Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 161-168
Statistique Canada, N° 12-001-XPB au catalogue

L'importance de la modélisation du plan d'échantillonnage dans l'imputation multiple pour les données manquantes

Jerome P. Reiter, Trivellore E. Raghunathan et Satkartar K. Kinney [1]

Résumé

La théorie de l'imputation multiple pour traiter les données manquantes exige que l'imputation soit faite conditionnellement du plan d'échantillonnage. Cependant, comme la plupart des progiciels standard utilisés pour l'imputation multiple fondée sur un modèle reposent sur l'hypothèse d'un échantillonnage aléatoire simple, de nombreux praticiens sont portés à ne pas tenir compte des caractéristiques des plans d'échantillonnage complexes, comme la stratification et la mise en grappes, dans leurs imputations. Or, la théorie prédit que l'analyse d'ensembles de données soumis de telle façon à une imputation multiple peut produire des estimations biaisées du point de vue du plan de sondage. Dans le présent article, nous montrons au moyen de simulations que i) le biais peut être important si les caractéristiques du plan sont reliées aux variables d'intérêt et que ii) le biais peu être réduit en tenant compte de l'effet des caractéristiques du plan dans les modèles d'imputation. Les simulations montrent aussi que l'introduction de caractéristiques non pertinentes du plan comme contraintes dans les modèles d'imputation peut donner lieu à des inférences conservatrices, à condition que les modèles contiennent aussi des variables explicatives pertinentes. Ces résultats portent à formuler la prescription qui suit à l'intention des imputeurs : le moyen le plus sûr de procéder consiste à inclure les variables du plan de sondage dans la spécification des modèles d'imputation. À l'aide de données réelles, nous donnons une démonstration d'une approche simple d'intégration des caractéristiques d'un plan de sondage complexe qui peut être suivie en utilisant certains progiciels standard pour créer des imputations multiples.

Mots clés : Plan de sondage complexe; imputation multiple; non-réponse; enquêtes.

1. Introduction

En général, dans les grandes enquêtes, les unités échantillonnées ne répondent pas toutes complètement au questionnaire. Certaines n'y répondent pas du tout et d'autres ne répondent qu'à certaines questions. Une approche pour traiter ce genre de non-réponse est l'imputation multiple des données manquantes (Rubin 1987). Elle a été utilisée, par exemple, dans le Fatality Analysis Reporting System (Heitjan et Little 1991), la Consumer Expenditures Survey (Raghunathan et Paulin 1998), la National Health and Nutrition Examination Survey (Schafer, Ezzati-Rice, Johnson, Khare, Little et Rubin 1998), la Survey of Consumer Finances (Kennickell 1998) et la National Health Interview Survey (Schenker, Raghunathan, Chiu, Makuc, Zhang et Cohen 2005). L'imputation multiple a également été proposée pour assurer la protection des renseignements personnels dans les fichiers de données à grande diffusion (Rubin 1993; Little 1993; Raghunathan, Reiter et Rubin 2003; Reiter 2003, 2004, 2005). Pour une revue d'autres applications, voir Rubin (1996), ainsi que Barnard et Meng (1999).

En théorie, lors de l'établissement de méthodes d'inférence d'après des ensembles de données ayant subi une imputation multiple, cette dernière est rendue conditionnelle au plan d'échantillonnage (Rubin 1987). Toutefois, les imputeurs tiennent rarement compte des caractéristiques des plans d'échantillonnage complexes, comme la stratification et la mise en grappes, lorsqu'ils utilisent les progiciels disponibles pour construire des modèles d'imputation. Ils se servent plutôt de modèles normaux ou de modèles de localisation généraux multivariés (par exemple, le logiciel NORM rédigé par Joe Schafer), ou de modèles de régression séquentielle (Raghunathan, Lepkowshi, van Hoewyk et Solenberger 2001). Bien que ces méthodes puissent être modifiées afin d'intégrer les caractéristiques du plan, cela se fait rarement.

L'objectif du présent article est double. En premier lieu, nous illustrons le biais qui peut se produire lorsque les imputeurs omettent de tenir compte des caractéristiques du plan de sondage complexe dans les modèles d'imputation. Pour cela, nous simulons une imputation multiple dans des échantillons à deux degrés, stratifiés et mis en grappes. Les simulations indiquent que le biais peut être important, même si l'on applique des estimateurs fondés sur le plan de sondage à des ensembles de données soumis à l'imputation multiple ne présentant qu'une quantité modérée de données manquantes. En deuxième lieu, nous proposons deux approches simples en vue de tenir compte des caractéristiques du plan dans les modèles d'imputation. La première, qui est relativement facile à mettre en œuvre, comprend des variables nominales pour les effets de strate ou de grappes dans les modèles d'imputation. La deuxième, qui requiert des calculs plus compliqués que la première,

1. Jerome P. Reiter et Satkartar K. Kinney, Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708, États-Unis; Trivellore E. Raghunathan, Department of Biostatistics and Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, États-Unis.

s'appuie sur des modèles hiérarchiques où i) les effets de la mise en grappes sont intégrés en utilisant des effets aléatoires et ii) les effets de la stratification sont intégrés en utilisant des effets fixes. Les simulations montrent que tenir compte du plan de sondage de cette façon peut réduire le biais. Elles illustrent aussi le fait qu'introduire des caractéristiques du plan qui ne sont pas reliées aux variables de l'enquête peut donner lieu à des inférences inefficaces, mais prudentes, comparativement à celles faites d'après des modèles dans lesquels ce genre de caractéristiques ne sont pas intégrées comme contraintes, à condition que les modèles incluent les variables explicatives requises pour que l'hypothèse selon laquelle les données manquent au hasard (Rubin 1976) soit plausible. Nous démontrons la première approche d'intégration des caractéristiques du plan en imputant des données manquantes dans le cas de la National Health and Nutrition Examination Survey selon une méthode de régression séquentielle.

2. Inférences d'après des ensembles de données multi-imputés

Afin de décrire la construction d'ensembles de données multi-imputés et les inférences d'après ces derniers, nous utilisons la notation de Rubin (1987). Pour une population finie de taille N, soit $I_j = 1$ si l'unité j est sélectionnée dans l'enquête originale, et $I_j = 0$ autrement, où $j = 1, 2, \ldots, N$. Soit $I = (I_1, \ldots, I_N)$. Soit n la taille d'échantillon obtenue au moyen d'un plan d'échantillonnage complexe. Pour simplifier la notation, supposons qu'une seule variable de l'enquête est sujette à la non-réponse. Soit $R_j = 1$, si l'unité j répond à l'enquête originale, et $R_j = 0$, autrement. Soit $R = (R_1, \ldots, R_N)$. La notation peut être étendue afin de traiter la non-réponse partielle multivariée, mais ce genre de complication n'est pas nécessaire aux fins de notre exposé.

Soit Y la matrice de données d'enquête de dimensions $N \times p$ pour toutes les unités de la population. Soit $Y_{\text{inc}} = (Y_{\text{obs}}, Y_{\text{mis}})$ la matrice de données d'enquête de dimensions $n \times p$ pour les unités pour lesquelles $I_j = 1$; Y_{obs} est la portion de Y_{inc} qui est observée, et Y_{mis} est la portion de Y_{inc} qui manque à cause de la non-réponse. Soit Z la matrice de variables du plan de dimensions $N \times d$ pour les N unités de la population, par exemple, des indicateurs de strates ou de grappes ou des mesures de taille. Nous supposons que ce genre d'information sur le plan est connue au moins approximativement, par exemple d'après les dossiers du recensement ou les bases de sondage.

Les valeurs de Y_{mis} sont habituellement construites d'après des tirages à partir d'une approximation de la loi bayésienne prédictive a posteriori de $(Y_{\text{mis}} | Z, Y_{\text{obs}}, I, R)$. Ces tirages sont répétés indépendamment $l = 1, \ldots, M$ fois pour obtenir M ensembles de données complets, $D^{(l)} = (Z, Y_{\text{obs}}, Y_{\text{mis}}^{(l)}, I, R)$.

D'après ces ensembles de données multi-imputés, un utilisateur des données veut faire des inférences au sujet d'un paramètre $Q = Q(Z, Y)$. Par exemple, Q pourrait être une moyenne de population ou un coefficient de régression de population. Dans chaque ensemble de données imputé, $D^{(l)}$, l'analyste estime Q au moyen d'un estimateur q, et la variance de q, au moyen d'un estimateur u. Nous supposons qu'il spécifie q et u en agissant comme si chaque $D^{(l)}$ était, en fait, formé de données recueillies auprès d'un échantillon aléatoire de (Z, Y) en se fondant sur le plan d'échantillonnage original I, c'est-à-dire que q et u sont des estimateurs sur données complètes.

Pour $l = 1, \ldots, M$, posons que $q^{(l)}$ et $u^{(l)}$ sont, respectivement, les valeurs de q et de u dans l'ensemble de données $D^{(l)}$. Sous les hypothèses décrites dans (Rubin 1987), l'analyste peut obtenir les inférences valides pour le scalaire Q en combinant les $q^{(l)}$ et $u^{(l)}$. Plus précisément, les quantités qui suivent sont nécessaires pour les inférences :

$$\bar{q}_M = \sum_{l=1}^{M} q^{(l)} / M \tag{1}$$

$$b_M = \sum_{l=1}^{M} (q^{(l)} - \bar{q}_M)^2 / (M - 1) \tag{2}$$

$$\bar{u}_M = \sum_{l=1}^{M} u^{(l)} / M. \tag{3}$$

L'analyste peut alors utiliser $\bar{q}_M$ pour estimer Q et $T_M = (1 + \frac{1}{M}) b_M + \bar{u}_M$ pour estimer la variance de $\bar{q}_M$. Quand n et M sont grands, les inférences pour le scalaire Q peuvent être fondées sur des lois normales, de sorte qu'un intervalle de confiance à $(1 - \alpha)\,\%$ pour Q est $\bar{q}_M \pm z(\alpha/2)\sqrt{T_M}$. Pour une valeur modérée de M, les inférences peuvent être fondées sur des lois t avec $v_M = (M - 1)(1 + r_M^{-1})^2$ degrés de liberté, où $r_M = (1 + M^{-1}) b_M / \bar{u}_M$, de sorte qu'un intervalle de confiance à $(1 - \alpha)\,\%$ pour Q est $\bar{q}_M \pm t_{vM}(\alpha/2)\sqrt{T_M}$. Des perfectionnements de ces règles de combinaison fondamentales ont été proposés par plusieurs auteurs, y compris Li, Raghunathan et Rubin (1991a), Li, Meng et Rubin (1991b), Raghunathan et Siscovick (1996), ainsi que Barnard et Rubin (1999).

3. Simulations illustratives

À la présente section, nous utilisons des simulations pour illustrer les biais/inefficacités associés à l'intégration des caractéristiques du plan dans les modèles d'imputation. Nous simulons trois populations cibles de $N = 100\,000$ unités, qui sont stratifiées et mises en grappes dans les strates. Dans la première population, Y dépend à la fois des effets de strate et de grappe. Dans la deuxième population, Y dépend des effets de strate, mais non des effets de grappe.

Dans la troisième population, Y n'est relié ni aux indicateurs de strate ni aux indicateurs de grappe. Nous utilisons la première population pour démontrer qu'il importe d'inclure toutes les variables du plan pertinentes, et les deuxième et troisième populations, pour examiner l'effet de l'inclusion de variables du plan non pertinentes. Les populations simulées sont stylisées afin d'illustrer l'importance de la modélisation du plan de sondage; par conséquent, la grandeur des biais/inefficacités n'est pas nécessairement généralisable à d'autres conditions.

Chaque population est divisée en cinq strates de taille égale comprenant chacune $N_h = 200$ grappes, pour $h = 1, \ldots, 5$. Chaque grappe c dans la strate h comprend N_{hc} unités. Dans chaque strate, 10 grappes ont $N_{hc} = 300$, 20 grappes ont $N_{hc} = 200$, 60 grappes ont $N_{hc} = 100$, 60 grappes ont $N_{hc} = 75$, et cinquante grappes ont $N_{hc} = 50$. Nous faisons varier les tailles de grappe afin de grossir les effets du plan lors du tirage d'échantillons en grappes à plusieurs degrés. Pour chaque population cible, il existe deux variables d'enquête, X et Y. Dans les trois populations, par souci de simplicité, nous générons chaque X_{hcj}, où l'indice j indique une unité dans la strate et la grappe hc, à partir de $X_{hcj} \sim N(0, 10^2)$. Pour générer Y, nous utilisons différentes méthodes pour chaque population, comme nous le décrirons aux sections qui suivent.

Nous échantillonnons aléatoirement les unités à partir de chaque population en utilisant un plan d'échantillonnage en grappes à plusieurs degrés. Pour commencer, nous tirons un échantillon aléatoire simple de $n_1 = 40$ grappes à partir de la strate 1, $n_2 = 20$ grappes à partir de la strate 2, $n_3 = 30$ grappes à partir de la strate 3, $n_4 = 10$ à partir de la strate 4 et $n_5 = 15$ grappes à partir de la strate 5. Les tailles des échantillons en grappes varient selon la strate afin de grossir les effets de plan comparativement à l'échantillonnage uniforme. Puis, nous tirons un échantillon aléatoire simple de 20 unités dans chaque grappe échantillonnée. Donc, nous obtenons 2 300 unités pour lesquelles $I_{hcj} = 1$.

Dans chaque population, les paramètres estimés sont $Q = \overline{Y}$, la moyenne de population de Y, et les coefficients de population de la régression de Y sur X. L'estimateur en données complètes de $\overline{Y}$ est l'estimateur sans biais fondé sur le plan de sondage habituel,

$$q = \frac{1}{100\,000}\left(\sum_{h=1}^{5} \frac{200}{n_h} \sum_{c \in h} \hat{y}_{hc}\right),$$

où $\hat{y}_{hc} = N_{hc}\overline{y}_{hc}$ est le total estimé dans la grappe hc. L'estimateur en données complètes de la variance de q est

$$u = \frac{1}{100\,000^2}\left(\begin{array}{l} \sum_{h=1}^{5} 200^2 \left(1 - \frac{n_h}{200}\right) s_h^2 / n_h \\ + \sum_{h=1}^{5} \frac{200}{n_h} \sum_{c \in h} N_{hc}^2 \left(1 - \frac{20}{N_{hc}}\right) s_{hc}^2 / 20 \end{array}\right),$$

où s_h^2 est la variance d'échantillon de $\hat{y}_{hc}$ et s_{hc}^2 est la variance d'échantillon de Y dans la grappe hc. Les estimateurs des coefficients dans la régression de Y sur X sont les estimateurs approximativement sans biais fondés sur le plan de sondage habituels, qui sont calculés en utilisant les routines « Survey » (Lumley 2004) du progiciel R. Ces routines estiment les variances selon des techniques de linéarisation par développement en série de Taylor. Ces estimateurs sont utilisés pour tous les ensembles de données multi-imputés dans toutes les simulations.

Pour chaque échantillon, nous posons que X est entièrement observée et que, pour Y, des données manquent pour environ 30 % des unités échantillonnées.

Pour chaque unité, la variable de réponse binaire, R_{hcj}, est tirée à partir d'une loi de Bernoulli :

$$\Pr(R_{hcj} = 1) = \frac{\exp(-0,847 - 0,1 X_{hcj})}{(1 + \exp(-0,847 - 0,1 X_{hcj}))} \qquad (4)$$

Ici, $R_{hcj} = 1$ signifie que la valeur de Y manque pour l'unité en question. L'équation 4 implique que Y_{mis} manque au hasard (Rubin 1976). Nous pouvons ignorer le mécanisme de création des données manquantes à condition que les imputations pour ces données soient conditionnelles à X. Délibérément, nous ne permettons pas que l'absence de données dépende de l'appartenance à la strate ou à la grappe, afin d'illustrer que le biais peut être dû au fait de ne pas tenir compte du plan de sondage, même si le mécanisme de création de données manquantes ignorable ne dépend pas du plan d'échantillonnage. Naturellement, si le plan d'échantillonnage est relié au fait que des données manquent, comme cela est le cas dans de nombreux ensembles de données réels, il faut introduire les contraintes du plan d'échantillonnage afin que le mécanisme de création des données manquantes soit ignorable.

Nous examinons trois stratégies d'imputation de Y_{mis} s'appuyant sur différentes utilisations de l'information sur le plan de sondage. Ces stratégies sont résumées au tableau 1. La première, dénotée EAS, omet entièrement de tenir compte du plan d'échantillonnage. La deuxième, dénotée FX, intègre la stratification et la mise en grappes grâce à l'utilisation d'effets fixes pour chaque grappe dans la strate. La troisième stratégie, dénotée HM, consiste à utiliser des modèles normaux à effets aléatoires dans lesquels sont intégrées la stratification et la mise en grappes. Pour EAS, un modèle est ajusté à l'ensemble de données complet. Pour FX et HM, les modèles sont ajustés séparément à chaque strate. Les trois stratégies comportent la régression sur X, parce que cette variable fait partie du mécanisme de création des données manquantes; ne pas conditionner à X violerait l'ignorabilité et causerait un biais.

Tableau 1
Stratégies d'imputation

Étiquette	Modèle d'imputation pour Y_{hcj} manquante
EAS	$N(\beta_0 + \beta_1 X_{hcj}, \sigma^2)$
FX	$N(\beta_{0h} + \beta_{1h} X_{hcj} + \omega_{hc}, \sigma_h^2)$
HM	$N(\beta_{0h} + \beta_{1h} X_{hcl} + \omega_{hc}, \sigma_h^2), \omega_{hc} \sim N(0, \tau^2)$

Toutes les imputations sont tirées d'après les lois bayésiennes prédictives a posteriori appropriées. Premièrement, nous sélectionnons les paramètres des modèles d'imputation à partir des lois a posteriori sachant les composantes des données observées, $(Z, X, Y_{\text{obs}}, I, R)$, qui sont incluses dans les modèles. Deuxièmement, nous sélectionnons les valeurs des données manquantes à partir des lois données au tableau 1. Nous utilisons des lois a priori diffusés pour tous les paramètres. Pour la stratégie HM, nous tirons les valeurs des paramètres en utilisant un échantillonneur de Gibbs (Gelfand et Smith 1990). Nous exécutons l'échantillonneur pendant une période de rodage pour obtenir la convergence approximative, puis nous utilisons chaque dixième tirage pour les imputations. Enfin, nous utilisons $M = 5$ imputations tirées indépendamment dans chaque ensemble de données pour chaque stratégie.

3.1 Simulation A : Illustration de la non-prise en compte des caractéristiques pertinentes du plan

Dans cette simulation, nous générons une population dans laquelle la distribution de Y diffère selon la strate et la grappe. Nous l'appelons « Population 1 ». Plus précisément, pour l'unité j dans la strate h et la grappe c, nous construisons la valeur de population de Y_{hcj} d'après

$$Y_{hcj} = 10 X_{hcj} + \beta_{0h} + \omega_{hc} + \in_{hcj} \quad (5)$$

où β_{0h} est une constante scalaire pour la strate h, ω_{hc} est une constante scalaire pour la grappe hc, et $\in_{hcj}$ est un terme d'erreur aléatoire tiré à partir de $N(0, 200^2)$. Les valeurs des effets de strate sont $\beta_{01} = 500, \beta_{02} = -250, \beta_{03} = 0, \beta_{04} = 250,$ et $\beta_{05} = -500$. Les valeurs de ω_{hc} sont obtenues en tirant cinq ensembles de $N_h = 200$ valeurs à partir de lois $N(0, 70^2)$ indépendantes. Les effets de strate et de grappe sont fortement dispersés afin de grossir les effets de plan comparativement à l'échantillonnage aléatoire simple, qui, à son tour, grossit les effets de la non-prise en compte du plan d'échantillonnage dans les imputations. Puis, nous tirons un échantillon à partir de cette population selon le plan d'échantillonnage en grappes stratifié décrit antérieurement. Nous créons l'indicateur de données manquantes R en utilisant l'équation 4.

Le tableau 2 montre les résultats de 1 000 répétitions des trois stratégies d'imputation décrites au tableau 1. La ligne supplémentaire annotée « Données complètes » donne les résultats en utilisant les données pour toutes les unités échantillonnées, c'est-à-dire en supposant qu'aucune unité pour laquelle $I_{hcj} = 1$ n'a $R_{hcj} = 0$. La colonne étiquetée « Couv. IC à 95 % » contient le pourcentage des 1 000 intervalles de confiance simulés qui contiennent le paramètre de population. La colonne étiquetée « Est. ponc. » contient les moyennes des 1 000 estimations ponctuelles de Q. La colonne étiquetée « Var. » contient les variances des 1 000 estimations ponctuelles de Q. La colonne étiquetée « Var. est. » contient les moyennes sur les 1 000 répétitions des variances estimées des estimations ponctuelles. Les colonnes étiquetées « Var(var. est.) » et « EQM(var. est.) » donnent la variance et l'erreur quadratique moyenne des 1 000 variances estimées.

Tableau 2
Propriétés des procédures d'imputation lorsque les caractéristiques du plan sont reliées à la variable d'enquête d'intérêt
La moyenne de population est égale à 3,2 et les coefficients de régression de population sont égaux à 3,0 et 10,1

	Méthode	Couv. IC à 95 %	Est. ponc.	Var.	Var. est.	Var. (var. est.)	EQM (var. est.)
Moyenne de Y	Données complètes	94,2	2,0	544,91	527,31	31 626,19	31 936,07
	EAS	38,0	45,8	327,79	360,74	11 927,97	13 013,35
	FX	94,8	2,4	554,09	579,92	37 474,82	38 141,70
	HM	94,5	2,3	551,02	553,16	34 056,39	34 060,99
Ordonnée à l'origine	Données complètes	93,0	2,4	529,51	499,73	18 543,13	19 430,21
	EAS	39,5	46,8	340,09	365,50	9 351,15	9 996,99
	FX	94,5	2,8	539,19	551,68	21 529,16	21 685,33
	HM	93,9	2,7	536,82	524,82	19 256,24	19 400,11
Pente	Données complètes	93,3	10,1	1,24	1,15	0,14	0,15
	EAS	64,8	7,6	2,10	2,20	0,55	0,56
	FX	94,5	10,1	1,45	1,44	0,18	0,18
	HM	95,7	10,1	1,53	1,65	0,29	0,30

Les imputations fondées sur la méthode EAS produisent des estimations gravement biaisées et une couverture très médiocre des intervalles de confiance dans cette population. Ces problèmes existent même si peu d'information manque et malgré le fait que nous utilisons des estimateurs sans biais par rapport au plan de sondage pour les inférences. Les méthodes FX et HM produisent toutes deux des estimations ponctuelles qui concordent approximativement avec les estimations ponctuelles basées sur les données complètes et donnent toutes deux des taux de couverture qui correspondent approximativement aux taux obtenus pour l'inférence d'après les données complètes. FX et HM ont des profils similaires, parce que les modèles à effets fixes et les modèles hiérarchiques produisent des estimations similaires des paramètres dans l'équation 5.

Lors de l'estimation de la moyenne de population, la variance associée à FX ou à HM n'est que légèrement plus grande que celle associée à l'estimateur d'après des données complètes. Il en est ainsi à cause des grands effets de grappe, qui font de la variance dans les cellules d'imputation un facteur dominant relativement à la variance entre cellules d'imputation. Autrement dit, la fraction d'information manquante due aux données manquantes est relativement faible comparativement à l'effet de la mise en grappes.

3.2 Simulation B : Illustration de l'inclusion de variables explicatives non pertinentes

La modélisation des caractéristiques du plan est essentielle quand ces dernières sont reliées aux variables d'enquête d'intérêt. Quelle est l'incidence de la modélisation de caractéristiques non pertinentes du plan sur les inférences? À la présente section, nous présentons les résultats de deux études par simulation réalisées en vue d'étudier cette question.

Premièrement, nous générons la « Population 2 » dans laquelle la distribution de Y diffère selon la strate, mais ne dépend pas des grappes. Pour cela, nous utilisons la même méthode que celle donnée par l'équation 5, en fixant ω_{hc} à zéro. Les $\in_{hcj}$ sont tirées à partir de $N(0, 100^2)$. Nous sélectionnons un échantillon à partir de la Population 2 et générons des données manquantes en utilisant les scénarios décrits antérieurement. Les résultats pour 1 000 répétitions sont présentés au tableau 3.

La méthode EAS continue de produire un biais important et une couverture médiocre des intervalles de confiance, parce qu'elle ne tient pas compte de la stratification. Pour les méthodes FX et HM, les moyennes des estimations ponctuelles se situent dans la marge d'erreur de simulation de la moyenne des estimations ponctuelles pour les données complètes. En outre, les taux de couverture des intervalles de confiance correspondent approximativement aux taux de couverture pour les intervalles obtenus d'après des données complètes. Donc, les méthodes FX et HM sont raisonnables pour ces populations, même si les caractéristiques de grappe non pertinentes sont incluses dans les modèles d'imputation.

Ensuite, nous générons la « Population 3 » dans laquelle la distribution de Y est indépendante des indicateurs d'appartenance aux strates et aux grappes. Plus précisément, pour générer Y, nous soustrayons β_{0h} des valeurs de Y générées dans la Population 2. Ensuite, nous tirons un échantillon à partir de la Population 3 en utilisant le plan d'échantillonnage en grappes stratifié et en créant des données manquantes selon les méthodes décrites antérieurement. Les résultats pour 1 000 répétitions sont présentés au tableau 4.

Tableau 3
Propriétés des procédures d'imputation lorsque la population comprend des effets de strate, mais non des effets de grappe
La moyenne de population est égale à 0,34 et les coefficients de régression de population sont égaux à 0,14 et 10,13

	Méthode	Couv. IC à 95 %	Est. ponc.	Var.	Var. est.	Var. (var. est.)	EQM (var. est.)
Moyenne de Y	Données complètes	93,6	0,37	468,97	461,88	29 301,77	29 352,04
	EAS	31,1	42,90	259,46	303,46	10 228,40	12 164,74
	FX	93,7	0,32	473,86	474,21	30 408,95	30 409,07
	HM	93,4	0,34	476,03	465,53	29 406,61	29 516,85
Ordonnée à l'origine	Données complètes	93,0	0,72	451,46	432,74	14 955,20	15 305,73
	EAS	31,5	43,10	275,22	311,36	8 134,04	9 440,57
	FX	93,2	0,66	456,08	444,88	15 539,21	15 664,64
	HM	92,3	0,68	457,48	436,25	14 941,00	15 391,75
Pente	Données complètes	93,1	10,09	0,99	0,91	0,09	0,10
	EAS	59,0	7,72	1,67	1,77	0,35	0,36
	FX	93,4	10,10	1,03	0,98	0,10	0,10
	HM	93,3	10,10	1,03	0,96	0,10	0,10

Tableau 4
Propriétés des procédures d'imputation lorsque les variables du plan sont entièrement non corrélées à la variable d'enquête d'intérêt
La moyenne de population est égale à 0,34 et les coefficients de régression de population sont égaux à 0,14 et 10,04

	Méthode	Couv. IC à 95 %	Est. ponc.	Var.	Var. est.	Var. (var. est.)	EQM (var. est.)
Moyenne de Y	Données complètes	94,7	0,35	14,61	14,73	32,65	32,66
	EAS	95,7	0,12	16,45	19,22	40,65	48,31
	FX	97,8	0,40	19,64	28,29	97,66	172,38
	HM	95,1	0,26	18,77	19,16	47,29	47,44
Ordonnée à l'origine	Données complètes	93,7	0,12	7,13	7,20	5,31	5,32
	EAS	96,8	−0,10	8,97	11,72	13,59	21,10
	FX	98,6	0,17	12,23	20,62	39,84	110,24
	HM	96,2	0,03	10,45	11,61	15,09	16,45
Pente	Données complètes	94,5	10,04	0,07	0,07	0,001	0,001
	EAS	96,4	10,07	0,10	0,13	0,002	0,003
	FX	96,4	10,04	0,12	0,15	0,003	0,004
	HM	95,2	10,05	0,11	0,12	0,002	0,002

La méthode EAS produit enfin des estimations ponctuelles dont les moyennes sont comprises dans la marge d'erreur de simulation de l'estimation ponctuelle moyenne d'après des données complètes. Il en est ainsi parce que l'imputation sous EAS reflète raisonnablement bien la structure de population. Il semble donc que ne pas tenir compte du plan d'échantillonnage dans les modèles d'imputation peut fournir des inférences acceptables lorsque les variables du plan ne sont que faiblement corrélées aux résultats de l'enquête. Comme dans les simulations antérieures, les méthodes FX et HM produisent des estimations ponctuelles moyennes comprises dans la marge d'erreur de simulation de l'estimation ponctuelle moyenne d'après des données complètes. Si nous comparons les trois stratégies d'imputation, nous voyons que FX et HM sont inefficaces comparativement à EAS, parce que les modèles d'imputation pour les deux premières méthodes estiment des paramètres qui sont approximativement nuls dans la population, tandis que dans la méthode EAS, leur valeur est fixée à zéro. La variance est plus faible pour HM que pour FX, parce que le modèle d'imputation hiérarchique lisse les effets de grappe estimés vers zéro.

Pour la méthode FX, le pourcentage d'intervalles de confiance qui couvrent Q est plus grand que le pourcentage observé pour les intervalles calculés d'après des données complètes et pour la méthode HM. Cela tient au fait que la variance estimée pour FX a tendance à être plus grande que la variance réelle. Ce biais par excès apparent dans T_M existe également dans le cas de la méthode EAS, ce qui donne un pourcentage de couverture plus grand que celui calculé pour les données complètes et la méthode HM.

4. Exemple fondé sur des données réelles

Nous allons maintenant examiner l'effet de la prise en compte de la stratification et de la mise en grappes lors de l'imputation pour traiter les données manquantes dans un ensemble de données réelles. Les données proviennent du fichier à grande diffusion des National Health and Nutrition Examination Surveys réalisées de 1999 à 2002. Les individus sont groupés en 56 grappes réparties entre 28 strates. De 5 % à 10 % de données manquantes sont relevées pour de nombreuses variables.

Nous avons imputé les données manquantes selon deux stratégies, l'une ne tenant pas compte des variables du plan (comme EAS) et l'autre intégrant les variables du plan au moyen d'effets fixes pour les indicateurs de grappes (comme FX). Dans le modèle d'imputation, nous avons inclus 27 variables nominales pour représenter 28 strates et une variable nominale dans chaque strate pour représenter les deux grappes emboîtées dans chaque strate. Autrement dit, nous avons inclus, en tout, 55 variables nominales à titre de variables explicatives. Nous avons utilisé une procédure de sélection séquentielle des variables pour repérer les interactions statistiquement significatives entre les variables nominales et les variables d'enquête, et nous avons également inclus ces interactions comme variables explicatives dans le modèle d'imputation. Nous avons imputé les valeurs suivant la méthode de régression séquentielle implémentée dans le progiciel IVEWARE (www.isr.umich.edu/src/smp/ive). Nous avons généré $M = 10$ ensembles de données pour chaque stratégie.

Nous considérons l'estimation de trois paramètres. Le premier est le pourcentage des personnes dans la population qui ont déjà fait vérifier leur taux de cholestérol (BPQ060). La proportion de données manquantes pour cette variable est d'environ 15 %. Les deuxième et troisième sont les coefficients de régression de population dans une régression logistique de BPQ060 sur le ratio revenu-seuil de pauvreté de la famille (INDFMPIR), variable continue pour laquelle la proportion de valeurs manquantes est d'environ 12 %. Ces paramètres sont estimés par des méthodes fondées sur

le plan de sondage au moyen des routines « Survey » du progiciel R.

Le tableau 5 donne les résultats pour les deux stratégies d'imputation. Pour toutes les analyses, les deux ensembles d'estimations sont fort semblables. Dans ce cas, l'intégration des variables du plan dans le modèle d'imputation n'a presque pas d'effets sur les résultats. Cela tient, en partie, aux faibles fractions d'information manquante et à l'insignifiance relative des effets de strate et de grappe. Cependant, la pénalité pour l'inclusion des caractéristiques du plan dans le modèle d'imputation est minime. À la lumière des résultats des simulations présentés à la section 3, nous intégrerions les caractéristiques du plan dans ce modèle d'imputation.

Tableau 5
Comparaison des résultats des données réelles lorsque les caractéristiques du plan sont incluses dans le modèle d'imputation et lorsque les caractéristiques du plan sont ignorées

	Est. ponc.	E.-t.	IC à 95 %
Moyenne de BPQ060			
Variables du plan	0,319	0,010	(0,299, 0,339)
Pas de variable du plan	0,319	0,011	(0,296, 0,341)
Ordonnée à l'origine : régression logistique			
Variables du plan	0,362	0,054	(0,256, 0,467)
Pas de variable du plan	0,352	0,052	(0,251, 0,454)
Pente : régression logistique			
Variables du plan	−0,409	0,020	(−0,449, −0,369)
Pas de variable du plan	−0,407	0,019	(−0,444, −0,371)

5. Conclusion

Quoique limitées, les études par simulation donnent à penser que ne pas tenir compte du plan d'échantillonnage dans l'imputation multiple peut être une pratique risquée. Lorsque les variables du plan sont corrélées aux variables d'enquête, comme dans notre simulation A, omettre de les inclure peut donner lieu à un biais important. Par ailleurs, l'inclusion de variables du plan non pertinentes, comme dans notre simulation B et dans l'exemple des enquêtes NHANES, produit, au pire, des inférences inefficaces et prudentes, lorsque les modèles d'imputation sont par ailleurs spécifiés correctement.

Inclure des variables nominales pour les effets de grappe réduit considérablement le biais comparativement à la non-prise en compte totale du plan. Cependant, l'introduction aveugle de variables nominales n'est pas une solution automatique. Lorsque la pente de la régression ou les variances diffèrent selon la grappe, l'utilisation de la méthode FX ou HM peut produire des estimations biaisées, puisque des caractéristiques importantes du plan sont omises. Les imputeurs qui soupçonnent l'existence de relations de ce genre devraient inclure les interactions appropriées avec les variables nominales pour les caractéristiques du plan, comme nous l'avons fait dans l'exemple des enquêtes NHANES. Dans le cas de certaines enquêtes, le plan peut être si complexe qu'il est impossible d'inclure des variables nominales pour chaque grappe. Le cas échéant, les imputeurs peuvent simplifier le modèle en ce qui concerne les variables du plan, par exemple en regroupant des catégories de grappes ou en incluant des variables de substitution (par exemple, taille de grappe) qui sont corrélées à la variable d'enquête d'intérêt.

Les simulations donnent à penser qu'il pourrait être avantageux d'utiliser des modèles hiérarchiques plutôt que des modèles à effets fixes pour l'imputation des données manquantes, particulièrement lorsque les effets de grappe sont semblables. Toutefois, les modèles hiérarchiques sont plus difficiles à ajuster que les modèles à effets fixes. Ainsi, l'ajustement de modèles hiérarchiques dans le cas de plans d'échantillonnage complexes lorsque des données manquent pour plusieurs variables continues et catégoriques est une tâche redoutable. Des modèles hiérarchiques séquentiels pourraient peut-être être ajustés dans un esprit semblable aux imputations par régression séquentielle de Raghunathan et coll. (2001). Il s'agit d'un domaine dans lequel devraient se poursuivre les travaux de recherche. Un autre inconvénient des modèles hiérarchiques est qu'il est plus facile de les spécifier incorrectement que les modèles à effets fixes. Ainsi, si les effets de grappe suivent une loi non normale, le modèle hiérarchique normal utilisé dans le présent article pourrait donner des imputations non plausibles.

Dans le cas de l'imputation multiple, la clé du succès réside dans la spécification d'un modèle d'imputation qui décrit raisonnablement la loi conditionnelle des valeurs manquantes sachant les valeurs observées. Souvent, les caractéristiques du plan sont corrélées aux variables d'enquête, de sorte que leur inclusion dans les modèles d'imputation réduit les risques d'erreur de spécification du modèle. Nous pensons que, dans de nombreux cas, les biais que peut causer l'exclusion de variables importantes, du plan ou d'autres variables reliées au mécanisme de création des données manquantes, surpassent les inefficacités qui pourraient résulter de l'estimation de petits coefficients. Cela renforce le conseil général fréquemment donné concernant l'imputation multiple : inclure toutes les variables qui sont reliées aux données manquantes dans les modèles d'imputation afin de rendre ignorable le mécanisme de création des données manquantes (par exemple, Meng 1994; Little et Raghunathan 1997; Schafer 1997; Collins, Schafer et Kam 2001).

Remerciements

La présente étude a été financée par la bourse ITR-0427889 de la National Science Foundation. Les auteurs remercient le rédacteur adjoint et les examinateurs de leurs commentaires et suggestions.

Bibliographie

Barnard, J., et Meng, X. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8, 17-36.

Barnard, J., et Rubin, D.B. (1999). Small-sample degrees of freedom with multiple-imputation. *Biometrika*, 86, 948-955.

Collins, L.M., Schafer, J.L. et Kam, C.K. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, 6, 330-351.

Gelfand, A.E., et Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Heitjan, D.F., et Little, R.J.A. (1991). Multiple imputation for the Fatal Accident Reporting System. *Applied Statistics*, 40, 13-29.

Kennickell, A.B. (1998). Multiple imputation in survey of consumer finances. Dans *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association, 11-20.

Li, K.H., Raghunathan, T.E. et Rubin, D.B. (1991a). Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86, 1065-1073.

Li, K.H., R.T.E., Meng, X.L. et Rubin, D.B. (1991b). Significance levels from repeated $p-$values with multiply-imputed data. *Statistica Sinica*, 1, 65-92.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Little, R.J.A., et Raghunathan, T.E. (1997). Should imputation of missing data condition on all observed variables? Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 617-622.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9, 8.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science*, 9, 538-558.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. et Solenberger, P. (2001). Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression. *Techniques d'enquête*, 27, 91-103.

Raghunathan, T.E., and Paulin, G.S. (1998). Multiple imputation of income in the Consumer Expenditure Survey: Evaluation of statistical inference. Dans *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association, 1-10.

Raghunathan, T.E., Reiter, J.P. et Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.

Raghunathan, T.E., et Siscovick, D.S. (1996). A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics*, 45, 335-352.

Reiter, J.P. (2003). Inférence pour les ensembles de microdonnées à grande diffusion partiellement synthétiques. *Techniques d'enquête*, 29, 203-211.

Reiter, J.P. (2004). Utilisation simultanée de l'imputation multiple pour les données manquantes et le contrôle de la divulgation. *Techniques d'enquête*, 30, 263-271.

Reiter, J.P. (2005). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society*, Séries A, 168, 185-205.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-590.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York : John Wiley & Sons, Inc.

Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462-468.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A. et Rubin, D.B. (1998). The NHANES III multiple imputation project. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 28-37.

Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G. et Cohen, A.J. Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association*, a paraître.