Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Imputation multiple pour la non-réponse dans les enquêtes à l’aide de poids de sondage et de marges auxiliaires

par Kewei Xu et Jerome P. Reiter

Date de diffusion : le 29 juin 2026

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, juin 2026
Vol. 52, nº 1, pp. 287-299
Statistique Canada, nº 12-001-X au catalogue

Imputation multiple pour la non-réponse dans les enquêtes à l'aide de poids de sondage et de marges auxiliaires

Kewei Xu et Jerome P. Reiter[1]

Résumé

Habituellement, il manque des valeurs dans les données d'enquête en raison de la non-réponse totale et partielle. Parfois, les organismes d'enquête connaissent les distributions marginales de certaines variables catégoriques dans la population visée. Comme l'ont démontré des travaux antérieurs, les organismes d'enquête peuvent tirer profit de ces distributions dans une imputation multiple pour la non-réponse totale non ignorable, en générant des imputations qui donnent lieu à des estimations plausibles à partir de données complètes pour les variables dont les marges sont connues. Cependant, ces travaux antérieurs ne font pas appel aux poids de sondage pour les unités non répondantes. Nous prolongeons ces travaux antérieurs pour utiliser les poids de sondage pour toutes les unités échantillonnées. Nous illustrons l'approche à l'aide d'études par simulation.

Mots-clés : Non-réponse non ignorable; non-réponse partielle; non-réponse totale; valeurs manquantes.

1. Introduction

Les données d'enquête sont généralement touchées par la non-réponse partielle et totale. Pour cette raison, les organismes d'enquête doivent avancer des hypothèses robustes concernant les raisons du manque de données, par exemple elles sont manquantes de façon aléatoire. Utiliser les renseignements tirés de sources de données auxiliaires constitue un moyen de réduire la dépendance aux hypothèses. À titre d'exemple et d'élément pertinent pour le contexte de notre travail, les organismes d'enquête pourraient connaître les pourcentages ou les totaux de population de certaines variables catégoriques qui se trouvent dans l'enquête grâce à de récents recensements, à des bases de données administratives ou à des enquêtes de grande qualité (Sadinle et Reiter, 2017). En effet, des méthodes, comme le calage et le ratissage, font souvent appel à ces renseignements pour gérer la non-réponse à l'enquête.

Nous examinons des contextes dans lesquels les analystes cherchent à intégrer de l'information auxiliaire sur les distributions marginales dans une imputation multiple pour la non-réponse (Rubin, 1987) dans les enquêtes. Nous nous appuyons sur le cadre de données manquantes avec marges auxiliaires (DMMA), introduit par Akande, Madson, Hillygus et Reiter (2021) et étendu aux enquêtes par Akande et Reiter (2022), Tang, Hillygus et Reiter (2024), et Yang et Reiter (2025). Dans les trois derniers travaux, on impute des valeurs manquantes pour faire en sorte que les inférences à partir de données complètes, pondérées par des poids d'enquête soient plausibles compte tenu des marges connues. Ces travaux imposent cependant une condition simplificatrice aux unités non répondantes, à savoir que l'analyste ne doit pas utiliser ses poids de sondage et doit plutôt les remplacer par une constante pratique.

Dans cette communication brève, nous proposons des modèles de DMMA qui permettent d'utiliser des poids fondés sur le plan d'échantillonnage pour toutes les unités échantillonnées. Ces modèles visent plus

1. Kewei Xu et Jerome P. Reiter, Department of Statistical Science, case postale 90251, Duke University, Durham, Caroline du Nord, 27708-0251. Courriel : jreiter@duke.edu.

particulièrement les échantillons à probabilités inégales dans lesquels on veut incorporer des liens entre les variables du plan d'échantillonnage et les variables d'enquête dans l'imputation. Dans ces modèles, on présume que des poids d'échantillonnage sont disponibles pour toutes les unités échantillonnées, ce qui devrait être vrai pour l'organisme qui a recueilli les données. La section 2 présente un court examen du modèle hybride de DMMA (Tang et coll., 2024; Yang et Reiter, 2025), que nous élargissons pour intégrer des poids d'échantillonnage pour les unités non répondantes. La section 3 présente la méthodologie. La section 4 illustre les méthodes à l'aide d'études par simulation. Le code et des résultats supplémentaires se trouvent à l'adresse suivante : https://github.com/kevinxu47/MDAM (en anglais).

2. Examen de la modélisation hybride de données manquantes avec marges auxiliaires

Lors de l'examen du modèle hybride de DMMA, nous respectons scrupuleusement la notation présente dans Yang et Reiter (2025). Supposons que $\mathcal{P}$ est une population finie comprenant N unités. Pour $i = 1,\ldots,N$, supposons que $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})$ désigne les p variables d'enquête pour l'unité i; supposons que $I_i = 1$ si l'unité i est sélectionnée pour être incluse dans l'enquête et que $I_i = 0$ dans les autres cas; supposons que $\pi_i = \Pr(I_i = 1)$ représente la probabilité que l'unité i soit sélectionnée dans l'enquête; supposons que $w_i^d = 1/\pi_i$ est son poids de sondage; et supposons que $\mathbf{z}_i$ correspond à ses variables du plan de sondage, par exemple les mesures de la taille ou les indicateurs de strate. Supposons que $n = \sum_{i=1}^{N} I_i$ est la taille d'échantillon visée. Nous désignons les unités échantillonnées par $\mathcal{S}$.

Supposons que U est l'indicateur de non-réponse totale de sorte que, pour chaque unité $i \in \mathcal{S}$, $U_i = 1$ si l'unité ne fournit pas de réponse à toute question d'enquête et $U_i = 0$ autrement. Supposons que $\mathbf{R} = (R_1,\ldots,R_p)$ représente les indicateurs de non-réponse partielle correspondant à $\mathbf{X} = (X_1,\ldots,X_p)$, de sorte que pour toute unité i pour laquelle $U_i = 0$ et toute X_j, $R_{ij} = 1$ si l'unité i ne répond pas à la question pour X_j et $R_{ij} = 0$ autrement. Lorsque $U_i = 1$, $(R_{i1},\ldots,R_{ij})$ ne sont pas définis.

L'organisme d'enquête dispose d'information auxiliaire $\mathcal{A}$ sur les distributions marginales d'un sous-ensemble de $\mathbf{X}$. Nous écrivons $X_j \in \mathcal{A}$ dès que X_j a une marge connue dans $\mathcal{A}$, que nous désignons par $\mathcal{A}_j$. Pour des raisons pratiques, nous présumons que chaque $\mathcal{A}_j$ se compose de totaux de population; incorporer des pourcentages constitue une modification simple. Pour toute variable catégorique X_j prenant les niveaux $c = 1,\ldots,m_j$, supposons que $T_{jc} = \sum_{i=1}^{N} I(x_{ij} = c)$ est le nombre total d'unités dans $\mathcal{P}$ au niveau c. Dans ce cas-ci, $I(\cdot) = 1$ lorsque la condition à l'intérieur des parenthèses est vraie et $I(\cdot) = 0$ autrement.

Comme l'ont décrit Tang et coll. (2024) et Yang et Reiter (2025), le modèle hybride de DMMA précise une distribution conjointe de $(\mathbf{X},\mathbf{R},U)$. Le modèle présume plus particulièrement $U \sim \text{Bernoulli}(\pi_u)$, où $\pi_u = \Pr(U = 1)$ est la probabilité marginale de non-réponse totale. Pour $j = 1,\ldots,p$, supposons que $g_j(-)$ représente le modèle pour X_j compte tenu de $X_1,\ldots,X_{j-1}$, et supposons que Ω_j et θ_j sont les paramètres de modèle correspondants. Par exemple, $g_j(-)$ pourrait être une régression logistique de X_j pour une

certaine fonction de $X_1,\ldots,X_{j-1},$ comprenant possiblement des interactions, de même qu'un effet principal pour U lorsque $X_j \in \mathcal{A}.$ Ainsi, pour $\mathbf{X}|U,$ le modèle hybride de DMMA utilise

$$X_1 \,|\, U \;\sim\; g_1(\Omega_1, \theta_1 UI(X_1 \in \mathcal{A}) \tag{2.1}$$

$$X_j \,|\, X_1,\ldots,X_{j-1},U \;\sim\; g_j(X_1,\ldots,X_{j-1},\Omega_j,\theta_j UI(X_j \in \mathcal{A}), \text{ pour } j = 2,\ldots,p. \tag{2.2}$$

Puisque la distribution de toute $X_j \in \mathcal{A}$ diffère pour les unités non répondantes et les répondants lorsque $\theta_j \neq 0,$ le modèle encode les mécanismes de données potentiellement manquantes de façon non aléatoire pour la non-réponse totale. L'hypothèse déterminante est que la fonction de prédiction pour toute $X_j \in \mathcal{A}$ ne comprend pas d'interactions entre U et les éléments de $(X_1,\ldots,X_{j-1}).$ Cette hypothèse est une version du mécanisme additif de données manquantes non ignorables (Hirano, Imbens, Ridder et Rubin, 2001; Sadinle et Reiter, 2019).

Dans le modèle pour chaque $R_j,$ supposons que $h_j(-)$ est une certaine fonction de prédiction qui exclut la X_j correspondante, mais pourrait inclure des effets majeurs et des effets d'interaction touchant d'autres variables dans $\mathbf{X}.$ Pour $j = 1,\ldots,p,$ nous avons

$$R_j \,|\, X_1,\ldots,X_p,U = 0 \;\sim\; \text{Bernoulli}(\pi_{R_j}),\; \text{logit}(\pi_{R_j}) = h_j(X_1,\ldots,X_{j-1},X_{j+1},\ldots X_p,\Phi_j). \tag{2.3}$$

Les modèles dans (2.3) encodent des mécanismes de non-réponse conditionnellement indépendante au niveau détaillé (Sadinle et Reiter, 2017), qui sont connus pour avoir des paramètres repérables.

Les algorithmes d'imputation et d'estimation décrits dans Tang et coll. (2024) et Yang et Reiter (2025) n'utilisent pas de poids de sondage pour les unités non répondantes. Ils remplacent plutôt les poids de sondage pour les unités non répondantes par une constante qui fait en sorte que $\sum_{i=1}^{N} w_i = N;$ c'est-à-dire qu'ils fondent les imputations et les inférences à partir de données complètes, pour toutes les unités $i \in \mathcal{S},$ sur

$$w_i \;=\; \begin{cases} w_i^d & \text{si } U_i = 0 \\ \dfrac{N - \sum_{k\in\mathcal{S}} w_k^d}{\sum_{k\in\mathcal{S}} U_k} & \text{si } U_i = 1. \end{cases} \tag{2.4}$$

Pour toute unité $i \in \mathcal{S}$ et tout $j = 1,\ldots,p,$ supposons que $x_{ij}^* = x_{ij}$ lorsque $(R_{ij} = 0, U_i = 0),$ et que x_{ij}^* est une valeur imputée lorsque $R_{ij} = 1$ ou $U_i = 1.$ Pour toute $X_j \in \mathcal{A},$ Akande et Reiter (2022), Tang et coll. (2024) et Yang et Reiter (2025) supposent que les imputations se conforment à (2.1) à (2.3), mais en ayant la contrainte supplémentaire selon laquelle

$$\hat{T}_{jc}^* \;=\; \sum_{i\in\mathcal{S}} w_i\, I(x_{ij}^* = c) \;\sim\; N(T_{jc}, V_{jc}), \tag{2.5}$$

où les poids proviennent de (2.4). Dans ce cas-ci, l'analyste établit V_{jc} et détermine la mesure dans laquelle $\hat{T}_{jc}^*$ correspond à T_{jc} dans tout ensemble de données complet.

L'utilisation de (2.5) pour les imputations a des points communs avec l'imputation aléatoire équilibrée (Chauvet, Deville et Haziza, 2011; Chauvet et Do Paco, 2018), même si les buts diffèrent. Dans l'imputation

aléatoire équilibrée, les valeurs imputées doivent toujours satisfaire certaines contraintes imposées, par exemple la moyenne pondérée par les poids d'enquête des valeurs imputées doit être égale à la moyenne pondérée par les poids d'enquête des valeurs observées. Dans les travaux sur l'imputation aléatoire équilibrée, on tend à envisager des imputations simples pour la non-réponse partielle, à présumer que les valeurs manquantes sont (complètement) attribuables au hasard et à fixer des contraintes selon les quantités de données observées. En revanche, dans le modèle hybride de DMMA, on envisage l'imputation multiple pour la non-réponse totale attribuable au fait que les données sont manquantes de façon non aléatoire et l'on utilise l'information auxiliaire marginale pour établir des contraintes de distribution sur les imputations.

Même si Tang et coll. (2024) ou Yang et Reiter (2025) ne l'ont pas fait, le modèle pour chaque X_j ou chaque R_j peut comprendre les variables du plan de sondage $\mathbf{Z}$ en tant que prédicteurs. Inclure les $\mathbf{Z}$ peut accroître le pouvoir explicatif et ainsi, ultimement, la qualité de l'imputation et l'exactitude de l'estimation, lorsqu'elles sont associées aux variables de l'enquête ou aux indicateurs de non-réponse (Reiter, Raghunathan et Kinney, 2006). Sinon, les modèles peuvent inclure une certaine fonction de W en tant que prédicteur. Il s'agit d'un conseil fréquent pour la modélisation d'une imputation (par exemple Kim, Brick, Fuller et Kalton, 2006; Quartagno, Carpenter et Goldstein, 2020) qui se révèle particulièrement important lorsque l'analyste faisant des imputations dispose des poids de sondage, mais pas des variables du plan de sondage parce qu'elles sont confidentielles, par exemple. Dans les modèles de DMMA de Tang et coll. (2024) et de Yang et Reiter (2025), cependant, chaque unité non répondante affiche la même valeur de w_i donnée dans (2.4). Alors, inclure W dans les modèles n'influerait pas sur les probabilités d'imputation et, par conséquent, ne contribuerait pas à l'intégration du plan au processus d'imputation pour les unités non répondantes.

3. Méthodologie

Nous modifions maintenant le modèle hybride de DMMA pour utiliser les poids de sondage pour toutes les personnes échantillonnées, y compris les unités non répondantes. Par souci de concision, nous supposons maintenant que W renvoie aux poids de sondage w_i^d plutôt qu'aux poids dans (2.4). Nous suivons la stratégie générale mentionnée dans Yang et Reiter (2025). Nous imputons d'abord des valeurs pour la non-réponse partielle. Nous utilisons ensuite $\mathcal{A}$ pour imputer des valeurs de variables ayant des marges pour les unités non répondantes. Enfin, nous imputons les valeurs des variables restantes pour les unités non répondantes. Notre innovation consiste à modifier la deuxième étape, de sorte que les imputations et les ensembles de données complets reposent sur les poids de sondage au lieu des poids de (2.4) pour les unités non répondantes. Cette modification exige de nouveaux algorithmes d'imputation, que nous présentons à la section 3.2. Pour les première et troisième étapes, nous appliquons des techniques présentées dans Yang et Reiter (2025) et résumées aux sections 3.1 et 3.3. Pour des raisons pratiques, nous présumons que $X_j \in \mathcal{A}$ pour $j = 1,\ldots,k < p$ et que les $p-k$ variables restantes n'ont pas de marges auxiliaires.

3.1 Variables pour lesquelles il y a une non-réponse partielle

Nous créons d'abord des imputations multiples pour toutes les valeurs de x_{ij} manquantes en raison d'une non-réponse partielle pour les unités pour lesquelles $U_i = 0$. Nous mettons en œuvre une imputation multiple au moyen d'équations en séries (MICE) en utilisant seulement les unités pour lesquelles $U_i = 0$ et comprenant possiblement $\mathbf{Z}$ en tant que prédicteurs. La spécification de l'algorithme de MICE n'exige pas de considérations extraordinaires. Dans nos stimulations, nous utilisons le paquet « mice » dans R (Van Buuren, 2018) et gardons ses paramètres par défaut, y compris le classement des variables. Nous exécutons la procédure de MICE pour créer L ensembles de données complets pour les unités pour lesquelles $U_i = 0$.

3.2 Variables ayant des marges pour les unités non répondantes

Après l'imputation de valeurs pour la non-réponse partielle, nous imputons, dans chaque ensemble de données complet, les valeurs des unités non répondantes pour toutes les $X_j \in \mathcal{A}$. L'analyste classe ces variables en fonction de la facilité de modélisation. Pour des raisons pratiques, nous supposons que X_1 est imputée en premier, que X_2 est imputée en deuxième, et ainsi de suite de façon séquentielle jusqu'à X_k.

Tout d'abord, pour chaque unité i pour laquelle $U_i = 1$ et dans chacun des L ensembles de données complets, l'analyste établit une distribution de probabilités initiale pour imputer x_{i1}. Par exemple, dans chaque ensemble de données complet, l'analyste estime une régression logistique (multinomiale) de X_1 sur une certaine fonction de $\mathbf{Z}$ (ou de W) et utilise les probabilités prédites tirées du modèle estimé. L'analyste pourrait aussi choisir de régresser X_1 sur une ordonnée à l'origine seulement ou de calculer les probabilités marginales de X_1 à l'aide d'estimateurs par quotient pondérés par les poids de sondage. Pour ces deux solutions, les probabilités initiales sont identiques pour toutes les unités non répondantes. Cette approche pourrait être particulièrement appropriée dans les cas où l'analyste ne s'attend pas à des associations entre les poids de sondage et les valeurs manquantes de X_1 pour les unités non répondantes.

Quelle que soit la méthode de calcul, nous appelons la distribution initiale pour l'imputation de x_{i1} « distribution fonctionnelle » et l'écrivons comme suit : $\{p_{i1c} : c = 1,\ldots,m_1; U_i = 1; i \in \mathcal{S}\}$. En raison de l'imputation pour la non-réponse partielle, la distribution fonctionnelle pour toute unité peut varier dans les différents ensembles de données complets; cependant, pour des raisons pratiques, nous renonçons à la notation désignant les ensembles de données complets. L'analyste effectue les calculs de la présente section pour chacun des L ensembles de données complets.

L'imputation de chaque x_{i1} manquante à l'aide de sa distribution fonctionnelle ne repose pas sur l'information dans $\mathcal{A}_1$. Les imputations résultantes peuvent permettre de générer des estimations de Horvitz et Thompson (1952) à partir de données complètes qui sont loin des totaux connus pour certains niveaux de X_1, surtout quand les valeurs manquantes ne sont pas aléatoires. Nous modifions donc $\{p_{i1c}\}$ afin que les imputations résultantes procurent des estimations raisonnables de ces totaux fondées sur le plan et sur des données complètes. Il s'agit de l'objectif premier et de la motivation de notre élargissement du cadre de DMMA.

Pour ce faire, nous simulons d'abord des valeurs plausibles de $\hat{T}_{1c}$ pour chaque niveau c en faisant appel aux théorèmes de la limite centrale pour de grands échantillons. Pour $c = 1,\ldots,(m_1 - 1)$, nous échantillonnons $\hat{T}_{1c}$ à partir de $N(T_{1c}, V_{1c})$ et posons $\hat{T}_{1m_1} = N - \sum_{c=1}^{m_1-1} \hat{T}_{1c}$. Nous trouvons ensuite des probabilités ajustées

$\{\tilde{p}_{i1c} : c = 1,\ldots,m_1; U_i = 1; i \in \mathcal{S}\}$ de sorte que, pour $c = 1,\ldots,m_1,$ l'espérance de chaque estimateur basé sur des données complètes $\hat{T}^*_{1c}$ provenant de (2.5) par rapport aux imputations pour la non-réponse totale soit approximativement égale à $\hat{T}_{1c}$. Autrement dit, en reconnaissant que

$$\begin{aligned}\mathbb{E}\left[\sum_{i\in\mathcal{S}} w_i\, I(x^*_{ij} = c)\right] &= \sum_{i\in\mathcal{S}} w_i\, I(x^*_{ij} = c)\, I(U_i = 0) + \mathbb{E}\left[\sum_{i\in\mathcal{S}} w_i\, I(x^*_{ij} = c)\, I(U_i = 1)\right] \\ &= \sum_{i\in\mathcal{S}} w_i\, I(x^*_{ij} = c)\, I(U_i = 0) + \sum_{i\in\mathcal{S}} w_i \tilde{p}_{i1c}\, I(U_i = 1),\end{aligned}$$

nous voulons que

$$\sum_{i\in\mathcal{S}} w_i \tilde{p}_{i1c}\, I(U_i = 1) \;=\; \hat{T}_{1c} - \sum_{i\in\mathcal{S}} w_i\, I(x^*_{i1} = c)\, I(U_i = 0).$$

Afin que les probabilités d'imputation demeurent liées à la distribution fonctionnelle, pour $c = 1,\ldots,$ $m_1 - 1,$ nous posons $\tilde{p}_{i1c} = f_{1c} p_{i1c},$ où chaque constante $f_{1c} > 0$ est obtenue par

$$f_{1c} \;=\; \frac{\hat{T}_{1c} - \sum_{i\in\mathcal{S}} w_i\, I(x^*_{i1} = c, U_i = 0)}{\sum_{i\in\mathcal{S}} p_{i1c} w_i\, I(U_i = 1)}. \tag{3.1}$$

Nous imputons x^*_{i1} pour chaque unité pour laquelle $U_i = 1$ en procédant à tirage aléatoire avec la distribution de probabilités ajustée, $\{\tilde{p}_{i1c} : c = 1,\ldots,m_1\},$ où $\tilde{p}_{i1m_1} = 1 - \sum_{c=1}^{m_1-1} f_{1c} p_{i1c}$. Si $\sum_{c=1}^{m_1-1} f_{1c} p_{i1c} > 1,$ nous calculons (3.1) pour $c = 1,\ldots,m_1$ et établissons que $\tilde{p}_{i1c} = f_{1c} p_{i1c} \Big/ \sum_{c=1}^{m_1} f_{1c} p_{i1c}$. Dans le cas où certaines valeurs $f_{1c} p_{i1c} < 0,$ nous posons les $\tilde{p}_{i1c} = 0.$ Des analystes pourraient envisager de réduire V_{1c} dans ce cas.

Nous passons maintenant à l'imputation de $X_2 \in \mathcal{A}$ pour les unités non répondantes. Nous adoptons une stratégie semblable : dans chaque ensemble de données complet, nous commençons par établir chaque distribution fonctionnelle d'unité non répondante pour l'imputation de x^*_{i2} compte tenu de x^*_{i1} et de $(\mathbf{z}_i, w_i),$ puis nous ajustons les probabilités pour que $(\hat{T}^*_{21},\ldots,\hat{T}^*_{2m_2})$ corresponde approximativement à une valeur échantillonnée de $(\hat{T}_{21},\ldots,\hat{T}_{2m_2})$. Nous présentons deux options pour l'imputation de X_2 : une basée sur l'application d'un ajustement multiplicateur aux probabilités fonctionnelles (section 3.2.1) et l'autre, sur la résolution d'un système d'équations faisant appel à $\mathcal{A}$ (section 3.2.2).

3.2.1 Méthode d'ajustement multiplicateur

Supposons que $p_{i2c} = \Pr(X_{i2} = c \mid x^*_{i1} = d, \mathbf{z}_i, w_i)$ est la probabilité fonctionnelle que $X_2 = c$ pour une unité $i,$ étant donné la valeur imputée $x^*_{i1} = d$ et $(\mathbf{z}_i, w_i)$. Nous les déterminons à partir d'une régression logistique (multinomiale) de X_2 sur X_1 et, possiblement, d'une certaine fonction de $\mathbf{Z}$ ou de $W,$ dont les coefficients sont estimés à partir de données complètes pour les unités pour lesquelles $U_i = 0$. Il convient de mentionner que si nous ne tenons pas compte de $\mathbf{Z}$ et de $W,$ $p_{i2c} = \Pr(X_2 = c \mid X_1 = d)$ pour $c = 1,\ldots,m_2$ et $d = 1,\ldots,m_1;$ c'est-à-dire que la probabilité conditionnelle est la même pour toutes les unités pour lesquelles $x^*_{i1} = d.$ Dans ce cas (que nous utilisons à la section 3.2.2), afin de simplifier la notation, nous laissons tomber l'indice i pour les unités individuelles et ajoutons un indice inférieur d pour la valeur de $X_1,$ ce qui donne $p_{i2c} = p_{2cd}$.

Pour $c = 1,\ldots,m_2 - 1,$ nous échantillonnons une valeur plausible de $\hat{T}_{2c}$ à partir d'une $N(T_{2c},V_{2c})$ et posons le total estimé pour le niveau final comme suit : $\hat{T}_{2m_2} = N - \sum_{c=1}^{m_2-1}\hat{T}_{2c}$. Comme dans (3.1), nous trouvons des constantes appropriées $f_{2c} > 0$ afin d'ajuster chaque p_{i2c}. Nous obtenons

$$f_{2c} \;=\; \frac{\hat{T}_{2c} - \sum_{i\in\mathcal{S}} w_i\, I(x_{i2}^* = c, U_i = 0)}{\sum_{i\in\mathcal{S}} p_{i2c} w_i\, I(U_i = 1)}. \tag{3.2}$$

Nous imputons chaque x_{i2}^* compte tenu de x_{i1}^* et de $\mathbf{z}_i$ (ou de w_i) pour les unités pour lesquelles $U_i = 1,$ en procédant à un tirage aléatoire à partir de la fonction de masse de probabilité ajustée $\{\tilde{p}_{i2c} : c = 1,\ldots,m_2\},$ où $\tilde{p}_{i2c} = f_{2c}p_{i2c}$ pour $c = 1,\ldots,m_2 - 1$ et $\tilde{p}_{i2m_2} = 1 - \sum_{c=1}^{m_2-1} f_{2c}p_{i2c}.$ Si $\sum_{c=1}^{m_2-1} f_{2c}p_{i2c} > 1,$ nous calculons (3.2) pour $c = 1,\ldots,m_2$ et posons $\tilde{p}_{i2c} = f_{2c}p_{i2c} \Big/ \sum_{c=1}^{m_2} f_{2c}p_{i2c}$. Quant à $X_1,$ si certaines valeurs $f_{2c}p_{i2c} < 0,$ nous posons ces $\tilde{p}_{i2c} = 0.$ Ce processus fait en sorte que les estimations de Horvitz et Thompson (1952) à partir des données complètes correspondent approximativement en espérance aux totaux $(\hat{T}_{21},\ldots,\hat{T}_{2m_2})$ sélectionnés, tout en intégrant aussi des liens entre $X_2,$ X_1 et $\mathbf{Z}$ (ou W) sous-entendus dans les probabilités fonctionnelles. Nous répétons le processus d'estimation et d'imputation de façon indépendante dans chaque ensemble de données complet. Nous appelons cette méthode « données manquantes avec marges auxiliaires — ajustement (DMMA-aj.) ».

Lorsque $\{p_{i2c}\}$ découle d'une régression logistique de X_2 sur X_1 et $\mathbf{Z}$ (ou W), utiliser $\{\tilde{p}_{i2c}\}$ équivaut à ajuster l'ordonnée à l'origine dans cette régression, tout en laissant tomber d'autres coefficients. Cela suppose un modèle dans lequel les logarithmes des rapports pour X_1 sont les mêmes pour les unités répondantes et non répondantes, ce qui concorde avec l'absence d'interaction entre U et X_1 dans (2.2).

3.2.2 Méthode des systèmes d'équations

La méthode de DMMA-aj. permet de révéler le lien entre X_2 et X_1 grâce à la régression logistique utilisée dans les probabilités fonctionnelles. Une autre approche consiste à imposer des contraintes aux imputations grâce à un système d'équations sous-entendues par les hypothèses qui sous-tendent (2.2) et (2.5), comme nous le décrivons à l'instant.

Comme dans la section 3.2.1, nous commençons par échantillonner des valeurs plausibles de $(\hat{T}_{21},\ldots,\hat{T}_{2m_2})$. Pour le moment, nous présumons que les probabilités fonctionnelles suivent $p_{i2c} = p_{2cd}$ pour toutes les unités. Dans tout ensemble de données complet, nous encodons l'hypothèse du logarithme des rapports constants dans (2.2) en tant qu'ensemble d'équations $(m_1 - 1)(m_2 - 1).$ Pour $c = 2,\ldots,m_2$ et $d = 2,\ldots,m_1,$ nous obtenons

$$\log\left[\frac{p_{2cd}}{p_{21d}}\right] - \log\left[\frac{p_{2c1}}{p_{211}}\right] \;=\; \log\left[\frac{\Pr(X_2 = c \mid X_1 = d, U = 0)}{\Pr(X_2 = 1 \mid X_1 = d, U = 0)}\right] - \log\left[\frac{\Pr(X_2 = c \mid X_1 = 1, U = 0)}{\Pr(X_2 = 1 \mid X_1 = 1, U = 0)}\right]. \tag{3.3}$$

Nous estimons les probabilités conditionnelles pour les unités répondantes dans (3.3) à l'aide d'estimateurs par le quotient pondérés par les poids d'enquête reposant sur les données complètes; par exemple

$$\Pr(X_2 = c \mid X_1 = d, U = 0) \;=\; \frac{\sum_{i\in S} w_i\, I(x_{i1}^* = d, x_{i2}^* = c, U_i = 0)}{\sum_{i\in S} w_i\, I(x_{i1}^* = d, U_i = 0)}. \tag{3.4}$$

Nous encodons les conditions sur $(\hat{T}^*_{21},\ldots,\hat{T}^*_{2m_2})$ dans (2.5) en tant qu'ensemble d'équations m_2. Nous obtenons

$$\sum_{d=1}^{m_1}\sum_{i\in\mathcal{S}} w_i p_{2cd}\, I(x^*_{i1}=d, x^*_{i2}=c, U_i=1) = \hat{T}_{2c} - \sum_{i\in\mathcal{S}} w_i\, I(x^*_{i2}=c, U_i=0). \tag{3.5}$$

Nous résolvons ces équations pour les $m_1(m_2-1)$ probabilités conditionnelles, $\{p_{2cd} : c=2,\ldots,m_2; d=1,\ldots m_1\}$. Nous imputons x^*_{i2} étant donné x^*_{i1} pour les unités non répondantes en utilisant les prélèvements de Bernoulli avec les probabilités $\{p_{2cd}\}$. Nous appelons cette méthode « données manquantes avec marges auxiliaires – système (DMMA-sys.) ».

Il est possible de modifier la méthode de DMMA-sys. pour utiliser des probabilités fonctionnelles $\{p_{i2c}\}$ qui varient pour les unités ayant la même valeur de x^*_{i1}. Pour ce faire, nous effectuons une imputation à l'aide de tirages bernoulliens avec des probabilités $\{\tilde{p}_{i2c}\}$, où $\tilde{p}_{i2c} = f_{2cd}\, p_{i2c}$ et

$$f_{2cd} = \frac{\sum_{i\in\mathcal{S}} w_i p_{2cd}}{\sum_{i\in\mathcal{S}} w_i p_{i2c}} \tag{3.6}$$

pour chaque $(X_2 = c, X_1 = d)$. L'ajustement dans (3.6) est motivé par l'égalité $\sum_{i\in\mathcal{S}} w_i p_{2c} = \sum_{i\in\mathcal{S}} w_i f_{2cd}\, p_{i2c}$.

3.2.3 Prise en compte de variables supplémentaires ayant des marges

Lorsque $k > 2$, pour la méthode de DMMA-aj., nous pouvons appliquer le processus utilisé dans (3.2) à chaque $X_j \in \mathcal{A}$. Si, par exemple, $X_3 \in \mathcal{A}$, l'analyste précise un ensemble de probabilités fonctionnelles, $p_{i3c} = \Pr(X_{i3} = c \mid x^*_{i1}, x^*_{i2}, \mathbf{z}_i, w_i)$ pour $c = 1,\ldots,m_3$ à l'aide d'une régression logistique de X_3 sur X_1, X_2, et $\mathbf{Z}$ (ou W). L'analyste échantillonne des valeurs de $(\hat{T}_{31},\ldots,\hat{T}_{3m_3})$ et calcule les valeurs de f_{3c} à l'aide d'expressions analogues à (3.2), en remplaçant les quantités basées sur X_2 par celles basées sur X_3. Pour la méthode de DMMA-sys., l'analyste peut résoudre un système d'équations semblables à celles de (3.3) et de (3.5). Cette méthode devient toutefois de plus en plus compliquée à mesure que des variables sont ajoutées à $\mathcal{A}$.

3.3 Variables sans marges pour les unités non répondantes

Après avoir créé L ensembles de données partiellement complets à partir des sections 3.1 et 3.2, nous imputons des valeurs x^*_{ij} pour les unités non répondantes de toutes les $X_j \notin \mathcal{A}$. Nous appliquons la procédure d'imputation par hot deck aléatoire mise au point par Yang et Reiter (2025). En paraphrasant leur présentation, pour toute unité $i \in \mathcal{S}$ pour laquelle $U_i = 0$, dans tout ensemble de données complet, nous supposons que $\mathbf{x}_i^{\mathcal{A}} = \{x_{ij} : X_j \in \mathcal{A}; j = 1,\ldots,p; U_i = 0\}$ correspond aux valeurs des variables (possiblement imputées) de l'enquête dans les données complètes pour les $X_j \in \mathcal{A}$. De même, pour toute unité $i' \in \mathcal{S}$ pour laquelle $U_{i'} = 1$, supposons que $\mathbf{x}_{i'}^{\mathcal{A}*} = \{x^*_{i'j} : X_j \in \mathcal{A}; j = 1,\ldots,p; U_i = 1\}$. Pour chaque unité i' pour laquelle $U_{i'} = 1$, dans chaque ensemble de données complet, nous construisons son ensemble de donneurs, $\mathcal{D}_{i'} = \{(x_{i1},\ldots,x_{ip}) : \mathbf{x}_i^{\mathcal{A}} = \mathbf{x}_{i'}^{\mathcal{A}*}, U_i = 0, i \in \mathcal{S}\}$. Nous échantillons ensuite aléatoirement un enregistrement i de $\mathcal{D}_{i'}$ et ajoutons son $\{(x_{i1},\ldots,x_{ik}) : X_j \notin \mathcal{A}\}$ à $\mathbf{x}_{i'}^{\mathcal{A}*}$ pour effectuer l'imputation totale $\mathbf{x}^*_{i'}$ pour l'unité i'. Nous

appliquons cette procédure à toutes les unités dans $\mathcal{S}$ pour lesquelles $U_{i'} = 1$; il en résulte un ensemble de données complet pour l'ensemble des n unités. Nous répétons ce processus dans chacun des L ensembles de données.

4. Études par simulation

Nous construisons une population $\mathcal{P}$ composée de $N = 3\,405\,809$ personnes à partir des fichiers de données à grande diffusion de 2023 de l'American Community Survey consultables sur IPUMS. Chaque personne dispose des six variables décrites dans le tableau 4.1. Nous comblons toute valeur manquante en utilisant une seule exécution sur mesure du paquet « mice ». Nous traitons le poids d'enquête dans le fichier comme étant une mesure de taille connue, $Z = (z_1, \ldots, z_N)$. Pour générer tout $\mathcal{S}$, nous échantillonnons les enregistrements tirés de $\mathcal{P}$ indépendamment avec des probabilités $\pi_i = 1/10z_i$, où $i = 1, \ldots, N$. Cet échantillonnage de Poisson donne environ $n = 7\,000$ personnes dans chacun des $\mathcal{S}$. Nous prenons 500 échantillons de Poisson indépendants.

Tableau 4.1
Description des variables de l'étude par simulation

Variable	Notation	Intervalle
Sexe	X_1	1 = Homme, 0 = Femme
État matrimonial	X_2	1 = Marié, 0 = Autre
Toute couverture d'assurance maladie	X_3	1 = Oui, 0 = Non
Fréquentation scolaire	X_4	1 = Oui, 0 = Non
Durée du trajet pour le travail	X_5	0 à 888
Score du revenu de travail	X_6	0 à 80

Pour chaque $\mathcal{S}$, nous générons une non-réponse totale en échantillonnant U_i pour toutes les $i \in \mathcal{S}$ à partir de

$$U \sim \text{Bernoulli}(\pi_U) \quad \text{logit}(\pi_U) = \omega_0 + \omega_1 X_1 + \omega_2 X_2. \tag{4.1}$$

Nous fixons $(\omega_0, \omega_1, \omega_2) =$ (-1,6; 0,5; 0,5) pour un taux de non-réponse totale d'environ 25 %. Nous faisons en sorte que toutes les données autres que Z et W soient complètement manquantes pour chaque unité où $U_i = 1$. Afin de générer une non-réponse partielle pour toute unité i où $U_i = 0$, nous échantillonnons R_{ij} pour $j = 1, \ldots, 6$ en utilisant

$$\begin{aligned} &R_j \mid X_1, X_2, X_3, X_4, X_5, X_6, Z, U = 0 \sim \text{Bernoulli}(\pi_{R_j}), \\ &\text{logit}(\pi_{R_j}) = \phi_{j0} + \sum_{k \neq j} \phi_{jk} X_k + \phi_Z Z. \end{aligned} \tag{4.2}$$

Nous fixons des paramètres dans (4.2) de sorte qu'environ 20 % de chaque variable de l'enquête est manquante pour les unités pour lesquelles $U_i = 0$. Plus précisément, $\phi_Z =$ 0,00001; $\phi_{j0} =$ -1,5 quand $j \leq 4$ et $\phi_{j0} =$ -1,6 quand $j = 5, 6$; et $\phi_{jk} =$ 0,1 quand $k \leq 4$ et $\phi_{jk} =$ 0,0001 quand $k = 5, 6$, pour chaque j. Nous posons comme manquantes toutes les x_{ij} pour lesquelles le R_{ij} échantillonné est égal à 1.

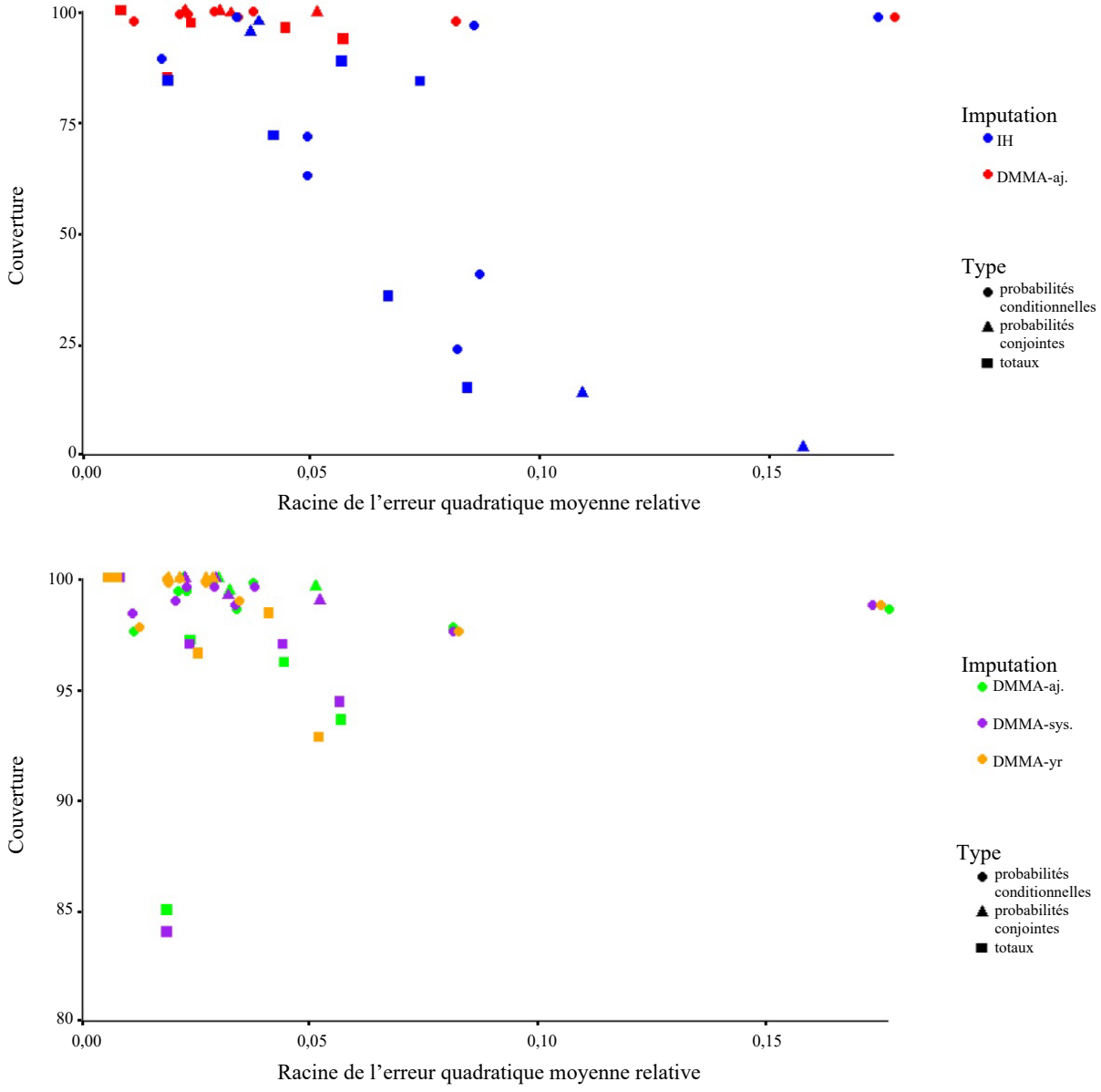
Nous supposons que $\mathcal{A}$ comprend $T_{X_1} = \sum_{i=1}^{N} x_{i1}$ et $T_{X_2} = \sum_{i=1}^{N} x_{i2}$. À l'aide de la section 3.1, nous générons $L = 10$ ensembles de données complets pour une non-réponse partielle en faisant appel à une mise en œuvre sur mesure du paquet « mice ». À l'aide de la section 3.2, nous mettons en œuvre les méthodes « DMMA-aj. » et « DMMA-sys. » avec des probabilités fonctionnelles dérivées de régressions logistiques de X_1 sur Z et de X_2 sur (X_1, Z). Pour les V_{X_1} et V_{X_2} utilisées pour échantillonner $\hat{T}_{X_1} \sim N(T_{X_1}, V_{X_1})$ et $\hat{T}_{X_2} \sim N(T_{X_2}, V_{X_2})$, nous avons recours au paquet « mice » en vue de créer un ensemble de données complet pour tous les $\mathcal{S}$ et de calculer les expressions pour les estimateurs de la variance non biaisés relatifs à $\hat{T}_{X_1}$ et à $\hat{T}_{X_2}$ dans le cadre d'un échantillonnage de Poisson. À l'aide de la section 3.3, nous créons des groupes de donneurs pour le hot deck en effectuant une correspondance sur (X_1, X_2).

Après avoir généré $L = 10$ ensembles de données complets, nous appliquons les méthodes inférentielles de Rubin (1987) pour les totaux marginaux de $(X_1, \ldots, X_6)$ et plusieurs probabilités conditionnelles et conjointes; consultez le site https://github.com/kevinxu47/MDAM (en anglais) pour obtenir la liste complète. Dans chaque ensemble de données complet, nous calculons les estimateurs ponctuels et les estimateurs de la variance de Horvitz et Thompson (1952) dans le cadre d'un échantillonnage de Poisson en utilisant les poids de sondage pour tous les enregistrements échantillonnés. À titre de comparaison, nous utilisons aussi la méthode présentée dans Yang et Reiter (2025) avec les poids de (2.4). Nous appelons cette méthode « données manquantes avec marges auxiliaires de Yang et Reiter (DMMA-yr) ». Enfin, nous employons une mise en œuvre sur mesure de « mice » pour imputer les éléments manquants, et échantillonner aléatoirement et avec remise des enregistrements entiers en tant qu'imputations pour la non-réponse totale. Il s'agit du « modèle reposant sur un mécanisme de non-réponse conditionnellement indépendante au niveau détaillé pour la non-réponse partielle et sur un hot deck pour la non-réponse totale (IH) ». Il ne repose pas sur $\mathcal{A}$.

La figure 4.1 résume les racines des erreurs quadratiques moyennes relatives (REQMr) des estimations ponctuelles et des taux de couverture empirique des intervalles de confiance à 95 % de l'imputation multiple relative aux méthodes pour les 500 échantillons. Les résultats sous forme totalisée sont présentés sur le site https://github.com/kevinxu47/MDAM (en anglais). La méthode de DMMA-aj. tend à produire des REQMr plus petites et des taux de couverture plus élevés que la méthode « IH ». Les taux de couverture des modèles de DMMA tendent à dépasser 95 % en raison de la surestimation des variances. Comme l'ont mentionné Yang et Reiter (2025), cela s'explique par le fait que les règles de combinaison de Rubin (1987) ne tiennent pas compte de l'utilisation des totaux connus dans l'imputation. Les méthodes « DMMA-aj. » et « DMMA-sys. » procurent des résultats semblables sur le plan qualitatif. Dans les deux méthodes, seul l'intervalle pour T_{X_3} est plus petit que le taux de couverture nominal (environ 85 %). Ce taux correspond approximativement au taux de couverture des intervalles de confiance fondés sur le plan qui ont été estimés avant l'introduction de toute donnée manquante (consultez les résultats totalisés), ce qui indique que la couverture inférieure à la valeur nominale est un trait de l'estimateur basé sur des données complètes, non une conséquence de la modélisation de DMMA. La méthode de DMMA-yr donne des résultats similaires à ceux obtenus avec les méthodes « DMMA-aj. » et « DMMA-sys. ». Comme l'indiquent de façon évidente les résultats totalisés, la méthode de DMMA-yr tend à avoir des variances plus faibles, ce qui est attribuable au remplacement de tous les poids des unités non répondantes par une constante.

Dans l'ensemble, les résultats indiquent que les nouveaux modèles de DMMA permettent aux analystes d'intégrer des totaux marginaux connus à une imputation multiple lorsqu'ils utilisent les poids de sondage de toutes les unités échantillonnées.

Figure 4.1 Simulation de racines des erreurs quadratiques moyennes relatives et de taux de couverture pour les totaux de population, les probabilités conditionnelles bidirectionnelles et les probabilités conjointes



Notes : - La partie du haut compare les modèles « DMMA-aj. » et « IH ». La partie du bas affiche les modèles « DMMA-aj. », « DMMA-sys. » et « DMMA-yr ».
- modèle de données manquantes avec marges auxiliaires – ajustement (DMMA-aj.); modèle de données manquantes avec marges auxiliaires – système (DMMA-sys.); modèle de données manquantes avec marges auxiliaires de Yang et Reiter (DMMA-yr). Imputation Hot-deck (IH), modèle reposant sur un mécanisme de non-réponse conditionnellement indépendante au niveau détaillé pour la non-réponse partielle et sur un hot deck pour la non-réponse totale.

Bibliographie

Akande, O., Madson, G., Hillygus, D.S. et Reiter, J.P. (2021). Leveraging auxiliary information on marginal distributions in nonignorable models for item and unit nonresponse. *Journal of the Royal Statistical Society Series A*, 184, 643-662.

Akande, O. et Reiter, J.P. (2022). Multiple imputations for nonignorable item nonresponse in complex surveys using auxiliary margins. Dans *Statistics in the Public Interest: In Memory of Stephen E. Fienberg*, 289-306. Cham: Springer International Publishing.

Chauvet, G., Deville, J.-C. et Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, 98, 459-471.

Chauvet, G. et Do Paco, W. (2018). Exact balanced random imputation for sample survey data. *Computational Statistics & Data Analysis*, 128, 1-16.

Hirano, K., Imbens, G.W., Ridder, G. et Rubin, D.B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69(6), 1645-1659.

Horvitz, D.G. et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Kim, J.K., Brick, J.M., Fuller, W.A. et Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society: Series B*, 68, 509-521.

Quartagno, M., Carpenter, J.R. et Goldstein, H. (2020). Multiple imputation with survey weights: A multilevel approach. *Journal of Survey Statistics and Methodology*, 8, 965-989.

Reiter, J.P., Raghunathan, T.E. et Kinney, S.K. (2006). L'importance de la modélisation du plan d'échantillonnage dans l'imputation multiple pour les données manquantes. *Techniques d'enquête*, 32(2), 161-168. Accessible à l'adresse : https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006002/article/9548-fra.pdf.

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

Sadinle, M. et Reiter, J.P. (2017). Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika*, 104, 207-220.

Sadinle, M. et Reiter, J.P. (2019). Sequentially additive nonignorable missing data modelling using auxiliary marginal information. *Biometrika*, 106, 889-911.

Tang, J., Hillygus, D.S. et Reiter, J.P. (2024). Using auxiliary marginal distributions in imputations for nonresponse while accounting for survey weights, with application to estimating voter turnout. *Journal of Survey Statistics and Methodology*, 12, 155-182.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press.

Yang, Y. et Reiter, J.P. (2025). Imputation de données manquantes non ignorables dans des enquêtes à l'aide de marges auxiliaires par imputations hot deck et séquentielle. *Techniques d'enquête*, 51(1), 273-301. Accessible à l'adresse : https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2025001/article/00004-fra.pdf.