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

Techniques d’enquête

Nouvelle méthode d’imputation hot deck double pour données manquantes bornées

par Yousung Park et Tae Yeon Kwon

Date de diffusion : le 30 juin 2020

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 STATCAN.infostats-infostats.STATCAN@canada.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

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

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 2020
Vol. 46, nº 1, pp. 129-149
Statistique Canada, nº 12-001-X au catalogue

Nouvelle méthode d'imputation hot deck double pour données manquantes bornées

Yousung Park et Tae Yeon Kwon[1]

Résumé

Dans les enquêtes, les bornes logiques entre variables ou entre vagues d'enquêtes compliquent l'imputation des valeurs manquantes. Nous proposons une nouvelle méthode d'imputation multiple par la régression pour traiter les non-réponses d'enquête avec bornes logiques bilatérales. La méthode d'imputation proposée satisfait automatiquement aux conditions de bornes sans procédure supplémentaire d'acceptation ou de rejet et utilise l'information sur les bornes pour dériver une valeur imputée et déterminer la pertinence de la valeur imputée. Les résultats de la simulation montrent que notre nouvelle méthode d'imputation surpasse les méthodes d'imputation actuelles pour les estimations de la moyenne et des quantiles, quels que soient les taux de valeurs manquantes, les distributions d'erreurs et les mécanismes de valeurs manquantes. Nous appliquons notre méthode pour imputer la variable du « nombre d'années de tabagisme » autodéclaré dans les dépistages médicaux successifs de la population coréenne.

Mots-clés : Hot deck; conditions de bornes bilatérales; imputation multiple; non-réponse partielle.

1 Introduction

La non-réponse d'enquête (ou la non-réponse partielle) survient dans de nombreux recensements ou enquêtes-échantillons, et plusieurs méthodes pour remplacer les éléments manquants ont été proposées. Dans une enquête, certaines variables manquantes sont bornées logiquement. Par exemple, dans la *National Health Interview Survey* (NHIS, Enquête nationale sur la santé réalisée par interviews) des États-Unis, certaines familles n'ont pas déclaré de revenu exact, mais des catégories de revenu, ce qui donne les bornes des valeurs exactes du revenu familial. Quand, au sein d'une famille, les revenus personnels sont déclarés pour certains membres de la famille seulement, la somme des revenus personnels déclarés donne la borne inférieure du revenu familial (Schenker, Raghunathan, Chiu, Makuc, Zhang et Cohen, 2006). Geraci et McLain (2018) ont traité plusieurs exemples de variables manquantes bornées dans les enquêtes, notamment les échelles psychométriques, les résultats cliniques et les notes scolaires.

Les vagues des enquêtes par panel et les ensembles de données par panel fournissent souvent les contraintes logiques des variables manquantes. Dans les données de dépistage médical périodique publiées par le service national de la santé de la Corée, la période de tabagisme manquante d'un fumeur à la vague en cours a pour borne inférieure la période de tabagisme déclarée à la vague précédente et pour borne supérieure son âge. Après avoir examiné les données de dépistage médical de la Corée de 2011 et de 2013, nous avons observé que les données de 2013 comportaient jusqu'à 73,5 % de valeurs manquantes pour les périodes de tabagisme quand nous traitons les périodes de tabagisme enfreignant les contraintes logiques comme étant des valeurs manquantes. En particulier, l'âge moyen des répondants à la question sur les périodes de tabagisme est de 9 ans inférieur à celui des non-répondants, ce qui signifie que le mécanisme

1. Yousung Park, Professeur, Department of Statistics, Korea University, 145 Anam-ro, Anam-dong, Seongbuk-gu, Séoul, Republic of Korea. Courriel : yspark@korea.ac.kr; Tae Yeon Kwon, auteur à contacter pour cet article, Professeur adjoint, Department of International Finance, Hankuk University of Foreign Studies, 81 Oedae-ro, Wangsan-ri, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do, Republic of Korea . Courriel : tykwon@hufs.ac.kr.

de valeurs manquantes de la période de tabagisme n'est pas une donnée manquante entièrement au hasard (MCAR pour *missing completely at random*) et que, par conséquent, l'imputation est nécessaire.

Geraci et McLain (2018) ont proposé une méthode d'imputation fondée sur les quantiles pour les variables manquantes unilatérales ou bilatérales où les valeurs supérieures et inférieures sont des constantes fixes, et ils ont montré que leur méthode présentait des avantages, surtout quand la taille de l'échantillon est modérément grande et que le modèle vrai est strictement non linéaire. La manière la plus courante d'adapter les bornes logiques à une méthode d'imputation multiple consiste à adopter la troncature ou une étape d'acceptation/rejet. Cependant, nos études par simulations montrent que cette étape ajoutée aux méthodes d'imputation multiple actuelles introduit un biais si les bornes bilatérales sont asymétriques.

Nous proposons une nouvelle méthode d'imputation multiple par la régression sans procédure d'acceptation/rejet ni troncature. Cette nouvelle méthode utilise les bornes bilatérales pour imputer les valeurs manquantes, respecte automatiquement les contraintes de bornes et inclut la méthode d'imputation présentée par Kwon et Park (2015) comme cas particulier. Nous l'avons nommé la méthode du hot deck double avec tirage de résidus proportionnés et appariement qui tient compte de bornes (TRP-AB), parce qu'on utilise deux étapes hot deck pour réduire le nombre de candidats donneurs et pour choisir un résidu proportionnel approprié qui est défini par le résidu habituel divisé par la distance entre une observation et sa borne inférieure ou supérieure. Ce résidu proportionné a été utilisé dans Kwon et Park (2015).

L'imputation hot deck qui remplace une valeur manquante par une observation « similaire » peut améliorer la performance de l'imputation par rapport aux méthodes d'imputation qui sont dérivées seulement à partir de scénarios assistés par un modèle. Andridge et Little (2010) ont montré que, en particulier, quand on utilise un modèle pour définir les appariements, l'imputation hot deck est moins vulnérable à la spécification erronée de modèle que les méthodes assistées par un modèle.

L'imputation multiple incorpore de l'incertitude due à l'imputation dans l'inférence statistique en remplaçant les valeurs manquantes plusieurs fois. La méthode de base donnée dans Rubin (1978) consiste à imputer la valeur manquante au moyen d'une valeur échantillonnée à partir de la distribution a posteriori normale. Cette méthode a été étendue à l'imputation des valeurs manquantes avec une borne logique (Raghunathan, Lepkowski, Van Hoewyk et Solenberger, 2001) par l'utilisation d'une distribution a posteriori normale tronquée (T-NORM). Rubin et Schenker (1986) et Rubin (2004) ont adopté la distribution empirique des résidus normalisés observés fondée sur un modèle de régression ajusté. Ils ont proposé une méthode d'imputation ajustée pour tenir compte de l'incertitude de la moyenne et de la variance (MV), qui impute la valeur manquante avec sa moyenne prédictive plus les résidus choisis aléatoirement à partir de leur distribution empirique.

À partir des idées élémentaires de l'imputation hot deck et de l'imputation multiple, la plupart des méthodes actuelles supposent la distribution des données (généralement normale) et emploient une distribution tronquée (généralement normale tronquée) pour respecter la borne logique de la valeur manquante (van Buuren et Groothuis-Oudshoorn, 2010; Honaker, King et Blackwell, 2012; Su, Gelman,

Hill, Yajima, 2011; Raghunathan et coll., 2001; Raghunathan, Solenberger et Van Hoewyk, 2002). La méthode d'appariement d'après la moyenne prédictive (AMP) impute une valeur manquante avec une observation sélectionnée aléatoirement ayant une moyenne prédictive ressemblant à celle de la valeur manquante (Little, 1988). Schenker et Taylor (1996) ont proposé la méthode de tirage de résidu local (TRL), qui remplace chaque valeur manquante par sa moyenne prédictive plus un résidu tiré aléatoirement dont la moyenne prédictive est proche de celle de la valeur manquante. Au lieu du résidu dans le TRL, Kwon et Park (2015) ont utilisé le résidu proportionné dont la distance entre la moyenne prédite et sa valeur de borne est proche de celle de la valeur manquante afin de respecter une borne unilatérale imposée aux variables d'intérêt.

Essentiellement, la méthode TRP-AB rejoint celle de Kwon et Park (2015). Cependant, elle ajoute une procédure d'appariement de plus pour tenir compte des bornes bilatérales et pour résoudre l'information de borne asymétrique. Cet appariement supplémentaire se fonde sur la borne la plus proche de la moyenne prédictive de chaque valeur manquante. Entre-temps, le TRP-AB impute la valeur manquante avec sa moyenne prédictive plus un résidu proportionné multiplié par la distance entre la moyenne prédite et la borne supérieure ou inférieure correspondante. Bien que le TRP-AB soit une méthode mixte, car il emploie un modèle de régression dans la première étape, puis une imputation hot deck double des valeurs manquantes dans la seconde étape, la méthode TRP-AB est nouvelle en ce sens qu'elle s'ajuste directement l'information de borne au lieu de tronquer la distribution désignée et qu'elle emploie l'information de borne pour déterminer la similitude entre les observations et les valeurs manquantes.

L'article se divise en cinq sections. Nous décrivons notre nouvelle méthode d'imputation et ses propriétés à la section 2. À la section 3, au moyen d'études par simulations, nous comparons notre méthode aux méthodes T-NORM et MV, AMP et TRL, avec une procédure supplémentaire de troncature pour respecter les contraintes de bornes et des méthodes de la série de TRP pour examiner l'effet de l'étape hot deck double de notre méthode, le TRP-AB. Dans la section 4, nous appliquons la méthode TRP-AB et les méthodes d'imputation existantes aux données de dépistage médical de la Corée de 2013 pour les valeurs manquantes des périodes de tabagisme. Finalement, on trouve une brève conclusion à la section 5.

2 Hot deck double avec tirage de résidus proportionnés et appariement qui tient compte de bornes

Supposons que les données sont composées d'un vecteur de variable explicative $\mathbf{X}_i$ entièrement observé et de la variable réponse Y_i pour $i = 1, \ldots, n,$ pour lequel certaines des valeurs Y_i sont manquantes (c'est-à-dire données manquantes partielles). Quelles que soient les données manquantes, on suppose que Y_i est individuellement bornée et que les valeurs des bornes sont données par

$$C_{i,L} \leq Y_i \leq C_{i,U} \tag{2.1}$$

où $C_{i,U}$ et $C_{i,L}$ sont les bornes supérieures et inférieures de Y_i, $i = 1, \ldots, n_0, n_0 + 1, \ldots, n$ et les premières valeurs n_0 Y_i sont observées et les valeurs restantes $(n - n_0)$ Y_i sont manquantes.

D'après Rubin (1987), nous générons les coefficients de régression β^* et la variance σ^{*2} à partir des distributions a posteriori données par

$$\sigma^{*2} \sim \hat{\sigma}^2_{\text{MCO}} (n_{\text{obs}} - q) \big/ \chi^2_{n_{\text{obs}}-1}, \ \beta^* \sim N\left(\hat{\beta}_{\text{MCO}}, \sigma^{*2} \left(X^T X\right)^{-1}\right) \tag{2.2}$$

où X est les covariables entièrement observées q et $\hat{\beta}_{\text{MCO}}$ et $\hat{\sigma}^2_{\text{MCO}}$ sont les estimations par les moindres carrés ordinaires (MCO) des coefficients de régression et de la variance, respectivement, à partir du modèle de régression adapté aux observations. Nous obtenons ensuite les moyennes prédictives notées $\hat{Y}_i^{\text{obs}}$ pour les valeurs observées de Y_i et $\hat{Y}_j^{\text{manq}}$ pour les valeurs manquantes de Y_j. Ensuite, chaque $\hat{Y}_i^{\text{obs}}$ ou $\hat{Y}_j^{\text{manq}}$ est situé dans l'un des trois intervalles suivants S^k où $k = -, 0, +$:

$$S^- = (-\infty, C_{iL}), \quad S^0 = (C_{iL}, C_{iU}), \quad S^+ = (C_{iU}, \infty).$$

Pour la valeur observée Y_i (c'est-à-dire $i = 1, \ldots, n_0$), nous définissons les résidus proportionnés supérieurs et inférieurs $\tilde{r}_{i,U}$ et $\tilde{r}_{i,L}$:

$$\tilde{r}_{i,U} = \frac{Y_i - \hat{Y}_i^{\text{obs}}}{C_{iU} - \hat{Y}_i^{\text{obs}}} \quad \text{et} \quad \tilde{r}_{i,L} = \frac{Y_i - \hat{Y}_i^{\text{obs}}}{C_{iL} - \hat{Y}_i^{\text{obs}}} \tag{2.3}$$

où nous supposons qu'il n'y a pas de valeur $\hat{Y}$ exactement égale à sa borne supérieure ou inférieure.

Les valeurs $\tilde{r}_{i,U}$ et $\tilde{r}_{i,L}$ de l'équation (2.3) sont ensuite divisées en trois ensembles en fonction de l'intervalle S^k auquel appartient $\hat{Y}^{\text{obs}}$. Pour $k = -, 0, +$,

$$R_U^k = \left\{\tilde{r}_{i,U}; \hat{Y}_i^{\text{obs}} \in S^k\right\} \quad \text{et} \quad R_L^k = \left\{\tilde{r}_{i,L}; \hat{Y}_i^{\text{obs}} \in S^k\right\}.$$

Enfin, nous imputons la valeur manquante Y_j pour $j = n_0 + 1, \ldots, n$ au moyen de

$$Y^*_{j,U} = \hat{Y}_j^{\text{manq}} + \tilde{r}^*_{j,U}\left(C_{j,U} - \hat{Y}_j^{\text{manq}}\right) \quad \text{ou} \quad Y^*_{j,L} = \hat{Y}_j^{\text{manq}} + \tilde{r}^*_{j,L}\left(C_{j,L} - \hat{Y}_j^{\text{manq}}\right). \tag{2.4}$$

Afin de sélectionner $\tilde{r}^*_{j,U}$ ou $\tilde{r}^*_{j,L}$ dans (2.4), nous utilisons maintenant une méthode hot deck qui considère des valeurs similaires comme leurs variables candidates (c'est-à-dire comme les donneurs possibles). L'imputation hot deck est une méthode de traitement des données manquantes dans laquelle chaque valeur manquante est remplacée par une réponse observée sélectionnée aléatoirement à partir d'un donneur contenant des unités similaires (Andridge et Little, 2010). Nous utilisons le scénario d'imputation hot deck ci-dessous.

1. [**Premier hot deck**] Si $\hat{Y}_j^{\text{manq}} \in S^k$ pour $k = -, 0, +$ et $\hat{Y}_j^{\text{manq}}$ est plus proche de C_{jU} que C_{jL}, alors nous sélectionnons l'ensemble de résidu proportionné supérieur correspondant R_U^k comme ensemble de donneurs possibles pour l'échantillonnage $\tilde{r}^*_{jU}$. De même, si $\hat{Y}_j^{\text{manq}}$ est plus près de C_{jL}, nous sélectionnons R_L^k pour l'échantillonnage de $\tilde{r}^*_{jL}$.
2. [**Deuxième hot deck**] Nous construisons les donneurs possibles à partir de R_U^k ou R_L^k sélectionnés dans le premier hot deck. Les donneurs possibles pour l'échantillonnage de $\tilde{r}^*_{jU}$

sont des valeurs $\tilde{r}_{i,U}$ pour lesquelles $C_{iU} - \hat{Y}_i^{\text{obs}}$ est proche de $C_{jU} - \hat{Y}_j^{\text{manq}}$ pour R_U^k. Dans le même ordre d'idées, les donneurs possibles des valeurs $\tilde{r}_{i,L}$ pour lesquelles $C_{iL} - \hat{Y}_i^{\text{obs}}$ proche de $C_{jL} - \hat{Y}_j^{\text{manq}}$ pour R_L^k.

3. [**Imputation**] Ensuite on échantillonne aléatoirement $\tilde{r}_{i,U}$ ou $\tilde{r}_{i,L}$ à partir des donneurs possibles correspondants pour imputer la valeur manquante Y_j avec $Y_{j,U}^*$ ou $Y_{j,L}^*$, respectivement, et les $\tilde{r}_{i,U}$ et $\tilde{r}_{i,L}$ sélectionnés pour $i = 1, \ldots, n_0$ sont notés par $\tilde{r}_{j,U}^*$ ou $\tilde{r}_{j,L}^*$ pour $j = n_0 + 1, \ldots, n$. Ici, les cas avec $Y_{j,U}^* \leq C_{j,L}$ et/ou $Y_{j,L}^* \geq C_{j,U}$ sont exclus de l'ensemble de donneurs possibles. Cela est rare, mais possible.

Théorème 1 *Les valeurs $Y_{j,U}^*$ et $Y_{j,L}^*$ satisfont toujours leurs conditions de borne.*

$$C_{j,L} \leq Y_{j,U}^* \leq C_{j,U} \text{ et } C_{j,L} \leq Y_{j,L}^* \leq C_{j,U}.$$

La démonstration se trouve en annexe.

Selon le théorème 1, les conditions de borne de Y_j pour $j = n_0 + 1, \ldots, n$ sont toujours satisfaites, car le TRP-AB impute la valeur manquante Y_j avec $Y_{j,U}^*$ ou $Y_{j,L}^*$. Nous pouvons supposer qu'il n'y a qu'une valeur de borne supérieure telle que $C_{i,L} = -\infty$ pour $i = 1, \ldots, n$. Alors, le premier hot deck n'est pas nécessaire, car R_U^k est automatiquement sélectionné et $Y_{j,U}^* \geq C_{j,L} = -\infty$.

Corollaire 1.1 *La méthode de TRP-AB est réduite à la méthode d'appariement d'information de borne de Kwon et Park (2015) s'il n'y a qu'une borne supérieure ou inférieure.*

Pour examiner les procédures de hot deck double utilisées dans le TRP-AB, nous examinons trois variantes du TRP-AB. La première variante est une *méthode de tirage de résidu proportionné* (TRP) consistant à retirer les deux étapes hot decks du TRP-AB et la deuxième élimine la première procédure de hot deck notée STRP. Ainsi, dans le TRP, nous échantillonnons au hasard à partir de tous les éléments se trouvant dans R_U^k et R_L^k. Dans STRP, l'ensemble de donneurs possibles est fondé sur la distance minimale entre l'une des bornes et la moyenne prédictive. Parmi les donneurs possibles pour $\tilde{r}_{j,U}^*$ et $\tilde{r}_{j,L}^*$, nous sélectionnons et construisons des donneurs finaux en nous fondant uniquement sur l'ordre de distance, sans distinction entre les bornes supérieures et inférieures. La troisième variante, notée STRP_2, élimine aussi la première étape hot deck comme dans STRP et modifie en outre la méthode d'appariement dans le deuxième hot deck. Le donneur possible dans STRP_2 consiste en $\tilde{r}_{i,U}$ et $\tilde{r}_{i,L}$ dont la moyenne prédite $\hat{Y}_i^{\text{obs}}$ est proche de $\hat{Y}_j^{\text{manq}}$.

3 Simulation

Nous utilisons les abréviations suivantes pour les méthodes d'imputation examinées dans les sections 1 et 2 : OBS (cas disponibles), T-NORM (imputation normale tronquée dans Rubin (1978); Raghunathan

et coll. (2001)), MV (méthode corrigée pour l'incertitude de la moyenne et de la variance dans Rubin et Schenker (1986)), AMP (appariement de la moyenne prédictive dans Little (1988)), TRL (méthode de tirage de résidu local dans Schenker et coll. (2006)) et trois variantes du TRP-AB notées TRP, STRP et STRP_2. Nous comparons ces huit méthodes d'imputation à notre méthode de TRP-AB où une procédure de troncature est ajoutée dans MV, AMP et TRL pour tenir compte des contraintes de bornes; la troncature est notée T-MV, T-AMP et T-TRL, respectivement.

Nous considérons une taille d'échantillon de 1 000 et des taux de valeurs manquantes de 20 % ou 50 % dans le modèle linéaire suivant :

$$Y_i = X_i + \varepsilon_i, \quad \text{où} \quad C_{iL} \le Y_i \le C_{iU} \quad \text{pour} \quad i = 1, \ldots, n, \tag{3.1}$$

et les X_i sont générés indépendamment de $N(2, 2)$ et de variables i.i.d. Les valeurs ε_i sont simulées à partir de $N(0, \sigma_Y)$ ou de la distribution-t avec un degré de liberté t_{df}. Les valeurs de bornes $C_{i,U}$ et $C_{i,L}$ sont générées avec $Y_i + |Z_{i,U}|$ et $Y_i - |Z_{i,L}|$ où $Z_{i,U} \sim N(0, \sigma_U)$ et $Z_{i,L} \sim N(0, \sigma_L)$, respectivement. Nous établissons $\text{Cor}(X, Y)$ à 0,7 ou 0,9 en ajustant σ_Y (ou t_{df}), et nous établissons $\text{Cor}(Y, C_U)$ et $\text{Cor}(Y, C_L)$ entre 0 et 0,9 en ajustant σ_U et σ_L. La corrélation $\text{Cor}(Y, C_U)$ $(\text{Cor}(Y, C_L))$ notée ρ_{y,c_u} (ρ_{y,c_l}) indique que la borne supérieure C_U a une information plus robuste pour Y que la borne inférieure C_L quand ρ_{y,c_u} est supérieur à ρ_{y,c_L} en valeur absolue.

Deux types de mécanismes de valeurs manquantes sont examinés. Premièrement, 20 % des valeurs Y sont choisies aléatoirement et traitées comme manquantes pour refléter le mécanisme de valeurs manquantes « entièrement au hasard (MCAR) ». Deuxièmement, nous établissons 80 % des valeurs Y_i à manquantes quand la valeur correspondante X_i est supérieure à sa moyenne et 20 % des valeurs Y_i à manquantes quand la valeur correspondante X_i est inférieure à sa moyenne. Cela résulte en approximativement 50 % de valeurs manquantes pour Y_i globalement et reflète les « données manquantes au hasard (MAR) ». Notons qu'aucune imputation n'est nécessaire pour les valeurs manquantes en cas de valeurs manquantes entièrement au hasard (MCAR), alors que l'imputation des valeurs manquantes est requise en cas de données manquantes au hasard (MAR) (Scheffer, 2002).

Nous répétons chaque scénario de simulation 1 000 fois avec le nombre d'imputations M égal à 5 et le nombre de donneurs possibles dans le bassin de sélection pour l'imputation m_d égal à 6. Une taille de donneur possible m_d peut être inférieure à 6 quand l'échantillon ne suffit pas pour composer un donneur, mais cela ne se produit pas quand la taille de l'échantillon est de 1 000. Nous choisissons les nombres fixes couramment utilisés $M = 5$ et $m_d = 6$ (Geraci et McLain, 2018; Schafer, Ezzati-Rice, Johnson, Khare, Little et Rubin, 1996; Schenker et Taylor, 1996), car on sait que ce type de configuration n'affecte pas significativement les performances des méthodes d'imputation comme cela a été démontré dans Schafer (1999) et Schenker et Taylor (1996).

Les méthodes d'imputation sont comparées sur le plan de l'exactitude et de l'efficacité des estimations pour les quantités de population : moyenne (μ) et les 5e, 25e, 50e, 75e et 95e centiles. L'inférence statistique après l'imputation multiple est effectuée, d'après Rubin (2004) et Schafer et coll. (1996). Nous

employons l'erreur absolue moyenne (EAM), la racine de l'erreur quadratique moyenne (REQM), un taux de couverture d'intervalle de confiance de 95 % (TC) et une largeur moyenne d'intervalle de confiance de 95 % (LMIC) comme critères d'évaluation pour mesurer l'exactitude et l'efficacité de l'estimation (Yucel et Demirtas, 2010; Yucel, He et Zaslavsky, 2008; Gelman, Van Mechelen, Verbeke, Heitjan et Meulders, 2005).

3.1 Résultats de la simulation selon un scénario de données manquantes entièrement au hasard (MCAR)

La figure 3.1 montre la distribution de $\hat{\mu} - \mu$ (biais) dans 1 000 ensembles de données simulées, avec 20 % de valeurs manquantes entièrement au hasard (MCAR) sous $(\rho_{y,c_l}, \rho_{y,c_u}) = (0{,}8; 0)$ et $\rho_{x,y} = 0{,}7$. Étant donné qu'aucune imputation n'est nécessaire pour les valeurs manquantes selon un scénario de données manquantes entièrement au hasard dans l'estimation de la moyenne et de la variance de Y, OBS est sans biais pour la moyenne de Y, comme on pouvait s'y attendre. La figure 3.1 montre toutefois que toutes les méthodes d'imputation, sauf TRP-AB, révèlent un problème de surestimation. Notons que la borne inférieure a une information robuste pour Y ($\rho_{y,c_l} = 0{,}8$) mais que la borne supérieure n'a pas d'information ($\rho_{y,c_u} = 0$). À part OBS et TRP-AB, cette information de borne asymétrique pousse les valeurs imputées vers le haut dans les autres méthodes d'imputation. Pour connaître l'effet de l'information de borne asymétrique sur l'exactitude de l'imputation, différentes valeurs de $(\rho_{y,c_l}, \rho_{y,c_u})$ sont prises en compte dans le tableau 3.1.

Quand les bornes supérieures et inférieures fournissent de l'information de borne pour Y de façon symétrique (c'est-à-dire $(\rho_{y,c_l}, \rho_{y,c_u}) = (0{,}9; 0{,}9)$), toutes les méthodes d'imputation sont comparables et font concurrence à OBS. Cependant, en présence d'information de borne asymétrique ($(\rho_{y,c_l}, \rho_{y,c_u}) = (0{,}8; 0)$ ou $(\rho_{y,c_l}, \rho_{y,c_u}) = (0{,}5; 0{,}8)$), l'exactitude de l'estimation des T-NORM, T-MV, T-AMP et T-TRL existantes est nettement plus mauvaise que celle d'OBS et de TRP-AB. En particulier, le taux de couverture de 95 % de l'IC (TC) diminue considérablement à mesure que le degré d'asymétrie augmente. En revanche, les taux de couverture de la série TRP (c'est-à-dire TRP, STRP, TRP-AB) résistent à une telle asymétrie, ce qui indique que le tirage de résidu proportionné résiste à l'information de borne asymétrique. Dans la série TRP, le TRP-AB surpasse TRP et STRP et est même meilleur qu'OBS pour ce qui est de l'EAM et de la REQM.

Notons qu'à part dans OBS et TRP-AB, les valeurs imputées par toutes les autres méthodes d'imputation font que la distribution de Y tend vers la borne ayant une information de borne plus faible. Plus précisément, toutes les méthodes d'imputation, sauf OBS et TRP-AB, ont tendance à surestimer la moyenne véritable de Y ($E(Y) = 2$) pour $(\rho_{y,c_l}, \rho_{y,c_u}) = (0{,}8; 0)$ car $\rho_{y,c_u} < \rho_{y,c_l}$, mais aussi à sous-estimer la moyenne véritable de $(\rho_{y,c_l}, \rho_{y,c_u}) = (0{,}5; 0{,}8)$ car $\rho_{y,c_u} > \rho_{y,c_l}$. Cette dépendance s'observe également avec le mécanisme de valeurs manquantes au hasard (MAR), comme cela est montré dans la prochaine section.

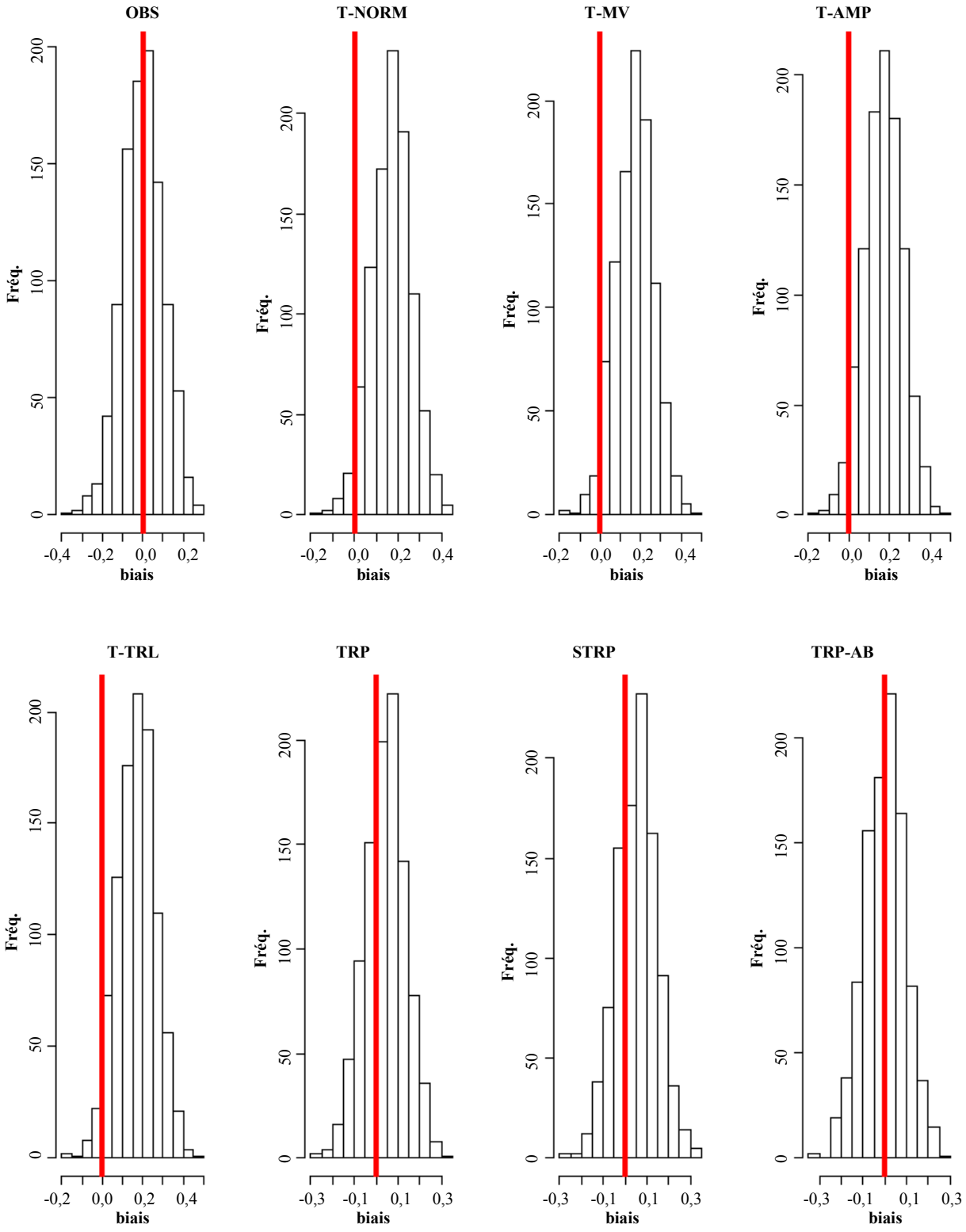

Figure 3.1 Distribution du biais, $\hat{\mu} - \mu$ dans l'estimation de la moyenne avec 20 % de valeurs manquantes entièrement au hasard avec erreur normale et $(\rho_{y,c_l}, \rho_{y,c_u}) = (0{,}8;\ 0)$ et $\rho_{x,y} = 0{,}7$.

Tableau 3.1
Résultats de simulation de l'estimation de la moyenne $(\mu = 2)$ avec 20 % de valeurs manquantes entièrement au hasard avec erreur normale

$\rho_{x,y}$	$(\rho_{y,c_l}, \rho_{y,c_u})$		OBS	T-NORM	T-MV	T-AMP	T-TRL	TRP	T-TRP	STRP	TRP-AB
0,9	(0,9; 0,9)	$\hat{\mu}$	2,003	2,003	2,003	2,003	2,003	2,003	2,003	2,003	2,003
		EAM	0,064	0,056	0,057	0,057	0,057	0,057	0,057	0,057	0,057
		REQM	0,081	0,071	0,071	0,071	0,072	0,071	0,071	0,071	0,071
		TC (%)	94,9	95,8	95,3	95,5	94,8	95,6	95,2	95,5	95,2
		LMIC	0,310	0,280	0,280	0,279	0,279	0,283	0,279	0,279	0,278
0,7	(0,8; 0)	$\hat{\mu}$	2,000	2,171	2,171	2,171	2,170	2,043	2,044	2,055	2,000
		EAM	0,080	0,174	0,174	0,174	0,173	0,083	0,084	0,088	0,075
		REQM	0,101	0,194	0,195	0,195	0,194	0,103	0,103	0,109	0,094
		TC (%)	94,9	54,1	54,3	52,1	53,7	92,3	91,4	89,8	94,2
		LMIC	0,393	0,367	0,366	0,362	0,363	0,362	0,358	0,358	0,359
0,7	(0,5; 0,8)	$\hat{\mu}$	2,000	1,906	1,906	1,906	1,907	1,927	1,973	1,979	1,983
		EAM	0,080	0,108	0,109	0,109	0,109	0,096	0,076	0,075	0,074
		REQM	0,102	0,131	0,132	0,132	0,132	0,118	0,096	0,095	0,094
		TC (%)	94,2	83,0	83,7	83,0	82,6	88,7	93,3	93,7	94,0
		LMIC	0,393	0,362	0,361	0,359	0,359	0,379	0,358	0,356	0,355

3.2 Résultats de simulation selon un scénario de données manquantes au hasard (MAR)

Le tableau 3.2 résume les résultats avec 50 % de valeurs manquantes au hasard (MAR) en cas de distribution $t\,(t_{df} = 3)$ et normale des erreurs. Comme on pouvait s'y attendre, OBS, qui utilise uniquement des valeurs observées dans l'estimation donne des résultats bien moins bons que toutes les méthodes d'imputation pour ce qui est de l'estimation de la moyenne véritable $\mu = 2$. Les résultats liés à l'information de borne asymétrique ressemblent aux résultats de simulation selon un scénario de données manquantes entièrement au hasard. L'exactitude et l'efficacité de T-NORM, T-MV, T-AMP et T-TRL sont nettement plus mauvaises que celles de la série TRP quand l'information de borne est asymétrique. Cela montre l'effet des deux étapes hot decks et du tirage de résidu proportionné sur l'exactitude et l'efficacité de l'estimation de la moyenne. Sauf pour l'information de borne symétrique selon les distributions t et normale, les taux de couverture de T-NORM, T-MV, T-AMP et T-TRL sont moins de 50 % plus petits que le taux visé de 95 %. Toutes les méthodes d'imputation, sauf TRP-AB, produisent la distribution empirique de Y biaisée à la borne avec une information de borne plus faible dans les distributions normale et t des erreurs. Cela signifie que seul TRP-AB résiste à l'information de borne asymétrique et aux distributions des erreurs, quel que soit le mécanisme de valeurs manquantes. Ainsi, TRP-AB surpasse les autres méthodes d'imputation dans tous les scénarios de simulation.

Afin d'examiner l'effet des procédures hot deck utilisées dans la méthode TRP-AB, nous l'avons comparée à STRP. On peut examiner l'effet de la première procédure hot deck en comparant TRP-AB et STRP, étant donné que cette première étape est supprimée dans STRP. On constate que TRP-AB donne toujours de meilleurs résultats que STRP, quelle que soit la mesure d'évaluation, tant que l'information de borne est asymétrique. Ainsi, plus l'information de borne est asymétrique d'un côté, plus le TC de STRP

est mauvais, si on le compare à celui de TRP-AB. On comprend que la première étape de hot deck joue un rôle important dans la résistance à l'asymétrie de l'information de borne.

On peut ensuite vérifier le rôle de la deuxième étape de hot deck en comparant STRP à TRP, qui n'utilise aucune des procédures hot deck. La méthode STRP est meilleure que TRP quand l'information de borne est moyennement asymétrique dans l'erreur normale ou l'erreur distribuée t, ce qui implique que la deuxième étape de hot deck fonctionne pour une distribution dont la queue est plus lourde que dans la distribution normale. Cette comparaison est examinée plus en détail dans l'évaluation suivante de l'estimation des centiles.

Comme nous l'avons décrit auparavant, STRP_2 est identique à STRP si ce n'est la méthode d'appariement aux fins de construction de donneurs possibles. Le donneur possible dans STRP_2 consiste en $\tilde{r}_{i,U}$ et $\tilde{r}_{i,L}$ dont la moyenne prédite $\hat{Y}_i^{\text{obs}}$ est proche de $\hat{Y}_j^{\text{manq}}$. Ainsi, en comparant STRP et STRP_2, nous examinons l'effet de l'appariement d'information de borne utilisé dans TRP-AB. Le tableau 3.2 montre que la méthode STRP est plus performante que STRP_2 quelles que soient l'information de borne et les distributions des erreurs, ce qui signifie que l'appariement d'information de borne fonctionne mieux que l'appariement d'après la moyenne habituel pour l'imputation de données manquantes bornées.

Tableau 3.2
Résultats de simulation de l'estimation de la moyenne $(\mu = 2)$ selon un scénario de 50 % de données manquantes au hasard

$(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u})$		OBS	T-NORM	T-MV	T-AMP	T-TRL	TRP	STRP	STRP_2	TRP-AB
erreur distribuée normale										
(0,9; 0,9; 0,9)	$\hat{\mu}$	1,045	2,002	2,002	1,995	2,002	2,001	2,001	2,000	2,001
	EAM	0,955	0,057	0,059	0,058	0,059	0,059	0,059	0,060	0,060
	REQM	0,958	0,072	0,075	0,074	0,074	0,074	0,075	0,076	0,075
	TC (%)	0,0	95,3	94,0	94,5	94,1	94,6	93,8	93,2	93,7
	LMIC	0,355	0,291	0,287	0,282	0,283	0,294	0,285	0,290	0,285
(0,7; 0,8; 0)	$\hat{\mu}$	1,043	2,426	2,425	2,417	2,424	2,097	2,103	2,098	1,993
	EAM	0,957	0,426	0,425	0,417	0,424	0,120	0,123	0,124	0,088
	REQM	0,963	0,44	0,442	0,434	0,441	0,148	0,150	0,152	0,109
	TC (%)	0,0	4,0	5,2	3,6	3,6	80,5	77,7	78,5	92,0
	LMIC	0,467	0,454	0,428	0,393	0,395	0,398	0,379	0,382	0,380
(0,7; 0,5; 0,8)	$\hat{\mu}$	1,045	1,77	1,771	1,761	1,771	1,822	1,952	1,828	1,961
	EAM	0,955	0,231	0,230	0,239	0,229	0,182	0,091	0,177	0,087
	REQM	0,961	0,249	0,251	0,259	0,249	0,205	0,114	0,203	0,110
	TC (%)	0,0	36,4	36,5	27,5	31,7	62,3	90,0	62,7	90,3
	LMIC	0,467	0,396	0,379	0,361	0,362	0,420	0,375	0,408	0,375
(0,55; 0,5; 0,5)	$\hat{\mu}$	1,041	1,992	1,990	1,983	1,991	1,989	1,991	1,990	1,991
	EAM	0,959	0,110	0,124	0,118	0,118	0,123	0,123	0,131	0,124
	REQM	0,971	0,138	0,155	0,148	0,149	0,155	0,155	0,165	0,156
	TC (%)	0,0	95,6	89,9	88,7	89,0	93,1	89,4	89,3	89,4
	LMIC	0,611	0,557	0,520	0,486	0,488	0,584	0,508	0,555	0,513
(0,55; 0,5; 0,8)	$\hat{\mu}$	1,042	1,613	1,613	1,604	1,613	1,754	1,902	1,761	1,906
	EAM	0,958	0,387	0,387	0,396	0,387	0,249	0,128	0,243	0,126
	REQM	0,969	0,404	0,406	0,414	0,406	0,276	0,158	0,271	0,156
	TC (%)	0,0	11,3	11,2	8,1	9,4	56,1	85,8	53,4	86,5
	LMIC	0,610	0,495	0,480	0,460	0,461	0,514	0,467	0,504	0,465

Tableau 3.2 (suite)
Résultats de simulation de l'estimation de la moyenne $(\mu = 2)$ selon un scénario de 50 % de données manquantes au hasard

$(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u})$		OBS	T-NORM	T-MV	T-AMP	T-TRL	TRP	STRP	$STRP_2$	TRP-AB
erreur distribuée t(3)										
(0,77; 0,9; 0,9)	$\hat{\mu}$	1,049	2,005	2,005	2,000	2,005	2,004	2,005	2,004	2,005
	EAM	0,951	0,067	0,069	0,067	0,067	0,069	0,069	0,069	0,07
	REQM	0,956	0,084	0,087	0,084	0,084	0,087	0,087	0,086	0,087
	TC (%)	0,0	96,2	95,2	95,5	96,0	95,5	95,1	95,5	95,4
	LMIC	0,428	0,341	0,335	0,328	0,329	0,341	0,333	0,336	0,334
(0,77; 0,9; 0)	$\hat{\mu}$	1,042	2,401	2,357	2,354	2,356	2,128	2,091	2,120	2,005
	EAM	0,958	0,401	0,357	0,354	0,356	0,138	0,108	0,131	0,076
	REQM	0,964	0,421	0,375	0,374	0,375	0,165	0,133	0,156	0,097
	TC (%)	0,0	2,7	6,5	4,9	5,2	71,3	80,1	72,3	93,2
	LMIC	0,429	0,407	0,391	0,37	0,37	0,36	0,338	0,348	0,342
(0,77; 0,5; 0,9)	$\hat{\mu}$	1,045	1,776	1,818	1,809	1,817	1,866	1,955	1,872	1,963
	EAM	0,955	0,224	0,185	0,192	0,185	0,142	0,086	0,137	0,083
	REQM	0,961	0,244	0,207	0,212	0,205	0,165	0,107	0,161	0,104
	TC (%)	0,0	31,7	44,0	37,1	42,1	67,7	87,6	67,3	88,7
	LMIC	0,431	0,355	0,342	0,326	0,329	0,359	0,338	0,351	0,338

Nous examinons également l'estimation des centiles en évaluant la performance de chaque méthode d'imputation dans l'estimation de la probabilité que Y soit plus grand que des quantiles de 5 %, 25 %, 50 %, 75 %, 95 %. Notons le quantile p^{e} par $y^{-1}(p)$ satisfaisant $P(Y > y^{-1}(p)) = 1 - p$. On choisit les cinq centiles pour savoir dans quelle mesure la distribution vraie de Y et la distribution estimée de Y diffèrent selon les différentes méthodes d'imputation. Le tableau 3.3 présente les résultats de l'estimation de centile selon un scénario de 50 % de données manquantes au hasard avec erreur normale quand $(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}) = (0{,}7;\ 0{,}5;\ 0{,}8)$ et (0,7; 0,8; 0). On voit dans la première ligne de chaque tableau que les méthodes existantes produisent visiblement des distributions asymétriques à droite quand $(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}) = (0{,}7;\ 0{,}5;\ 0{,}8)$ car $\rho_{y,c_u} > \rho_{y,c_l}$, alors qu'elles produisent des distributions asymétriques à gauche quand $(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}) = (0{,}7;\ 0{,}8;\ 0)$ car $\rho_{y,c_u} < \rho_{y,c_l}$.

Tableau 3.3
Résultats de simulation du centile $(P_k$, où $P_k = P(Y \geq y^{-1}(k))$ et $y^{-1}(p)$ satisfait l'estimation $P(Y > y^{-1}(p)) = 1 - p)$ selon un scénario de 50 % de données manquantes au hasard avec erreur normale

Critère	Paramètre	OBS	T-NORM	T-MV	T-AMP	T-TRL	TRP	STRP	$STRP_2$	TRP-AB
					$(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}) = (0{,}7;\ 0{,}5;\ 0{,}8)$					
moyenne	$P_{0,05}$	0,922	0,947	0,947	0,947	0,947	0,935	0,950	0,936	0,951
	$P_{0,25}$	0,637	0,732	0,732	0,731	0,732	0,733	0,750	0,734	0,752
	$P_{0,50}$	0,350	0,467	0,467	0,465	0,466	0,494	0,497	0,495	0,500
	$P_{0,75}$	0,139	0,216	0,216	0,216	0,217	0,232	0,240	0,233	0,241
	$P_{0,95}$	0,022	0,037	0,037	0,036	0,037	0,045	0,043	0,044	0,043

Tableau 3.3 (suite)
Résultats de simulation du centile (P_k, où $P_k = P(Y \geq y^{-1}(k))$ et $y^{-1}(p)$ satisfait l'estimation $P(Y > y^{-1}(p)) = 1 - p$) selon un scénario de 50 % de données manquantes au hasard avec erreur normale

Critère	Paramètre	OBS	T-NORM	T-MV	T-AMP	T-TRL	TRP	STRP	$STRP_2$	TRP-AB
					$(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}) = (0{,}7; 0{,}5; 0{,}8)$					
EAM	$P_{0,05}$	0,028	0,006	0,006	0,007	0,007	0,015	0,005	0,015	0,005
	$P_{0,25}$	0,113	0,019	0,019	0,021	0,021	0,018	0,011	0,018	0,011
	$P_{0,50}$	0,150	0,034	0,033	0,036	0,035	0,014	0,013	0,015	0,013
	$P_{0,75}$	0,111	0,034	0,034	0,035	0,034	0,020	0,015	0,021	0,014
	$P_{0,95}$	0,028	0,013	0,013	0,015	0,014	0,007	0,008	0,008	0,009
TC (%)	$P_{0,05}$	21,6	95,1	95,0	90,4	92,3	57,6	96,8	56,3	96,9
	$P_{0,25}$	0,0	81,9	80,3	72,9	73,7	82,5	96,3	82,4	96,6
	$P_{0,50}$	0,0	60,7	59,7	50,5	52,3	95,5	95,6	92,5	95,5
	$P_{0,75}$	0,0	51,9	47,1	43,4	45,4	83,3	93,0	77,8	92,2
	$P_{0,95}$	7,5	77,4	77,7	58,3	63,0	97,6	93,9	92,9	92,3
					$(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}) = (0{,}7; 0{,}8; 0)$					
moyenne	$P_{0,05}$	0,921	0,956	0,956	0,956	0,956	0,952	0,953	0,952	0,950
	$P_{0,25}$	0,637	0,780	0,781	0,781	0,781	0,761	0,764	0,762	0,750
	$P_{0,50}$	0,350	0,558	0,558	0,559	0,559	0,519	0,519	0,519	0,500
	$P_{0,75}$	0,137	0,314	0,314	0,313	0,313	0,257	0,260	0,257	0,249
	$P_{0,95}$	0,021	0,076	0,076	0,074	0,075	0,055	0,051	0,055	0,049
EAM	$P_{0,05}$	0,029	0,007	0,007	0,007	0,007	0,005	0,006	0,006	0,006
	$P_{0,25}$	0,113	0,031	0,031	0,031	0,031	0,015	0,017	0,015	0,012
	$P_{0,50}$	0,150	0,058	0,058	0,059	0,059	0,022	0,021	0,023	0,014
	$P_{0,75}$	0,113	0,064	0,064	0,063	0,063	0,015	0,016	0,018	0,012
	$P_{0,95}$	0,029	0,026	0,026	0,025	0,026	0,007	0,006	0,009	0,006
TC (%)	$P_{0,05}$	21,8	90,5	90,0	87,8	87,6	97,9	95,2	96,5	96,4
	$P_{0,25}$	0,0	48,1	45,9	43,8	43,7	89,3	84,0	86,6	95,9
	$P_{0,50}$	0,0	8,4	9,6	11,4	11,1	82,9	80,6	75,8	95,8
	$P_{0,75}$	0,0	7,0	6,6	11,0	9,9	93,6	88,2	83,8	96,7
	$P_{0,95}$	6,3	33,8	31,9	35,8	31,9	93,4	96,0	86,8	97,4

Le degré d'asymétrie est considérablement moindre dans la série TRP, le TRP-AB étant la méthode la plus performante dans l'estimation des centiles. Quand l'information de borne est moyennement asymétrique (c'est-à-dire $(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}) = (0{,}7; 0{,}5; 0{,}8)$, la deuxième procédure hot deck est importante, tandis que la première l'est moins car STRP est meilleur que TRP, mais comparable à TRP-AB. En revanche, quand $(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}) = (0{,}7; 0{,}8; 0)$, la première étape hot deck est plus importante pour le choix d'une borne correcte en raison d'une information de borne extrêmement asymétrique. TRP-AB est alors la méthode la plus efficace de la série TRP parce qu'elle est la seule à comporter la première étape de hot deck. De plus, la méthode STRP est plus efficace que $STRP_2$ dans l'estimation des centiles. En bref, les doubles procédures hot deck comprenant un appariement d'information de borne et un tirage de résidu proportionné sont essentielles non seulement pour les

restrictions de borne, mais aussi pour l'information de borne asymétrique et même pour l'information de borne symétrique et les distributions à queue lourde.

4 Analyse empirique

4.1 Données

En Corée, les services d'assurance maladie sont nationaux et obligatoires en vertu de la loi, et les données relatives à l'information médicale pour l'ensemble de la population coréenne sont enregistrées dans une base de données nationale sur la santé. Une base de données de cohortes d'échantillon est construite par échantillonnage aléatoire stratifié à partir de ces données nationales sur la santé à des fins de recherche. Elle conserve une structure de cohorte en vigueur depuis un échantillon de 2002 (Lee, Lee, Park, Shin et Kim, 2016). Plusieurs études médicales ont été menées à partir de ces mégadonnées médicales publiées récemment (Kwon, Lim et Park, 2017; Kim, Kwon, Yu, Kim, Choi, Baik, Park et Kim, 2017; Kim, Lee, Kim, Kim, Choi, Baik, Choi, Pop-Busui, Park et Kim, 2015; Ko, Yoon, Kim, Kim, Kim et Seo, 2016; Ko, Jo, Park, Kim, Kim et Park, 2016; Rim, Kim, Han et Chung, 2015).

Nous appliquons des méthodes d'imputation aux données de ces cohortes d'échantillon, en particulier aux valeurs manquantes des variables des données autodéclarées. Les dossiers de dépistage médical comportent des variables mesurées par les professionnels de la santé, comme la taille, le poids, la tension artérielle et la glycémie. Elles sont fiables et complètement observées. En revanche, d'autres variables de dépistage comme la période de tabagisme, la fréquence de l'exercice physique et les habitudes de consommation d'alcool sont autodéclarées. Elles ont une probabilité d'incomplétude et d'inexactitude, comme le montrent Crossley et Kennedy (2002), Cambois, Robine et Mormiche (2007), et Kwon et Park (2016).

À partir des données de dépistage médical des cohortes d'échantillon pour 2011 et 2013, nous imputons les valeurs manquantes des périodes de tabagisme (en années) des fumeurs de l'année 2013 dont l'âge se situait entre 20 et 84 ans et qui avaient des dossiers de dépistage en 2011 et en 2013. Soit $Y_{k,i}^{\text{auto}}$ la période autodéclarée de tabagisme et $\hat{\text{AGE}}_{k,i}$ la valeur minimale de l'âge catégorisée en classes de cinq ans pour la personne i l'année k, respectivement.

Il n'y a pas de données manquantes dans $Y_{2013,i}^{\text{auto}}$ car nous avons limité notre analyse aux personnes ayant répondu qu'elles fumaient en 2013. On considère les valeurs déraisonnables de périodes de tabagisme dans $Y_{2013,i}^{\text{auto}}$ comme manquantes, en comparant les périodes de tabagisme et l'âge en 2011 et 2013, c'est-à-dire si $Y_{2013,i}^{\text{auto}}$ satisfait $Y_{2013,i}^{\text{auto}} > \hat{\text{AGE}}_{2013,i} - a$, $Y_{2013,i}^{\text{auto}} < Y_{2011,i}^{\text{auto}} + 2 - b_1$, ou $Y_{2013,i}^{\text{auto}} > Y_{2011,i}^{\text{auto}} + 2 + b_2$ où a est un âge minimal auquel la personne a commencé à fumer, et b_1 et b_2 sont les tolérances en fonction de la mémoire humaine. Nous notons $Y_{2013,i}$ la nouvelle période de tabagisme en 2013 pour la distinguer de $Y_{2013,i}^{\text{auto}}$ sans valeurs manquantes.

Comme l'ont fait Raghunathan et coll. (2001), nous établissons la borne supérieure, C_{Ui}, à l'âge minimum $\hat{\text{A}}\text{GE}_{2013,i} -$ auquel la personne a commencé à fumer. Bien que Raghunathan et coll. (2001) fixent l'âge minimal pour fumer à 18 ans, nous considérons un âge minimum de 10 ans, soit l'âge de début du tabagisme le plus bas observé dans les données des cohortes de l'échantillon. Nous établissons $Y_{2011,i}^{\text{auto}}$ pour une borne inférieure C_{Li}. Quand $Y_{2011,i}^{\text{auto}}$ est une valeur manquante, ce qui représente 0,09 % des données, nous établissons C_{Li} à 0. Si $Y_{2011,i}^{\text{auto}}$ est identique à $Y_{2013,i}$, nous établissons la valeur C_{Li} à $Y_{2011,i}^{\text{auto}} - 0{,}001$ pour nous assurer qu'aucun dénominateur du résidu proportionné donné dans l'équation (2.3) soit nul.

En ajustant b_1 et b_2 qui montrent la mesure dans laquelle nous permettons l'erreur due à la mémoire humaine, le taux de valeurs manquantes varie de 44,0 % à 73,5 %. En supposant $b_1 = b_2 = 2$ quand nous pouvons autoriser jusqu'à deux ans d'erreur de mémoire humaine, le taux de données manquantes est de 44,0 %. Si nous ne permettons aucune erreur, c'est-à-dire que nous supposons $b_1 = 1$, $b_2 = 1$ pour ceux qui ont fumé en 2013 mais n'ont pas fumé en 2011 et $b_1 = 2$, $b_2 = 1$ pour ceux qui ont fumé en 2013 et en 2011, le taux de données manquantes est de 73,5 %.

Afin d'ajuster le modèle de régression pour les périodes de tabagisme, nous utilisons le sexe, l'âge et le niveau de revenu comme prédicteurs importants, le tableau 4.1 présente le résumé des données utilisées dans l'article. L'information sur le revenu individuel utilisée pour l'estimation des primes d'assurance maladie a été observée sous la forme d'une variable ordonnée catégorisée comportant 11 niveaux. Nous avons reclassé les groupes de revenu en trois groupes : élevé (30 % supérieurs), faible (30 % inférieurs) et moyen (autres), ce qui a permis d'améliorer l'ajustement.

Tableau 4.1
Résumé des données

		taux de données manquantes (%)	n	âge moyen	ratio d'hommes (%)	ratio de groupe de revenu (%)	
						30 % supérieurs	30 % inférieurs
avec ± tolérance de 2 ans	obs.	44,0	19 601	42,6	95,8	47,6	10,4
	manquantes		15 414	48,5	95,2	44,3	15,6
sans tolérance	obs.	73,5	9 266	38,7	95,9	49,1	7,6
	manquantes		25 749	47,6	95,4	45,1	14,5
	total		35 015	45,2	95,6	46,1	12,7
		$\text{cor}\left(Y_{2013}, \hat{\text{A}}\text{GE}_{2013}\right)$			$\text{cor}\left(Y_{2013}^{\text{auto}}, Y_{2011}^{\text{auto}}\right)$		
avec ± tolérance de 2 ans		0,79			0,99		
sans tolérance		0,71			0,99		

L'âge moyen des personnes qui n'ont pas répondu à la question sur le tabagisme en 2013 est d'environ 6 à 9 ans supérieur à celui des répondants. La distribution et la moyenne du revenu indiquent que le niveau de revenu des non-répondants est inférieur au niveau de revenu des répondants. Étant donné que l'âge et le niveau de revenu sont des prédicteurs importants de la période de tabagisme, il est difficile de supposer que le mécanisme de valeurs manquantes de la période de tabagisme est MCAR.

Comme le montre le tableau 4.1, la corrélation entre $Y_{2013,i}$ et $Y_{2011,i}^{\text{auto}}$ atteint la valeur élevée de 0,99 en raison du traitement de Y_{2013}^{auto} comme une valeur manquante lorsque les contraintes logiques ne sont pas satisfaites selon l'hypothèse que Y_{2011}^{auto} est correct. Cependant, il va de soi que Y_{2011}^{auto} a le même problème que Y_{2013}^{auto} et qu'il n'est pas fiable non plus. En dépit d'une corrélation aussi élevée, ceci est la raison pour laquelle $Y_{2011,i}^{\text{auto}}$ n'est pas inclus comme prédicteur. L'information de borne erronée affecte uniquement les valeurs imputées individuelles, mais une information de borne erronée comme prédicteur a des effets sur les estimations globales du modèle de régression, ce qui a une grande incidence sur la fiabilité globale de l'imputation. Par conséquent, nous utilisons seulement les variables mesurées comme le sexe, l'âge et le revenu qui sont recueillies par les pouvoirs publics comme base de la collecte des primes d'assurance maladie nationales. Il faut noter que la variable de l'âge est également utilisée comme information sur la borne supérieure.

Comme il est fort possible que les périodes de tabagisme autodéclarées en 2011 soient incorrectes, on pourrait critiquer le fait d'établir Y_{2011}^{auto} comme borne inférieure. C'est pourquoi nous considérons une autre borne inférieure avec $C_{Li} = 1$ qui est la plus petite période de tabagisme observée chez les fumeurs actuels.

4.2 Résultats

Le modèle de régression pour la période de tabagisme en 2013 $Y_{2013,i}$ est ajusté avec des cas entièrement observés, comme le montre le tableau 4.2.

Dans le tableau 4.2, le sexe (femme) est une variable indicatrice avec la valeur 1 désignant les femmes, l'âge est la valeur centrale de l'âge catégorisée en classes de cinq ans, et le revenu (faible) et le revenu (moyen) sont deux variables indicatrices avec la valeur 1 pour désigner l'appartenance à un groupe de revenu particulier.

Tableau 4.2
Modèle de régression pour la période de tabagisme en 2013

	avec ± tolérance de 2 ans		sans tolérance	
	estimation	**valeur-t**	**estimation**	**valeur-t**
ordonnée à l'origine	-9,42	-54,00	-6,70	-23,97
sexe (femme)	-8,38	-40,51	-8,58	-28,41
âge	0,69	182,34	0,64	96,56
revenu (faible)	-0,32	-2,25	-0,76	-3,21
revenu (moyen)	-0,50	-5,77	-0,99	-7,81
R^2	0,66		0,55	

Le tableau 4.3 montre la moyenne de $Y_{2013,i}$ estimée par chacune des cinq méthodes d'imputation avec $M = 5$ et $m_d = 6$. Une taille de donneur possible m_d peut être inférieure à 6 quand l'échantillon ne suffit pas pour composer un donneur, mais cela ne se produit pas dans nos données. Nous examinons quatre scénarios différents composés de deux réglages de bornes inférieures et de deux tolérances d'erreurs de mémoire humaine.

Contrairement à ce que montraient les résultats de simulation, la sous-estimation est plus importante par T-NORM que par OBS quand $C_{Li} = 1$. La distribution des périodes de tabagisme observées est légèrement asymétrique à droite, car la distance entre Q50 et Q95 est plus grande que celle entre Q50 et Q5. Cependant, T-NORM impute une moyenne prédictive plus un résidu aléatoire généré à partir d'une distribution normale tronquée, et non pas à partir de la distribution empirique des résidus, qui est asymétrique à droite. Étant donné que la borne inférieure $C_{Li} = 1$ est loin de la moyenne, la possibilité de sélectionner une erreur négative est plus élevée par rapport à la distribution normale tronquée que par rapport à la distribution empirique des résidus, asymétrique à droite. Cela explique la sous-estimation par T-NORM. Toutes les autres méthodes d'imputation estiment que la période moyenne de tabagisme est plus longue qu'OBS, car elles utilisent des résidus empiriques.

OBS produit une moyenne plus élevée de périodes de tabagisme avec une tolérance que sans tolérance, car le coefficient de régression de l'âge est plus élevé avec tolérance que sans tolérance, comme le montre le tableau 4.2. Sauf par la méthode T-NORM, les périodes de tabagisme estimées sont plus longues sans tolérance qu'avec tolérance, et l'écart est plus petit quand $C_{Li} = 1$ que quand $C_{Li} = Y_{2011,i}$.

La méthode TRP-AB est la plus robuste, quelle que soit la façon dont nous définissons les données manquantes et la borne inférieure. Il s'agit d'une propriété d'imputation souhaitable quand l'information sur les bornes n'est pas fiable. Par ailleurs, les résultats d'estimation des méthodes d'imputation actuelles (T-NORM, T-MV, T-TRL, T-AMP) dépendent visiblement du choix de borne et de tolérance de la mémoire humaine. Les distributions estimées de la période de tabagisme par les méthodes actuelles bougent substantiellement vers la droite quand $C_{Li} = Y_{2011,i}$ relativement à quand $C_{Li} = 1$ car $Y_{2011,i}$ est une borne plus informative que la borne constante. Cependant, la distribution avec TRP-AB n'est que marginalement modifiée pour des bornes différentes.

Tableau 4.3
Moyenne estimée, quantiles de 5, 25, 50, 75 et 95 % des années de tabagisme chez les fumeurs coréens en 2013

	avec ± tolérance de 2 ans en raison de la mémoire humaine (taux de valeurs manquantes = 44,0 %)											
	$C_{Li} = Y_{2011,i}$						$C_{Li} = 1$					
	moyenne	Q5	Q25	Q50	Q75	Q95	moyenne	Q5	Q25	Q50	Q75	Q95
OBS	19,55	7,00	12,00	20,00	25,00	40,00	19,55	7,00	12,00	20,00	25,00	40,00
T-NORM	21,46	8,00	14,00	20,00	27,98	40,22	17,66	3,92	10,00	16,99	22,91	35,00
T-MV	22,79	9,00	15,00	20,05	30,00	41,80	21,45	7,01	14,45	20,00	28,17	40,00
T-TRL	22,63	9,00	15,00	20,00	30,00	40,20	21,32	7,00	14,00	20,00	30,00	40,00
T-AMP	22,64	9,00	15,00	20,00	30,00	40,53	21,31	7,00	14,00	20,00	30,00	40,00
TRP-AB	20,61	6,31	13,00	20,00	26,90	40,00	21,33	7,00	14,00	20,00	30,00	40,00

Tableau 4.3 (suite)
Moyenne estimée, quantiles de 5, 25, 50, 75 et 95 % des années de tabagisme chez les fumeurs coréens en 2013

	sans aucune tolérance en raison de la mémoire humaine (taux de valeurs manquantes = 73,5 %)											
	$C_{Li} = Y_{2011,i}$						$C_{Li} = 1$					
	moyenne	Q5	Q25	Q50	Q75	Q95	moyenne	Q5	Q25	Q50	Q75	Q95
OBS	17,25	5,00	11,00	16,00	22,00	32,00	17,25	5,00	11,00	16,00	22,00	32,00
T-NORM	21,92	8,00	14,77	20,66	28,00	40,75	14,61	2,70	8,13	13,64	20,00	30,00
T-MV	24,01	10,00	16,75	22,85	30,34	42,67	21,63	7,19	15,00	20,95	27,68	37,87
T-TRL	24,40	10,00	16,00	23,00	30,00	47,40	21,44	5,00	13,00	20,00	28,80	42,00
T-AMP	24,41	10,00	16,00	23,00	30,00	47,45	21,47	5,00	13,00	20,00	28,80	42,00
TRP-AB	21,11	7,00	13,00	20,00	27,00	41,80	21,42	5,00	13,00	20,00	28,80	42,00

La figure 4.1 présente les estimations de la densité selon la méthode du noyau pour les périodes de tabagisme obtenues utilisant OBS et TRP-AB selon deux réglages de bornes inférieures et deux tolérances d'erreurs dues à la mémoire humaine. L'imputation par TRP-AB déplace la distribution de la période de tabagisme à la droite et la disperse largement, comparativement à la distribution construite seulement sur des observations (OBS).

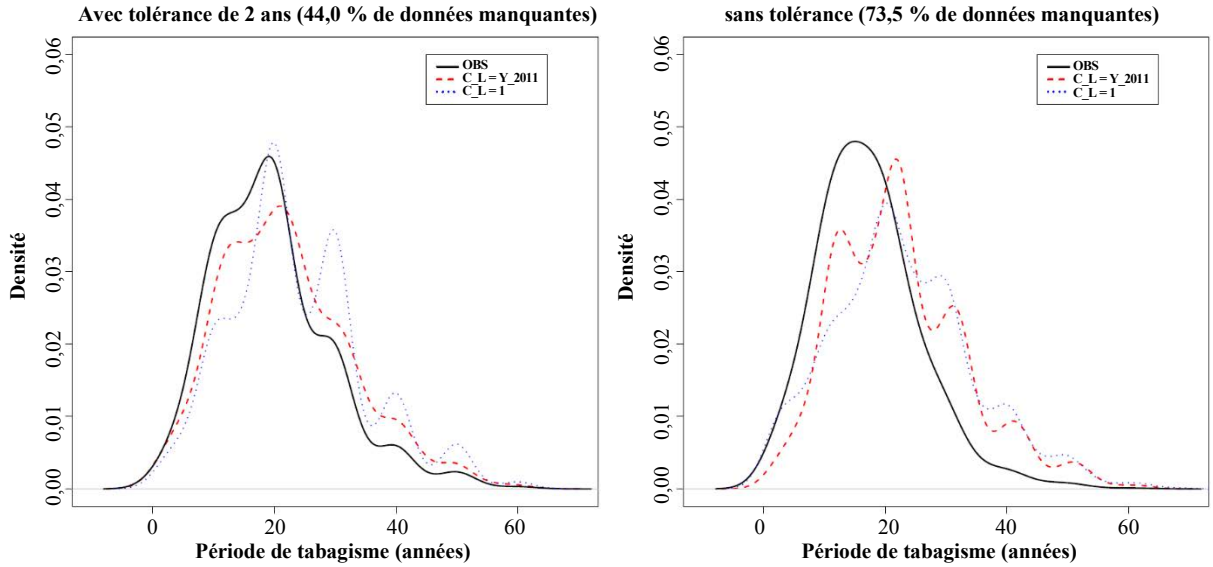

Figure 4.1 Estimations de la densité selon la méthode du noyau pour les périodes de tabagisme obtenues utilisant OBS et TRP-AB selon deux réglages de bornes inférieures et deux tolérances d'erreurs dues à la mémoire humaine.

5 Conclusion

Nous avons proposé une méthode d'imputation multiple des variables manquantes quand les valeurs manquantes sont bornées logiquement, ce qui est souvent le cas dans les recensements ou les

enquêtes-échantillons. Les méthodes d'imputation actuelles avec troncature supplémentaire ou une étape d'acceptation/rejet ont produit des estimations biaisées, selon l'étendue de l'asymétrie de l'information sur les bornes. Leurs valeurs d'imputation diminuent et se rapprochent de la borne en cas de corrélation plus faible avec la variable manquante. Toutefois, si l'on emploie un tirage de résidu proportionné, un appariement d'information de borne et une double procédure hot deck, notre méthode dite TRP-AB produit des estimations plus exactes et efficaces de la moyenne et des centiles, quels que soient les taux de variables manquantes, les mécanismes de valeurs manquantes et les distributions des variables manquantes.

De plus, notre méthode d'imputation TRP-AB résiste à l'information de borne asymétrique en ce sens que ses valeurs imputées ne dépendent pas de l'étendue de l'asymétrie de l'information de borne. En particulier, en présence de deux ou de plusieurs variables pour l'information de borne, ou quand la fiabilité de l'information sur la borne inférieure est suspecte, l'imputation par TRP-AB est un outil puissant permettant d'estimer précisément les paramètres d'intérêt.

La méthode TRP-AB fonctionne également pour une seule imputation. Il peut y avoir des cas où (surtout dans les statistiques officielles) un seul ensemble de données de sortie définitives est nécessaire et où les utilisateurs n'ont pas les moyens sophistiqués nécessaires pour traiter une imputation multiple.

Remerciements

Ces travaux ont été soutenus par une subvention de la *National Research Foundation of Korea* (NRF) financée par le *Ministry of Science and ICT* (MSIT) du gouvernement coréen (NRF-2018R1C1B5043739). Ces travaux ont aussi été financés par une subvention de la *Korea University* (K1910711) et le *Hankuk University of Foreign Studies Research Fund*. Nous remercions tous les participants de la *Korean Health Insurance Cohort study* et de la *National Health Insurance Service* (NHIS) qui forment une partie de la base de données de la NHIS (Numéro de gestion de recherche : NHIS- 2016-2-129).

Annexe

Démonstration du théorème 1.

Il suffit de montrer que $Y^*_{jU} \le C_{jU}$ et $Y^*_{jL} \ge C_{jL}$ en raison des contraintes de l'étape d'imputation.

1. Si $\hat{Y}_j^{\text{manq}} \in S^0$, alors $\tilde{r}^*_{jU}$ est échantillonné à partir de R_U^0 dont l'élément $\tilde{r}_{iU} \le 1$ pour tous les i car $Y_i \le C_{iU}$ et $C_{iU} - \hat{Y}_i > 0$ pour $\tilde{r}_{iU} \in R_U^0$. Étant donné que $\tilde{r}^*_{jU}$ est un de ces $\tilde{r}_{iU}$, nous avons $\tilde{r}^*_{jU} \le 1$. De plus $C_{jU} - \hat{Y}_j^{\text{manq}} > 0$ donne

$$Y^*_{j,U} = \hat{Y}_j^{\text{manq}} + \tilde{r}^*_{j,U}\left(C_{j,U} - \hat{Y}_j^{\text{manq}}\right) \le \hat{Y}_j^{\text{manq}} + 1 \times \left(C_{j,U} - \hat{Y}_j^{\text{manq}}\right) = C_{jU}. \qquad \text{(A.1)}$$

De même, si $\hat{Y}_j^{\text{manq}} \in S^0$, alors $\tilde{r}_{jL}^*$ est sélectionné aléatoirement à partir de R_L^0 dont l'élément $\tilde{r}_{iL} \leq 1$ pour toute valeur i car $Y_i \geq C_{iL}$ et $C_{iL} - \hat{Y}_i < 0$ pour $i \in R_L^0$. Étant donné que $\tilde{r}_{jL}^*$ est un de ces $\tilde{r}_{iL}$, $\tilde{r}_{jL}^* \leq 1$. En utilisant $C_{j,L} - \hat{Y}_j^{\text{manq}} < 0$ car $\hat{Y}_j^{\text{manq}} \in S^0$, nous obtenons

$$Y_{j,L}^* = \hat{Y}_j^{\text{manq}} + \tilde{r}_{jL}^* \left(C_{j,L} - \hat{Y}_j^{\text{manq}}\right) \geq \hat{Y}_j^{\text{manq}} + 1 \times \left(C_{j,L} - \hat{Y}_j^{\text{manq}}\right) = C_{jL}. \qquad \text{(A.2)}$$

2. Si $\hat{Y}_j^{\text{manq}} \in S^+$, alors $\tilde{r}_{jU}^*$ est échantillonné à partir de R_U^+ dont l'élément $\tilde{r}_{iU} \geq 1$ pour tous les i car $Y_i \leq C_{i,U}$ et $C_{iU} - \hat{Y}_i < 0$. Étant donné que $\tilde{r}_{jU}^*$ est un de ces $\tilde{r}_{iU}$, nous avons $\tilde{r}_{jU}^* \geq 1$. De plus $C_{jU} - \hat{Y}_j^{\text{manq}} < 0$ donne

$$Y_{j,U}^* = \hat{Y}_j^{\text{manq}} + \tilde{r}_{j,U}^* \left(C_{j,U} - \hat{Y}_j^{\text{manq}}\right) \leq \hat{Y}_j^{\text{manq}} + 1 \times \left(C_{j,U} - \hat{Y}_j^{\text{manq}}\right) = C_{jU}. \qquad \text{(A.3)}$$

De même, si $\hat{Y}_j^{\text{manq}} \in S^+$, alors $\tilde{r}_{jL}^*$ est sélectionné aléatoirement à partir de R_L^+ dont l'élément $\tilde{r}_{iL} \leq 1$ pour toute valeur i car $Y_i \geq C_{iL}$ et $C_{iL} - \hat{Y}_i < 0$ pour $i \in R_L^+$. Étant donné que $\tilde{r}_{jL}^*$ est un de ces $\tilde{r}_{iL}$, $\tilde{r}_{jL}^* \leq 1$. En utilisant $C_{j,L} - \hat{Y}_j^{\text{manq}} < 0$ car $\hat{Y}_j^{\text{manq}} \in S^+$, nous obtenons

$$Y_{j,L}^* = \hat{Y}_j^{\text{manq}} + \tilde{r}_{jL}^* \left(C_{j,L} - \hat{Y}_j^{\text{manq}}\right) \geq \hat{Y}_j^{\text{manq}} + 1 \times \left(C_{j,L} - \hat{Y}_j^{\text{manq}}\right) = C_{jL}. \qquad \text{(A.4)}$$

3. Si $\hat{Y}_j^{\text{manq}} \in S^-$, alors $\tilde{r}_{jU}^*$ est échantillonné à partir de R_U^- dont l'élément $\tilde{r}_{iU} \leq 1$ pour tous les i car $Y_i \leq C_{iU}$ et $C_{iU} - \hat{Y}_i > 0$ pour $\tilde{r}_{iU} \in R_U^-$. Étant donné que $\tilde{r}_{jU}^*$ est un de ces $\tilde{r}_{iU}$, nous avons $\tilde{r}_{jU}^* \leq 1$. De plus $C_{jU} - \hat{Y}_j^{\text{manq}} > 0$ donne

$$Y_{j,U}^* = \hat{Y}_j^{\text{manq}} + \tilde{r}_{j,U}^* \left(C_{j,U} - \hat{Y}_j^{\text{manq}}\right) \leq \hat{Y}_j^{\text{manq}} + 1 \times \left(C_{j,U} - \hat{Y}_j^{\text{manq}}\right) = C_{jU}. \qquad \text{(A.5)}$$

De même, si $\hat{Y}_j^{\text{manq}} \in S^-$, alors $\tilde{r}_{jL}^*$ est sélectionné aléatoirement à partir de R_L^- dont l'élément $\tilde{r}_{iL} \geq 1$ pour toute valeur i car $Y_i \geq C_{iL}$ et $C_{iL} - \hat{Y}_i > 0$ pour $i \in R_L^-$. Étant donné que $\tilde{r}_{jL}^*$ est un de ces $\tilde{r}_{iL}$, $\tilde{r}_{jL}^* \geq 1$. En utilisant $C_{j,L} - \hat{Y}_j^{\text{manq}} > 0$ car $\hat{Y}_j^{\text{manq}} \in S^-$, nous obtenons

$$Y_{j,L}^* = \hat{Y}_j^{\text{manq}} + \tilde{r}_{jL}^* \left(C_{j,L} - \hat{Y}_j^{\text{manq}}\right) \geq \hat{Y}_j^{\text{manq}} + 1 \times \left(C_{j,L} - \hat{Y}_j^{\text{manq}}\right) = C_{jL}. \qquad \text{(A.6)}$$

Bibliographie

Andridge, R.R., et Little, R.J.A. (2010). A review of hot deck imputation for survey non-response. *Revue Internationale de Statistique*, 78(1), 40-64.

Cambois, E., Robine, J.-M. et Mormiche, P. (2007). Did the prevalence of disability in France really fall in the 1990s? A discussion of questions asked in the French Health Survey. *Population-E*, 62(2), 313-337.

Crossley, T.F., et Kennedy, S. (2002). The reliability of self-assessed health status. *Journal of Health Economics*, 21, 4, 643-658.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D.F. et Meulders, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics*, 61(1), 74-85.

Geraci, M., et McLain, A. (2018). Multiple imputation for bounded variables. *Psychometrika*, 83(4), 919-940.

Honaker, J., King, G. et Blackwell, M. (2012). Amelia II: A program for missing data. R package, version 1.6.4.

Kim, N.H., Kwon, T.Y., Yu, S., Kim, N.H., Choi, K.M., Baik, S.H., Park, Y. et Kim, S.G. (2017). Increased vascular disease mortality risk in prediabetic Korean adults is mainly attributable to ischemic stroke. *Stroke*, 48, 4, 840-845.

Kim, N.H., Lee, J., Kim, T.J., Kim, N.H., Choi, K.M., Baik, S.H., Choi, D.S., Pop-Busui, R., Park, Y. et Kim, S.G. (2015). Body mass index and mortality in the general population and in subjects with chronic disease in Korea: A nationwide cohort study (2002-2010). *PLoS ONE*, 10(10).

Ko, M.J., Jo, A.J., Park, C.M., Kim, H.J., Kim, Y.J. et Park, D.W. (2016). Level of blood pressure control and cardiovascular events: SPRINT criteria versus the 2014 hypertension recommendations. *Journal of the American College of Cardiology*, 67(24), 2821-2831.

Ko, S., Yoon, S.J., Kim, D., Kim, A.R., Kim, E.J. et Seo, H.Y. (2016). Metabolic risk profile and cancer in Korean men and women. *Journal of Preventive Medicine and Public Health*, 49(3), 143-152.

Kwon, T.Y., et Park, Y. (2015). A new multiple imputation method for bounded missing values. *Statistics and Probability Letters*, 107, 204-209.

Kwon, T.Y., et Park, Y. (2016). Reliability of self-reported data for prevalence and health life expectancy studies: Comparison with sample cohort DB of National Health Insurance Services. *The Korean Journal of Applied Statistics*, 39(7), 1329-1346.

Kwon, T.Y., Lim, J. et Park, Y. (2017). Health life expectancy in Korea based on sample cohort database of National Health Insurance Services. *The Korean Journal of Applied Statistics*, 30(3), 475-486.

Lee, J., Lee, J.S., Park, S.H., Shin, S.A. et Kim, K. (2016). Cohort profile: The National Health Insurance Service-National Sample Cohort (NHIS-NSC), South Korea. *International Journal of Epidemiology*, doi: 10.1093/ije/dyv319.

Little, R. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-296.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. et Solenberger, P. (2001). Une technique multidimensionelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression. *Techniques d'enquête*, 27, 1, 91-103. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2001001/article/5857-fra.pdf.

Raghunathan, T.E., Solenberger, P.W. et Van Hoewyk, J. (2002). Iveware: Imputation and variance estimation software. Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.

Rim, T.H., Kim, D.W., Han, J.S. et Chung, E.J. (2015). Retinal vein occlusion and the risk of stroke development: A 9-year nationwide population-based study. *Ophthalmology*, 122(6), 1187-1194.

Rubin, D.B. (1978). Multiple imputation 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*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (2004). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B., et Schenker, N. (1986). Multiple imputations for interval estimation from simple random sampling with ignorable nonresponse. *Journal of American Statistical Association*, 81(394), 366-374.

Schafer, J.L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8(1), 3-15.

Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A. et Rubin, D.B. (1996). The NHANES III multiple imputation project. *Race/Ethnicity*, 60(21.2).

Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences*, 3, 153-160.

Schenker, N., Raghunathan, T.E., Chiu, P.L., Makuc, D.M., Zhang, G. et Cohen, A.J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association*, 101(475), 924-933.

Schenker, N., et Taylor, J.M.G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 22(4), 425-446.

Su, Y.-S., Gelman, A., Hill, J. et Yajima, M. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, 45(2), 1-31.

van Buuren, S., et Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67.

Yucel, R.M., et Demirtas, H. (2010). Impact of non-normal random effects on inference by multiple imputation: A simulation assessment. *Computational Statistics & Data Analysis*, 54(3), 790-801.

Yucel, R.M., He, Y. et Zaslavsky, A.M. (2008). Using calibration to improve rounding in imputation. *The American Statistician*, 62(2), 125-129.