\theta_{ij} = Pr(Y_1 = i, Y_2 = j)$.

Les données observées sont représentées dans le tableau ci-dessous : on distingue 4 types d'observations, celles indicées par A pour lesquelles Y_1 et Y_2 sont mesurées, celles indicées par B pour lesquelles Y_1 seulement est mesurée, celles indicées par C pour lesquelles Y_2 seulement est mesurée et celles indicées par D pour lesquelles Y_1 et Y_2 sont manquants.

$$Y_{\text{obs}} = \{n_{ij}^A, n_{i*}^B, n_{*j}^C, n_{**}^D, \;\; i, j = 1, 2\}.$$

Tableau croissant Y_1 et Y_2

		Y_2		
		1	2	manquant
Y_1	1	n_{11}^A	n_{12}^A	n_{1*}^B
	2	n_{21}^A	n_{22}^A	n_{2*}^B
	manquant	n_{*1}^C	n_{*2}^C	n_{**}^D
				n

Le nombre, inconnu, d'observations n_{ij} peut s'exprimer comme la somme des nombres d'observations provenant quatre types d'observations : $n_{ij} = n_{ij}^A + n_{ij}^B + n_{ij}^C + n_{ij}^D$. Le nombre n_{ij}^A est observé alors que les nombres n_{ij}^B, n_{ij}^C et n_{ij}^D ne le sont pas.

Les nombres d'observations (n_{i1}^B, n_{i2}^B) conditionnellement à Y_{obs} et θ suivent une distribution multinomiale :

$$(n_{i1}^B, n_{i2}^B)/Y_{\text{obs}}, \theta \propto M(n_{i*}^B, \theta_{i1}/\theta_{i*}, \; \theta_{i2}/\theta_{i*}).$$

où $\theta_{i*} = \theta_{i1} + \theta_{i2}$

D'où $E(n_{ij}^B/Y_{\rm obs},\theta) = n_{i*}^B\theta_{ij}/\theta_{i*}$

De même, les nombres d'observations (n_{1j}^C, n_{2j}^C) conditionnellement à $Y_{\rm obs}$ et θ suivent une distribution multinomiale :

$$(n_{1j}^C, n_{2j}^C)/Y_{\rm obs},\theta \propto M(n_{*j}^C, \theta_{1j}/\theta_{*j},\ \theta_{2j}/\theta_{*j}).$$

où $\theta_{*j} = \theta_{1j} + \theta_{2j}$

D'où : $E(n_{ij}^C/Y_{\rm obs},\theta) = n_{*j}^C\theta_{ij}/\theta_{*j}$.

De même, les nombres d'observations $(n_{11}^D, n_{12}^D, n_{21}^D, n_{22}^D)$ conditionnellement à $Y_{\rm obs}$ et θ suivent une distribution multinomiale :

$$(n_{11}^D, n_{12}^D, n_{21}^D, n_{22}^D)/Y_{\rm obs},\theta \propto M(n_{**}^D, \theta_{11}, \theta_{12},\ \theta_{21},\ \theta_{22}).$$

D'où : $E(n_{ij}^D/Y_{\rm obs},\theta) = n_{**}^D\theta_{ij}$.

Étape estimation (E)

On part d'une première valeur des paramètres θ^0 à estimer

θ_{11}^0	θ_{12}^0
θ_{21}^0	θ_{22}^0

Et on estime les nombres d'observations n_{ij} en remplaçant les n_{ij}^A, n_{ij}^B, n_{ij}^C, n_{ij}^D par leurs espérances conditionnelles :

$$E(n_{ij}/Y_{\rm obs},\theta^0) = n_{ij}^A + n_{i*}^B\theta_{ij}^0/\theta_{i*}^0 + n_{*j}^C\theta_{ij}^0/\theta_{*j}^0 + n_{**}^D\theta_{ij}^0$$

Étape maximisation (M)

On en déduit les estimations des paramètres θ du maximum de vraisemblance, soit :

$$\theta_{11}^1 = \frac{1}{n}E(n_{11}/Y_{\rm obs},\theta^0) \quad \theta_{12}^1 = \frac{1}{n}E(n_{12}/Y_{\rm obs},\theta^0)$$
$$\theta_{21}^1 = \frac{1}{n}E(n_{21}/Y_{\rm obs},\theta^0) \quad \theta_{22}^1 = \frac{1}{n}E(n_{22}/Y_{\rm obs},\theta^0)$$

On peut combiner les deux étapes de l'algorithme

$$\theta_{ij}^{t+1} = \frac{1}{n}\left[n_{ij}^A + n_{i*}^B\frac{\theta_{ij}^t}{\theta_{i*}^t} + n_{*j}^C\frac{\theta_{ij}^t}{\theta_{*j}^t} + n_{**}^D\theta_{ij}^t\right]$$

Et on itère à partir de ces nouvelles valeurs des paramètres jusqu'à la convergence.

A1.2 Algorithme «Data Augmentation» dans le cas de 2 variables binaires

Étape (I) : Imputation des valeurs manquantes

Calcul des probabilités $P(Y_{\text{manq}}|Y_{\text{obs}}, \theta)$

On part des valeurs des paramètres θ fournies par l'algorithme EM (valeurs initiales θ^0)

		Y_2 = 1	Y_2 = 2
Y_1	1	θ^0_{11}	θ^0_{12}
	2	θ^0_{21}	θ^0_{22}

Pour l'individu «i» calcul des probabilités des profils suivants :

y_{i1}	y_{i2}	$P(y_{i1}\mid y_{i2}, \theta^0)$
?	1	$P(y_{i1}=1\mid y_{i2}=1, \theta^0)= \theta^0_{11}/(\theta^0_{11}+\theta^0_{21})$
		$P(y_{i1}=2\mid y_{i2}=1, \theta^0)= \theta^0_{21}/(\theta^0_{11}+\theta^0_{21})$
?	2	$P(y_{i1}=1\mid y_{i2}=2, \theta^0)= \theta^0_{12}/(\theta^0_{12}+\theta^0_{22})$
		$P(y_{i1}=2\mid y_{i2}=2, \theta^0)= \theta^0_{22}/(\theta^0_{12}+\theta^0_{22})$
1	?	$P(y_{i2}=1\mid y_{i1}=1, \theta^0)= \theta^0_{11}/(\theta^0_{11}+\theta^0_{12})$
		$P(y_{i2}=2\mid y_{i1}=1, \theta^0)= \theta^0_{12}/(\theta^0_{11}+\theta^0_{12})$
2	?	$P(y_{i2}=1\mid y_{i1}=2, \theta^0)= \theta^0_{21}/(\theta^0_{21}+\theta^0_{22})$
		$P(y_{i2}=2\mid y_{i1}=2, \theta^0)= \theta^0_{22}/(\theta^0_{21}+\theta^0_{22})$
?	?	$P(y_{i1}=1 \text{ et } y_{i2}=1, \theta^0)= \theta^0_{11}$
		$P(y_{i1}=1 \text{ et } y_{i2}=2, \theta^0)= \theta^0_{12}$
		$P(y_{i1}=2 \text{ et } y_{i2}=1, \theta^0)= \theta^0_{21}$
		$P(y_{i1}=2 \text{ et } y_{i2}=2, \theta^0)= \theta^0_{22}$

On dispose ainsi de toute l'information permettant d'imputer toute valeur manquante pour Y_1 et/ou Y_2. Par exemple si $\theta^0_{12} = 0.10$ et $\theta^0_{22} = 0.06$: pour l'individu dont le profil (? , 2), on a au départ en utilisant les valeurs de θ fournies par l'algorithme EM :

$$P(y_{i1} = 1 | y_{i2} = 2, \theta^0) = \theta^0_{12}/(\theta^0_{12} + \theta^0_{22}) = \frac{1.10}{0.10 + 0.06} = 0.625$$

On tire alors au hasard un nombre (r) compris entre 0 et 1 : si $r < 0.625$ $y_{i1} = 1$ et si $r \geqslant 0.625$ $y_{i1} = 2$. Il en résulte un tableau complet (tableau imputé).

Étape (P) : distribution *a posteriori* des paramètres θ

Le tableau Y étant complet (données imputées), on peut déterminer la distribution de $\theta = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$ qui est :

$$P[\theta/\text{tableau } Y \text{ imputé}] = \frac{P(Y/\theta)P(\theta)}{P(Y)}$$

à condition de fixer une distribution *a priori* de θ : distribution de Dirichlet qui, conjuguée à la distribution multinomiale $P(Y/\theta)$, fournit une nouvelle distribution de Dirichlet. C'est la partie théorique délicate de la procédure

On tire au sort, dans cette distribution, un quadruple $\theta^0 = (\theta^0_{11}, \theta^0_{12}, \theta^0_{21}, \theta^0_{22})$ et on calcule la probabilité :

$$P(Y^0_{\text{manq}} | Y_{\text{obs}}, \theta^0)$$

Et on itère : obtention d'une chaîne de Markov (MCMC)

$$\theta^0 \longrightarrow (Y^0_{\text{manq}}) \longrightarrow \theta^1 \longrightarrow (Y^1_{\text{manq}}) \longrightarrow \theta^2 \longrightarrow (Y^2_{\text{manq}}) \longrightarrow \dots$$

jusqu'à la convergence en probabilité vers $P(Y_{\text{manq}}, \theta | Y_{\text{obs}})$

On obtient ainsi un tableau T1 final imputé ($m = 1$)

En pratique on construit $m > 1$ tableaux imputés (en général $m = 5$), qu'on obtient :

- soit en construisant une *chaîne simple* : on laisse pour cela passer plusieurs itérations (500 par exemple) pour oublier la dépendance entre les θ et on fixe le nombre total d'itérations (10 000 par exemple) et le nombre d'imputations ($m = 5$ par défaut). On impute alors les données manquantes au bout de 2 000, 4 000, 6 000, 8 000 et 10 000 itérations. On obtient ainsi 5 tableaux imputés T1, T2, T3, T4 et T5.
- soit en construisant une *chaîne multiple* : on laisse passer plusieurs itérations (500 par exemple) pour oublier la dépendance entre les θ et on impute les données manquantes au bout d'un nombre d'itérations fixé (1 000 par exemple). On obtient le premier tableau imputé T1. On repart du tout début pour obtenir, à chaque fois, un tableau imputé. On obtient ainsi une chaîne multiple constituée

de T1, T2, T3, T4 et T5 en partant de $m = 5$ différentes estimations des paramètres.

Annexe 2

A2.1 Algorithme EM dans le cas de 2 variables continues

Rappel : données bidimensionnelles complètes

Soit Y le tableau $(n, 2)$ contenant n individus (en ligne) et 2 variables (en colonne), de terme général y_{ij}. La i-ième ligne de Y, exprimée en terme de vecteur colonne est : $y_i = (y_{i1}, y_{i2})'$. On suppose que $y_1, y_2, \ldots, y_n$ sont n réalisations indépendantes d'un vecteur aléatoire $(Y_1, Y_2)'$ distribué normalement suivant $N(\mu, \Sigma)$, où $\theta = (\mu, \Sigma)$ est le paramètre inconnu à estimer.

Statistiques exhaustives

$$T_1 = \left(\sum_i y_{i1} \ \sum_i y_{i2} \right)' = Y'\mathbf{1} \qquad T_2 = \begin{pmatrix} \sum_i y_{i1}^2 & \sum_i y_{i1} y_{i2} \\ \sum_i y_{i2} y_{i1} & \sum_i y_{i2}^2 \end{pmatrix} = Y'Y$$

Les paramètres $\widehat{\mu}$ et $\widehat{\Sigma}$ ne dépendent que des statistiques exhaustives :

$$E(T_1) = n\mu \qquad \text{et} \qquad E(T_2) = n(\Sigma + \mu\mu')$$

Estimations de μ et Σ du maximum de vraisemblance

$$\widehat{\mu} = \frac{1}{n} Y'\mathbf{1} = \frac{T_1}{n} \qquad \text{et} \qquad \widehat{\Sigma} = \frac{1}{n} Y'Y - \widehat{\mu}\widehat{\mu}' = \frac{T_2}{n} - \frac{T_1^2}{n^2}$$

Dans le cas de données incomplètes, le tableau Y peut se mettre sous la forme $(Y_{\text{obs}}, Y_{\text{manq}})$. L'algorithme EM présente deux étapes :

Étape E : (*Estimation*)

Cette étape ne remplit pas les éléments du tableau Y manquants mais les parties des statistiques exhaustives manquantes. T_1 est remplacé par $E(T_1/Y_{\text{obs}}, \theta^{(t)})$ et T_2 est remplacé par $E(T_2/Y_{obs}, \theta^{(t)})$. On a à considérer les 4 profils suivants : profil 1 $(Y_{1\,\text{obs}}, Y_{2\,\text{obs}})$, profil 2 $(Y_{1\,\text{manq}}, Y_{2\,\text{obs}})$, profil 3 $(Y_{1\,\text{obs}}, Y_{2\,\text{manq}})$ et profil 4 $(Y_{1\,\text{manq}}, Y_{2\,\text{manq}})$.

Remarque. – Le profil 4 ne contribue pas à la vraisemblance et ralentit la convergence de l'algorithme.

L'objectif est de calculer $E(T_1/Y_{\text{obs}}, \theta)$ et $E(T_2/Y_{\text{obs}}, \theta)$. Pour cela on doit calculer des quantités différentes suivant les profils. Les individus du profil 1 ne

posent pas de problème puisque les données sont complètes pour ces individus. Sont requises :

- pour les individus du profil 2, $E(y_{i1}/y_{i2},\theta)$ pour le calcul de T_1 et $E(y_{i1}^2/y_{i2},\theta)$ et $E(y_{i1}y_{i2}/y_{i2},\theta)$ pour le calcul de T_2,
- pour les individus du profil 3, $E(y_{i2}/y_{i1},\theta)$ pour le calcul de T_1 et $E(y_{i2}^2/y_{i1},\theta)$ et $E(y_{i1}y_{i2}/y_{i1},\theta)$ pour le calcul de T_2,
- pour les individus du profil 4, $E(y_{i1}/\theta)$, $E(y_{i1}^2/\theta)$, $E(y_{i2}/\theta)$, $E(y_{i2}^2/\theta)$ et $E(y_{i1}y_{i2}/\theta)$.

  On donne ici des éléments de calcul :

$$E(T_1/Y_{\rm obs},\theta)=\left[E\left(\sum_i(y_{i1}/Y_{\rm obs},\theta)\right)E\left(\sum_i(y_{i2}/Y_{\rm obs},\theta)\right)\right]'$$

$$E\left(\sum_i(y_{i1}/Y_{\rm obs},\theta)\right)=\sum_{i\in P_1}y_{i1}+\sum_{i\in P_2}E(y_{i1}/y_{i2},\theta)+\sum_{i\in P_3}y_{i1}+\sum_{i\in P_4}E(y_{i1}/\theta)$$

idem pour $E\left(\sum_i(y_{i2}/Y_{\rm obs},\theta)\right)$

$$E(T_2/Y_{\rm obs},\theta)=\begin{bmatrix}E\left(\sum_i(y_{i1}^2/Y_{\rm obs},\theta)\right) & E\left(\sum_i(y_{i1}y_{i2}/Y_{\rm obs},\theta)\right)\\ E\left(\sum_i(y_{i2}y_{i1}/Y_{\rm obs},\theta)\right) & E\left(\sum_i(y_{i2}^2/Y_{\rm obs},\theta)\right)\end{bmatrix}$$

Pour le profil 1 ($Y_{1\,\rm obs}$ et $Y_{2\,\rm obs}$) :

$$\begin{aligned}&E(y_{i1}/Y_{\rm obs},\theta)=y_{i1} && E(y_{i2}/Y_{\rm obs},\theta)=y_{i2}\\ &E(y_{i1}^2/Y_{\rm obs},\theta)=y_{i1}^2 && E(y_{i2}^2/Y_{\rm obs},\theta)=y_{i2}^2\\ &E(y_{i1}y_{i2}/Y_{\rm obs},\theta)=y_{i1}y_{i2}\end{aligned}$$

Pour le profil 2 ($Y_{1\,\rm manq}$ et $Y_{2\,\rm obs}$) :

$$\begin{aligned}&E(y_{i1}/Y_{\rm obs},\theta)=\mu_1+(y_{i2}-\mu_2)\frac{\sigma_{12}}{\sigma_{22}}\,; && E(y_{i2}/Y_{\rm obs},\theta)=y_{i2}\\ &E(y_{i1}y_{i2}/Y_{\rm obs},\theta)=\left(\mu_1+(y_{i2}-\mu_2)\frac{\sigma_{12}}{\sigma_{22}}\right)y_{i2}\,; && E(y_{i2}^2/Y_{\rm obs},\theta)=y_{i2}^2\\ &E(y_{i1}^2/Y_{\rm obs},\theta)=\sigma_{11}-\frac{\sigma_{12}^2}{\sigma_{22}}+\left[\left(\mu_1-\mu_2\frac{\sigma_{12}}{\sigma_{22}}\right)+\frac{\sigma_{12}}{\sigma_{22}}y_{i2}\right]^2\end{aligned}$$

Pour le profil 3 ($Y_{1\,\mathrm{obs}}$ et $Y_{2\,\mathrm{manq}}$) , on obtient les espérances mathématiques correspondantes en remplaçant 1 par 2 et 2 par 1 dans les expressions des espérances du profil 2.

Pour le profil 4 ($Y_{1\,\mathrm{manq}}$ et $Y_{2\,\mathrm{manq}}$) :

$$\begin{aligned} &E(y_{i1}/Y_{\mathrm{obs}},\theta) = \mu_1 \quad ; \quad E(y_{i2}/Y_{\mathrm{obs}},\theta) = \mu_2 \\ &E(y_{i1}^2/Y_{\mathrm{obs}},\theta) = \sigma_{11} + \mu_1^2 \quad ; \quad E(y_{i2}^2/Y_{\mathrm{obs}},\theta) = \sigma_{22} + \mu_2^2 \\ &E(y_{i1}y_{i2}/Y_{\mathrm{obs}},\theta) = \sigma_{12} + \mu_1\mu_2 \end{aligned}$$

En effectuant l'étape E, on met à jour les statistiques exhaustives $E(T_j/Y_{\mathrm{obs}},\theta^t)$.

Étape M (*Maximisation*)

Elle consiste à trouver les estimations du maximum de vraisemblance des paramètres quand les données sont complètes :

$$\widehat{\mu}^{(t+1)} = \frac{T_1^{(t)}}{n} \qquad \text{et} \qquad \widehat{\Sigma}^{(t+1)} = \frac{T_2^{(t)}}{n} - \left(\frac{T_1^{(t)}}{n}\right)^2$$

A2.2 Algorithme « Data Augmentation » dans le cas de 2 variables continues

Étape (I) : Imputation

On part de $\theta^{(0)} = \left(\mu^{(0)}, \Sigma^{(0)}\right)$ de θ, obtenu en utilisant l'algorithme EM. Puisqu'on connaît toutes les distributions conditionnelles, on calcule les probabilités : $P(Y_{\mathrm{manq}}^{(t)}|Y_{\mathrm{obs}},\theta^{(t)})$. On tire :

- des valeurs manquantes du profil 2 en utilisant la distribution de probabilité $P(Y_{i1} = y_{i1}|Y_{i2} = y_{i2},\theta^{(t)})$ qui est une distribution normale connue.
- des valeurs manquantes du profil 3 en utilisant la distribution de probabilité $P(Y_{i2} = y_{i2}|Y_{i1} = y_{i1},\theta^{(t)})$ qui est une distribution normale connue
- des valeurs manquantes du profil 4 en utilisant la distribution de probabilité $P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}/\theta^{(t)}) \sim N\left[\begin{pmatrix}\mu_1^{(t)}\\ \mu_2^{(t)}\end{pmatrix}\begin{pmatrix}\sigma_{11}^{(t)} & \sigma_{12}^{(t)}\\ \sigma_{21}^{(t)} & \sigma_{22}^{(t)}\end{pmatrix}\right]$.

On remplit ainsi toutes les cases manquantes du tableau $Y = (Y_{\mathrm{obs}}, Y_{\mathrm{manq}}^{(0)})$ (tableau imputé) et on en déduit $\theta^{(t+1)}$.

Étape (P) : distribution *a posteriori* de θ

$$P(\theta^{(t+1)}/Y_{\mathrm{obs}}, Y_{\mathrm{manq}}^{(t)}) \qquad \text{avec} \qquad Y = (Y_{\mathrm{obs}}, Y_{\mathrm{manq}}^{(t)}) = \text{tableau imputé}$$

On utilise pour cela la formule $P[(\mu,\Sigma)/Y] = P(\Sigma/Y)P(\mu/\Sigma,Y)$ qui conduit à estimer les distributions de (Σ/Y) et de $(\mu/\Sigma,Y)$. On montre [Schafer, 2000] que :

– la distribution de (Σ/Y) est approximée par $W^{-1}[n-1,(nS)^{-1}]$, distribution de Wishart inversée où S est la matrice des variances-covariances observée, d'où on tire au sort une matrice $\Sigma^{(t)}$.

– distribution de $(\mu/\Sigma,Y)$ est approximée par une distribution $N[\overline{y},n^{-1}\Sigma^{(t)}]$, d'où l'on tire au sort de $\mu^{(t)}$.

On construit ainsi une chaîne MCMC :

$$\theta^0 \longrightarrow (Y^0_{\text{manq}}) \longrightarrow \theta^1 \longrightarrow (Y^1_{\text{manq}}) \longrightarrow \theta^2 \longrightarrow (Y^2_{\text{manq}}) \longrightarrow \ldots$$

jusqu'à la convergence de θ

On en déduit le premier tableau imputé T_1 et on procède comme dans le cas de deux variables binaires pour construire les $m > 1$ tableaux imputés.

Annexe 3

Étude de la convergence de l'algorithme «Data Augmentation»

La convergence de l'algorithme DA a été explorée théoriquement sous diverses conditions. Mais en pratique, la vérification de la convergence n'est pas simple. On peut néanmoins contrôler la convergence d'une chaîne simple en représentant graphiquement :

– les estimations des paramètres à la i-ième itération en fonction du n° de l'itération (time series plots) : $\theta^{(t)}$ en fonction de l'itération t et

– les liaisons entre séries d'estimations des paramètres successives. Il s'agit plus précisément de fonctions d'autocorrélations qui représentent les corrélations entre séries de valeurs de θ en fonction d'une variable décalage (*lag*).

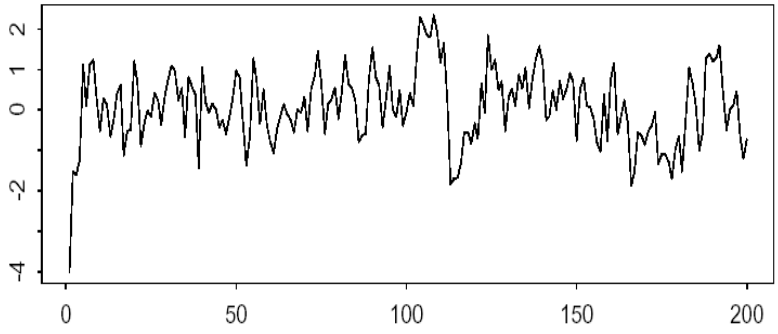
Comme dans la procédure EM, la convergence de la procédure DA est liée au taux d'information manquante : une importante proportion d'information manquante entraîne une convergence lente. Le sens de convergence est, cependant, assez différent car la procédure DA est une procédure stochastique qui converge en probabilité.

Quand EM converge, les paramètres estimés ne changent plus d'une itération à l'autre. Alors que quand DA converge, la distribution des paramètres ne change pas d'une itération à l'autre bien que les valeurs des paramètres aléatoires continuent de changer. Pour cette raison, évaluer la convergence de la procédure DA est beaucoup plus compliquée que celle de EM.

Les graphiques de la figure 6 représentent les estimations des paramètres à la i-ième itération en fonction du n° de l'itération.

La courbe (a) de cette figure 6 représente une bande de valeurs horizontales sans tendance ni vers le haut, ni vers le bas, ce qui indique une convergence rapide de l'algorithme. Par contre la courbe (b) représente des estimations avec une tendance vers le haut suivie d'une tendance vers le bas, ce qui indique une convergence lente de l'algorithme.

(a) Convergence rapide



(b) Convergence lente

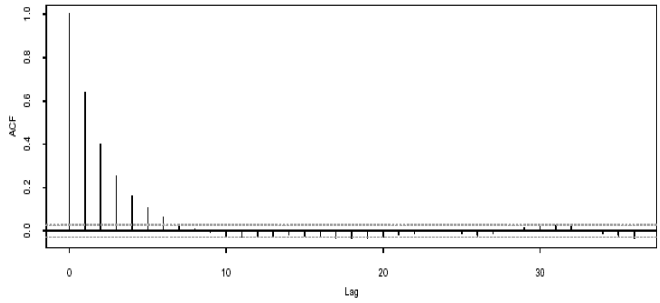
FIGURE 6
Estimations des paramètres à la i-ième itération en fonction du n° de l'itération [source : Schafer J.L., 2000]

Cette lenteur de la convergence peut être due soit à un nombre trop grand de paramètres à estimer, soit à un nombre d'itérations pas assez grand.

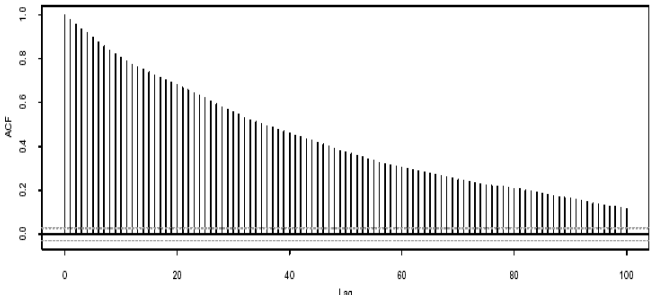
Les graphiques de la figure 7 fournissent les corrélations (en ordonnées) en fonction de $lag = 0, 1, 2, \ldots$ Ainsi la corrélation correspondant à $lag = 0$ représente la corrélation entre la série de valeurs $(\theta^1, \theta^2, \theta^3, \ldots, \theta^k)$ et elle même. Pour $lag = 1$ (décalage de 1), on a la corrélation entre la série $(\theta^1, \theta^2, \theta^3, \ldots, \theta^{k-1})$ et la série $(\theta^2, \theta^3, \theta^4, \ldots, \theta^k)$. Pour $lag = 2$ (décalage de 2), on a la corrélation entre la série $(\theta^1, \theta^2, \theta^3, \ldots, \theta^{k-2})$ et la série $(\theta^3, \theta^4, \theta^5, \ldots, \theta^k)$, ainsi de suite… Ces fonctions d'autocorrélations donnent une assez bonne idée de la convergence

de l'algorithme. La convergence de l'algorithme est rapide si les autocorrélations disparaissent rapidement (figure 7-a), ce qui n'est pas le cas dans la (figure 7-b).

(a) Convergence rapide



(b) Convergence lente



(c) FAC bruitée

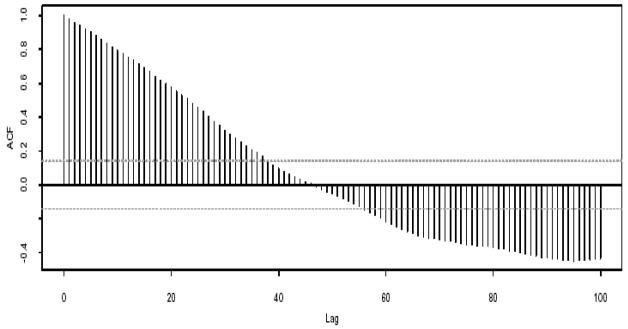


FIGURE 7
Fonctions d'autocorrélations entre séries d'estimations des paramètres successives [source : Schafer J.L., 2000]

Les corrélations négatives (qui n'ont pas de sens) notées dans la (figure 7-c) sont dues à des fluctuations d'échantillonnage que l'on observe quand le fichier à analyser est de petite taille.

Références

COX D.R. (1972), *Regression models and life-tables (with discussion)*. Journal of the Royal Statistical Society, Series B, 34 :187-220.

DEMPSTER A.P., LAIRD, N.M. and RUBIN D.B. (1977), Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). JRSS Series B, **39**, 1-38.

GILKS W.R., RICHARDSON S. and SPIEGELHALTER D.J. (1996), *Markov Chain Monte Carlo in practice*. Chapman & Hall, London.

GOLDBERG M. and LECLERC A. (1994), *Cohorte GAZEL, 20 000 volontaires d'EDF-GDF pour la recherche médicale. Bilan 1989-1993*. Les Éditions INSERM, Paris.

GREENLAND S. and FINKLE W.D. (1995), *A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses*. Amer. Journ. of Epidemiology;142 :1255-64.

HOSMER D.W. and LEMESHOW S. (1989), *Applied logistic regression*. New York : John Wiley & Sons, Inc.

LITTLE R.J.A. et RUBIN D.B. (1987), *Statistical Analysis with Missing Data*. New York : Wiley.

NAKACHE J.-P. et CONFAIS J. (2003), *Statistique explicative appliquée : analyse discriminante, modèle logistique, segmentation par arbre*. Éditions Technip Paris.

NAKACHE J.-P., GUÉGUEN A., ZINS M. et GOLDBERG M. (2004), *Analyse de Données de Survie Groupées avec Covariables dépendant du temps : Application à l'étude de l'effet prédictif de l'état de santé perçu sur le décès, chez les hommes de la cohorte Gazel observés dans la période 1989-1999*. Revue de Statistique Appliquée, vol. LII, n° 2, 27-49.

ROBERTS G.O. (1996), *Markov chain concepts related to sampling algorithms*. In Markov Chain Monte Carlo in Practice (eds Gilks W.R., Richardson S and Spiegelhalter D.J.). Chapman & Hall, London, 45-57.

RUBIN D.B. (1978), *Multiple Imputations in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse*. Proceedings of the Survey Research Methods Section, American Statistical Association, 20-34.

RUBIN D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*. Wiley & Sons, NewYork.

SAS/STAT SOFTWARE (2001), *Version 8.2* SAS Institute Inc., Cary, NC (USA).

SCHAFER J.L. and OLSEN M.K. (1998), *Multiple Imputation for Multivariate Missing-Data Problems : a Data Analyst's Perspective*. http ://www.stat.psu.edu/ jls/mbr.pdf

SCHAFER J.L. (2000), *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC.

TANNER M.A. and WONG W.H. (1987), *The calculation of posterior distributions by data augmentation (with discussion)*. JASA, **82**, 528-550.

VACH W. and BLETTNER M. (1991), *Biased Estimation of the Odds Ratio in Case-Control Studies due to the Use of Ad Hoc Methods of Correcting for Missing Values for Confounding Variables*. Amer. Journ. of Epidemiology;134 :895-907.