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

Techniques d'enquête

Imputation multiple de valeurs manquantes dans des données des ménages contenant des zéros structurels

par Olanrewaju Akande, Jerome Reiter et Andrés F. Barrientos

Date de diffusion : le 27 juin 2019

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 2019
Vol. 45, nº 2, pp. 289-315
Statistique Canada, nº 12-001-X au catalogue

Imputation multiple de valeurs manquantes dans des données des ménages contenant des zéros structurels

Olanrewaju Akande, Jerome Reiter et Andrés F. Barrientos[1]

Résumé

Nous exposons une méthode d'imputation de valeurs manquantes dans des données catégoriques multivariées emboîtées au sein des ménages. Cette méthode reposant sur un modèle à classes latentes (i) permet des variables au double niveau des ménages et des particuliers, (ii) attribue dans ce modèle une probabilité nulle aux configurations impossibles des ménages et (iii) peut préserver les distributions multivariées à la fois dans et entre les ménages. Nous présentons un échantillonneur de Gibbs pour l'estimation du modèle et la production des imputations. Nous décrivons en outre des stratégies d'amélioration de l'efficacité de calcul pour l'estimation du modèle. Nous illustrons enfin le rendement de la méthode à l'aide de données imitant les variables recueillies dans des recensements types de la population.

Mots-clés : Catégorique; recensement; vérification; latent; mélange; non-réponse.

1 Introduction

Dans une foule de recensements de la population et d'enquêtes démographiques, les organismes statistiques recueillent des données sur les particuliers formant un groupe dans un logement. Dans le recensement décennal aux États-Unis par exemple, le Census Bureau observe l'âge, la race, le sexe et le lien avec le chef du ménage de chaque membre du ménage. Il observe aussi si les occupants sont propriétaires ou non du logement. Après la collecte, les organismes partagent ces ensembles de données à des fins d'analyse secondaire sous forme de tableaux récapitulatifs, d'échantillons de microdonnées à grande diffusion ou de fichiers d'accès restreint.

Lorsqu'ils créent ces produits de données, ils ont normalement à traiter la non-réponse partielle pour les variables tant des particuliers que des ménages, ce qu'ils font d'habitude à l'aide d'une certaine procédure d'imputation. Idéalement, de telles procédures obéissent à trois desiderata. D'abord, les imputations préservent le mieux possible la distribution conjointe des variables, ce qui comprend la conservation des liens au sein des ménages. Ainsi, la race manquante du conjoint correspondra probablement (mais certainement non à coup sûr) à la race du chef du ménage, correspondance dont devrait tenir compte la procédure d'imputation. En deuxième lieu, les imputations respectent les zéros structurels. Pour prendre un exemple, l'âge de la fille ne peut dépasser l'âge de la mère biologique. Les imputations ne devraient pas créer de combinaisons impossibles de particuliers au sein d'un ménage. Enfin, la procédure permet de propager une incertitude appropriée dans les analyses ultérieures des données.

Les approches typiques d'imputation des valeurs manquantes des ménages exploitent une variante de l'imputation hot-deck (Kalton et Kasprzyk, 1986; Andridge et Little, 2010). Il reste que, selon la façon de

1. Olanrewaju M. Akande, Department of Statistical Science, Duke University, Durham, NC 27708. Courriel : olanrewaju.akande@duke.edu; Jerome P. Reiter est professeur de statistique, Duke University, Durham, NC 27708. Courriel : jerry@stat.duke.edu; Andrés F. Barrientos est associé postdoctoral, Department of Statistical Science, Duke University, Durham, NC 27708. Courriel : anfebar@stat.duke.edu.

mettre en œuvre le hot-deck, un ou plusieurs des desiderata pourraient ne pas être respectés. En fait, nous ne connaissons aucune procédure d'imputation hot-deck applicable aux données des ménages qui respecte expressément les trois desiderata. Une solution de rechange consiste à estimer un modèle décrivant la distribution conjointe de toutes les variables et à imputer les valeurs manquantes à partir des distributions prédictives implicites de ce modèle. Dans le cas des données des ménages, un de ces modèles est le mélange de Dirichlet à données emboîtées des produits de distributions multinomiales (MDPDM), qui a été conçu par Hu, Reiter et Wang (2018) et où on suppose que (i) chaque ménage est membre d'une classe latente au niveau des ménages et que (ii) chaque particulier est membre d'une classe latente au niveau des particuliers emboîtée dans sa classe latente au niveau des ménages. Dans ce modèle, on attribue une probabilité nulle aux combinaisons correspondant à des zéros structurels et on traite simultanément les variables au niveau des ménages et au niveau des particuliers. Le MDPDM est attrayant comme moteur d'imputation et il peut préserver les associations multivariées, tout en évitant les imputations qui créent des ménages impossibles. Il est apparenté aux modèles proposés par Vermunt (2003, 2008) et Bennink, Croon, Kroon et Vermunt (2016), mais ceux-ci servent à la régression plutôt qu'à l'imputation multivariée et ne s'occupent pas des zéros structurels.

Hu et coll. (2018) procèdent par MDPDM pour produire des jeux de données synthétiques (Rubin, 1993; Raghunathan et Rubin, 2001; Reiter et Raghunathan, 2007) à des fins de limitation de la divulgation statistique, mais sans décrire comment les utiliser en imputation des données manquantes, ce que nous allons précisément faire dans le présent article. Avec des zéros structurels dans le MDPDM, nous ne disposons pas en forme close de distributions conditionnelles des valeurs manquantes étant donné les valeurs observées. Nous devons donc ajouter une étape d'échantillonnage de rejet à l'échantillonneur de Gibbs utilisé par Hu et coll. (2018) de manière à produire des jeux de données complètes comme sous-produits des algorithmes de Monte Carlo à chaîne de Markov (MCMC) servant à l'estimation du modèle. Il est possible d'analyser ces jeux complets par inférence d'imputation multiple (Rubin, 1987). Nous exposons également deux nouvelles stratégies permettant d'accélérer les calculs MDPDM et qui consistent (i) à convertir les données du chef du ménage en variables au niveau des ménages par opposition au niveau des particuliers et (ii) à recourir à une approximation de la fonction de vraisemblance. Ces nouveautés extensibles sont nécessaires, puisque le MDPDM est plutôt vorace en calcul même sans données manquantes. On peut aussi faire appel aux stratégies d'accélération lorsqu'on se sert du MDPDM pour produire des données synthétiques.

Voici comment se présente le reste de cet article. À la section 2, nous examinons le modèle MDPDM en présence de zéros structurels et l'échantillonneur MCMC pour l'ajustement du modèle sans données manquantes. À la section 3, nous étendons aux données manquantes l'échantillonneur MCMC pour le modèle MDPDM. À la section 4, nous présentons les deux stratégies d'accélération de l'échantillonneur MCMC. À la section 5, nous livrons les résultats des études de simulation utilisées pour évaluer le rendement du MDPDM comme moteur d'imputation multiple, et ce, en employant les deux stratégies d'accélération de l'exécution. À la section 6, il est question des conclusions, des mises en garde et des travaux futurs.

2 Examen du modèle MDPDM

Hu et coll. (2018) présentent le modèle MDPDM en précisant notamment en quoi il peut préserver les associations entre variables et prendre en compte les zéros structurels. Ici, nous résumons le même modèle sans entrer dans le détail des raisons de son adoption et en nous contentant de renvoyer le lecteur à Hu et coll. (2018) pour un complément d'information. Nous commençons par la notation nécessaire à la compréhension du modèle et de l'échantillonneur de Gibbs avec données complètes. Notre exposé suit de près celui de Hu et coll. (2018).

2.1 Notation et spécification du modèle

Posons que les données visent n ménages. Chaque ménage $i = 1, \ldots, n$ contient n_i particuliers et, par conséquent, il y a $\sum_{i=1}^{n} n_i = N$ individus dans les données. Soit $X_{ik} \in \{1, \ldots, d_k\}$ la valeur de la variable catégorique k pour le ménage i. On la suppose identique pour tous les n_i membres du ménage i, où $k = p+1, \ldots, p+q$. Soit $X_{ijk} \in \{1, \ldots, d_k\}$ la valeur de la variable catégorique k pour le membre j du ménage i, où $j = 1, \ldots, n_i$ et $k = 1, \ldots, p$. Soit $\mathbf{X}_i = (X_{i(p+1)}, \ldots, X_{i(p+q)}, X_{i11}, \ldots, X_{in_i p})$ l'ensemble des variables au niveau des ménages et au niveau des particuliers pour les n_i membres du ménage i.

Soit $\mathcal{H}$ le jeu de toutes les tailles de ménage possibles dans la population. Pour tout $h \in \mathcal{H}$, soit $\mathcal{C}_h$ représentant le jeu de toutes les combinaisons de variables au double niveau des particuliers et des ménages pour les ménages de taille h, avec les combinaisons impossibles; en d'autres termes, $\mathcal{C}_h = \prod_{k=p+1}^{p+q} \{1, \ldots, d_k\} \prod_{j=1}^{h} \prod_{k=1}^{p} \{1, \ldots, d_k\}$. Soit $\mathcal{S}_h \subset \mathcal{C}_h$ représentant le jeu des combinaisons impossibles, c'est-à-dire des zéros structurels, pour les ménages de taille h. Sont comprises les combinaisons de variables dans tout particulier (une personne âgée de trois ans ne peut être un conjoint, par exemple) ou entre les particuliers d'un même ménage (quelqu'un ne peut être plus âgé que ses parents biologiques, par exemple). Soit $\mathcal{C} = \bigcup_{h \in \mathcal{H}} \mathcal{C}_h$ et $\mathcal{S} = \bigcup_{h \in \mathcal{H}} \mathcal{S}_h$.

Bien que le modèle MDPDM que nous employons limite le support de $\mathbf{X}_i$ à $\mathcal{C} - \mathcal{S}$, il est bon de comprendre le modèle sans restrictions au départ quant au support de $\mathbf{X}_i$. Chaque ménage i appartient à une des classes F représentant les types latents de ménages. Pour $i = 1, \ldots, n$, soit $G_i \in \{1, \ldots, F\}$ indiquant la classe de ménages pour le ménage i. Soit $\pi_g = \Pr(G_i = g)$ la probabilité que le ménage i appartienne à la classe g. Dans toute classe, toutes les variables au niveau des ménages suivent des distributions multinomiales indépendantes. Pour tout $k \in \{p+1, \ldots, p+q\}$ et tout $c \in \{1, \ldots, d_k\}$, soit $\lambda_{gc}^{(k)} = \Pr(X_{ik} = c \,|\, G_i = g)$ pour toute classe g, où $\lambda_{gc}^{(k)}$ est de la même valeur pour tout ménage de la classe g. Soit $\pi = \{\pi_1, \ldots, \pi_F\}$, et $\lambda = \left\{\lambda_{gc}^{(k)} : c = 1, \ldots, d_k;\ k = p+1, \ldots, p+q;\ g = 1, \ldots, F\right\}$.

Dans chaque classe de ménages, chaque particulier appartient à une des classes latentes S au niveau des particuliers. Pour $i = 1, \ldots, n$ et $j = 1, \ldots, n_i$, soit M_{ij} représentant la classe latente au niveau des particuliers pour le membre j du ménage i. Soit $\omega_{gm} = \Pr(M_{ij} = m \,|\, G_i = g)$ la probabilité que le membre j du ménage i appartienne au niveau des particuliers à la classe m emboîtée au niveau des

ménages au sein de la classe g. Dans toute classe au niveau des particuliers, toutes les variables de ce niveau suivent des distributions multinomiales indépendantes. Pour tout $k \in \{1, \ldots, p\}$ et tout $c \in \{1, \ldots, d_k\}$, soit $\phi_{gmc}^{(k)} = \Pr(X_{ijk} = c \,|\, (G_i, M_{ij}) = (g, m))$ pour la paire de classes (g, m), où $\phi_{gmc}^{(k)}$ est de la même valeur pour tout particulier dans la paire de classes (g, m). Soit $\omega = \{\omega_{gm} : g = 1, \ldots, F; m = 1, \ldots, S\}$, et $\phi = \{\phi_{gmc}^{(k)} : c = 1, \ldots, d_k; k = 1, \ldots, p; m = 1, \ldots, S; g = 1, \ldots, F\}$.

Pour les besoins de l'échantillonneur de Gibbs à la section 2.2, il est bon de distinguer les valeurs de $\mathbf{X}_i$ qui respectent toutes les contraintes de zéros structurels de celles qui ne les respectent pas. Soit l'exposant «1» indiquant qu'une variable aléatoire a un support seulement dans $\mathcal{C} - \mathcal{S}$. Ainsi, $\mathbf{X}_i^1$ représente les données d'un ménage dont les valeurs sont en restriction seulement dans $\mathcal{C} - \mathcal{S}$ (il ne s'agit pas d'un ménage impossible); $\mathbf{X}_i$ représente les données d'un ménage avec des valeurs dans $\mathcal{C}$. Soit $\mathcal{X}^1$ les données observées de n ménages, c'est-à-dire une réalisation de $(\mathbf{X}_1^1, \ldots, \mathbf{X}_n^1)$. Le noyau du MDPDM, $\Pr(\mathcal{X}^1 \,|\, \theta)$, est

$$L(\mathcal{X}^1 \mid \theta) = \prod_{i=1}^{n} \sum_{h \in \mathcal{H}} \mathbb{1}\{n_i = h\} \mathbb{1}\{\mathbf{X}_i^1 \notin \mathcal{S}_h\} \left[\sum_{g=1}^{F} \pi_g \prod_{k=p+1}^{p+q} \lambda_{gX_{ik}^1}^{(k)} \prod_{j=1}^{h} \sum_{m=1}^{S} \omega_{gm} \prod_{k=1}^{p} \phi_{gmX_{ijk}^1}^{(k)} \right], \tag{2.1}$$

où θ comprend tous les paramètres et $\mathbb{1}\{.\}$ correspond à l'unité lorsque la condition entre accolades $\{\}$ est vraie et à zéro dans les autres cas.

Pour tout $h \in \mathcal{H}$, soit $n_{1h} = \sum_{i=1}^{n} \mathbb{1}\{n_i = h\}$ le nombre de ménages de la taille h dans $\mathcal{X}^1$ et $\pi_{0h}(\theta) = \Pr(\mathbf{X}_i \in \mathcal{S}_h \,|\, \theta)$. Comme il est indiqué dans Hu et coll. (2018), la constante de normalisation dans la vraisemblance en (2.1) est $\prod_{h \in \mathcal{H}} (1 - \pi_{0h}(\theta))^{n_{1h}}$. Ainsi, la distribution postérieure est

$$\Pr(\theta \,|\, \mathcal{X}^1, T(\mathcal{S})) \propto \Pr(\mathcal{X}^1 \,|\, \theta) \Pr(\theta) = \frac{1}{\prod_{h \in \mathcal{H}} (1 - \pi_{0h}(\theta))^{n_{1h}}} L(\mathcal{X}^1 \,|\, \theta) \Pr(\theta) \tag{2.2}$$

où $T(\mathcal{S})$ fait voir une densité de probabilité pour MDPDM avec un support limité à $\mathcal{C} - \mathcal{S}$.

La vraisemblance dans (2.1) peut s'écrire comme modèle génératif de la forme

$$\begin{aligned} X_{ik} \,|\, G_i, \lambda \sim\ & \text{Discrète}\left(\lambda_{G_i 1}^{(k)}, \ldots, \lambda_{G_i d_k}^{(k)}\right) \\ & \forall i = 1, \ldots, n \ \text{ et } \ k = p+1, \ldots, p+q \end{aligned} \tag{2.3}$$

$$\begin{aligned} X_{ijk} \mid G_i, M_{ij}, \phi, n_i \sim\ & \text{Discrète}\left(\phi_{G_i M_{ij} 1}^{(k)}, \ldots, \phi_{G_i M_{ij} d_k}^{(k)}\right) \\ & \forall i = 1, \ldots, n\ ,\ j = 1, \ldots, n_i \ \text{ et } \ k = 1, \ldots, p \end{aligned} \tag{2.4}$$

$$\begin{aligned} G_i \,|\, \pi \sim\ & \text{Discrète}(\pi_1, \ldots, \pi_F) \\ & \forall i = 1, \ldots, n \end{aligned} \tag{2.5}$$

$$\begin{aligned} M_{ij} \,|\, G_i, \omega, n_i \sim\ & \text{Discrète}(\omega_{G_i 1}, \ldots, \omega_{G_i S}) \\ & \forall i = 1, \ldots, n \ \text{ et } \ j = 1, \ldots, n_i \end{aligned} \tag{2.6}$$

où la distribution discrète est la distribution multinomiale avec taille d'échantillon correspondant à l'unité. Nous limitons le support de chaque $\mathbf{X}_i$ pour être sûrs que le modèle attribue une probabilité nulle à toutes les combinaisons désirées dans $\mathcal{S}$. On peut utiliser le modèle dans (2.3) à (2.6) sans limiter le support à $\mathcal{C} - \mathcal{S}$. On ne tient pas compte alors de tous les zéros structurels. S'il convient peu à la distribution conjointe des données des ménages, un tel modèle se révèle utile dans le cas de l'échantillonneur de Gibbs. Nous prenons le modèle génératif dans (2.3) à (2.6) avec un support intégral dans tous $\mathcal{C}$ en tant que MDPDM non tronqué. En revanche, nous appelons le modèle en (2.1) MDPDM tronqué.

Pour les distributions antérieures, nous suivons les recommandations de Hu et coll. (2018). Nous utilisons des distributions uniformes et indépendantes de Dirichlet comme distributions antérieures pour λ et ϕ, et la représentation tronquée en découpe du processus de Dirichlet comme distributions antérieures pour π et ω (Sethuraman, 1994; Dunson et Xing, 2009; Si et Reiter, 2013; Manrique-Vallier et Reiter, 2014) :

$$\lambda_g^{(k)} = \left(\lambda_{g1}^{(k)}, \ldots, \lambda_{gd_k}^{(k)}\right) \sim \text{Dirichlet}\,(1, \ldots, 1) \tag{2.7}$$

$$\phi_{gm}^{(k)} = \left(\phi_{gm1}^{(k)}, \ldots, \phi_{gmd_k}^{(k)}\right) \sim \text{Dirichlet}\,(1, \ldots, 1) \tag{2.8}$$

$$\pi_g = u_g \prod_{f<g} (1 - u_f) \quad \text{pour} \quad g = 1, \ldots, F \tag{2.9}$$

$$u_g \sim \text{bêta}\,(1, \alpha) \quad \text{pour} \quad g = 1, \ldots, F-1, \; u_F = 1 \tag{2.10}$$

$$\alpha \sim \text{gamma}\,(0{,}25;\; 0{,}25) \tag{2.11}$$

$$\omega_{gm} = v_{gm} \prod_{s<m} (1 - v_{gs}) \quad \text{pour} \quad m = 1, \ldots S \tag{2.12}$$

$$v_{gm} \sim \text{bêta}\,(1, \beta_g) \quad \text{pour} \quad m = 1, \ldots, S-1, \; v_{gS} = 1 \tag{2.13}$$

$$\beta_g \sim \text{gamma}\,(0{,}25;\; 0{,}25)\,. \tag{2.14}$$

Nous fixons les paramètres des distributions de Dirichlet à $\mathbf{1}_{d_k}$ (vecteur d_k – dimensionnel des unités) et à 0,25 les paramètres des distributions gamma en (2.11) et (2.14) pour représenter les spécifications vagues des distributions antérieures. Nous posons également $\beta_g = \beta$ pour faciliter le calcul. Voir Hu et coll. (2018) pour mieux connaître les spécifications des distributions antérieures.

D'un point de vue conceptuel, nous pouvons interpréter les classes latentes au niveau des ménages comme des grappes de ménages d'une composition homogène (ménages avec enfants, ménages dont les membres ne sont pas apparentés, etc.). De même, il est possible d'interpréter les classes latentes au niveau des particuliers comme des grappes d'individus aux caractéristiques homogènes (conjoint plus âgé de sexe masculin, jeunes enfants de sexe féminin, etc.). Toutefois, notre propos n'est pas d'interpréter ces classes

dans une optique d'imputation, celles-ci servant avant tout à l'induction d'une dépendance entre variables et particuliers dans la distribution conjointe.

Il importe de choisir F et S de sorte qu'ils soient assez grands pour une estimation précise de la distribution conjointe. Il ne s'agit cependant pas de faire F et S si grands qu'ils produisent une abondance de classes vides dans l'estimation du modèle. Créer un grand nombre de classes vides ne peut qu'accroître la durée du calcul sans gain correspondant de précision de l'estimation. Cela peut poser tout particulièrement un problème avec l'échantillonneur de Gibbs dans un MDPDM tronqué, ces classes mettant de la masse de probabilité dans des régions de l'espace où des combinaisons impossibles risquent de se produire, d'où une convergence plus lente de l'échantillonneur de Gibbs.

Nous recommandons donc la stratégie de Hu et coll. (2018) au moment de poser (F, S). Les analystes peuvent prendre des valeurs modérées des deux, disons entre 10 et 15, pour les premiers passages de rodage. Après convergence, ils examinent les échantillons postérieurs des classes latentes et vérifient ainsi combien de ces classes sont occupées au niveau des particuliers et au niveau des ménages. Ces contrôles prédictifs postérieurs peuvent démontrer le besoin de disposer de valeurs plus grandes pour F et S. Si le nombre de classes occupées au niveau des ménages atteint F, nous suggérons d'accroître F. Si le nombre de classes occupées au niveau des particuliers atteint S, nous suggérons d'accroître F d'abord et S ensuite, et ce, peut-être en sus de F, s'il ne peut suffire de relever F seul. Si les contrôles prédictifs postérieurs n'indiquent en rien que des valeurs supérieures de F et S sont nécessaires, les analystes n'ont pas à augmenter le nombre de classes, car on ne peut s'attendre à ce qu'il en résulte une amélioration de la précision de l'estimation. À noter qu'une logique semblable vaut pour d'autres contextes de modèles de mélange (Walker, 2007; Si et Reiter, 2013; Manrique-Vallier et Reiter, 2014; Murray et Reiter, 2016).

2.2 Échantillonneur MCMC pour le MDPDM

Hu et coll. (2018) recourent à une stratégie d'augmentation des données (Manrique-Vallier et Reiter, 2014) pour estimer la distribution postérieure en (2.2). Ils posent que les données observées $\mathcal{X}^1$, qui visent tous les ménages possibles, forment un sous-ensemble d'un échantillon hypothétique $\mathcal{X}$ de $(n + n_0)$ ménages directement tiré du MDPDM non tronqué. En d'autres termes, $\mathcal{X}$ est produit sur le support $\mathcal{C}$ où toutes les combinaisons sont possibles et où les règles des zéros structurels ne sont pas appliquées, mais nous observons uniquement l'échantillon de n ménages $\mathcal{X}^1$ qui respectent ces mêmes règles sans observer l'échantillon de n_0 ménages $\mathcal{X}^0 = \mathcal{X} - \mathcal{X}^1$ qui ne les respectent pas.

Nous employons la stratégie de Hu et coll. (2018) et augmentons les données. Pour chaque $h \in \mathcal{H}$, nous simulons $\mathcal{X}$ à partir du MDPDM non tronqué et nous arrêtons lorsque le nombre de ménages possibles simulés dans $\mathcal{X}$ correspond directement à n_{1h} pour tout $h \in \mathcal{H}$. Nous remplaçons par $\mathcal{X}^1$ les ménages possibles simulés dans $\mathcal{X}$ et supposons donc que $\mathcal{X}$ contient déjà $\mathcal{X}^1$, d'où le simple besoin de produire la partie $\mathcal{X}^0$ de $\mathcal{S}$. Dans un tirage de $\mathcal{X}$, nous prenons θ dans la distribution postérieure définie par le

MDPDM non tronqué et traitons $\mathcal{X}$ comme données observées. Il est possible d'estimer cette distribution postérieure à l'aide d'un échantillonneur de Gibbs en bloc (Ishwaran et James, 2001; Si et Reiter, 2013).

Nous présenterons maintenant l'échantillonneur MCMC intégral pour l'ajustement du modèle MDPDM tronqué. Soit $\mathbf{G}^0$ et $\mathbf{M}^0$ des vecteurs des indicateurs d'appartenance aux classes latentes pour les ménages dans $\mathcal{X}^0$ et soit n_{0h} le nombre de ménages de taille h dans $\mathcal{X}^0$ avec $n_0 = \sum_h n_{0h}$. Dans chaque distribution conditionnelle intégrale, soit « – » représentant le conditionnement par l'ensemble des autres variables et paramètres du modèle. À chaque itération MCMC, nous procédons ainsi :

S1. Poser $\mathcal{X}^0 = \mathbf{G}^0 = \mathbf{M}^0 = \varnothing$. Pour chaque $h \in \mathcal{H}$, répéter ce qui suit :

(a) Poser $t_0 = 0$ et $t_1 = 0$.

(b) Échantillonner $G_i^0 \in \{1, \ldots, F\} \sim \text{Discrète}\left(\pi_1^{**}, \ldots, \pi_F^{**}\right)$, où $\pi_g^{**} \propto \lambda_{gh}^{(k)} \pi_g$ et où k est l'indice de « taille du ménage » pour les variables au niveau des ménages.

(c) Pour $j = 1, \ldots, h$, échantillonner $M_{ij}^0 \in \{1, \ldots, S\} \sim \text{Discrète}\left(\omega_{G_i^0 1}, \ldots, \omega_{G_i^0 S}\right)$.

(d) Poser $X_{ik}^0 = h$, où X_{ik}^0 correspond à la variable pour la taille du ménage. Échantillonner le reste des valeurs au niveau des ménages et au niveau des particuliers à l'aide des vraisemblances en (2.3) et (2.4). Fixer à $\mathbf{X}_i^0$ la valeur simulée du ménage.

(e) Si $\mathbf{X}_i^0 \in \mathcal{S}_h$, poser $t_0 = t_0 + 1$, $\mathcal{X}^0 = \mathcal{X}^0 \cup \mathbf{X}_i^0$, $\mathbf{G}^0 = \mathbf{G}^0 \cup G_i^0$ et $\mathbf{M}^0 = \mathbf{M}^0 \cup \{M_{i1}^0, \ldots, M_{ih}^0\}$, sinon poser $t_1 = t_1 + 1$.

(f) Si $t_1 < n_{1h}$, retourner à (b), sinon poser $n_{0h} = t_0$.

S2. Pour les observations en $\mathcal{X}^1$,

(a) Échantillonner $G_i \in \{1, \ldots, F\} \sim \text{Discrète}\left(\pi_1^*, \ldots, \pi_F^*\right)$ pour $i = 1, \ldots, n$, où

$$\pi_g^* = \Pr\left(G_i = g \,|-\right) = \frac{\pi_g \left[\prod_{k=p+1}^{q} \lambda_{gX_{ik}^1}^{(k)} \left(\prod_{j=1}^{n_i} \sum_{m=1}^{S} \omega_{gm} \prod_{k=1}^{p} \phi_{gmX_{ijk}^1}^{(k)}\right)\right]}{\sum_{f=1}^{F} \pi_f \left[\prod_{k=p+1}^{q} \lambda_{fX_{ik}^1}^{(k)} \left(\prod_{j=1}^{n_i} \sum_{m=1}^{S} \omega_{gm} \prod_{k=1}^{p} \phi_{fmX_{ijk}^1}^{(k)}\right)\right]}$$

pour $g = 1, \ldots, F$. Poser $G_i^1 = G_i$.

(b) Échantillonner $M_{ij} \in \{1, \ldots, S\} \sim \text{Discrète}\left(\omega_{G_i^1 1}^*, \ldots, \omega_{G_i^1 S}^*\right)$ pour $i = 1, \ldots, n$ et $j = 1, \ldots, n_i$, où

$$\omega_{G_i^1 m}^* = \Pr\left(M_{ij} = m \,|-\right) = \frac{\omega_{G_i^1 m} \prod_{k=1}^{p} \phi_{G_i^1 m X_{ijk}^1}^{(k)}}{\sum_{s=1}^{S} \omega_{G_i^1 s} \prod_{k=1}^{p} \phi_{G_i^1 s X_{ijk}^1}^{(k)}}$$

pour $m = 1, \ldots, S$. Poser $M_{ij}^1 = M_{ij}$.

S3. Poser $u_F = 1$. Échantillonner

$$u_g \,|- \sim \text{bêta}\left(1 + U_g, \alpha + \sum_{f=g+1}^{F} U_f\right), \quad \pi_g = u_g \prod_{f<g} (1 - u_f)$$

où

$$U_g = \sum_{i=1}^{n} 1\left(G_i^1 = g\right) + \sum_{i=1}^{n_0} 1\left(G_i^0 = g\right)$$

pour $g = 1, \ldots, F-1$.

S4. Poser $v_{gM} = 1$ pour $g = 1, \ldots, F$. Échantillonner

$$v_{gm} \,|- \sim \text{bêta}\left(1 + V_{gm},\ \beta + \sum_{s=m+1}^{S} V_{gs}\right),\quad \omega_{gm} = v_{gm}\prod_{s<m}(1 - v_{gs})$$

où

$$V_{gm} = \sum_{i=1}^{n} 1\left(M_{ij}^1 = m,\, G_i^1 = g\right) + \sum_{i=1}^{n_0} 1\left(M_{ij}^0 = m,\, G_i^0 = g\right)$$

pour $m = 1, \ldots, S-1$ et $g = 1, \ldots, F$.

S5. Échantillonner

$$\lambda_g^{(k)} \,|- \sim \text{Discrète}\left(1 + \eta_{g1}^{(k)}, \ldots, 1 + \eta_{gd_k}^{(k)}\right)$$

où

$$\eta_{gc}^{(k)} = \sum_{i\,|G_i^1 = g}^{n} 1\left(X_{ik}^1 = c\right) + \sum_{i\,|G_i^0 = g}^{n_0} 1\left(X_{ik}^0 = c\right)$$

pour $g = 1, \ldots, F$ et $k = p+1, \ldots, q$.

S6. Échantillonner

$$\phi_{gm}^{(k)} \,|- \sim \text{Discrète}\left(1 + v_{gm1}^{(k)}, \ldots, 1 + v_{gmd_k}^{(k)}\right)$$

où

$$v_{gmc}^{(k)} = \sum_{i,\,j\,|G_i^1 = g,\ M_{ij}^1 = m}^{n} 1\left(X_{ijk}^1 = c\right) + \sum_{i,\,j\,|G_i^0 = g,\ M_{ij}^0 = m}^{n_0} 1\left(X_{ijk}^0 = c\right)$$

pour $g = 1, \ldots, F$, $m = 1, \ldots, S$ et $k = 1, \ldots, p$.

S7. Échantillonner

$$\alpha \,|- \sim \text{gamma}\left(a_\alpha + F - 1,\ b_\alpha - \sum_{g=1}^{F-1}\log\left(1 - u_g\right)\right).$$

S8. Échantillonner

$$\beta \,|- \sim \text{gamma}\left(a_\beta + F \times (S-1),\ b_\beta - \sum_{m=1}^{S-1}\sum_{g=1}^{F}\log\left(1 - v_{gm}\right)\right).$$

Cet échantillonneur de Gibbs est implanté dans le progiciel en R « NestedCategBayesImpute » (Wang, Akande, Hu, Reiter et Barrientos, 2016). Ce logiciel peut servir à produire des versions synthétiques des données initiales, mais encore faut-il que le jeu de données soit complet.

3 Traitement des données manquantes à l'aide du MDPDM

Nous modifions l'échantillonneur de Gibbs pour le modèle MDPDM tronqué en vue d'y incorporer les données manquantes. Pour $i = 1, \ldots, n$, soit $\mathbf{a}_i = (a_{i(p+1)}, \ldots, a_{i(p+q)})$ un vecteur avec $a_{ik} = 1$ lorsque la variable $k \in \{p+1, \ldots, p+q\}$ au niveau des ménages est manquante dans $\mathbf{X}_i^1$ et avec $a_{ik} = 0$ dans les autres cas. Pour $i = 1, \ldots, n$ et $j = 1, \ldots, n_i$, soit $\mathbf{b}_{ij} = (b_{ij1}, \ldots, b_{ijp})$ un vecteur avec $b_{ijk} = 1$ lorsque la variable $k \in \{1, \ldots, p\}$ au niveau des particuliers pour l'individu $j \in \{1, \ldots, n_i\}$ est manquante dans $\mathbf{X}_i^1$ et avec $b_{ijk} = 0$ dans les autres cas. Pour chaque ménage i, soit $\mathbf{X}_i^1 = (\mathbf{X}_i^{\text{obs}}, \mathbf{X}_i^{\text{mis}})$, où $\mathbf{X}_i^{\text{obs}}$ comprend toutes les valeurs des données correspondant à $a_{ik} = 0$ et $b_{ijk} = 0$, et où $\mathbf{X}_i^{\text{mis}}$ comprend toutes les valeurs des données correspondant à $a_{ik} = 1$ et $b_{ijk} = 1$. Nous supposons que les données sont manquantes au hasard (Rubin, 1976).

Pour incorporer les valeurs manquantes à l'échantillonneur de Gibbs, nous devons échantillonner la distribution conditionnelle intégrale de chaque variable dans $\mathbf{X}_i^{\text{mis}}$, conditionnellement aux variables pour lesquelles $a_{ik} = 0$ et $b_{ijk} = 0$, à chaque itération. Nous ajoutons alors une neuvième étape

S9. Pour $i = 1, \ldots, n$, échantillonner $\mathbf{X}_i^{\text{mis}}$ à partir de sa distribution conditionnelle intégrale

$$\Pr\left(\mathbf{X}_i^{\text{mis}} \,|-\right) \propto 1\{\mathbf{X}_i^1 \notin \mathcal{S}_h\}\left(\pi_{G_i^1} \prod_{k\,|\,a_{ik}=1}^{p+q} \lambda_{G_i^1 X_{ik}^1}^{(k)} \prod_{j=1}^{n_i} \omega_{G_i^1 M_{ij}^1} \prod_{k\,|\,b_{ijk}=1}^{p} \phi_{G_i^1 M_{ij}^1 X_{ijk}^1}^{(k)}\right).$$

Il est non trivial d'échantillonner à partir de cette distribution conditionnelle en raison de la dépendance entre les variables qu'induisent les règles des zéros structurels dans chaque $\mathcal{S}_h$. À cause de cette dépendance, il est impossible de simplement échantillonner chaque variable indépendamment à l'aide des vraisemblances en (2.3) et (2.4). Si nous pouvions produire le jeu de données complètes pour tous les ménages à entrées manquantes conditionnellement aux valeurs observées, il serait facile de calculer la probabilité dans chaque cas et d'échantillonner à partir de cet ensemble. Malheureusement, un tel traitement est peu pratique lorsque la taille de chaque $\mathcal{S}_h$ est grande. Même avec une taille modeste, chaque ménage pourrait présenter différents jeux de données complètes, ce qui nécessiterait un calcul, un stockage et un espace de mémoire importants.

Il reste que la distribution conditionnelle intégrale dans S9 prend une forme semblable à celle du noyau du MDPDM tronqué en (2.1), de sorte qu'on puisse produire les échantillons désirés par un second plan d'échantillonnage de rejet. Pour l'essentiel, il s'agit d'échantillonner à partir d'une version non tronquée de la distribution conditionnelle intégrale $P_{\mathbf{X}_i^{\text{mis}}}^* = \pi_{G_i^1} \prod_{k\,|\,a_{ik}=1}^{p+q} \lambda_{G_i^1 X_{ik}^1}^{(k)} \left(\prod_{j=1}^{n_i} \omega_{G_i^1 M_{ij}^1} \prod_{k\,|\,b_{ijk}=1}^{p} \phi_{G_i^1 M_{ij}^1 X_{ijk}^1}^{(k)}\right)$, jusqu'à l'obtention d'un échantillon valide satisfaisant à la relation $\mathbf{X}_i^1 \notin \mathcal{S}_h$; voir en annexe la preuve que ce plan d'échantillonnage de rejet crée un échantillonneur valide de Gibbs. À noter aussi que, comme $P_{\mathbf{X}_i^{\text{mis}}}^*$ lui-même n'est pas tronqué, il est possible de procéder par échantillonnage indépendant de chaque variable en (2.3) et (2.4). Nous remplaçons donc l'étape S9 par S9'.

S9'. Pour $i = 1, \ldots, n,$ échantillonner $\mathbf{X}_i^{\text{mis}}$ de la manière suivante :

(a) Pour chaque variable manquante au niveau des ménages, c'est-à-dire pour chaque variable où $k \in \{p+1, \ldots, p+q\}$ avec $a_{ik} = 1,$ échantillonner X_{ik}^{1} par (2.3).

(b) Pour chaque variable manquante au niveau des particuliers, c'est-à-dire pour chaque variable où $j = 1, \ldots, n_i$ et $k \in \{1, \ldots, p\}$ avec $b_{ijk} = 1,$ échantillonner X_{ijk}^{1} par (2.4).

(c) Fixer à $\mathbf{X}_i^{\text{mis}*}$ les valeurs échantillonnées au niveau des ménages et au niveau des particuliers.

(d) Combiner les $\mathbf{X}_i^{\text{mis}*}$ aux $\mathbf{X}_i^{\text{obs}}$ observés, c'est-à-dire poser $\mathbf{X}_i^{1*} = (\mathbf{X}_i^{\text{obs}}, \mathbf{X}_i^{\text{mis}*})$. Si $\mathbf{X}_i^{1*} \notin \mathcal{S}_h,$ poser $\mathbf{X}_i^{\text{mis}} = \mathbf{X}_i^{\text{mis}*},$ sinon retourner à l'étape (9'a).

Pour amorcer chaque $\mathbf{X}_i^{\text{mis}},$ nous suggérons d'échantillonner à partir de la distribution marginale empirique de chaque variable k en se reportant aux cas disponibles pour chacune et en exigeant que le ménage satisfasse à la relation $\mathbf{X}_i^{1} \notin \mathcal{S}_h$.

4 Stratégies d'accélération de l'échantillonneur MCMC

L'étape de l'échantillonnage de rejet dans l'échantillonneur de Gibbs à la section 2.2 peut se révéler inefficace si $\mathcal{S}$ est grand (Manrique-Vallier et Reiter, 2014; Hu et coll., 2018), parce que l'échantillonneur tend à créer un grand nombre de ménages impossibles avant de donner un nombre suffisant de ménages possibles. Il faut aussi prévoir du temps de calcul pour vérifier si chaque ménage échantillonné respecte ou non toutes les règles des zéros structurels. Les coûts de calcul s'alourdissent quand l'échantillonneur s'incorpore aussi les valeurs manquantes. Dans cette section, nous présentons deux stratégies propres à réduire le nombre de ménages impossibles que crée l'algorithme, d'où une accélération de l'échantillonneur. On trouvera en annexe des études de simulation démontrant que les stratégies en question sont de nature à accélérer nettement l'échantillonneur MCMC.

4.1 Passage du chef du ménage au niveau des ménages

Nombreux sont les jeux de données comprenant une variable du lien de chaque membre du ménage avec le chef du ménage. Ce dernier doit être unique dans un ménage, restriction qui peut rendre compte d'une forte proportion de combinaisons dans $\mathcal{S}$. Comme simple exemple pratique, considérons un jeu de données de $n = 1\,000$ ménages de taille deux pour un total de $N = 2\,000$ particuliers. Posons que les données ne contiennent aucune variable au niveau des ménages et qu'elles comportent deux variables au niveau des particuliers, à savoir l'âge et le lien avec le chef du ménage. Posons également 100 niveaux d'âge et 13 niveaux de lien avec le chef du ménage (chef du ménage, conjoint du chef du ménage, etc.). Ainsi, $\mathcal{C}$ renferme $13^2 \times 100^2 = 1{,}69 \times 10^6$ combinaisons. Posons également que la règle « le chef du ménage doit être unique dans un ménage » est la seule règle de zéros structurels définie pour cet ensemble de données. Ainsi, $\mathcal{S}$ contient $1{,}45 \times 10^6$ combinaisons impossibles, soit environ 86 % de la taille de $\mathcal{C}$. Si le modèle,

par exemple, attribue une probabilité uniforme à toutes les combinaisons dans $\mathcal{C}$, nous nous attendrions à échantillonner quelque $(0{,}86/0{,}14) * 1\,000 \approx 6\,143$ ménages impossibles à chaque itération pour augmenter le n ménages possibles.

Nous traitons plutôt les variables pour le chef du ménage comme caractéristiques au niveau des ménages, ce qui permet d'écarter les règles des zéros structurels définies pour le seul chef du ménage. Dans cet exemple pratique, le passage du chef du ménage au niveau des ménages crée une variable au niveau des ménages, soit l'âge du chef du ménage à 100 niveaux. On peut ne pas tenir compte pour les chefs de ménage de la variable du lien avec le chef du ménage. Pour les autres membres du ménage, la variable du lien avec le chef du ménage compte maintenant 12 niveaux, puisque le niveau correspondant à « chef du ménage » est écarté. Ainsi, $\mathcal{C}$ contient $12 \times 100^2 = 1{,}20 \times 10^5$ combinaisons et $\mathcal{S}$ ne renferme aucune combinaison impossible. Nous n'aurions même plus à échantillonner les ménages impossibles avec l'échantillonneur de Gibbs à la section 2.2.

En règle générale, cette stratégie peut réduire grandement la taille de $\mathcal{S}$, mais sans la ramener d'ordinaire à zéro comme dans l'exemple simple que nous avons cité, puisque $\mathcal{S}$ contient normalement des combinaisons issues d'autres types de règles des zéros structurels. Cette stratégie ne remplace pas non plus l'échantillonneur de rejet à la section 2.2; c'est plutôt une technique de recodage des données qui peut se combiner à l'échantillonneur.

4.2 Fixation d'une borne supérieure au nombre de ménages impossibles en échantillonnage

Pour réduire la durée du calcul, nous pouvons fixer une borne supérieure au nombre de cas échantillonnés dans $\mathcal{X}^0$. Une façon d'y parvenir est de remplacer n_{1h} à l'étape S1(f) dans la section 2.2 par $\lceil n_{1h} \times \psi_h \rceil$ pour un certain ψ_h de sorte que $1/\psi_h$ soit un entier positif. Nous échantillons donc seulement $\lceil n_{0h} \times \psi_h \rceil$ ménages impossibles approximativement pour chaque $h \in \mathcal{H}$. Ce faisant, nous sous-estimons cependant la masse de probabilité réelle attribuée à $\mathcal{S}$ par le modèle. Une illustration possible est l'exemple simple à la section 4.1. Posons que le modèle attribue une probabilité uniforme à toutes les combinaisons dans $\mathcal{C}$ comme plus haut. Nous posons $\psi_2 = 0{,}5$, et, par conséquent, nous échantillonnons environ $3\,072 = \lceil 6\,143 \times 0{,}5 \rceil$ ménages impossibles à chaque itération de l'échantillonneur MCMC. La probabilité de créer un ménage impossible est de $3\,072/(1\,000 + 3\,072) = 0{,}75$; c'est moins que la valeur réelle de 0,86. Nous nous trouverions donc à sous-estimer la contribution réelle de $\{\mathcal{X}^0, \mathbf{G}^0, \mathbf{M}^0\}$ à la vraisemblance.

Pour employer la méthode de plafonnement-pondération, nous devons apporter une correction qui repondère la contribution de $\{\mathcal{X}^0, \mathbf{G}^0, \mathbf{M}^0\}$ à la vraisemblance conjointe intégrale. Nous appliquons à cette fin des idées semblables à celles de Chambers et Skinner (2003) et de Savitsky et Toth (2016), qui approchent la vraisemblance des données inobservées intégrales par une pseudovraisemblance en pondération (les $1/\psi_h$). Les ménages impossibles contribuent uniquement à la vraisemblance conjointe intégrale par les distributions discrètes dans (2.3) à (2.6). Les statistiques suffisantes pour l'estimation des

paramètres des distributions discrètes en question sont les valeurs de dénombrement observées pour les variables correspondantes de l'ensemble $\{\mathcal{X}^1, \mathbf{G}^1, \mathbf{M}^1, \mathcal{X}^0, \mathbf{G}^0, \mathbf{M}^0\}$, et ce, dans chaque classe latente pour les variables au niveau des ménages et dans chaque paire de classes latentes pour les variables au niveau des particuliers. Pour chaque $h \in \mathcal{H}$, nous pouvons donc repondérer la contribution des ménages impossibles en multipliant par $1/\psi_h$ les valeurs observées de dénombrement des ménages de taille h dans $\{\mathcal{X}^0, \mathbf{G}^0, \mathbf{M}^0\}$ pour la variable correspondante et les classes latentes. On élève de la sorte la contribution en vraisemblance des ménages impossibles de taille h à la puissance de $1/\psi_h$. Bien sûr, $1/\psi_h$ n'a pas à être un entier positif, nous avons cette exigence seulement pour que la valeur de sa multiplication avec les valeurs observées de dénombrement soit sans décimales. Nous modifions l'échantillonneur de Gibbs et y incorporons la méthode de plafonnement-pondération en remplaçant les étapes S1, S3, S4, S5 et S6 (voir les étapes modifiées en annexe).

Poser chaque $\psi_h = 1$ correspond à poser l'échantillonneur de rejet au départ et, par là, les deux méthodes devraient donner des résultats très ressemblants quand ψ_h approche de 1. Selon notre expérience, les résultats de la méthode de plafonnement-pondération deviennent nettement moins précis que l'échantillonneur de rejet proprement dit quand $\psi_h < 1/4$. Le temps épargné grâce à cette technique d'accélération par rapport à l'échantillonneur classique dépend des caractéristiques des données et des valeurs spécifiées pour la pondération $\{\psi_h : h \in \mathcal{H}\}$. Nous suggérons pour la sélection des ψ_h d'essayer différentes valeurs en commençant par des valeurs proches de l'unité et en faisant porter les passages initiaux de l'échantillonneur MCMC sur un petit échantillon aléatoire des données. Les analystes devraient examiner la convergence et le comportement de mélange des chaînes à comparer à la chaîne avec tous les ψ_h correspondant à l'unité. Ils devraient choisir des valeurs assurant une accélération raisonnable, tout en préservant la convergence et le mélange. C'est ce que l'on peut faire rapidement en comparant les tracés temporels d'un jeu aléatoire de paramètres du modèle hors commutation d'étiquettes (comme α et β) ou en examinant les probabilités marginales, bivariées et trivariées estimées à l'aide de données synthétiques produites par l'échantillonneur MCMC.

5 Étude empirique

Pour évaluer le rendement du MDPDM comme méthode d'imputation, ainsi que les stratégies d'accélération, nous utilisons les fichiers de microdonnées à grande diffusion de l'*American Community Survey* (ACS) de 2012, lesquels peuvent être téléchargés du site du Census Bureau (http://www2.census.gov/acs2012_1yr/pums/). Nous construisons une population de 764 580 ménages de taille $\mathcal{H} = \{2, 3, 4\}$ sur laquelle nous échantillonnons $n = 5\,000$ ménages comprenant $N = 13\,181$ particuliers. Nous travaillons avec les variables décrites au tableau 5.1 qui imitent celles du recensement décennal des États-Unis. Les zéros structurels sont ceux de l'âge et des liens entre les particuliers partageant un logement (voir en annexe toute la liste des règles que nous avons appliquées). Nous faisons passer le chef du ménage au niveau des ménages comme à la section 4.1 pour réaliser des gains de calcul.

Nous introduisons les valeurs manquantes à l'aide du scénario qui suit. Nous mettons la taille du ménage et l'âge du chef du ménage en observation intégrale. Nous mettons en blanc aléatoirement et indépendamment 30 % de chaque variable dans le cas des variables restantes au niveau des ménages. Pour les particuliers autres que le chef du ménage, nous mettons en blanc aléatoirement et indépendamment 30 % des valeurs du sexe, de la race et de l'origine hispanique. Nous mettons l'âge en valeurs manquantes à des taux de 50 %, 20 %, 40 % et 30 % pour la variable des liens dans les ensembles {2}, {3, 4, 5, 10}, {7, 9} et {6, 8, 11, 12, 13} respectivement. Nous mettons en valeurs manquantes la variable des liens à des taux de 40 %, 25 %, 10 % et 55 % pour la variable de l'âge dans les ensembles $\{x: x \leq 20\}$, $\{x: 20 < x \leq 50\}$, $\{x: 50 < x \leq 70\}$ et $\{x: x > 70\}$ respectivement. Nous obtenons ainsi une proportion approximative de 30 % de valeurs manquantes pour les deux variables. Les valeurs sont manquantes pour environ 8 % des particuliers de l'échantillon dans la double variable de l'âge et des liens et pour 2 % dans les variables du sexe, de l'âge et des liens conjointement. Ce mécanisme produit des données qui, d'un point de vue technique, ne sont pas manquantes au hasard, mais nous employons le modèle MDPDM de toute manière pour voir les possibilités dans le cas d'un mécanisme compliqué de valeurs manquantes. Les taux réels de non-réponse partielle dans les données du recensement sont généralement inférieurs à ceux que nous utilisons ici, mais nous employons des taux élevés pour soumettre le MDPDM à un test éprouvant de résistance. Nous introduisons aussi les valeurs manquantes à l'aide d'un scénario de données manquantes entièrement au hasard avec des taux de l'ordre de 10 % pour toutes les variables. Bref, les résultats seront semblables à ceux que nous aurons ici, mais en plus exacts en raison de proportions inférieures de valeurs manquantes. On peut examiner ces résultats en annexe.

Tableau 5.1
Description des variables de l'étude (où CM est l'abréviation de chef du ménage)

Description des variables		Catégories
Variables au niveau des ménages	Propriété du logement	1 = propriété ou achat, 2 = location
	Taille du ménage	2 = 2 membres, 3 = 3 membres, 4 = 4 membres
	Sexe du CM	1 = homme, 2 = femme
	Race du CM	1 = blanc, 2 = noir 3 = Amérindien ou natif de l'Alaska 4 = Chinois, 5 = Japonais 6 = autre Asiatique/insulaire du Pacifique, 7 = autre race 8 = deux races principales 9 = trois races principales ou plus
	Origine hispanique du CM	1 = non-Hispanique, 2 = Mexicain 3 = Portoricain, 4 = Cubain, 5 = autre
	Âge du CM	1 = moins d'un an, 2 = 1 an 3 = 2 ans... 96 = 95 ans
Variables au niveau des particuliers	Sexe	Même que « sexe du CM »
	Race	Même que « race du CM »
	Origine hispanique	Même que « origine hispanique du CM »
	Âge	Même que « âge du CM »
	Lien avec le chef du ménage	1 = conjoint, 2 = enfant biologique 3 = enfant adoptif, 4 = enfant par alliance, 5 = frère ou sœur 6 = parent, 7 = petit-enfant, 8 = parent par alliance 9 = enfant par alliance, 10 = autre apparenté 11 = pensionnaire, colocataire ou partenaire 12 = autre enfant non apparenté ou en foyer nourricier

Nous estimons le MDPDM au moyen de deux méthodes avec dans chaque cas l'étape d'échantillonnage de rejet S9' à la section 3. Dans la première, nous considérons $\psi_2 = \psi_3 = \psi_4 = 1$ sans employer la méthode de plafonnement-pondération et, dans la seconde, nous prenons $\psi_2 = \psi_3 = 1/2$ et $\psi_4 = 1/3$. Dans chacune, nous exécutons 10 000 itérations de l'échantillonneur MCMC, écartant les 5 000 premières comme rodage et élaguant les échantillons restants à toutes les cinq itérations, ce qui donne 1 000 itérations MCMC après rodage. Nous posons $F = 30$ et $S = 15$ pour chaque méthode en fonction des passages initiaux de rodage. Pour les deux méthodes, le nombre effectif de grappes occupées au niveau des ménages va ordinairement de 13 à 16 avec 25 comme maximum; le nombre effectif de grappes occupées au niveau des particuliers va de 3 à 5 avec 10 comme maximum à l'échelle des grappes des ménages. Pour le contrôle de convergence, nous avons examiné les tracés temporels de α, β, et les moyennes pondérées d'un échantillon aléatoire des probabilités multinomiales en (2.3) et (2.4) (puisque les probabilités multinomiales mêmes sont sujettes à commutation d'étiquettes).

Avec les deux méthodes, nous produisons $L = 50$ jeux de données complètes, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(50)})$, au moyen de la distribution prédictive postérieure du MDPDM par laquelle nous estimons toutes les distributions marginales, les distributions bivariées de toutes les paires possibles de variables et les distributions trivariées de tous les trios possibles de variables. Nous estimons également plusieurs probabilités qui dépendent des liens au sein du ménage et du chef du ménage pour juger du rendement du MDPDM dans l'estimation de relations complexes. Nous obtenons des intervalles de confiance par inférence d'imputation multiple (Rubin, 1987). Reprenons brièvement la marche à suivre. Soit q l'estimateur ponctuel de données complètes pour un certain paramètre Q et soit u l'estimateur de variance lié à q. Pour $l = 1, \ldots, L$, soit $q^{(l)}$ et $u^{(l)}$ les valeurs de q et u dans le jeu de données complètes $\mathbf{Z}^{(l)}$. Nous utilisons $\bar{q}_L = \sum_{l=1}^{L} q^{(l)} \big/ L$ comme estimation ponctuelle de Q. Nous employons $T_L = (1 + 1/L)\, b_L + \bar{u}_L$ comme variance estimée de $\bar{q}$, où $b_L = \sum_{l=1}^{L} (q^{(l)} - \bar{q}_L)^2 \big/ (L-1)$ et $\bar{u}_L = \sum_{l=1}^{L} u^{(l)} \big/ L$. Nous inférons au sujet de Q par $(\bar{q}_L - Q) \sim t_v (0, T_L)$, où t_v est une distribution t avec $v = (L-1)\,(1 + \bar{u}_L / [(1 + 1/L)\, b_L])^2$ degrés de liberté.

Les figures 5.1 et 5.2 présentent la valeur de $\bar{q}_{50}$ pour les diverses probabilités marginales, bivariées et trivariées estimées et tracées par rapport à leur estimation correspondante du jeu de données initiales sans valeurs manquantes. La figure 5.1 indique les résultats du MDPDM avec l'échantillonneur de rejet et la figure 5.2, les résultats correspondants avec la méthode de plafonnement-pondération. Dans les deux cas, les estimations ponctuelles sont proches de celles des données avant introduction des valeurs manquantes, d'où l'impression que le MDPDM réussit vraiment à appréhender les caractéristiques importantes de la distribution conjointe des variables. La figure 5.2 en particulier montre aussi que la méthode de plafonnement-pondération ne vient pas dégrader les estimations.

Le tableau 5.2 fait état des intervalles de confiance à 95 % de plusieurs probabilités concernant les liens au sein du ménage, ainsi que de la valeur de la population entière de 764 580 ménages. Il s'agit des deux intervalles venant des moteurs d'imputation MDPDM et de l'intervalle venant des données avant

introduction des valeurs manquantes. Dans ce dernier cas, nous utilisons l'intervalle habituel de Wald, $\hat{p} \pm 1{,}96 \sqrt{\hat{p}(1-\hat{p})/n}$, où $\hat{p}$ est le pourcentage de l'échantillon correspondant. Le plus souvent, les intervalles du modèle MDPDM avec échantillonnage de rejet intégral sont proches des intervalles des données sans valeurs manquantes. Ils comprennent généralement la quantité vraie de la population. Le moteur d'imputation MDPDM crée un biais par défaut perceptible pour les pourcentages des ménages de race homogène et le biais s'accroît avec la taille du ménage. C'est là un paramètre difficile à estimer exactement par imputation, plus particulièrement dans le cas des ménages de plus grande taille. Hu et coll. (2018) relèvent des biais par défaut lorsqu'on utilise le MDPDM (avec traitement des données du chef du ménage comme variables au niveau des particuliers) pour produire des données synthétiques intégrales en faisant observer que le biais diminue à mesure qu'augmente la taille de l'échantillon. Le MDPDM s'ajuste de mieux en mieux à la distribution conjointe à mesure que progresse la taille de l'échantillon. Nous nous attendons, par conséquent, à ce que le moteur d'imputation MDPDM gagne en précision à mesure qu'augmente cette taille et que diminue la proportion de valeurs manquantes.

Les estimations par intervalle sont généralement semblables dans la méthode de plafonnement-pondération et dans l'échantillonnage de rejet intégral avec une certaine dégradation en particulier pour les pourcentages de ménages de race homogène selon la taille du ménage. Cette dégradation ne va pas sans un côté positif : à en juger par les passages de l'échantillonneur MCMC dans un portable standard, le MDPDM est environ de 42 % plus rapide avec la méthode de plafonnement-pondération et le passage du chef du ménage au niveau des ménages qu'avec le seul traitement des données du chef du ménage au niveau des ménages.

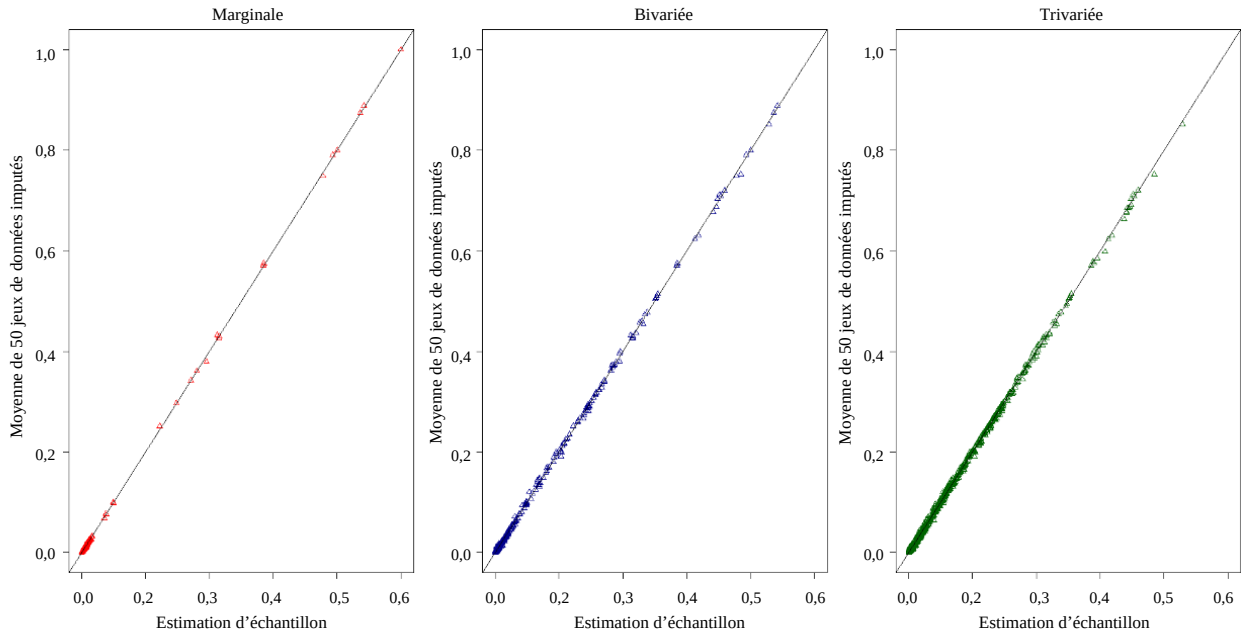

Figure 5.1 Probabilités marginales, bivariées et trivariées calculées dans l'échantillon et jeux de données imputées par le MDPDM tronqué avec échantillonneur de rejet et passage des données du chef du ménage au niveau des ménages.

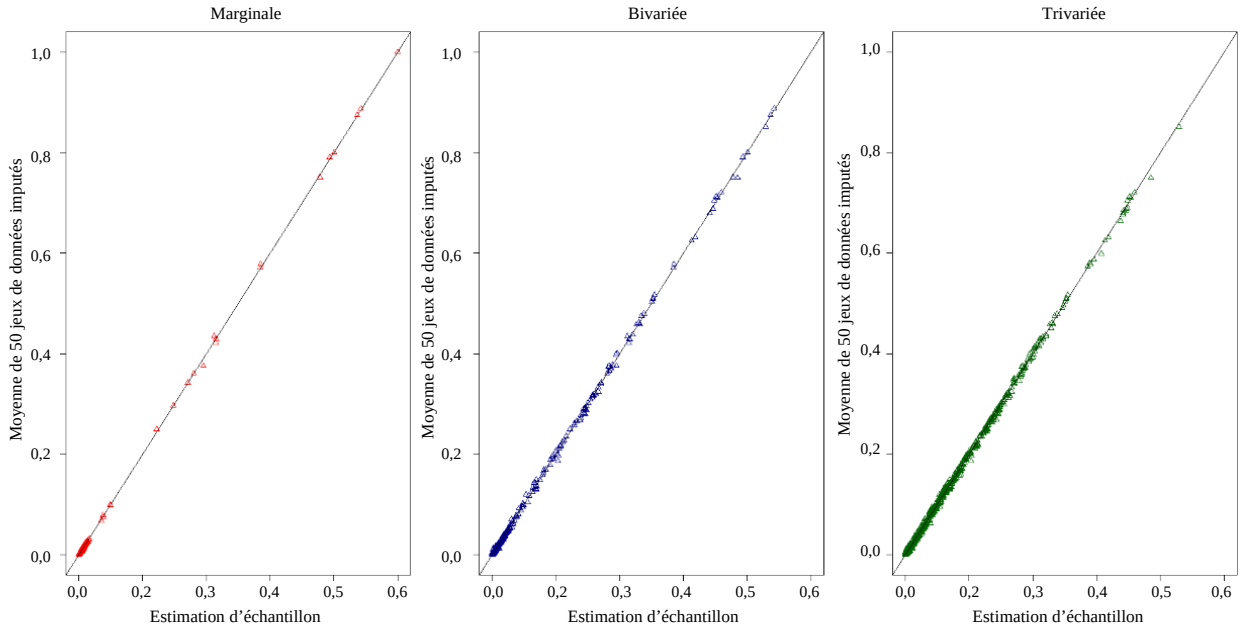

Figure 5.2 Probabilités marginales, bivariées et trivariées calculées dans l'échantillon et jeux de données imputées par le MDPDM tronqué avec méthode de plafonnement-pondération et passage des données du chef du ménage au niveau des ménages.

Tableau 5.2
Intervalles de confiance pour certaines probabilités qui dépendent des liens au sein du ménage dans les jeux de données initiales et imputées. La mention « Sans données manquantes » vise les données échantillonnées avant l'introduction des valeurs manquantes. La mention « MDPDM » vise le MDPDM tronqué avec passage des données du chef du ménage au niveau des ménages. La mention « MDPDM plafonné » vise le MDPDM tronqué avec méthode de plafonnement-pondération et passage des données du chef du ménage au niveau des ménages. La mention « *Q* » est la valeur de toute la population de 764 580 ménages

	Q	Sans données manquantes	MDPDM	MDPDM plafonné
Ménage de race homogène : $n_i = 2$	0,942	(0,932; 0,949)	(0,891; 0,917)	(0,884; 0,911)
$n_i = 3$	0,908	(0,907; 0,937)	(0,843; 0,890)	(0,821; 0,870)
$n_i = 4$	0,901	(0,879; 0,917)	(0,793; 0,851)	(0,766; 0,828)
Conjoint présent	0,696	(0,682; 0,707)	(0,695; 0,722)	(0,695; 0,722)
Couple de même race	0,656	(0,641; 0,668)	(0,640; 0,669)	(0,634; 0,664)
Conjoint présent, chef du ménage blanc	0,600	(0,589; 0,616)	(0,603; 0,632)	(0,604; 0,634)
Couple blanc	0,580	(0,569; 0,596)	(0,577; 0,606)	(0,574; 0,604)
Couple avec différence d'âge de moins de cinq ans	0,488	(0,465; 0,492)	(0,341; 0,371)	(0,324; 0,355)
Chef du ménage de sexe masculin, propriétaire du logement	0,476	(0,456; 0,484)	(0,450; 0,479)	(0,451; 0,480)
Chef du ménage de plus de 35 ans, aucun enfant présent	0,462	(0,441; 0,468)	(0,442; 0,470)	(0,443; 0,471)
Au moins un enfant biologique présent	0,437	(0,431; 0,458)	(0,430; 0,459)	(0,428; 0,456)
Chef du ménage plus âgé que le conjoint, blanc	0,322	(0,309; 0,335)	(0,307; 0,339)	(0,311; 0,343)
Femme d'âge adulte avec au moins un enfant de moins de 5 ans	0,078	(0,070; 0,085)	(0,062; 0,078)	(0,061; 0,077)
Chef du ménage blanc d'origine hispanique	0,066	(0,064; 0,078)	(0,062; 0,079)	(0,062; 0,078)
Couple non blanc, propriétaire du logement	0,058	(0,050; 0,063)	(0,038; 0,052)	(0,037; 0,051)
Deux générations présentes, chef du ménage noir	0,057	(0,053; 0,066)	(0,052; 0,066)	(0,052; 0,067)
Chef du ménage noir, propriétaire du logement	0,052	(0,046; 0,058)	(0,044; 0,058)	(0,044; 0,059)
Conjoint présent, chef du ménage noir	0,039	(0,032; 0,042)	(0,032; 0,044)	(0,031; 0,043)
Couple formé d'un Blanc et d'un non-Blanc	0,034	(0,029; 0,039)	(0,038; 0,053)	(0,043; 0,059)
Chef du ménage d'origine hispanique de plus de 50 ans, propriétaire du logement	0,029	(0,025; 0,034)	(0,023; 0,034)	(0,024; 0,034)
Un petit-enfant présent	0,028	(0,023; 0,033)	(0,024; 0,035)	(0,023; 0,035)
Femme noire d'âge adulte avec au moins un enfant de moins de 18 ans	0,027	(0,028; 0,038)	(0,025; 0,036)	(0,025; 0,036)
Au moins deux générations présentes, couple d'origine hispanique	0,027	(0,022; 0,031)	(0,022; 0,032)	(0,023; 0,033)

Tableau 5.2 (suite)
Intervalles de confiance pour certaines probabilités qui dépendent des liens au sein du ménage dans les jeux de données initiales et imputées. La mention « Sans données manquantes » vise les données échantillonnées avant l'introduction des valeurs manquantes. La mention « MDPDM » vise le MDPDM tronqué avec passage des données du chef du ménage au niveau des ménages. La mention « MDPDM plafonné » vise le MDPDM tronqué avec méthode de plafonnement-pondération et passage des données du chef du ménage au niveau des ménages. La mention « *Q* » est la valeur de toute la population de 764 580 ménages

	Q	Sans données manquantes	MDPDM	MDPDM plafonné
Couple d'origine hispanique avec au moins un enfant biologique	0,025	(0,020; 0,028)	(0,019; 0,029)	(0,020; 0,030)
Au moins trois générations présentes	0,023	(0,020; 0,028)	(0,017; 0,026)	(0,017; 0,026)
Un seul parent	0,020	(0,016; 0,024)	(0,013; 0,021)	(0,013; 0,021)
Au moins un enfant par alliance	0,019	(0,018; 0,026)	(0,019; 0,030)	(0,019; 0,030)
Homme adulte d'origine hispanique avec au moins un enfant de moins de 10 ans	0,018	(0,017; 0,025)	(0,014; 0,022)	(0,014; 0,022)
Au moins un enfant adoptif, couple blanc	0,008	(0,005; 0,010)	(0,004; 0,010)	(0,004; 0,011)
Couple noir avec au moins deux enfants biologiques	0,006	(0,003; 0,007)	(0,003; 0,007)	(0,003; 0,007)
Chef du ménage noir de moins de 40 ans, propriétaire du logement	0,005	(0,005; 0,009)	(0,006; 0,013)	(0,007; 0,013)
Trois générations présentes, couple blanc	0,005	(0,004; 0,008)	(0,004; 0,010)	(0,004; 0,009)
Chef du ménage blanc de moins de 25 ans, propriétaire du logement	0,003	(0,002; 0,005)	(0,003; 0,007)	(0,003; 0,007)

6 Analyse

L'étude empirique fait voir que le MDPDM peut permettre une imputation de grande qualité de données catégoriques emboîtées dans les ménages. Autant que nous sachions, nous avons là le premier moteur d'imputation paramétrique applicable à des données catégoriques multivariées emboîtées. L'étude fait également voir que les organismes ne devraient pas s'attendre avec des tailles d'échantillon modestes à ce que le MDPDM préserve toutes les caractéristiques de la distribution conjointe. Bien sûr, tel est le cas pour tout moteur d'imputation. Dans le cas du MDPDM, la possibilité existe pour les organismes d'améliorer la précision des quantités visées en recodant les données servant à l'ajustement du modèle. Par exemple, on peut établir au niveau des ménages une nouvelle variable égale à un quand la race est homogène et à zéro autrement et remplacer la variable de la race individuelle par une nouvelle variable de niveau 1 si la race est la même que celle du chef du ménage, 2 si la race est blanche et diffère de la race du chef du ménage et 3 si la race est noire et diffère de la race du chef du ménage, et ainsi de suite. On estimerait le modèle MDPDM avec la même variable de la race au niveau des ménages et la nouvelle variable correspondante au niveau des particuliers. Le MDPDM pourrait ainsi estimer avec précision les proportions d'une même race, puisqu'il s'agirait là seulement d'une autre variable au niveau des ménages comme la propriété du logement. On ajouterait aussi au calcul des zéros structurels pour la race. Un thème possible pour de futures recherches serait l'évaluation de l'équilibre à trouver entre la fidélité et les coûts en calcul de tels recodages.

Le MDPDM peut être vorace en calcul même avec les mesures d'accélération que nous avons présentées. La voracité de l'algorithme tient aux étapes d'échantillonnage de rejet. Heureusement, ces étapes peuvent facilement s'accomplir dans un traitement informatique en parallèle. Nous pourrions commander, par exemple, à chaque processeur de produire une fraction des cas impossibles de la section 2.2. Nous pourrions aussi étaler sur un grand nombre de processeurs les étapes de rejet pour l'imputation. Ces mesures devraient réduire la durée du calcul par un facteur correspondant en gros au nombre de processeurs disponibles.

L'étude empirique a porté sur les ménages jusqu'à la taille de quatre membres. Nous avons exécuté le modèle dans un temps raisonnable (en quelques heures sur un portable standard) avec des ménages jusqu'à la taille de sept membres. Les résultats de précision sont semblables qualitativement. À mesure qu'augmente la taille des ménages, le modèle peut produire des centaines, voire des milliers de fois de plus de ménages impossibles que de ménages possibles, ce qui ralentit l'algorithme. Dans ce cas, la méthode de plafonnement-pondération est essentielle au caractère pratique des applications.

Remerciements

Cette étude a reçu des subventions de la *National Science Foundation* (NSF SES 1131897) et de l'*Alfred P. Sloan Foundation* (G-2-15-20166003).

Annexe

Voici une annexe où nous démontrons que l'étape S9' d'échantillonnage de rejet à la section 3 produit des échantillons à partir de la bonne distribution postérieure. Nous y présentons aussi l'échantillonneur modifié de Gibbs pour la méthode de plafonnement-pondération, ainsi qu'une liste des règles de zéros structurels servant à l'ajustement du modèle MDPDM. Nous livrons enfin les résultats empiriques des mesures d'accélération mentionnées dans notre étude avec des données synthétiques, ainsi que des résultats supplémentaires dans le traitement des données manquantes par le MDPDM selon un scénario de valeurs manquantes entièrement au hasard.

A.1 Preuve que l'étape S9' d'échantillonnage de rejet à la section 3 tire des échantillons de la bonne distribution postérieure

Les valeurs X_{ik}^{1} et X_{ijk}^{1} produites par l'échantillonneur de rejet à l'étape S9' viennent de distributions postérieures intégrales, d'où la validité de l'échantillonneur de Gibbs. La preuve découle des propriétés de l'échantillonnage de rejet (ou d'un simple mode acceptation-rejet). La distribution visée est la distribution conditionnelle intégrale pour $\mathbf{X}_i^{\text{mis}}$. Elle peut être ainsi reformulée :

$$p\left(\mathbf{X}_i^{\text{mis}}\right) = \frac{1\left\{\mathbf{X}_i^{1} \notin \mathcal{S}_h\right\}}{\Pr\left(\mathbf{X}_i \notin \mathcal{S}_h \mid \theta\right)} g\left(\mathbf{X}_i^{\text{mis}}\right)$$

où

$$g\left(\mathbf{X}_i^{\text{mis}}\right) = \pi_{G_i^1} \prod_{k \mid a_{ik}=1}^{p+q} \lambda_{G_i^1 X_{ik}^1}^{(k)} \left(\prod_{j=1}^{n_i} \omega_{G_i^1 M_{ij}^1} \prod_{k \mid b_{ijk}=1}^{p} \phi_{G_i^1 M_{ij}^1 X_{ijk}^1}^{(k)} \right).$$

Dans notre plan d'échantillonnage de rejet, nous employons $g\left(\mathbf{X}_i^{\text{mis}}\right)$ comme proposition pour $p\left(\mathbf{X}_i^{\text{mis}}\right)$. Pour bien montrer que les tirages viennent vraiment de $p\left(\mathbf{X}_i^{\text{mis}}\right)$, nous devons vérifier que

$w(\mathbf{X}_i^{\text{mis}}) = p(\mathbf{X}_i^{\text{mis}})/g(\mathbf{X}_i^{\text{mis}}) < M$, où $1 < M < \infty$, et que nous acceptons chaque échantillon avec la probabilité $w(\mathbf{X}_i^{\text{mis}})/M$. Dans notre cas,

1. $w(\mathbf{X}_i^{\text{mis}}) = p(\mathbf{X}_i^{\text{mis}})/g(\mathbf{X}_i^{\text{mis}}) = 1\{\mathbf{X}_i^1 \notin \mathcal{S}_h\}/\Pr(\mathbf{X}_i \notin \mathcal{S}_h \mid \theta) \leq 1/\Pr(\mathbf{X}_i \notin \mathcal{S}_h \mid \theta)$, et $0 < \Pr(\mathbf{X}_i \notin \mathcal{S}_h \mid \theta) < 1 \Rightarrow 1 < 1/\Pr(\mathbf{X}_i \notin \mathcal{S}_h \mid \theta) < \infty$ nécessairement.
2. En prélevant jusqu'à obtention d'un échantillon valide satisfaisant à la relation $\mathbf{X}_i^1 \notin \mathcal{S}_h$, nous échantillonnons en réalité avec la probabilité $w(\mathbf{X}_i^{\text{mis}})/M = 1\{\mathbf{X}_i^1 \notin \mathcal{S}_h\}$.

A.2 Échantillonneur modifié de Gibbs pour la méthode de plafonnement-pondération

L'échantillonneur modifié de Gibbs pour la méthode de plafonnement-pondération remplace de la manière suivante les étapes S1, S3, S4, S5 et S6 dans le corps du texte.

S1*. Pour chaque $h \in \mathcal{H}$, reprendre les étapes S1(a) à S1(e) comme plus haut, mais modifier l'étape S1(f) : si $t_1 < \lceil n_{1h} \times \psi_h \rceil$, retourner à l'étape (b), sinon poser $n_{0h} = t_0$.

S3*. Poser $u_F = 1$. Échantillonner

$$u_g \mid - \sim \text{bêta}\left(1 + U_g, \alpha + \sum_{f=g+1}^{F} U_f\right), \quad \pi_g = u_g \prod_{f<g}(1 - u_f)$$

où

$$U_g = \sum_{i=1}^{n} 1(G_i^1 = g) + \sum_{h \in \mathcal{H}} \frac{1}{\psi_h} \sum_{i \mid n_i^0 = h} 1(G_i^0 = g)$$

pour $g = 1, \ldots, F - 1$.

S4*. Poser $v_{gM} = 1$ pour $g = 1, \ldots, F$. Échantillonner

$$v_{gm} \mid - \sim \text{bêta}\left(1 + V_{gm}, \beta + \sum_{s=m+1}^{S} V_{gs}\right), \quad \omega_{gm} = v_{gm} \prod_{s<m}(1 - v_{gs})$$

où

$$V_{gm} = \sum_{i=1}^{n} 1\left(M_{ij}^1 = m, G_i^1 = g\right) + \sum_{h \in \mathcal{H}} \frac{1}{\psi_h} \sum_{i \mid n_i^0 = h} 1\left(M_{ij}^0 = m, G_i^0 = g\right)$$

pour $m = 1, \ldots, S - 1$ et $g = 1, \ldots, F$.

S5*. Échantillonner

$$\lambda_g^{(k)} \mid - \sim \text{Discrète}\left(1 + \eta_{g1}^{(k)}, \ldots, 1 + \eta_{gd_k}^{(k)}\right)$$

où

$$\eta_{gc}^{(k)} = \sum_{i \mid G_i^1 = g}^{n} 1\left(X_{ik}^1 = c\right) + \sum_{h \in \mathcal{H}} \frac{1}{\psi_h} \sum_{i \mid n_i^0 = h,\ G_i^0 = g} 1\left(X_{ik}^0 = c\right)$$

pour $g = 1, \ldots, F$ et $k = p+1, \ldots, q$.

S6*. Échantillonner

$$\phi_{gm}^{(k)} \,\big|\, - \sim \text{Discrète}\left(1 + \nu_{gm1}^{(k)}, \ldots, 1 + \nu_{gmd_k}^{(k)}\right)$$

où

$$\nu_{gmc}^{(k)} = \sum_{i\,|\,G_i^1 = g,\ M_{ij}^1 = m}^{n} 1\left(X_{ijk}^{\,1} = c\right) + \sum_{h \in \mathcal{H}} \frac{1}{\psi_h} \sum_{i\,|\,n_i^0 = h,\ G_i^0 = g,\ M_{ij}^0 = m} 1\left(X_{ijk}^{\,0} = c\right)$$

pour $g = 1, \ldots, F$, $m = 1, \ldots, S$ et $k = 1, \ldots, p$.

A.3 Liste des zéros structurels

Nous ajustons le modèle MDPDM avec des zéros structurels pour l'âge et les liens entre les particuliers partageant le logement. La liste complète des règles appliquées figure au tableau A.1. Celles-ci sont tirées de l'ACS de 2012 par relevé des combinaisons relatives à la variable des liens ne figurant pas dans la population construite. Il ne faut pas y voir une liste « réelle » des combinaisons impossibles dans les données du recensement.

Tableau A.1
Liste des zéros structurels

Description
Règles communes à la production des jeux de données tant synthétiques qu'imputées.
1. Le chef du ménage doit être unique dans chaque ménage et âgé d'au moins 16 ans.
2. Chaque ménage ne peut avoir plus d'un conjoint et celui-ci doit être âgé d'au moins 16 ans.
3. Les membres des couples mariés sont de sexe opposé et leur différence d'âge ne peut être de plus de 49 ans.
4. Le parent le plus jeune doit être plus âgé que le chef du ménage d'au moins 4 ans.
5. Le parent par alliance le plus jeune doit être plus âgé que le chef du ménage d'au moins 4 ans.
6. La différence d'âge entre le chef du ménage et les frères et sœurs ne peut être de plus de 37 ans.
7. Le chef du ménage doit être âgé d'au moins 31 ans pour être grand-père ou grand-mère et son conjoint doit être âgé d'au moins 17 ans. Il faut aussi que le conjoint soit plus âgé que le petit-enfant le plus vieux d'au moins 26 ans.
Règles de production des jeux de données synthétiques.
8. Le chef du ménage doit être plus âgé que l'enfant le plus vieux d'au moins 7 ans.
Règles de production des jeux de données imputées.
9. Le chef du ménage doit être plus âgé que l'enfant biologique le plus vieux d'au moins 7 ans.
10. Le chef du ménage doit être plus âgé que l'enfant adoptif le plus vieux d'au moins 11 ans.
11. Le chef du ménage doit être plus âgé que l'enfant par alliance le plus vieux d'au moins 9 ans.

A.4 Étude empirique des méthodes d'accélération

Nous évaluons le rendement des deux méthodes d'accélération mentionnées dans le corps du texte à l'aide de données synthétiques. Nous prenons les fichiers de microdonnées à grande diffusion de l'ACS de 2012, qui sont téléchargeables du site du Census Bureau http://www2.census.gov/acs2012_1yr/pums/, pour construire une population de 857 018 ménages de taille $\mathcal{H} = \{2, 3, 4, 5, 6\}$, sur laquelle nous échantillonnons $n = 10\,000$ ménages comprenant $N = 29\,117$ particuliers. Nous travaillons avec les

variables décrites au tableau A.2. Nous évaluons les méthodes par des probabilités qui dépendent des liens au sein du ménage et du chef du ménage.

Tableau A.2
Description des variables utilisées dans l'illustration par données synthétiques

Description des variables		Catégories
Variables au niveau des ménages	Propriété du logement	1 = propriété ou achat, 2 = location
	Taille du ménage	2 = 2 membres, 3 = 3 membres, 4 = 4 membres 5 = 5 membres, 6 = 6 membres
Variables au niveau des particuliers	Sexe	1 = homme, 2 = femme
	Race	1 = Blanc, 2 = Noir 3 = Amérindien ou natif de l'Alaska 4 = Chinois, 5 = Japonais 6 = autre Asiatique/insulaire du Pacifique, 7 = autre race 8 = deux races principales 9 = trois races principales ou plus
	Origine hispanique	1 = Non-Hispanique, 2 = Mexicain 3 = Portoricain, 4 = Cubain, 5 = autre
	Âge	1 = moins d'un an, 2 = 1 an 3 = 2 ans... 96 = 95 ans
	Lien avec le chef du ménage	1 = chef du ménage, 2 = conjoint, 3 = enfant 4 = enfant par alliance, 5 = parent, 6 = parent par alliance 7 = frère ou sœur, 8 = frère ou sœur par alliance, 9 = petit-enfant 10 = autre apparenté, 11 = partenaire/ami/visiteur 12 = autre non apparenté

Nous considérons le MDPDM avec deux méthodes et, dans chaque cas, passage des données du chef du ménage au niveau des ménages comme à la section 4.1 dans le corps du texte, de même qu'avec la méthode de plafonnement-pondération à la section 4.2. Pour la première méthode, nous prenons $\psi_2 = \psi_3 = \psi_4 = \psi_5 = \psi_6 = 1$ et, pour la seconde, $\psi_2 = \psi_3 = 1/2$ et $\psi_4 = \psi_5 = \psi_6 = 1/3$. Nous comparons ces méthodes au MDPDM présenté dans Hu et coll., 2018. Dans chaque cas, nous créons $L = 50$ jeux de données synthétiques, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(50)})$. Nous produisons ces jeux de sorte que le nombre de ménages de taille $h \in \mathcal{H}$ dans chaque $\mathbf{Z}^{(l)}$ corresponde exactement à n_h dans les données observées. Ainsi, $\mathbf{Z}$ comprend des données partiellement synthétiques (Little, 1993; Reiter, 2003), bien que chaque Z_{ijk} produit soit une valeur simulée. Nous combinons les estimations en employant la méthode dans Reiter (2003). Pour résumer, soit q l'estimateur ponctuel d'un certain paramètre Q et soit u l'estimateur de variance lié à q. Pour $l = 1, \ldots, L$, soit q_l et u_l les valeurs de q et u dans le jeu de données synthétiques $\mathbf{Z}^{(l)}$. Nous utilisons $\bar{q} = \sum_{l=1}^{L} q_l / L$ comme estimation ponctuelle de Q et $T = \bar{u} + b/L$ comme variance estimée de $\bar{q}$, où $b = \sum_{l=1}^{L} (q_l - \bar{q})^2 / (L-1)$ et $\bar{u} = \sum_{l=1}^{L} u_l / L$. Nous inférons au sujet de Q par $(\bar{q} - Q) \sim t_v(0, T)$, où t_v est une distribution t avec $v = (L-1)(1 + L\bar{u}/b)^2$ degrés de liberté.

Pour chaque méthode, nous exécutons 20 000 itérations de l'échantillonneur MCMC en écartant les 10 000 premières comme rodage et en élaguant les échantillons restants à toutes les cinq itérations, ce qui donne 2 000 itérations MCMC après rodage. Nous créons les $L = 50$ jeux de données synthétiques par

échantillonnage aléatoire à partir de ces 2 000 itérations. Nous posons $F = 40$ et $S = 15$ pour chaque méthode en fonction des passages initiaux de rodage. Pour un contrôle de convergence, nous avons examiné les tracés temporels de α, β et les moyennes pondérées d'un échantillon aléatoire des probabilités multinomiales en vraisemblance pour le MDPDM. Pour les deux méthodes, le nombre effectif de grappes occupées au niveau des ménages va d'ordinaire de 20 à 33 avec 38 comme maximum et le nombre correspondant de grappes occupées au niveau des particuliers pour toutes les grappes au niveau des ménages va de 5 à 9 avec 12 comme maximum.

Si on exécute l'échantillonneur MCMC dans un portable standard, le passage des données du chef du ménage au niveau des ménages crée à lui seul une accélération d'environ 63 % de l'échantillonneur de rejet par défaut comparativement à 40 % environ pour la seule méthode de plafonnement-pondération.

Le tableau A.3 présente les intervalles de confiance à 95 % pour chacune des méthodes. Somme toute, les trois ont des intervalles de confiance qui se ressemblent, d'où l'impression que les mesures d'accélération ne font guère perdre de précision. Pour la plupart, les intervalles sont relativement assimilables aux intervalles des données initiales sauf pour le pourcentage de couples d'âge homogène. La dernière ligne représente un contrôle rigoureux de la mesure dans laquelle chaque méthode peut établir une probabilité plutôt difficile à estimer exactement. Dans ce cas, la probabilité que le chef du ménage et le conjoint soient du même âge peut être d'une estimation peu facile, car l'âge de chacun peut prendre 96 valeurs. Les trois méthodes s'éloignent donc de l'estimation par les données initiales en pareil cas. Ces résultats semblent indiquer la possibilité d'accélérer nettement l'échantillonneur avec une perte minimale de précision de l'estimation et des intervalles de confiance en conséquence pour les paramètres de population.

Tableau A.3
Intervalles de confiance pour certaines probabilités qui dépendent des liens au sein du ménage dans les jeux de données initiales et synthétiques. La mention « Données initiales » vise les données échantillonnées. La mention « MDPDM » vise l'échantillonneur MCMC par défaut à la section 2.2 dans le corps du texte. La mention « MDPDM avec passage du chef du ménage » est l'échantillonneur par défaut avec passage des données du chef du ménage au niveau des ménages. La mention « MDPDM plafonné avec passage du chef du ménage » vise la méthode de plafonnement-pondération et le passage des données du chef du ménage au niveau des ménages

		Données initiales	MDPDM	MDPDM avec passage du chef du ménage	MDPDM plafonné avec passage du chef du ménage
Race homogène	$n_i = 2$	(0,939; 0,951)	(0,918; 0,932)	(0,912; 0,928)	(0,910; 0,925)
	$n_i = 3$	(0,896; 0,920)	(0,859; 0,888)	(0,845; 0,875)	(0,844; 0,874)
	$n_i = 4$	(0,885; 0,912)	(0,826; 0,860)	(0,813; 0,848)	(0,817; 0,852)
	$n_i = 5$	(0,879; 0,922)	(0,786; 0,841)	(0,786; 0,841)	(0,777; 0,834)
	$n_i = 6$	(0,831; 0,910)	(0,701; 0,803)	(0,718; 0,819)	(0,660; 0,768)
Conjoint présent		(0,693; 0,711)	(0,678; 0,697)	(0,676; 0,695)	(0,677; 0,695)
Conjoint avec chef du ménage blanc		(0,589; 0,608)	(0,577; 0,597)	(0,576; 0,595)	(0,575; 0,595)
Conjoint avec chef du ménage noir		(0,036; 0,043)	(0,035; 0,043)	(0,034; 0,042)	(0,034; 0,042)

Tableau A.3 (suite)
Intervalles de confiance pour certaines probabilités qui dépendent des liens au sein du ménage dans les jeux de données initiales et synthétiques. La mention « Données initiales » vise les données échantillonnées. La mention « MDPDM » vise l'échantillonneur MCMC par défaut à la section 2.2 dans le corps du texte. La mention « MDPDM avec passage du chef du ménage » est l'échantillonneur par défaut avec passage des données du chef du ménage au niveau des ménages. La mention « MDPDM plafonné avec passage du chef du ménage » vise la méthode de plafonnement-pondération et le passage des données du chef du ménage au niveau des ménages

	Données initiales	MDPDM	MDPDM avec passage du chef du ménage	MDPDM plafonné avec passage du chef du ménage
Couple blanc	(0,570; 0,589)	(0,560; 0,580)	(0,553; 0,573)	(0,552; 0,572)
Couple blanc, propriétaire du logement	(0,495; 0,514)	(0,468; 0,488)	(0,461; 0,481)	(0,463; 0,483)
Couple de race homogène	(0,655; 0,673)	(0,636; 0,655)	(0,626; 0,645)	(0,625; 0,644)
Couple formé d'un Blanc et d'un non-Blanc	(0,028; 0,035)	(0,028; 0,035)	(0,034; 0,041)	(0,036; 0,044)
Couple non blanc, propriétaire du logement	(0,057; 0,067)	(0,047; 0,056)	(0,045; 0,053)	(0,045; 0,054)
Mère présente seulement	(0,017; 0,022)	(0,014; 0,019)	(0,014; 0,019)	(0,013; 0,018)
Un parent présent seulement	(0,021; 0,026)	(0,026; 0,032)	(0,026; 0,033)	(0,027; 0,033)
Enfants présents	(0,507; 0,527)	(0,493; 0,512)	(0,517; 0,537)	(0,511; 0,531)
Frères et sœurs présents	(0,022; 0,028)	(0,027; 0,034)	(0,027; 0,033)	(0,027; 0,033)
Petit-enfant présent	(0,041; 0,049)	(0,051; 0,060)	(0,049; 0,058)	(0,050; 0,059)
Trois générations présentes	(0,036; 0,044)	(0,037; 0,045)	(0,042; 0,050)	(0,040; 0,048)
Chef du ménage blanc, plus âgé que le conjoint	(0,309; 0,327)	(0,283; 0,301)	(0,294; 0,313)	(0,302; 0,321)
Chef du ménage d'origine non hispanique	(0,882; 0,894)	(0,875; 0,888)	(0,879; 0,891)	(0,876; 0,889)
Chef du ménage blanc, d'origine hispanique	(0,071; 0,082)	(0,074; 0,085)	(0,072; 0,082)	(0,073; 0,084)
Couple d'âge homogène	(0,087; 0,098)	(0,027; 0,034)	(0,023; 0,029)	(0,024; 0,031)

A.5 Étude empirique de l'imputation des données manquantes entièrement au hasard

Nous évaluons aussi le rendement du MDPDM comme méthode d'imputation dans un scénario de données manquantes entièrement au hasard. Nous utilisons les mêmes données qu'à la section 5 dans le corps du texte. Rappelons que les données visent $n = 5\,000$ ménages de taille $\mathcal{H} = \{2, 3, 4\}$, comprenant $N = 13\,181$ particuliers. Nous introduisons les valeurs manquantes dans un scénario de données manquantes entièrement au hasard. Nous sélectionnons aléatoirement 80 % des ménages comme cas complets pour toutes les variables. Dans le cas des 20 % de ménages restants, nous mettons la variable « taille du ménage » en observation intégrale et nous mettons en blanc aléatoirement – et indépendamment – 50 % de chaque variable restante au niveau des ménages et au niveau des particuliers. Nous employons ces faibles taux pour imiter les taux réels de non-réponse partielle dans les données du recensement.

Comme dans le corps du texte, nous estimons le modèle MDPDM à l'aide de deux méthodes en combinant dans chaque cas l'étape de rejet à la section 4.1 et la méthode de plafonnement-pondération à la section 4.2 dans le texte. Dans la première méthode, nous considérons $\psi_2 = \psi_3 = \psi_4 = 1$ et, dans la seconde, nous prenons $\psi_2 = \psi_3 = 1/2$ et $\psi_4 = 1/3$. Dans chaque cas, nous exécutons 10 000 itérations de l'échantillonneur MCMC en écartant les 5 000 premières comme rodage et en élaguant les échantillons restants à toutes les cinq itérations, ce qui donne 1 000 itérations MCMC après rodage. Nous posons $F = 30$ et $S = 15$ pour chaque méthode en fonction des passages initiaux de rodage. Nous contrôlons la convergence comme dans le corps du texte. Pour les deux méthodes, nous produisons $L = 50$ jeux de

données complètes, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(50)})$, au moyen de la distribution prédictive postérieure du MDPDM, ce qui nous permet d'estimer les mêmes probabilités que dans le texte.

Les figures A.1 et A.2 présentent les diverses probabilités marginales, bivariées et trivariées estimées $(\overline{q}_{50})$ et tracées par rapport à l'estimation correspondante des données initiales sans valeurs manquantes. La figure A.1 livre les résultats du MDPDM avec l'échantillonneur de rejet et la figure A.2, les mêmes résultats avec la méthode de plafonnement-pondération. Dans les deux cas, le MDPDM réussit à appréhender les caractéristiques importantes de la distribution conjointe des variables, les estimations ponctuelles étant très proches de celles des données avant introduction des valeurs manquantes. Bref, les résultats ressemblent fort aux résultats dans le corps du texte, mais en plus précis.

Le tableau A.4 fait état des intervalles de confiance à 95 % de certaines probabilités concernant les liens au sein du ménage, ainsi que de la valeur de la population entière de 764 580 ménages. Il s'agit des deux intervalles venant des moteurs d'imputation MDPDM et de l'intervalle venant des données avant introduction des valeurs manquantes. Les intervalles sont généralement plus précis que ceux qui figurent dans le corps du texte. Il fallait s'y attendre, puisque nous employons des proportions inférieures de valeurs manquantes dans le scénario des données manquantes entièrement au hasard. Le plus souvent, les intervalles du MDPDM avec les deux méthodes tendent à comprendre la quantité vraie de la population. Là encore, le moteur d'imputation MDPDM engendre un biais par défaut pour les pourcentages de ménages de race homogène. Comme nous le mentionnons dans le texte, c'est là un paramètre difficile à estimer avec précision par imputation, plus particulièrement dans le cas des ménages de plus grande taille.

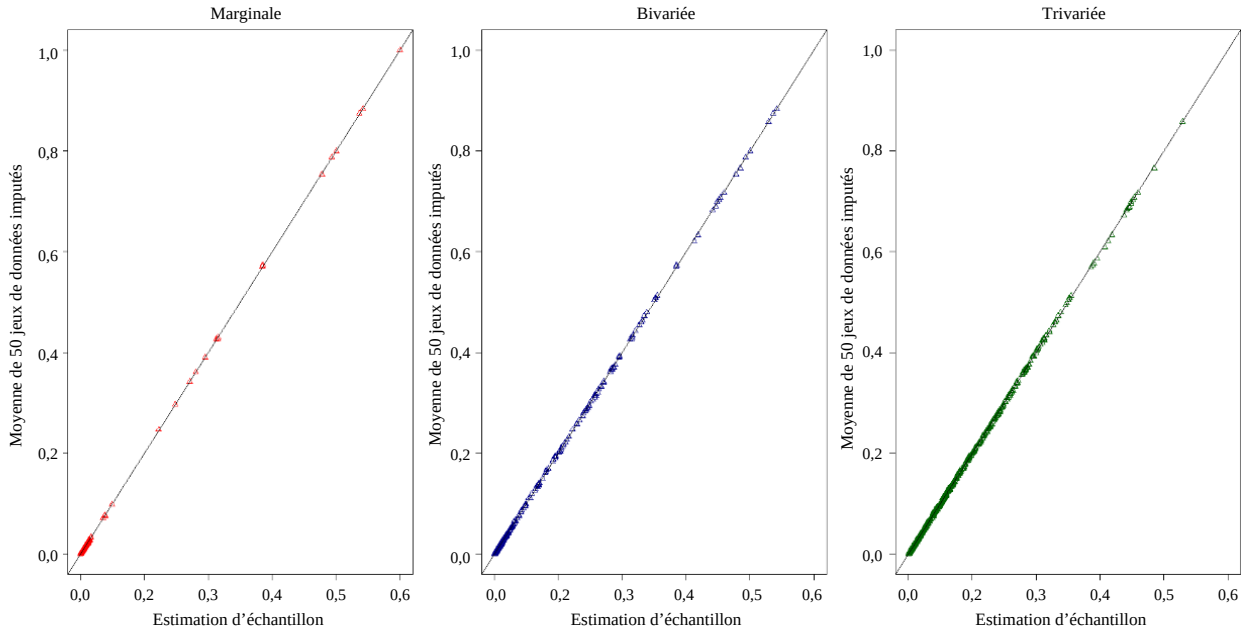

Figure A.1 Probabilités marginales, bivariées et trivariées calculées dans l'échantillon et jeux de données imputées dans un scénario de valeurs manquantes entièrement au hasard et avec MDPDM tronqué, échantillonneur de rejet et passage des données du chef du ménage au niveau des ménages.

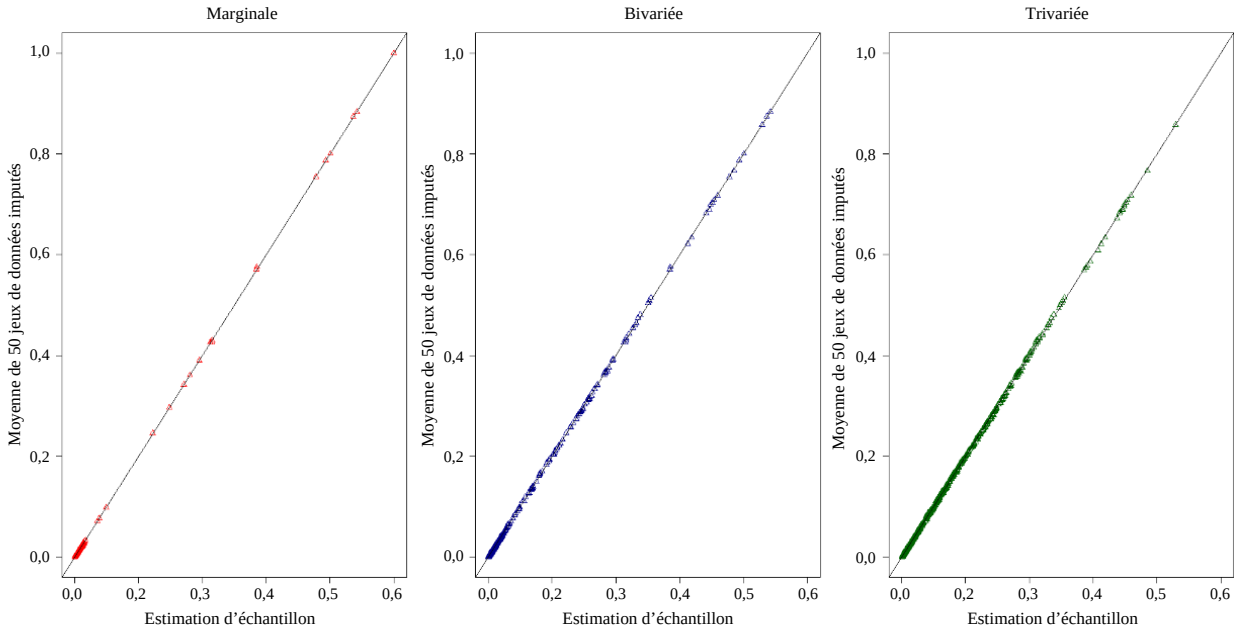

Figure A.2 Probabilités marginales, bivariées et trivariées calculées dans l'échantillon et jeux de données imputées dans un scénario de valeurs manquantes entièrement au hasard avec MDPDM tronqué, méthode de plafonnement-pondération et passage des données du chef du ménage au niveau des ménages.

Tableau A.4
Intervalles de confiance pour certaines probabilités qui dépendent des liens au sein du ménage dans les jeux de données initiales et imputées selon un scénario de valeurs manquantes entièrement au hasard. La mention « Sans données manquantes » vise les données échantillonnées avant introduction des valeurs manquantes. La mention « MDPDM » vise le MDPDM tronqué avec passage des données du chef du ménage au niveau des ménages. La mention « MDPDM plafonné » vise le MDPDM tronqué avec méthode de plafonnement-pondération et passage des données du chef du ménage au niveau des ménages. La mention « *Q* » est la valeur de la population entière de 764 580 ménages

		Q	Sans données manquantes	MDPDM	MDPDM plafonné
Ménage de race homogène :	$n_i = 2$	0,942	(0,932; 0,949)	(0,924; 0,944)	(0,925; 0,946)
	$n_i = 3$	0,908	(0,907; 0,937)	(0,887; 0,924)	(0,890; 0,925)
	$n_i = 4$	0,901	(0,879; 0,917)	(0,854; 0,900)	(0,855; 0,900)
Conjoint présent		0,696	(0,682; 0,707)	(0,683; 0,709)	(0,683; 0,709)
Couple de race homogène		0,656	(0,641; 0,668)	(0,637; 0,664)	(0,638; 0,665)
Conjoint présent; chef du ménage blanc		0,600	(0,589; 0,616)	(0,590; 0,618)	(0,590; 0,618)
Couple blanc		0,580	(0,569; 0,596)	(0,568; 0,596)	(0,568; 0,597)
Couple avec différence d'âge de moins de 5 ans		0,488	(0,465; 0,492)	(0,422; 0,451)	(0,422; 0,450)
Chef du ménage de sexe masculin, propriétaire du logement		0,476	(0,456; 0,484)	(0,455; 0,483)	(0,456; 0,485)
Chef du ménage de plus de 35 ans, aucun enfant présent		0,462	(0,441; 0,468)	(0,438; 0,466)	(0,438; 0,466)
Au moins un enfant biologique présent		0,437	(0,431; 0,458)	(0,432; 0,460)	(0,432; 0,460)
Chef du ménage plus âgé que le conjoint, chef du ménage blanc		0,322	(0,309; 0,335)	(0,308; 0,335)	(0,306; 0,333)
Femme d'âge adulte avec au moins un enfant de moins de 5 ans		0,078	(0,070; 0,085)	(0,068; 0,084)	(0,067; 0,083)
Chef du ménage blanc d'origine hispanique		0,066	(0,064; 0,078)	(0,064; 0,079)	(0,064; 0,079)
Couple non blanc, propriétaire du logement		0,058	(0,050; 0,063)	(0,048; 0,061)	(0,048; 0,061)
Deux générations présentes, chef du ménage noir		0,057	(0,053; 0,066)	(0,053; 0,066)	(0,053; 0,067)
Chef du ménage noir, propriétaire du logement		0,052	(0,046; 0,058)	(0,046; 0,059)	(0,046; 0,059)
Conjoint présent, chef du ménage noir		0,039	(0,032; 0,042)	(0,032; 0,043)	(0,032; 0,042)
Couple formé d'un Blanc et d'un non-Blanc		0,034	(0,029; 0,039)	(0,032; 0,044)	(0,032; 0,044)

Tableau A.4 (suite)
Intervalles de confiance pour certaines probabilités qui dépendent des liens au sein du ménage dans les jeux de données initiales et imputées selon un scénario de valeurs manquantes entièrement au hasard. La mention « Sans données manquantes » vise les données échantillonnées avant introduction des valeurs manquantes. La mention « MDPDM » vise le MDPDM tronqué avec passage des données du chef du ménage au niveau des ménages. La mention « MDPDM plafonné » vise le MDPDM tronqué avec méthode de plafonnement-pondération et passage des données du chef du ménage au niveau des ménages. La mention « *Q* » est la valeur de la population entière de 764 580 ménages

	Q	Sans données manquantes	MDPDM	MDPDM plafonné
Chef du ménage d'origine hispanique de plus de 50 ans, propriétaire du logement	0,029	(0,025; 0,034)	(0,025; 0,035)	(0,025; 0,035)
Un petit-enfant présent	0,028	(0,023; 0,033)	(0,024; 0,034)	(0,024; 0,034)
Femme noire d'âge adulte avec au moins un enfant de moins de 18 ans	0,027	(0,028; 0,038)	(0,027; 0,037)	(0,027; 0,037)
Au moins deux générations présentes, couple d'origine hispanique	0,027	(0,022; 0,031)	(0,022; 0,031)	(0,022; 0,031)
Couple d'origine hispanique avec au moins un enfant biologique	0,025	(0,020; 0,028)	(0,019; 0,028)	(0,019; 0,028)
Au moins trois générations présentes	0,023	(0,020; 0,028)	(0,019; 0,028)	(0,019; 0,028)
Un seul parent	0,020	(0,016; 0,024)	(0,016; 0,024)	(0,016; 0,024)
Au moins un enfant par alliance	0,019	(0,018; 0,026)	(0,018; 0,027)	(0,018; 0,027)
Homme adulte d'origine hispanique avec au moins un enfant de moins de 10 ans	0,018	(0,017; 0,025)	(0,016; 0,025)	(0,016; 0,025)
Au moins un enfant adoptif, couple blanc	0,008	(0,005; 0,010)	(0,005; 0,010)	(0,005; 0,010)
Couple noir avec au moins deux enfants biologiques	0,006	(0,003; 0,007)	(0,003; 0,007)	(0,003; 0,007)
Chef du ménage noir de moins de 40 ans, propriétaire du logement	0,005	(0,005; 0,009)	(0,005; 0,010)	(0,005; 0,011)
Trois générations présentes, couple blanc	0,005	(0,004; 0,008)	(0,004; 0,010)	(0,004; 0,009)
Chef du ménage blanc de moins de 25 ans, propriétaire du logement	0,003	(0,002; 0,005)	(0,004; 0,009)	(0,004; 0,009)

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.

Bennink, M., Croon, M.A., Kroon, B. et Vermunt, J.K. (2016). Micro-macro multilevel latent class models with multiple discrete individual-level variables. *Advances in Data Analysis and Classification*.

Chambers, R., et Skinner, C. (2003). Analysis of Survey Data, Wiley Series in Survey Methodology, Wiley.

Dunson, D.B., et Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104, 1042-1051.

Hu, J., Reiter, J.P. et Wang, Q. (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Analysis*, 13, 183-200.

Ishwaran, H., et James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 161-173.

Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1, 1-17. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/1986001/article/14404-fra.pdf.

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

Manrique-Vallier, D., et Reiter, J.P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *Journal of Computational and Graphical Statistics*, 23, 1061-1079.

Murray, J.S., et Reiter, J.P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence (à venir). *Journal of the American Statistical Association*.

Raghunathan, T.E., et Rubin, D.B. (2001). Multiple imputation for statistical disclosure limitation. Rapport technique.

Reiter, J.P. (2003). Inférence pour les ensembles de microdonnées à grande diffusion partiellement synthétiques. *Techniques d'enquête*, 29, 2, 203-211. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2003002/article/6785-fra.pdf.

Reiter, J.P., et Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462-1471.

Rubin, D.B. (1976). Inference and missing data (with discussion). *Biometrika*, 63, 581-592.

Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys, New York: John Wiley & Sons, Inc.

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

Savitsky, T.D., et Toth, D. (2016). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10.1, 1677-1708.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-650.

Si, Y., et Reiter, J.P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38.5, 199-521.

Vermunt, J.K. (2003). Multilevel latent class models. *Sociological Methodology*, 213-239.

Vermunt, J.K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, 33-51.

Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 1, 45-54.

Wang, Q., Akande, O., Hu, J., Reiter, J. et Barrientos, A. (2016). NestedCategBayesImpute: Modeling and Generating Synthetic Versions of Nested Categorical Data in the Presence of Impossible Combinations. *The Comprehensive R Archive Network*.