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

Techniques d’enquête

Les modèles d’apprentissage profond sont-ils plus efficaces pour l’imputation de données manquantes dans les enquêtes ? Une comparaison empirique fournit des éléments de preuve

par Zhenhua Wang, Olanrewaju Akande, Jason Poulos et Fan Li

Date de diffusion : le 15 décembre 2022

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2022
Vol. 48, nº 2, pp. 407-437
Statistique Canada, nº 12-001-X au catalogue

Les modèles d'apprentissage profond sont-ils plus efficaces pour l'imputation de données manquantes dans les enquêtes ? Une comparaison empirique fournit des éléments de preuve

Zhenhua Wang, Olanrewaju Akande, Jason Poulos et Fan Li[1]

Résumé

L'imputation multiple est une approche populaire pour traiter les données manquantes découlant de la non-réponse dans les enquêtes-échantillons. L'imputation multiple au moyen d'équations en séries (MICE) est l'un des algorithmes d'imputation multiple les plus utilisés pour les données multivariées, mais son fondement théorique est insuffisant et elle exige beaucoup de calculs. Récemment, des méthodes d'imputation des données manquantes fondées sur des modèles d'apprentissage profond ont été élaborées, ce qui a donné des résultats encourageants dans de petites études. Cependant, peu de recherches ont été menées sur l'évaluation de leur rendement dans des contextes réalistes par rapport à la MICE, en particulier dans le cadre de grandes enquêtes. Nous menons de vastes études de simulation fondées sur un sous-échantillon de l'*American Community Survey* afin de comparer les propriétés d'échantillonnage répété de quatre méthodes d'apprentissage automatique fondées sur l'imputation multiple : MICE avec arbres de classification; MICE avec forêts aléatoires; réseaux antagonistes génératifs pour l'imputation; et imputation multiple à l'aide d'autoencodeurs débruiteurs. Nous constatons que les méthodes d'imputation fondées sur des modèles d'apprentissage profond sont plus efficaces que la MICE en ce qui a trait au temps de calcul. Cependant, étant donné le choix par défaut des hyperparamètres dans les progiciels communs, la MICE avec arbres de classification dépasse constamment, souvent de loin, les méthodes d'imputation fondées sur l'apprentissage profond quant au biais, à l'erreur quadratique moyenne et à la couverture dans une gamme de paramètres réalistes.

Mots-clés : Apprentissage profond; données d'enquête; données manquantes; études par simulation; imputation multiple au moyen d'équations en séries; sélection des hyperparamètres.

1. Introduction

De nombreuses enquêtes-échantillons comportent des données manquantes découlant de la non-réponse totale, lorsqu'un sous-ensemble de participants ne répond pas à l'enquête, ou de la non-réponse partielle, lorsque les valeurs manquantes se limitent à des questions particulières. Dans les sondages d'opinion, la non-réponse peut indiquer le refus de révéler une préférence ou l'absence de préférence (De Leeuw, Hox et Huisman, 2003). Si elles ne sont pas traitées adéquatement, les données manquantes peuvent mener à des analyses statistiques biaisées, surtout lorsqu'il y a des différences systématiques entre les données observées et les données manquantes (Rubin, 1976; Little et Rubin, 2019). L'analyse de cas complets d'unités pour lesquelles les données sont entièrement observées est souvent impossible et peut entraîner un biais important dans la plupart des situations (Little et Rubin, 2019). Par conséquent, de nombreux analystes tiennent compte des données manquantes en imputant les valeurs manquantes, puis en procédant comme si les valeurs imputées étaient des valeurs réelles.

1. Zhenhua Wang est étudiant au doctorat au Department of Statistics, University of Missouri, Columbia, MO 65211. Courriel : zhenhua.wang@mail.missouri.edu; Olanrewaju Akande est chercheur chez Meta Platforms, Inc. Courriel : akandelanre13@gmail.com; Jason Poulos est associé de recherche postdoctorale au Department of Health Care Policy, Harvard Medical School, Boston, MA. Courriel : poulos@hcp.med.harvard.edu; Fan Li est professeur au Department of Statistical Science, Box 90251, Duke University, Durham, Caroline du Nord 27708, C.P. 90251. Courriel : fl35@duke.edu.

L'imputation multiple (Rubin, 1987) est une approche répandue pour traiter les valeurs manquantes. Dans le cadre de l'imputation multiple, un analyste crée $L>1$ ensembles de données complets en remplaçant les valeurs manquantes dans les données de l'échantillon par des tirages plausibles générés à partir de la distribution prédictive de modèles probabilistes s'appuyant sur les données observées. Pour chaque ensemble de données complet, l'analyste peut ensuite calculer les estimations de l'échantillon pour les paramètres de la population d'intérêt et combiner les estimations de l'échantillon pour tous les L ensembles de données à l'aide des méthodes d'inférence de l'imputation multiple mises au point dans Rubin (1987) et, plus récemment, dans Rubin (1996), Barnard et Meng (1999), Reiter et Raghunathan (2007) et Harel et Zhou (2007). Dans l'imputation multiple, la variance estimée d'un paramètre consiste en des variances intra-imputation et inter-imputation, et tient donc compte de la variabilité inhérente des valeurs imputées. Il convient de prendre note que dans les études fondées sur des données d'enquête, l'imputation unique, par exemple par couplage ou régression, demeure courante pour le traitement des données manquantes, où la variance est estimée au moyen de la méthode delta ou des méthodes de rééchantillonnage (Chen et Haziza, 2019; Haziza et Vallée, 2020).

1.1 Imputation par modèle

Il existe deux stratégies générales de modélisation pour l'imputation multiple. La première stratégie, connue sous le nom de *modélisation conjointe*, consiste à préciser une distribution conjointe pour toutes les variables des données, puis à générer des imputations à partir des distributions (prédictives) conditionnelles implicites des variables pour lesquelles il manque des valeurs (Schafer, 1997). La stratégie de modélisation conjointe s'aligne sur le fondement théorique de Rubin (1987), mais il peut être difficile de spécifier un modèle conjoint ayant des variables de grande dimension de différents types. En effet, les méthodes de modélisation conjointe les plus reconnues, comme « PROC MI » dans SAS (Yuan, 2011), et « AMELIA » (Honaker, King et Blackwell, 2011) et « norm » dans `R` (Schafer, 1997), formulent une hypothèse simplificatrice selon laquelle les données suivent des distributions gaussiennes multivariées, même pour les variables catégoriques, ce qui peut entraîner un biais (Horton, Lipsitz et Parzen, 2003). Des études récentes ont permis de mettre au point des modèles de modélisation conjointe souples fondés sur des modèles bayésiens non paramétriques avancés comme le modèle par mélange selon le processus de Dirichlet (Manrique-Vallier et Reiter, 2014; Murray et Reiter, 2016). Cependant, ces méthodes sont coûteuses sur le plan des calculs et s'adaptent souvent mal aux cas ayant des variables de grande dimension.

La deuxième stratégie est la *spécification entièrement conditionnelle* (van Buuren, Brand, Groothuis-Oudshoorn et Rubin, 2006), où l'on précise séparément une distribution conditionnelle univariée pour chaque variable ayant des valeurs manquantes, compte tenu de toutes les autres variables, et où l'on impute les valeurs manquantes variable par variable de façon itérative, d'une manière s'apparentant à

l'échantillonneur de Gibbs. La méthode de la spécification entièrement conditionnelle la plus répandue est l'imputation multiple au moyen d'équations en séries (MICE) (van Buuren et Groothuis-Oudshoorn, 2011), habituellement mise en œuvre par la spécification des modèles linéaires généralisés (GLM) pour les distributions conditionnelles univariées (Raghunathan, Lepkowski, Van Hoewyk et Solenberger, 2001; Royston et White, 2011; Su, Gelman, Hill et Yajima, 2011). Des études récentes indiquent que la spécification des modèles conditionnels par arbres de classification et de régression (CART) (Breiman, Friedman, Olshen et Stone, 1984; Burgette et Reiter, 2010) dépasse largement la MICE par les GLM (Akande, Li et Reiter, 2017). Une extension naturelle de la MICE avec CART consiste à utiliser des méthodes s'appuyant sur un ensemble d'arbres, comme les forêts aléatoires, plutôt d'utiliser un seul arbre (Breiman, 2001; Doove, van Buuren et Dusseldorp, 2014).

La MICE est intéressante dans les données d'enquête à grande échelle parce qu'elle permet l'imputation de différents types de variables de manière simple et souple. Cependant, la MICE présente un inconvénient théorique clé, à savoir l'incompatibilité possible des distributions conditionnelles spécifiées, c'est-à-dire que les distributions ne correspondent pas à une distribution conjointe (Arnold et Press, 1989; Gelman et Speed, 1993; Li, Yu et Rubin, 2012). Malgré cet inconvénient, la MICE fonctionne remarquablement bien dans des applications réelles, et de nombreuses simulations ont démontré qu'elle est plus efficace que de nombreuses méthodes de modélisation conjointe théoriquement solides; voir van Buuren (2018) pour obtenir des études de cas. Cependant, la MICE nécessite aussi de nombreux calculs (White, Royston et Wood, 2011) et ne peut généralement pas être mise en parallèle. De plus, les progiciels reconnus pour la mise en œuvre de la MICE par les GLM, par exemple `mice` dans `R` (van Buuren et Groothuis-Oudshoorn, 2011), boguent souvent dans des contextes comportant des variables non continues de grande dimension, comme des variables catégoriques comportant de nombreuses catégories (Akande et coll., 2017).

1.2 Imputation au moyen de modèles d'apprentissage profond

Les récents progrès de l'apprentissage profond élargissent considérablement la portée des modèles complexes pour les données de grande dimension. Ces progrès apportent l'espoir qu'une nouvelle génération de méthodes d'imputation des données manquantes fondées sur des modèles d'apprentissage profond puisse tenir compte des limites théoriques et computationnelles des méthodes statistiques existantes. Par exemple, les modèles génératifs profonds comme les réseaux antagonistes génératifs (GAN) (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville et Bengio, 2014) sont naturellement adaptés à la production d'imputations multiples parce qu'ils sont conçus pour générer des données qui ressemblent le plus possible aux données observées. Une méthode dans ce volet est le réseau antagoniste génératif pour l'imputation (GAIN) de Yoon, Jordon et Schaar (2018). L'imputation multiple à l'aide d'autoencodeurs débruiteurs (MIDA) (Gondara et Wang, 2018; Lu, Perrone et Unpingco, 2020)

est une autre méthode générative basée sur des réseaux neuronaux profonds formés à traiter des données d'entrée corrompues afin de forcer les réseaux à apprendre une représentation de faible dimension utile des données d'entrée plutôt que leur fonction d'identité (Vincent, Larochelle, Bengio et Manzagol, 2008; Vincent, Larochelle, Lajoie, Bengio, Manzagol et Bottou, 2010). Plusieurs méthodes ont été proposées pour l'imputation des valeurs manquantes dans les données chronologiques au moyen d'autoencodeurs variationnels (Fortuin, Baranchuk, Rätsch et Mandt, 2020) ou de réseaux neuronaux récurrents (Lipton, Kale et Wetzel, 2016; Monti, Bronstein et Bresson, 2017; Cao, Wang, Li, Zhou, Li et Li, 2018; Che, Purushotham, Cho, Sontag et Liu, 2018; Yoon, Zame et van der Schaar, 2018).

Les méthodes d'imputation multiple fondées sur l'apprentissage profond présentent plusieurs avantages, du moins théoriquement, par rapport aux modèles statistiques conventionnels, notamment : i) elles permettent d'éviter de faire des hypothèses de distribution; ii) elles permettent de facilement traiter des types de données mixtes; iii) elles permettent de modéliser des relations non linéaires entre les variables; iv) elles sont censées bien fonctionner dans des paramètres de grande dimension; et v) elles peuvent tirer parti de la puissance du processeur graphique pour accélérer le calcul. Plusieurs études font état d'un rendement encourageant des méthodes d'imputation multiple fondées sur l'apprentissage profond par rapport à la MICE (par exemple Yoon, Jordon et Schaar, 2018). Toutefois, les conclusions de ces études sont fondées sur des données probantes limitées. Premièrement, les études sont habituellement fondées sur de petites simulations ou sur plusieurs ensembles de données de « référence » publics bien étudiés, comme ceux décrits à la section 5, qui ne ressemblent pas aux données d'enquête. Deuxièmement, les évaluations sont habituellement fondées sur quelques mesures de rendement globales, comme l'erreur quadratique moyenne de prédiction globale ou l'exactitude. Il est possible que ces mesures ne donnent pas un tableau complet des comparaisons, et elles peuvent même être parfois trompeuses, comme nous le verrons plus loin. Troisièmement, étant donné l'incertitude du processus de données manquantes, il est essentiel d'examiner les propriétés d'échantillonnage répété des méthodes d'imputation, mais celles-ci ont rarement été évaluées. Enfin, le réglage des hyperparamètres est crucial pour les modèles d'apprentissage automatique, et différents réglages peuvent donner des résultats radicalement différents, par contre peu de renseignements sont fournis sur le réglage des hyperparamètres et ses conséquences sur le rendement des méthodes d'imputation.

Motivés par ces limites, nous effectuons, dans la présente étude, des simulations exhaustives fondées sur des données d'enquête réelles afin d'évaluer les méthodes d'imputation multiple au moyen d'une gamme de mesures de rendement. Plus précisément, nous effectuons des simulations fondées sur un sous-échantillon de l'*American Community Survey* pour comparer les propriétés d'échantillonnage répété de quatre méthodes d'imputation multiple mentionnées précédemment : MICE avec CART (MICE-CART), MICE avec des forêts aléatoires (MICE-RF), GAIN et MIDA. Nous avons découvert que les méthodes d'imputation multiple fondées sur l'apprentissage profond sont plus efficaces que la MICE en ce qui a

trait au temps de calcul. Cependant, la MICE-CART dépasse constamment, souvent de loin, les méthodes d'apprentissage profond quant au biais, à l'erreur quadratique moyenne et à la couverture dans une gamme de paramètres réalistes. Cette observation contredit les constatations antérieures dans la littérature sur l'apprentissage automatique et soulève des questions sur les mesures appropriées pour évaluer les méthodes d'imputation. Elle fait également ressortir l'importance d'évaluer les propriétés d'échantillonnage répété des méthodes d'imputation. Bien que nous mettions l'accent sur l'imputation multiple dans la présente étude, nous remarquons que les méthodes d'imputation multiple mentionnées précédemment sont facilement applicables pour générer une imputation unique lorsque L est établi à 1. Des preuves empiriques exhaustives indiquent que la variance intra-imputation domine habituellement la variance inter-imputation dans l'imputation multiple. Par conséquent, nous nous attendons à ce que les schémas entre les différentes méthodes d'imputation observées soient également valables si ces méthodes sont utilisées pour une seule imputation.

Le reste de la présente étude est organisé comme suit. À la section 2, nous examinons les quatre méthodes d'imputation multiple utilisées dans notre évaluation. À la section 3, nous décrivons un cadre comportant plusieurs mesures pour évaluer les méthodes d'imputation. À la section 4, nous décrivons le plan de simulation et les résultats à l'aide de données d'enquête à grande échelle, et à la section 5, nous résumons les résultats d'évaluation des ensembles de données de référence utilisés dans la documentation sur l'apprentissage automatique. Enfin, à la section 6, nous concluons en présentant un guide pratique pour la mise en œuvre dans des applications réelles.

2. Méthode d'imputation des données manquantes

Nous présentons d'abord la notation. Prenons un échantillon ayant n unités, dont chacune est associée à p variables. Si Y_{ij} correspond à la valeur de la variable j pour chaque i, où $j = 1, \ldots, p$ et $i = 1, \ldots, n$. Ici, Y peut être continu, binaire, catégorique ou binaire-continu. Pour chaque i, supposons que $\mathbf{Y}_i = \left(Y_{i1}, \ldots, Y_{ip}\right)$. Pour chaque variable j, supposons que $\mathbf{Y}_j = \left(Y_{1j}, \ldots, Y_{nj}\right)$. Supposons que $\mathbf{Y} = \left(\mathbf{Y}_1, \ldots, \mathbf{Y}_n\right)$ correspond à la matrice $n \times p$ comprenant les données pour tous les enregistrements inclus dans l'échantillon. Nous écrivons $\mathbf{Y} = \left(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}\right)$, où $\mathbf{Y}_{\text{obs}}$ et $\mathbf{Y}_{\text{mis}}$ sont les parties observées et manquantes de $\mathbf{Y}$, respectivement. Nous écrivons $\mathbf{Y}_{\text{mis}} = \left(\mathbf{Y}_{\text{mis},1}, \ldots, \mathbf{Y}_{\text{mis},\,p}\right)$, où $\mathbf{Y}_{\text{mis},\,j}$ représente toutes les valeurs manquantes pour la variable j, et $j = 1, \ldots, p$. De même, nous écrivons $\mathbf{Y}_{\text{obs}} = \left(\mathbf{Y}_{\text{obs},1}, \ldots, \mathbf{Y}_{\text{obs},\,p}\right)$ pour les données observées correspondantes.

Lors de l'imputation multiple, l'analyste génère les valeurs des données manquantes $\mathbf{Y}_{\text{mis}}$ au moyen de modèles prédéfinis estimés avec $\mathbf{Y}_{\text{obs}}$, ce qui donne un ensemble de données complet. L'analyste répète ensuite le processus pour générer L ensembles de données complets, $\left\{\mathbf{Y}^{(l)} : l = 1, \ldots, L\right\}$, qui sont accessibles aux fins d'inférence ou de diffusion. Pour l'inférence, l'analyste peut calculer des estimations

de l'échantillon pour les paramètres de la population dans chaque ensemble de données complet $\mathbf{Y}^{(l)}$ et il peut les combiner au moyen des règles d'inférence de l'imputation multiple élaborées par Rubin (1987), qui seront examinées à la section 3.

2.1 L'imputation multiple au moyen d'équations en séries et les modèles d'arbres de classification

Dans la MICE, l'analyste commence par préciser un modèle conditionnel univarié distinct pour chaque variable pour laquelle il manque des valeurs. L'analyste précise ensuite un ordre d'itération dans la séquence des modèles conditionnels, au moment de l'imputation. La liste ordonnée des variables s'écrit $\left(\mathbf{Y}_{(1)},\ldots,\mathbf{Y}_{(p)}\right)$. Ensuite, l'analyste initialise chaque $\mathbf{Y}_{\text{mis},(j)}$. Les options les plus populaires consistent à échantillonner à partir de : i) la distribution marginale de la variable $\mathbf{Y}_{\text{obs},(j)}$ correspondante; ou ii) la distribution conditionnelle de la variable $\mathbf{Y}_{(j)}$, compte tenu de toutes les autres variables, construites uniquement à l'aide des cas disponibles.

Après l'initialisation, l'algorithme de la MICE suit un processus itératif qui parcourt la séquence des modèles univariés. Pour chaque variable j à chaque itération t, on fait correspondre le modèle conditionnel $(\mathbf{Y}_{(j)} \,|\, \mathbf{Y}_{\text{obs},(j)}, \{\mathbf{Y}_{(k)}^{(t)}\colon k<j\}, \{\mathbf{Y}_{(k)}^{(t-1)}\colon k>j\})$. Ensuite, on remplace $\mathbf{Y}_{\text{mis},(j)}^{(t)}$ par des tirages à partir du modèle implicite $(\mathbf{Y}_{\text{mis},(j)}^{(t)} \,|\, \mathbf{Y}_{\text{obs},(j)}, \{\mathbf{Y}_{(k)}^{(t)}\colon k<j\}, \{\mathbf{Y}_{(k)}^{(t-1)}\colon k>j\})$. Le processus itératif se poursuit pendant T itérations totales jusqu'à la convergence, et les valeurs de l'itération finale constituent un ensemble de données complet $\mathbf{Y}^{(l)} = \left(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(T)}\right)$. L'ensemble du processus est ensuite répété L fois pour créer les L ensembles de données complets. Nous fournissons un pseudocode détaillant chaque étape de l'algorithme de la MICE dans la documentation supplémentaire.

Dans la MICE-CART, l'analyste utilise un CART (Breiman et coll., 1984) pour les modèles conditionnels univariés dans l'algorithme de la MICE. Le CART suit une structure d'arbre décisionnel qui s'appuie sur des parties fractionnées binaires récursives pour répartir l'espace de prévision en régions distinctes sans chevauchement. La partie supérieure de l'arbre représente souvent sa racine et chaque partie fractionnée binaire successive divise l'espace de prévision en deux nouvelles branches au fur et à mesure que l'on descend l'arbre. Le critère de fractionnement à chaque feuille est habituellement choisi pour réduire au minimum une mesure d'entropie théorique de l'information. Les parties fractionnées qui ne diminuent pas l'ajustement insuffisant d'une quantité raisonnable par rapport à un seuil fixe sont élaguées. L'arbre est ensuite construit jusqu'à ce qu'un critère d'arrêt soit satisfait; par exemple un nombre minimal d'observations dans chaque feuille.

Une fois l'arbre entièrement construit, on génère $\mathbf{Y}_{\text{mis},(j)}^{(t)}$ en faisant descendre l'arbre jusqu'à la feuille appropriée au moyen des combinaisons dans $(\{\mathbf{Y}_k^{(t)}\colon k<j\}, \{\mathbf{Y}_k^{(t-1)}\colon k>j\})$, puis en échantillonnant à partir des $Y_{(j)}^{\text{obs}}$ valeurs de cette feuille. Autrement dit, pour toute combinaison dans $(\{\mathbf{Y}_k^{(t)}\colon k<j\}, \{\mathbf{Y}_k^{(t-1)}\colon k>j\})$, on utilise la proportion de valeurs de $\mathbf{Y}_j^{\text{obs}}$ dans la feuille correspondante pour estimer la

distribution conditionnelle $(\mathbf{Y}_{(j)} \,|\, \mathbf{Y}_{\text{obs},(j)}, \{\mathbf{Y}_{(k)}^{(t)}: k < j\}, \{\mathbf{Y}_{(k)}^{(t-1)}: k > j\})$. Le processus itératif se poursuit à nouveau pendant T itérations totales, et les valeurs de l'itération finale constituent un ensemble de données complet.

Dans le cadre de la MICE-RF, on utilise plutôt des forêts aléatoires pour les modèles conditionnels univariés dans la MICE (par exemple Stekhoven et Bühlmann, 2012; Shah, Bartlett, Carpenter, Nicholas et Hemingway, 2014). Les forêts aléatoires (Ho, 1995; Breiman, 2001) constituent une méthode s'appuyant sur un ensemble d'arbres qui construit des arbres décisionnels multiples à partir des données, plutôt qu'un seul arbre comme le CART. Plus précisément, les forêts aléatoires construisent plusieurs arbres décisionnels au moyen d'échantillons bootstrap de l'arbre original, et ne s'appuient que sur un échantillon des facteurs prédictifs pour les partitions récursives dans chaque arbre. Cette approche peut réduire considérablement la prévalence des arbres instables ainsi que la corrélation entre les arbres individuels, puisqu'elle empêche les mêmes variables de dominer le processus de partitionnement dans l'ensemble des arbres. En théorie, cette décorrélation devrait se traduire par des prévisions ayant moins de variance (Hastie, Tibshirani et Friedman, 2009).

Pour l'imputation, l'analyste entraîne d'abord un modèle de forêts aléatoires pour chaque $\mathbf{Y}_{(j)}$ à l'aide des cas disponibles et compte tenu de toutes les autres variables. Ensuite, l'analyste génère des prévisions pour $\mathbf{Y}_{\text{mis},j}$ dans ce modèle. Plus précisément, pour toute valeur $\mathbf{Y}_{(j)}$ catégorique, et compte tenu de toute combinaison particulière dans $(\{\mathbf{Y}_k^{(t)}: k < j\}, \{\mathbf{Y}_k^{(t-1)}: k > j\})$, l'analyste génère d'abord des prévisions pour chaque arbre en fonction des valeurs $\mathbf{Y}_j^{\text{obs}}$ de la feuille correspondante pour cet arbre, puis il utilise le niveau majoritaire le plus courant de toutes les prévisions de tous les arbres. Pour une valeur $\mathbf{Y}_{(j)}$ continue, l'analyste utilise plutôt la moyenne de toutes les prédictions de tous les arbres. Le processus itératif se poursuit de nouveau pour toutes les variables, pendant T itérations totales, et les valeurs de l'itération finale constituent un ensemble de données complet. Un hyperparamètre particulièrement important dans les forêts aléatoires est le nombre maximum d'arbres d.

Pour nos évaluations, nous utilisons le progiciel `mice` R pour mettre en œuvre la MICE-CART et la MICE-RF, et nous conservons le réglage d'hyperparamètres par défaut dans le progiciel pour imiter la pratique courante dans des applications réelles. Plus précisément, nous avons fixé le nombre minimum d'observations dans chaque feuille terminale à 5 et le seuil d'élagage à 0,0001 dans la MICE-CART. Dans la MICE-RF, le nombre maximum d'arbres d est fixé à 10.

2.2 Réseau antagoniste génératif pour l'imputation

Le GAIN (Yoon, Jordon et Schaar, 2018) est une méthode d'imputation fondée sur les GAN (Goodfellow et coll., 2014), qui consiste en une fonction génératrice G et une fonction discriminante D. Pour toute matrice de données $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$, nous remplaçons $\mathbf{Y}_{\text{mis}}$ par du bruit aléatoire, Z_{ij}, échantillonné à partir d'une distribution uniforme. La fonction génératrice G permet d'entrer ces données

initialisées et une matrice de masque $\mathbf{M}$, dans laquelle $M_{ij} \in \{0,1\}$ indique les valeurs observées de $\mathbf{Y}$, et elle génère les valeurs estimées des données observées et des données manquantes, $\hat{\mathbf{Y}}$. La fonction discriminante D utilise $\hat{\mathbf{Y}} = \left(\mathbf{Y}_{\text{obs}}, \hat{\mathbf{Y}}_{\text{mis}}\right)$ et une matrice d'indice $\mathbf{H}$ de la même dimension pour déterminer les valeurs observées ou imputées par G, ce qui donne une matrice de masque prédite $\hat{\mathbf{M}}$. La matrice d'indice, échantillonnée à partir de la distribution de Bernoulli dans laquelle p est égal à un hyperparamètre de « taux d'indice », révèle à D des renseignements partiels à propos de $\mathbf{M}$ afin d'aider à guider G dans l'apprentissage de la distribution sous-jacente de $\mathbf{Y}$.

Nous entraînons d'abord D à réduire au minimum la fonction de perte, $L_D\left(\mathbf{M}, \hat{\mathbf{M}}\right)$, pour chaque mini-lot de taille n_i :

$$L_D\left(\mathbf{M}, \hat{\mathbf{M}}\right) = \sum_{i=1}^{n_i}\sum_{j=1}^{J} M_{ij} \log\left(\hat{M}_{ij}\right) + \left(1 - M_{ij}\right) \log\left(1 - \hat{M}_{ij}\right). \tag{2.1}$$

Ensuite, G est entraîné pour réduire au minimum la fonction de perte (2.2), qui se compose d'une perte de la fonction génératrice, $L_G\left(\mathbf{M}, \hat{\mathbf{M}}\right)$, et d'une perte de la fonction de reconstruction, $L_M\left(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M}\right)$. La perte de la fonction génératrice (2.3) est réduite au minimum lorsque D définit incorrectement les valeurs imputées en valeurs observées. La perte de la fonction de reconstruction (2.4) est réduite au minimum lorsque les valeurs prédites sont semblables aux valeurs observées, et elle est pondérée par l'hyperparamètre β :

$$L\left(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M}, \hat{\mathbf{M}}\right) = L_G\left(\mathbf{M}, \hat{\mathbf{M}}\right) + \beta L_M\left(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M}\right), \tag{2.2}$$

$$L_G\left(\mathbf{M}, \hat{\mathbf{M}}\right) = \sum_{i=1}^{n_i}\sum_{j=1}^{J} M_{ij} \log\left(1 - \hat{M}_{ij}\right), \tag{2.3}$$

$$L_M\left(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M}\right) = \sum_{i=1}^{n_i}\sum_{j=1}^{J} \left(1 - M_{ij}\right) L_{\text{rec}}\left(Y_{ij}, \hat{Y}_{ij}\right), \tag{2.4}$$

où

$$L_{\text{rec}}\left(Y_{ij}, \hat{Y}_{ij}\right) = \begin{cases} \left(\hat{Y}_{ij} - Y_{ij}\right)^2 & \text{si } Y_{ij} \text{ est continu} \\ -Y_{ij} \log \hat{Y}_{ij} & \text{si } Y_{ij} \text{ est catégorique.} \end{cases} \tag{2.5}$$

Dans nos expériences, nous modélisons à la fois G et D en tant que réseaux neuronaux entièrement connectés, chaque fonction ayant trois couches cachées, et θ unités cachées par couche cachée. Les poids des couches cachées sont initialisés uniformément au hasard selon la méthode d'initialisation de Xavier (Glorot et Bengio, 2010). Nous utilisons une fonction d'activation Leaky ReLU (Maas, Hannun et Ng, 2013) pour chaque couche cachée, et une fonction d'activation Softmax pour la couche de sortie de G dans le cas de variables catégoriques, ou une fonction d'activation sigmoïde dans le cas de variables

numériques et pour la couche de sortie de D. Nous facilitons ce choix de couche de sortie pour les variables numériques en transformant toutes les variables continues pour qu'elles soient dans la fourchette de $(0, 1)$ au moyen de la normalisation MinMax : $Y_{ij}^{*} = \{Y_{ij} - \min(Y_{\cdot j})\} \big/ \{\max(Y_{\cdot j}) - \min(Y_{\cdot j})\}$, où $\min(Y_{\cdot j})$ et $\max(Y_{\cdot j})$ sont respectivement le minimum et le maximum de la variable j. Après l'imputation, nous ramenons chaque valeur à son échelle d'origine. Nous générons des imputations multiples à l'aide de plusieurs exécutions du modèle et d'une imputation initiale variable des valeurs manquantes.

Pour mettre en œuvre le GAIN dans nos évaluations, nous utilisons la même architecture que celle de Yoon, Jordon et Schaar (2018). Nous établissons $\beta = 100$, θ étant égal au nombre de caractéristiques des données d'entrée, et nous ajustons le taux d'indice dans une seule simulation. Conformément à la pratique courante trouvée dans la littérature sur les GAN (Berthelot, Schumm et Metz, 2017; Ham, Jun et Kim, 2020), nous suivons l'évolution des pertes des fonctions génératrices et discriminantes du GAIN, et nous ajustons manuellement le taux d'indice de sorte que les deux pertes soient qualitativement similaires. Plus précisément, nous sélectionnons d'abord sommairement le taux d'indice parmi {0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9}. Nous déterminons ensuite la valeur finale en réalisant une étape d'ajustement supplémentaire. Dans le scénario des données manquantes au hasard (DMH), par exemple, après avoir observé que la valeur optimale se situe dans la fourchette (0,1; 0,2), nous effectuons une recherche parmi les valeurs {0,11; 0,12; 0,13; 0,14; 0,15; 0,16; 0,17; 0,18; 0,19}. Enfin, nous attribuons une valeur de 0,3 et 0,13 au taux d'indice optimal des scénarios de données manquantes complètement au hasard (DMCH) et de DMH, respectivement. Nous entraînons les réseaux pour 200 périodes au moyen de la descente par gradient stochastique et des mini-lots de taille 512 pour apprendre les poids des paramètres. Nous utilisons l'optimiseur Adam pour adapter le taux d'apprentissage, lequel se situe initialement à 0,001 (Kingma et Ba, 2014).

2.3 L'imputation multiple au moyen d'autoencodeurs débruiteurs

L'imputation multiple au moyen d'autoencodeurs débruiteurs (MIDA) (Gondara et Wang, 2018; Lu et coll., 2020) étend l'utilisation d'une classe de réseaux neuronaux, les encodeurs débruiteurs, à l'imputation multiple. Un autoencodeur est un modèle de réseau neuronal entraîné à apprendre la fonction de lien identité des données d'entrée. Les autoencodeurs débruiteurs corrompent intentionnellement les données d'entrée afin d'empêcher les réseaux d'apprendre la fonction de lien identité, pour qu'ils apprennent plutôt une représentation de faible dimension utile des données d'entrée. L'architecture de la MIDA se compose d'un encodeur et d'un décodeur, chacun étant modélisé comme un réseau neuronal entièrement connecté ayant trois couches cachées et θ unités cachées par couche cachée. Nous effectuons d'abord une imputation initiale des valeurs manquantes en utilisant la moyenne pour les variables continues et l'étiquette la plus fréquente pour les variables catégoriques, ce qui donne un ensemble de

données complet $\mathbf{Y}_0$. L'encodeur introduit $\mathbf{Y}_0$ et corrompt les données d'entrée en éliminant aléatoirement la moitié des variables. Les données d'entrée corrompues sont appliquées à une représentation dimensionnelle plus grande par l'ajout de Θ unités cachées à chaque couche cachée successive de l'encodeur. Le décodeur reçoit les données de sortie de l'encodeur et ajuste symétriquement le codage à la dimension d'entrée d'origine. Toutes les couches cachées s'appuient sur une fonction d'activation, la tangente hyperbolique (tanh), tandis que la couche de sortie du décodeur repose sur une fonction d'activation Softmax (sigmoïde) dans le cas de variables catégoriques (numériques). Les imputations multiples sont générées par plusieurs exécutions dans lesquelles les poids des couches cachées sont initialisés sous forme de variable aléatoire gaussienne.

Selon Lu et coll. (2020), nous entraînons la MIDA en deux phases : une phase primaire et une phase de mise au point. Dans la phase primaire, nous transmettons les données initialement imputées à la MIDA et nous l'entraînons pendant N_{prime} périodes. Dans la phase de mise au point, la MIDA est entraînée en fonction des données de sortie de la phase primaire pendant N_{tune} périodes, et elle produit le résultat. La fonction de perte est utilisée dans les deux phases et ressemble étroitement à la perte de la fonction de reconstruction dans le GAIN :

$$L\left(Y_{ij_0}, \hat{Y}_{ij}, M_{ij}\right) = \begin{cases} \left(1-M_{ij}\right)\left(Y_{ij_0}-\hat{Y}_{ij}\right)^2 & \text{si } Y_{ij} \text{ est continu} \\ -\left(1-M_{ij}\right) Y_{ij_0} \log \hat{\mathbf{Y}}_{ij} & \text{si } Y_{ij} \text{ est catégorique.} \end{cases} \tag{2.6}$$

Pour mettre en œuvre la MIDA dans nos évaluations, nous utilisons la même architecture et nous ajustons les hyperparamètres dans une seule simulation, comme dans Lu et coll. (2020). Nous traçons l'évolution de la fonction de perte L et sélectionnons le nombre d'unités supplémentaires Θ parmi {1; 2; 3; 4; 5; 6; 7; 8; 9; 10} pour réduire la perte. Dans nos expériences, nous établissons θ comme étant équivalent au nombre de caractéristiques des données d'entrée et nous ajoutons $\Theta = 7$ unités cachées à chacune des trois couches cachées de l'encodeur. Nous entraînons le modèle pendant $N_{\text{prime}} = 100$ périodes dans la phase primaire et pendant $N_{\text{tune}} = 2$ périodes dans la phase de mise au point. Comme dans le GAIN, nous apprenons les paramètres du modèle à l'aide de la descente par gradient stochastique comprenant des mini-lots de taille 512, et nous utilisons l'optimiseur Adam pour adapter le taux d'apprentissage au taux initial de 0,001.

3. Évaluation par simulation des méthodes d'imputation

Les méthodes d'imputation des données manquantes sont habituellement évaluées au moyen de simulations fondées sur des données réelles (van Buuren, 2018). Plus précisément, on crée des valeurs manquantes à partir d'un ensemble de données complet selon un mécanisme de données manquantes (Little et Rubin, 2014), on impute les valeurs manquantes à l'aide d'une méthode particulière, puis on compare ces valeurs imputées aux « vraies » valeurs originales en fonction de certaines mesures.

Nous allons d'abord passer rapidement en revue les règles de Rubin sur les combinaisons de l'imputation multiple. Supposons que Q soit le paramètre cible dans la population, et que $q^{(l)}$ et $u^{(l)}$ soient, respectivement, l'estimateur ponctuel et l'estimateur de la variance de Q en fonction du l^{e} ensemble de données imputé. L'estimateur ponctuel de l'imputation multiple de Q est $\bar{q}_L = \sum_{l=1}^{L} q^{(l)} / L$, et l'estimateur correspondant de la variance est égale à $T_L = (1 + 1/L) b_L + \bar{u}_L$, où $b_L = \sum_{l=1}^{L} \left(q^{(l)} - \bar{q}_L\right)^2 / (L-1)$, et $\bar{u}_L = \sum_{l=1}^{L} u^{(l)} / L$. L'intervalle de confiance de Q est construit à l'aide de $(\bar{q}_L - Q) \sim t_v(0, T_L)$, où t_v est une distribution t ayant $v = (L-1)\left(1 + \bar{u}_L / \left[(1+1/L) b_L\right]\right)^2$ degrés de liberté.

Premièrement, notre procédure d'évaluation par simulation consiste à choisir un ensemble de données ayant toutes les valeurs observées, qui est considéré comme la « population ». Nous choisissons ensuite un ensemble de paramètres cibles Q et nous calculons leurs valeurs à partir de ces données sur la population, qui sont considérées comme la « réalité du terrain ». Les paramètres sont habituellement des statistiques sommaires des variables ou des paramètres dans un modèle d'analyse en aval, par exemple un coefficient dans un modèle de régression (Tang, Song, Belin et Unützer, 2005; Huque, Carlin, Simpson et Lee, 2018). Deuxièmement, nous tirons au hasard sans remise H échantillons de taille n à partir des données de la population et, dans chacun des échantillons $(h = 1, \ldots, H)$, nous créons des données manquantes en fonction d'un mécanisme de données manquantes précis et d'une proportion préétablie de valeurs manquantes. Troisièmement, pour chaque échantillon simulé pour lequel il manque des données, nous créons L ensembles de données imputés à l'aide de la méthode d'imputation à l'étude et construisons l'estimateur ponctuel et l'estimateur par intervalles de chaque paramètre au moyen des règles de Rubin. Enfin, nous calculons les indicateurs de rendement de chaque paramètre à partir des quantités obtenues à l'étape précédente.

Dans l'application empirique, nous sélectionnons un grand sous-échantillon complet à partir de l'*American Community Survey* (ACS), une enquête nationale qui porte sur les caractéristiques distinctives de nombreuses grandes données d'enquête et qui constitue notre population. Étant donné que les variables discrètes sont répandues dans l'ACS, ainsi que dans la plupart des données d'enquête, nous nous concentrons sur les probabilités marginales des variables binaires et catégoriques, par exemple une variable catégorique ayant K catégories possède $K-1$ paramètres. Afin d'évaluer dans quelle mesure les méthodes d'imputation préservent les propriétés de distribution multivariée, comme dans Akande et coll. (2017), nous tenons également compte des probabilités bivariées de toutes les combinaisons bidirectionnelles de catégories dans les variables binaires et catégoriques. Une autre mesure utile est la corrélation par paires d'échantillons finis entre des variables continues. Pour les variables continues, les paramètres communs sont la moyenne, la médiane ou la variance. Pour faciliter des comparaisons importantes des résultats entre les variables catégoriques et continues, nous proposons de discrétiser chaque variable continue en K catégories en fonction des quantiles de l'échantillon. Nous évaluons

ensuite ces variables continues regroupées en classes en tant que variables catégoriques en fonction des paramètres des probabilités marginales et bivariées mentionnés précédemment.

Pour chaque paramètre Q, nous considérons trois mesures. La première mesure met l'accent sur le biais. Pour tenir compte des paramètres proches de zéro qui prévalent dans les probabilités des variables catégoriques, nous considérons le biais standardisé absolu (BSA) de chaque paramètre Q comme suit :

$$\text{BSA} = \sum_{h=1}^{H} \left| \bar{q}_L^{(h)} - Q \right| \Big/ \left(H \cdot Q \right), \tag{3.1}$$

où $\bar{q}_L^{(h)}$ est l'estimateur ponctuel de l'imputation multiple de Q dans la simulation h.

La deuxième mesure est l'erreur quadratique moyenne relative (EQMR), qui correspond au ratio entre l'EQM de l'estimation de Q à partir des données imputées et celle de l'estimation à partir des données échantillonnées avant l'introduction des données manquantes :

$$\text{EQMR} = \frac{\sum_{h=1}^{H} \left(\bar{q}_L^{(h)} - Q \right)^2}{\sum_{h=1}^{H} \left(\tilde{Q}^{(h)} - Q \right)^2}, \tag{3.2}$$

où $\bar{q}_L^{(h)}$ a été défini plus tôt, et $\tilde{Q}^{(h)}$ est l'estimateur prototype de Q, c'est-à-dire l'estimateur ponctuel de toutes les données échantillonnées dans la simulation h.

La troisième mesure est le taux de couverture, qui correspond à la proportion des intervalles de confiance $\alpha\,\%$ (par exemple $95\,\%$), indiquée par $\text{IC}_h^{\alpha} \left(h = 1, \ldots, H \right)$, dans les simulations H qui contiennent le vrai Q:

$$\text{Couverture} = \sum_{h=1}^{H} \mathbf{1} \left\{ Q \in \text{IC}_h^{\alpha} \right\} \Big/ H. \tag{3.3}$$

Nous recommandons d'effectuer un grand nombre de simulations (par exemple $H \geq 100$) pour obtenir des estimations fiables de l'EQM et de la couverture. Cela ne poserait pas de problème pour les algorithmes d'apprentissage profond, qui peuvent généralement être exécutés en quelques secondes, même avec des échantillons de grande taille. Cependant, cela peut être prohibitif sur le plan des calculs d'utiliser des algorithmes de la MICE lorsque chacune des données simulées est grande (par exemple $n = 100\,000$ dans certaines de nos simulations). Dans les cas où l'on ne doit se fier qu'à quelques simulations, voire une seule, pour réaliser l'évaluation, nous proposons une mesure modifiée du biais. Plus précisément, pour chaque variable catégorique ou chaque variable regroupée en classes j, nous définissons le biais absolu pondéré comme la somme du biais absolu pondéré par la probabilité marginale réelle dans chaque catégorie :

$$\text{Biais absolu pondéré} = \sum_{k=1}^{K} Q_{jk} \left| \bar{q}_{jk}^{(h)} - Q_{jk} \right|, \tag{3.4}$$

où K est le nombre total de catégories, Q_{jk} est la probabilité marginale de la population de la catégorie k dans la variable j, et $\bar{q}_{jk}^{(h)}$ est son estimateur ponctuel correspondant dans la simulation h. Nous pouvons également calculer la moyenne du biais absolu pondéré en fonction d'un certain nombre d'échantillons simulés à plusieurs reprises.

La procédure et les mesures détaillées ci-dessus diffèrent de la pratique courante dans la littérature sur l'apprentissage automatique. Par exemple, dans le cadre de nombreuses études sur l'apprentissage automatique qui portent sur l'imputation des données manquantes, des simulations sont effectuées sur des ensembles de données de référence, mais ces données ont souvent une structure et des caractéristiques très différentes des données d'enquête et sont donc moins informatives selon l'objectif de la présente étude. L'un de ces ensembles de données est l'ensemble de données sur le cancer du sein qui se trouve dans le *Machine Learning Repository* de l'Université d'Irvine en Californie (Dua et Graff, 2017), qui ne compte que 569 unités d'échantillonnage et aucune variable catégorique. De plus, ces simulations sont habituellement fondées sur la création aléatoire et répétitive de valeurs manquantes d'un seul ensemble de données plutôt que sur le fait de tirer des échantillons d'une population à plusieurs reprises, ce qui ne tient pas compte du mécanisme d'échantillonnage. Ces évaluations reposent sur des mesures axées sur l'exactitude des prévisions individuelles plutôt que sur les caractéristiques de distribution. Plus précisément, les mesures les plus couramment utilisées sont la racine carrée de l'erreur quadratique moyenne (REQM) et l'exactitude (Gondara et Wang, 2018; Yoon, Jordon et Schaar, 2018; Lu et coll., 2020). Les deux mesures peuvent être définies d'une manière globale ou propre à une variable, mais la littérature sur l'apprentissage automatique est habituellement axée sur la version globale. La REQM globale est définie comme suit :

$$\text{REQM} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j} M_{ij}\left(\hat{Y}_{ij} - Y_{ij}\right)^2}{\sum_{i=1}^{n}\sum_{j} M_{ij}}}, \tag{3.5}$$

où Y_{ij} est la valeur de la variable continue j pour une personne i dans les données complètes avant l'introduction des données manquantes, et $\hat{Y}_{ij}$ est la valeur imputée correspondante. Pour les valeurs non manquantes (c'est-à-dire $M_{ij} = 1$), $Y_{ij} = \hat{Y}_{ij}$. L'exactitude (globale) est définie pour les variables catégoriques, c'est-à-dire la proportion des valeurs imputées étant égale à la « vraie » valeur originale correspondante :

$$\text{Exactitude} = \frac{\sum_{i=1}^{n}\sum_{j\in S_{\text{cat}}} M_{ij}\mathbf{1}\left(\hat{Y}_{ij} = Y_{ij}\right)}{\sum_{i=1}^{n}\sum_{j\in S_{\text{cat}}} M_{ij}}, \tag{3.6}$$

où S_{cat} est l'ensemble des variables catégoriques.

Un certain nombre de mises en garde s'imposent pour la REQM et les mesures d'exactitude. Premièrement, elles sont habituellement calculées selon un seul échantillon imputé comme mesure globale

d'une méthode d'imputation, mais cela ne tient pas compte de l'incertitude des imputations. Deuxièmement, la REQM et l'exactitude sont des sommaires de valeurs uniques et ne saisissent pas la caractéristique de distribution multivariée des données. Troisièmement, la REQM ne tient pas compte des différentes échelles de variables et peut facilement être dominée par quelques valeurs aberrantes. De plus, elle est souvent calculée sans différenciation entre les variables continues et catégoriques. Enfin, lorsqu'il y a plusieurs (L) données imputées, une méthode courante consiste à utiliser la moyenne de la valeur imputée L en tant que $\hat{Y}_{ij}$ dans l'équation (3.5), mais la signification statistique des mesures résultantes n'est pas claire. Cela est particulièrement problématique pour les variables catégoriques. Pour ces raisons, nous exprimons une mise en garde contre l'utilisation de la REQM globale et l'exactitude comme seules mesures pour comparer les méthodes d'imputation. Il faut faire preuve de prudence lorsqu'on les interprète.

4. Évaluation fondée sur l'*American Community Survey*

Dans la présente section, nous évaluons les quatre méthodes d'imputation décrites à la section 2 en suivant la procédure et les mesures décrites à la section 3. Par souci de simplicité, dans les analyses suivantes, nous utilisons l'arbre de classification et de régression (CART) et l'algorithme de forêt aléatoire (RF) pour désigner respectivement la MICE-CART et la MICE-RF.

4.1 Les données sur la « population »

Nous utilisons l'échantillon de microdonnées à grande diffusion d'une durée d'un an de l'ACS de 2018 pour construire notre population. Les données de l'ACS de 2018 contiennent des variables au niveau du ménage, par exemple, si une maison est possédée ou louée, et des variables au niveau individuel, par exemple, l'âge, le revenu et le sexe des personnes dans chaque ménage. Étant donné que les personnes qui vivent dans un ménage sont souvent dépendantes et que les méthodes d'imputation que nous évaluons supposent généralement l'indépendance de toutes les observations, nous avons établi notre unité d'observation au niveau du ménage, où l'indépendance est plus susceptible de tenir. Nous retirons d'abord les unités correspondant aux maisons vacantes. Ensuite, nous supprimons les unités pour lesquelles il manque des valeurs, afin de ne conserver que les cas complets. Dans chaque ménage, nous conservons également les données au niveau individuel correspondant uniquement au chef de ménage et nous les fusionnons avec les variables au niveau du ménage, ce qui donne un ensemble riche de variables comportant des relations conjointes potentiellement complexes.

Il est souvent difficile de générer des imputations plausibles pour des variables ordinales à plusieurs niveaux lorsqu'il y a une très faible masse aux niveaux les plus élevés, comme c'est le cas pour certaines variables dans les données de l'ACS. Selon Li, Baccini, Mealli, Zell, Frangakis et Rubin (2014), nous

traitons les variables ordinales de plus de 10 niveaux comme des variables continues. Nous suivons également l'approche d'Akande et coll. (2017) qui consiste à exclure les variables binaires où les probabilités marginales ne correspondent pas à $np > 10$ ou à $n(1-p) > 10$; ceci permet d'éliminer les paramètres là où le théorème de la limite centrale risque de ne pas de tenir. Pour chaque variable catégorique comportant plus de deux niveaux, mais moins de 10 niveaux où cela pourrait également poser un problème, nous fusionnons les niveaux avec un petit nombre d'observations dans les données sur la population. Par exemple, pour la variable de la langue du ménage, nous reprogrammons les niveaux de cinq à trois (anglais, espagnol et autres), parce que la probabilité de ne parler ni l'anglais ni l'espagnol dans l'ensemble de la population est inférieure à 8,8 %.

Les données finales sur la population contiennent 1 257 501 unités, avec 18 variables binaires, 20 variables catégoriques ayant de 3 à 9 niveaux et 8 variables continues. Nous décrivons les variables plus en détail dans la documentation supplémentaire. Nous calculons les valeurs de population des paramètres Q décrits à la section 3, y compris toutes les probabilités marginales et bivariées de variables continues discrètes et regroupées en classes. Nous faisons varier la taille des échantillons simulés de 10 000 à 100 000 et simulons les données manquantes soit selon le mécanisme de données manquantes complètement au hasard (DMCH), soit selon le mécanisme de données manquantes au hasard (DMH) dans chacun de ces scénarios.

4.2 Simulations où n = 10 000

Nous tirons d'abord au hasard $H = 100$ échantillons d'une taille $n = 10\,000$ et nous déterminons qu'il manque 30 % de chaque échantillon selon le mécanisme de DMCH ou selon le mécanisme de DMH. Le CART ou la RF prend environ 2,8 et 9,2 heures, respectivement, pour créer $L = 10$ ensembles de données imputés qui comportent des paramètres par défaut sur un ordinateur de bureau standard ayant une seule unité centrale de traitement. Les méthodes d'apprentissage en profondeur sont beaucoup plus rapides parce qu'elles tirent parti de la puissance de calcul du processeur graphique lorsqu'elles sont mises en œuvre dans le cadre logiciel de TensorFlow optimisé pour le processeur graphique (Abadi, Agarwal, Barham, Brevdo, Chen, Citro, Corrado, Davis, Dean, Devin, Ghemawat, Goodfellow, Harp, Irving, Isard, Jia, Jozefowicz, Kaiser, Kudlur, Levenberg, Mané, Monga, Moore, Murray, Olah, Schuster, Shlens, Steiner, Sutskever, Talwar, Tucker, Vanhoucke, Vasudevan, Viégas, Vinyals, Warden, Wattenberg, Wicke, Yu et Zheng, 2015). Le GAIN prend environ 1,5 minute et la MIDA prend environ 4 minutes pour créer $L = 10$ ensembles de données complets au moyen d'un processeur graphique GeForce GTX 1660 Ti. Il convient de mentionner qu'il est impossible d'ajuster manuellement l'hyperparamètre dans chacune des 100 simulations de chaque scénario pour les modèles d'apprentissage profond. Donc, pour chaque scénario, nous avons choisi au hasard une simulation et nous avons ajusté les hyperparamètres en suivant la procédure décrite à la section 2. Nous avons ensuite appliqué ces hyperparamètres sélectionnés à toutes les simulations.

4.2.1 Scénario des données manquantes complètement au hasard

Pour créer le scénario de DMCH, nous établissons au hasard 30 % des valeurs de chaque variable en tant que valeurs manquantes de façon indépendante. Le tableau 4.1 présente les distributions du BSA estimé et de l'EQMR de toutes les probabilités marginales et bivariées dans les données imputées selon les quatre méthodes d'imputation.

Dans l'ensemble, pour les paramètres des probabilités marginales et bivariées des variables continues catégoriques et regroupées en classes, la MICE avec CART dépasse de façon importante les trois autres méthodes, produisant constamment le plus petit BSA et la plus petite EQMR. La RF se classe au deuxième rang, dépassant constamment les méthodes d'apprentissage profond. L'avantage des algorithmes de la MICE est particulièrement prononcé dans les quantiles supérieurs (par exemple 75 % et 90 %), ce qui indique que les imputations GAIN et MIDA présentent de grandes variations par rapport à des échantillons et des variables répétés. En effet, la MIDA et le GAIN mènent à des queues très longues dans l'estimation des statistiques sommaires des variables. Par exemple, pour les probabilités bivariées de variables continues regroupées en classes, le centile de 90 % du BSA de la MIDA et du GAIN est environ 20 et 27 fois plus grand que celui du CART, respectivement. L'écart est encore plus grand pour l'EQMR. Il n'y a pas de tendance cohérente lorsqu'on compare la MIDA et le GAIN. Plus précisément, pour les variables continues, la MIDA dépasse généralement le GAIN, mais la différence est faible, sauf pour les centiles supérieurs, où la GAIN a tendance à produire une EQMR et un biais très importants. Pour les variables catégoriques, le GAIN dépasse la MIDA la moitié du temps, mais, encore une fois, entraîne la plus grande variation des imputations entre les variables. De plus, une observation intéressante et quelque peu étonnante est que la MICE avec CART dépasse constamment la RF, parfois de beaucoup, quel que soit le choix du paramètre ou de la mesure.

Tableau 4.1
Distributions du biais standardisé absolu (BSA) (×100) et de l'erreur quadratique moyenne relative (EQMR) de toutes les probabilités marginales et bivariées selon les imputations découlant des quatre méthodes d'imputation multiple, lorsque *n* = 10 000 et que 30 % des valeurs sont manquantes complètement au hasard

	Quantiles		Probabilités marginales				Probabilités bivariées			
			CART	RF	GAIN	MIDA	CART	RF	GAIN	MIDA
BSA (×100)	Cat.	10 %	0,05	0,47	0,76	0,98	0,15	1,14	1,21	1,54
		25 %	0,13	1,25	1,48	2,22	0,40	2,83	3,08	3,93
		50 %	0,27	2,80	3,22	4,69	1,05	6,74	7,14	8,47
		75 %	0,64	5,86	7,18	8,86	2,51	13,59	17,03	15,23
		90 %	1,14	10,01	19,55	14,41	5,34	22,33	26,92	21,90
	Cont.	10 %	0,06	0,24	7,25	2,73	0,19	1,30	6,05	4,80
		25 %	0,10	1,05	12,86	8,36	0,43	3,24	17,61	12,01
		50 %	0,21	3,59	27,30	18,51	1,02	6,61	34,29	24,07
		75 %	0,43	5,43	30,21	26,84	1,90	11,76	49,38	39,54
		90 %	0,81	8,49	46,41	31,36	3,42	20,79	90,90	64,65

« Cat. » désigne les variables catégoriques et « Cont. » désigne les variables continues regroupées en classes.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

Tableau 4.1(suite)
Distributions du biais standardisé absolu (BSA) (×100) et de l'erreur quadratique moyenne relative (EQMR) de toutes les probabilités marginales et bivariées selon les imputations découlant des quatre méthodes d'imputation multiple, lorsque *n* = 10 000 et que 30 % des valeurs sont manquantes complètement au hasard

	Quantiles		Probabilités marginales				Probabilités bivariées			
			CART	RF	GAIN	MIDA	CART	RF	GAIN	MIDA
EQMR	Cat.	10 %	1,05	1,67	2,50	3,38	0,96	1,11	2,75	2,98
		25 %	1,16	2,40	4,97	9,03	1,08	1,61	4,33	4,75
		50 %	1,37	5,99	10,37	14,89	1,25	3,35	7,40	8,16
		75 %	1,49	10,25	27,73	26,16	1,48	9,07	14,87	15,80
		90 %	1,62	16,22	97,33	40,16	1,89	23,91	36,37	27,92
	Cont.	10 %	1,19	1,50	44,06	4,35	0,82	0,86	7,40	2,05
		25 %	1,30	1,77	74,42	13,82	0,92	1,11	14,80	4,90
		50 %	1,44	3,31	139,24	72,57	1,07	1,90	32,26	13,76
		75 %	1,55	6,71	284,00	150,35	1,26	4,09	88,78	47,56
		90 %	1,64	19,69	603,38	451,44	1,54	10,80	282,29	127,15

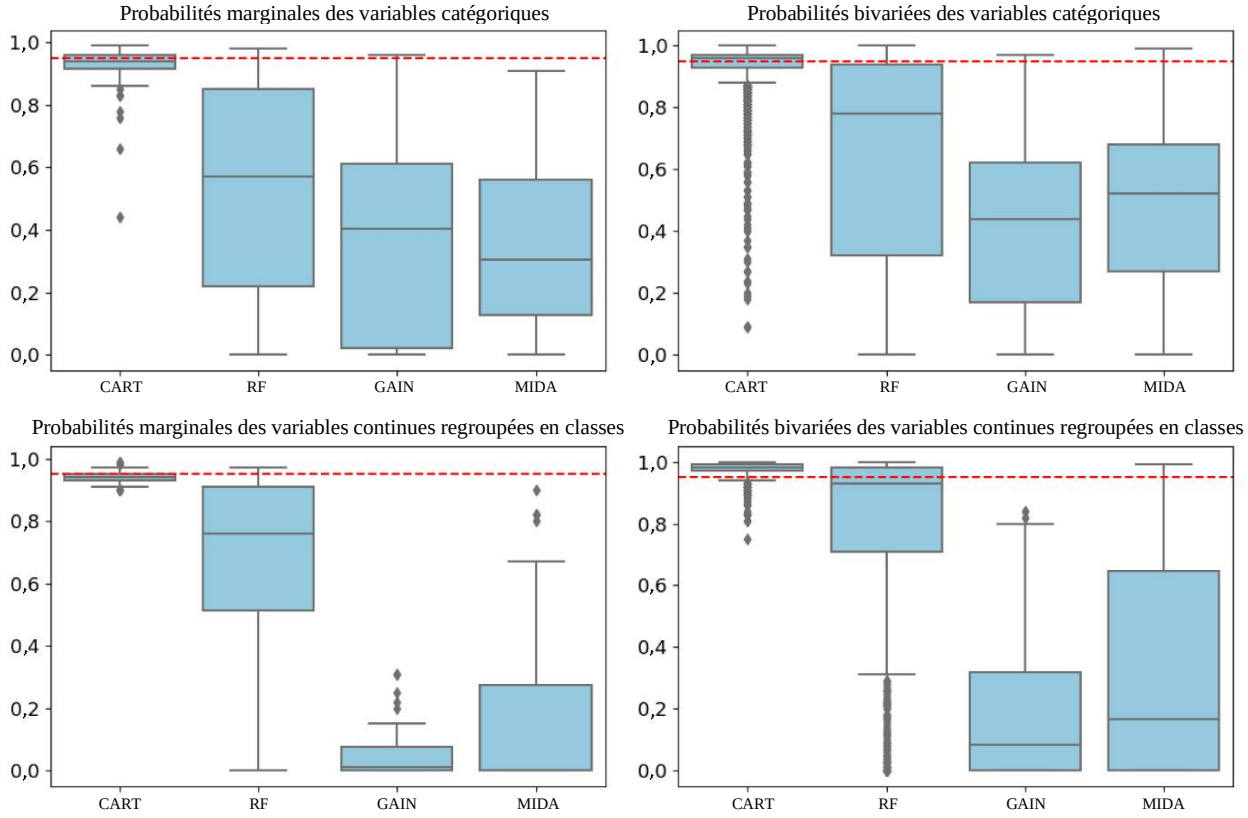
« Cat. » désigne les variables catégoriques et « Cont. » désigne les variables continues regroupées en classes.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

Toutes les méthodes donnent généralement des estimations moins biaisées (c'est-à-dire de plus petits BSA) des probabilités marginales que des probabilités bivariées. Ceci montre que la préservation des caractéristiques de distribution multivariée est plus difficile que celle des caractéristiques de distribution univariée. L'avantage du CART par rapport aux autres méthodes est comparativement plus important lorsqu'on estime des paramètres bivariés par rapport aux paramètres univariés. Fait intéressant, l'EQMR a tendance à être plus élevée pour les probabilités marginales que pour les probabilités bivariées. Cela est probablement dû au fait que le dénominateur de la définition de l'EQMR vu dans l'équation (3.2) correspond à l'EQM des données échantillonnées avant l'introduction des données manquantes, qui ont tendance à être plus faibles pour les probabilités marginales que pour les probabilités bivariées. Le CART donne des EQM qui sont très proches des EQM correspondantes à partir des données échantillonnées avant l'introduction des données manquantes; c'est-à-dire que l'EQMR est proche de 1. Au contraire, les deux méthodes d'apprentissage profond, et le GAIN en particulier, peuvent entraîner une EQMR extrêmement importante pour plusieurs paramètres.

Les figures 4.1 présentent les taux de couverture estimés des intervalles de confiance de 95 % pour les probabilités marginales et bivariées. Les tendances relatives à la couverture entre différentes méthodes sont semblables à celles relatives au biais et à l'EQM. Plus précisément, le CART tend à produire des taux de couverture qui se rapprochent du niveau nominal de 95 %, la médiane étant constamment autour de 95 % et correspondant à un écart interquartile serré. En revanche, les taux de couverture de la RF, du GAIN et de la MIDA sont beaucoup plus éloignés du niveau nominal de 95 %. Par exemple, les taux de couverture médians pour le GAIN et la MIDA sont tous inférieurs à 0,60, et même inférieurs à 0,30 pour les variables continues. Un examen plus approfondi de l'exactitude des prévisions de chaque variable révèle que le GAIN et la MIDA ont tendance à générer des imputations biaisées vers les niveaux les plus fréquents, et que le GAIN, en particulier, produit généralement des intervalles plus étroits que les autres

méthodes. Cela prouve une fois de plus que les méthodes d'apprentissage profond produisent des biais importants. Toutes les méthodes ont tendance à produire des taux de couverture médians plus élevés pour les probabilités bivariées que pour les probabilités marginales, bien que les queues à gauche soient généralement plus longues pour les probabilités bivariées que pour les probabilités marginales.

Figure 4.1 Taux de couverture de l'intervalle de confiance de 95 % pour toutes les probabilités marginales et bivariées obtenues à partir des quatre méthodes d'imputation dans les simulations comprenant *n* = 10 000 et 30 % de valeurs manquantes complètement au hasard.



La ligne pointillée rouge correspond à 0,95.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network; MIDA = Multiple imputation using denoising autoencoders.

4.2.2 Scénario de données manquantes au hasard

Nous considérons également un scénario de DMH, qui est plus plausible que le scénario de DMCH en pratique. Nous avons établi six variables à observer en entier : l'âge, le sexe, l'état matrimonial, la race, le niveau de scolarité et la catégorie de travailleur. Il serait fastidieux de spécifier un mécanisme de DMH différent pour chacune des 40 variables restantes, alors nous les divisons aléatoirement en trois groupes, dont un composé de 10 variables et deux composés de 15 variables chacun. Nous précisons ensuite un modèle de non-réponse distinct que nous utiliserons pour générer les données manquantes pour les variables de chaque groupe. Plus précisément, nous postulons un modèle logistique par groupe,

conditionnel aux six variables entièrement observées, sur la base desquelles nous générons des indicateurs binaires de données manquantes pour chaque variable de ce groupe. Ce processus entraîne un taux de données manquantes d'environ 30 % pour chacune des 40 variables. Nous décrivons les modèles plus en détail dans la documentation supplémentaire.

Le tableau 4.2 présente les distributions du BSA et de l'EQMR de toutes les probabilités marginales et bivariées obtenues à partir des quatre méthodes. Toutes les méthodes donnent un BSA et une EQMR plus importants selon le scénario de DMH que selon le scénario de DMCH précédent. C'est normal, car le mécanisme de DMH correspond à une hypothèse plus forte que le mécanisme de DMCH, lequel exige le conditionnement d'un plus grand nombre de renseignements. Néanmoins, les tendances globales du rendement relatif entre les méthodes demeurent les mêmes que celles du mécanisme de DMCH. Plus précisément, le CART produit une fois de plus des estimations ayant les plus faibles BSA et EQMR, ce qui représente une marge encore plus grande que la méthode selon le mécanisme de DMCH parmi les quatre méthodes. Le CART est suivi de la RF, puis de la MIDA et du GAIN. Une observation notable est la détérioration du rendement des méthodes d'apprentissage profond, en particulier du GAIN, lors de l'imputation de variables continues, ce qui aboutit parfois à une multiplication par plusieurs centaines de l'EQMR par rapport au CART. Cela dénote les énormes incertitudes associées au GAIN lors de l'imputation de variables continues.

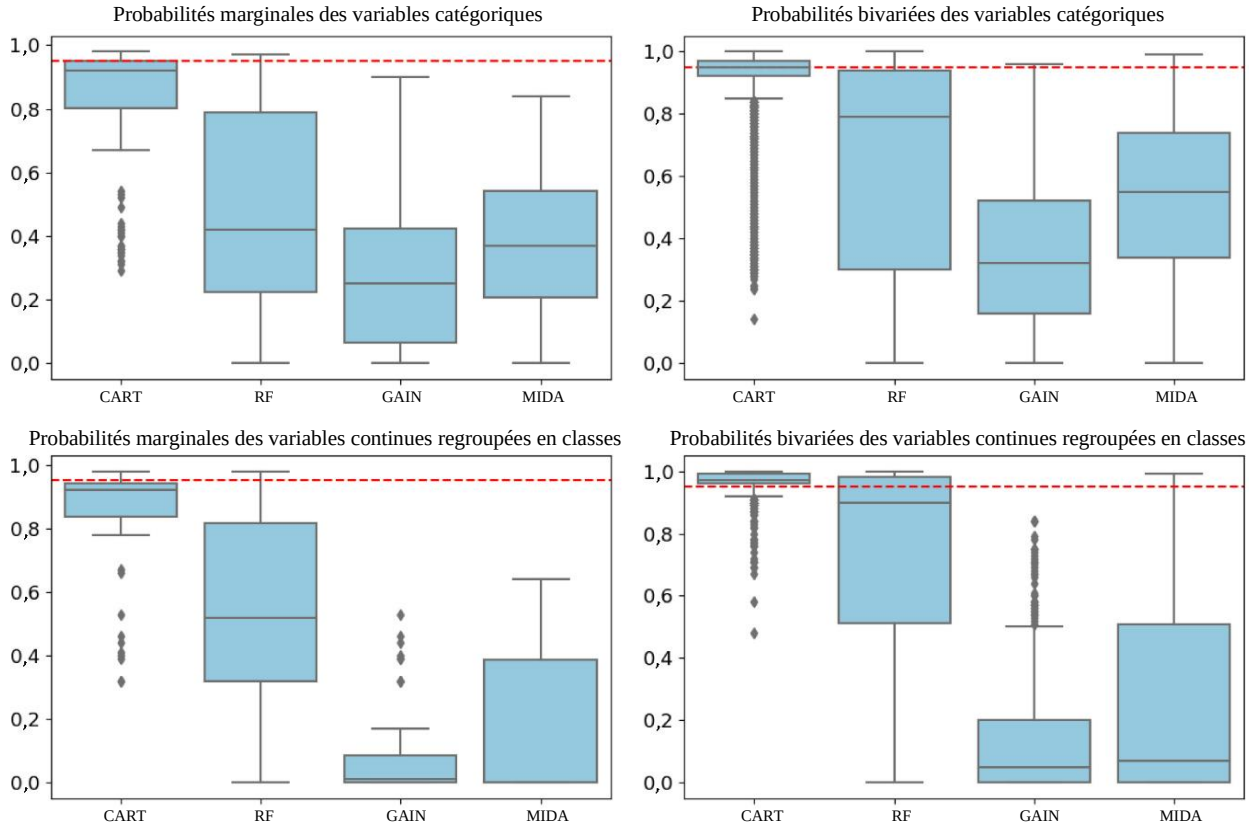
Tableau 4.2
Distributions du biais standardisé absolu (BSA) (×100) et de l'erreur quadratique moyenne relative (EQMR) pour toutes les méthodes, lorsque *n* = 10 000 et que 30 % des valeurs ne sont pas manquantes complètement au hasard, selon la totalité des probabilités marginales et bivariées possibles

	Quantiles		Probabilités marginales				Probabilités bivariées			
			CART	RF	GAIN	MIDA	CART	RF	GAIN	MIDA
BSA (×100)	Cat.	10 %	0,05	0,13	0,15	0,14	0,15	0,71	0,76	0,89
		25 %	0,11	0,44	0,62	0,61	0,40	2,23	2,55	3,20
		50 %	0,29	2,13	3,05	4,55	1,08	6,06	6,85	8,14
		75 %	1,04	4,98	6,63	10,22	2,49	13,43	16,78	16,19
		90 %	1,80	10,49	18,91	17,00	5,68	24,06	28,04	25,36
	Cont.	10 %	0,07	0,29	0,33	0,33	0,27	1,17	10,87	6,18
		25 %	0,17	1,07	9,64	3,13	0,69	3,49	23,67	16,26
		50 %	0,67	3,14	32,86	23,85	1,58	7,83	38,52	31,17
		75 %	1,20	6,95	39,57	36,09	3,40	15,20	53,59	47,34
		90 %	3,40	12,39	63,45	41,99	5,94	25,16	97,47	85,44
EQMR	Cat.	10 %	1,00	1,00	1,00	1,00	0,97	1,00	1,53	1,93
		25 %	1,08	1,82	2,56	4,75	1,04	1,39	3,78	4,03
		50 %	1,33	4,33	19,03	15,13	1,25	3,00	10,42	8,38
		75 %	1,72	13,08	55,07	33,36	1,59	9,56	27,45	16,95
		90 %	2,27	18,70	101,91	48,44	2,23	27,44	64,01	32,85
	Cont.	10 %	1,00	1,00	1,00	1,00	0,88	0,90	11,19	2,96
		25 %	1,38	1,83	90,98	8,49	1,00	1,16	20,15	6,87
		50 %	1,70	4,57	207,58	96,08	1,18	2,29	45,25	21,33
		75 %	2,12	11,47	692,67	239,69	1,50	6,95	125,39	70,90
		90 %	3,12	50,56	1342,23	806,43	2,12	18,07	459,78	205,14

« Cat. » désigne les variables catégoriques et « Cont. » désigne les variables continues regroupées en classes.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

Les figures 4.2 présentent les taux de couverture estimés des intervalles de confiance de 95 % pour les probabilités marginales et bivariées, selon chaque méthode. Comme dans le cas du biais et de l'EQM, toutes les méthodes entraînent généralement des taux de couverture inférieurs à ceux du mécanisme de DMH par rapport au mécanisme de DMCH. Les queues à gauche sont visiblement plus longues dans certains cas, mais les tendances globales, lorsqu'on compare les méthodes, demeurent les mêmes. Plus précisément, le CART a toujours tendance à donner des taux de couverture supérieurs à 90 %, tandis que les trois autres méthodes ont toujours un taux de couverture inférieur. En particulier, le GAIN et la MIDA se traduisent par des taux de couverture médians extrêmement bas, en dessous de 7 %, pour les variables continues. Ceci est étroitement lié à l'observation précédente de la grande incertitude des méthodes d'apprentissage profond dans l'imputation de variables continues.

Figure 4.2 Taux de couverture des intervalles de confiance de 95 % pour toutes les probabilités marginales et bivariées obtenues à partir des quatre méthodes dans les simulations comportant n = 10 000 et 30 % de valeurs manquantes au hasard.



La ligne pointillée rouge correspond à 0,95.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network; MIDA = Multiple imputation using denoising autoencoders.

Enfin, pour illustrer le fait qu'il peut être trompeur d'évaluer seulement la REQM et les mesures d'exactitude globales, nous présentons les erreurs-types moyennes et empiriques de la REQM globale et de l'exactitude globale des 100 simulations au tableau 4.3, où les DMCH se trouvent dans la partie supérieure et les DMH, dans la partie inférieure. Dans le cas des deux mécanismes de données manquantes, pour les variables continues, la MIDA permet d'obtenir la plus petite REQM globale, suivie du CART ainsi que de la RF et du GAIN qui viennent en dernier. Pour les variables catégoriques, le CART et le GAIN permettent d'obtenir l'exactitude globale la plus élevée, la MIDA étant tout juste derrière et la RF apparaissant en dernier. Ces tendances, qui ne sont pas étonnantes, diffèrent de celles déclarées précédemment en fonction de probabilités marginales et bivariées et de mesures différentes. Comme nous avons vu à la section 3, la REQM et l'exactitude globales ne saisissent pas les caractéristiques de distribution des données multivariées ou les propriétés d'échantillonnage répété des méthodes d'imputation.

Tableau 4.3
Moyennes selon 100 simulations de la racine carrée de l'erreur quadratique moyenne (REQM) globale des variables continues et l'exactitude globale des variables catégoriques

Mécanisme	Mesure	CART	RF	GAIN	MIDA
DMCH	REQM	0,128 (0,002)	0,159 (0,003)	0,161 (0,008)	0,112 (0,002)
	Exactitude	0,785 (0,001)	0,658 (0,003)	0,782 (0,002)	0,752 (0,004)
DMH	REQM	0,130 (0,003)	0,154 (0,004)	0,145 (0,009)	0,110 (0,002)
	Exactitude	0,819 (0,001)	0,704 (0,003)	0,820 (0,002)	0,780 (0,007)

Les erreurs-types empiriques sont indiquées entre parenthèses.
Les données manquantes complètement au hasard (DMCH) se trouvent dans la partie supérieure et les données manquantes au hasard (DMH) se trouvent dans la partie inférieure, et 30 % de toutes les données sont manquantes.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

4.3 Simulations où $n = 100\,000$ et 30 % des valeurs sont manquantes complètement au hasard

Les modèles d'apprentissage profond exigent habituellement un échantillon de grande taille pour l'entraînement. Par conséquent, pour donner à la MIDA et au GAIN un cadre plus favorable et pour étudier la sensibilité de nos résultats aux variations de la taille de l'échantillon, nous générons un scénario de simulation de $H = 10$ échantillons et de $n = 100\,000$ selon le mécanisme de DMCH. Autrement dit, nous établissons au hasard 30 % des valeurs de chaque variable en tant que valeurs manquantes indépendamment. Dans la présente étude, nous ne générons que 10 simulations en raison des énormes coûts liés au calcul selon la MICE pour des échantillons de cette taille. Dans ce scénario, nous omettons la RF, car les résultats précédents de la section 4.2 ont montré que la RF est toujours inférieure au CART en ce qui a trait au rendement et au calcul. Nous utilisons le CART, le GAIN et la MIDA pour créer $L = 10$ ensembles de données complets.

Comme il faut habituellement un nombre beaucoup plus grand de simulations pour calculer de façon fiable l'EQM et la couverture, nous nous concentrons sur la mesure du biais absolu pondéré (3.4). Le tableau 4.4 présente les distributions du biais absolu pondéré estimé (dont la moyenne est établie sur 10 simulations) des probabilités marginales des variables catégoriques et des variables continues regroupées en classes. Dans l'ensemble, les tendances, lorsqu'on compare les quatre méthodes, demeurent conformes à celles observées à la section 4.2. Plus précisément, le CART entraîne de nouveau la plus petite différence absolue pondérée dans les variables catégoriques et continues, et l'avantage est particulièrement prononcé pour les variables continues. Par exemple, pour les variables catégoriques, la MIDA et le GAIN donnent une médiane du biais absolu pondéré au moins 9 et 11 fois plus élevée que le CART, respectivement. L'avantage du CART augmente à environ 30 et 60 fois par rapport aux variables continues avec la MIDA et le GAIN, respectivement. De plus, le CART donne de bons résultats pour l'ensemble des variables, comme en témoigne la faible variation du biais absolu pondéré, par exemple 0,07 pour le centile de 10 % et 0,33 pour le centile de 90 % des variables catégoriques. En revanche, les deux modèles d'apprentissage profond entraînent une variation beaucoup plus grande entre les variables, par exemple 0,57 pour le centile de 10 % et 2,92 pour le centile de 90 % des variables catégoriques selon la MIDA, et encore plus pour le GAIN. En résumé, en dehors du temps de calcul, la MICE avec CART dépasse de façon considérable la MIDA et le GAIN en ce qui concerne le biais et la variance, indépendamment de la taille de l'échantillon.

Tableau 4.4
Distributions du biais absolu pondéré ($\times 100$) et de la moyenne sur 10 échantillons simulés, où pour chacun n = 10 000 et 30 % des valeurs sont manquantes complètement au hasard

Quantiles	Variables catégoriques			Variables continues regroupées par classes		
	CART	GAIN	MIDA	CART	GAIN	MIDA
10 %	0,07	0,43	0,57	0,10	5,52	1,98
25 %	0,11	1,11	1,02	0,11	6,65	2,78
50 %	0,15	1,74	1,40	0,12	7,36	4,04
75 %	0,24	3,77	2,07	0,13	9,40	6,50
90 %	0,33	4,63	2,92	0,15	11,31	7,72

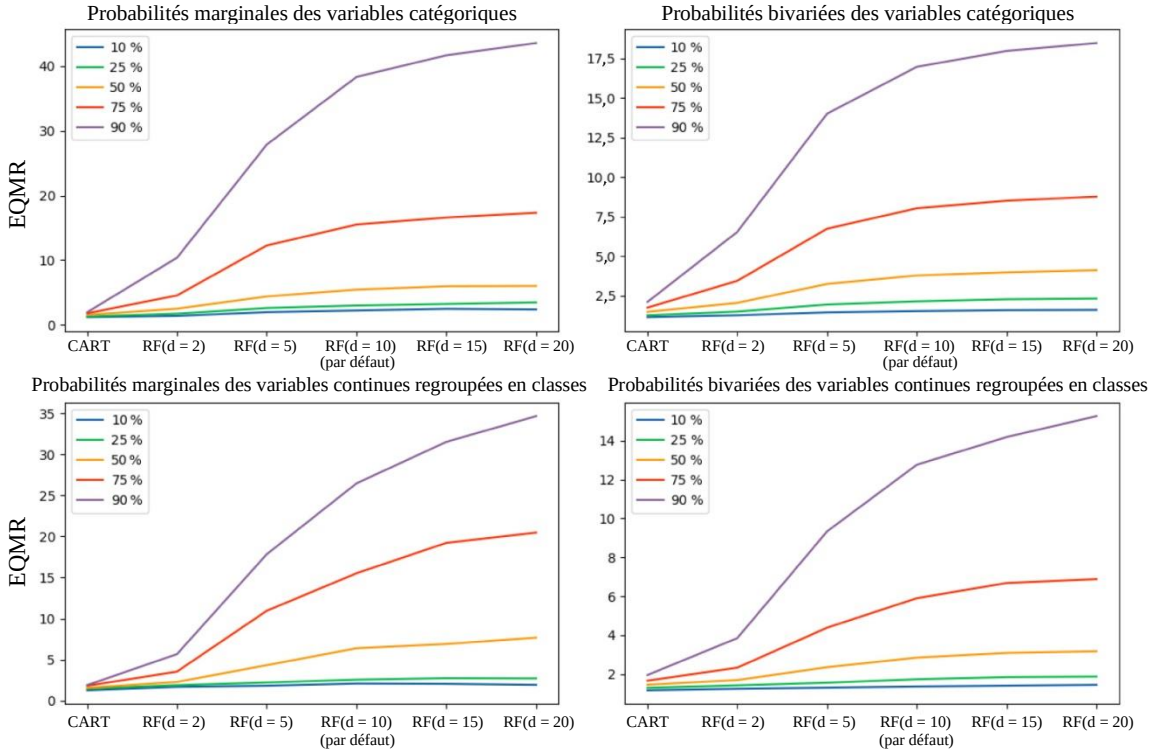
CART = Classification and regression trees; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

4.4 Rôle des hyperparamètres dans l'imputation multiple au moyen d'équations en séries fondée sur l'arborescence

La tendance selon laquelle le CART dépasse la RF est étonnante, car on sait que les méthodes d'ensemble sont habituellement supérieures aux méthodes à arbre unique. Cependant, la même tendance a également été observée dans une autre étude récente (Wongkamthong et Akande, 2021). Nous étudions le rôle de l'hyperparamètre clé dans la RF, à savoir le nombre maximum d'arbres d, dans les simulations. Nous avons sélectionné au hasard des données simulées de taille $n = 10\,000$ et 30 % des entrées manquantes complètement au hasard. Nous utilisons le progiciel `mice` pour ajuster la RF en fonction des

différents nombres d'arbres : $d = 2, 5, 10, 15, 20,$ où $d = 10$ est le paramètre par défaut. L'EQMR des variables catégoriques imputées, ajustée à l'aide de chaque valeur d, ainsi que celle reposant sur le CART, est présentée sous forme de trajectoires à la figure 4.3. Ces trajectoires révèlent une tendance constante : les quantiles supérieurs de l'EQMR, en particulier ceux de plus de 50 %, se détériorent rapidement à mesure que le nombre maximal d'arbres de la RF augmente, tandis que les quantiles inférieurs, par exemple ceux de 10 % et de 25 %, demeurent stables. Nous avons constaté une tendance semblable en ce qui a trait à la mesure du biais standardisé et aux variables continues; par conséquent, les résultats sont omis dans la présente étude. Cela laisse entendre qu'un plus grand nombre d'arbres dans la RF (du moins tel que cela est mis en œuvre dans le progiciel `mice`) entraîne une queue beaucoup plus longue dans la distribution du biais et des EQM. Cela est probablement dû à un surajustement. Nous ne pouvons pas exclure la possibilité qu'un ajustement plus personnalisé des hyperparamètres de la RF puisse dépasser le CART dans certaines applications. Toutefois, un tel ajustement précis propre à un cas de l'algorithme MICE n'est généralement pas accessible à la grande majorité des utilisateurs de l'imputation multiple qui se fient à l'ajustement par défaut des progiciels populaires comme `mice`.

Figure 4.3 Quantiles de l'erreur quadratique moyenne relative (EQMR) pour l'ensemble des probabilités marginales et bivariées des variables catégoriques et continues regroupées en classes, selon les modèles par arbres de classification et de régression (CART) et l'algorithme de forêt aléatoire (RF) comprenant différents nombres d'arbres, pour un échantillon de simulation ayant *n* = 10 000 et 30 % des valeurs manquantes complètement au hasard.



5. Évaluation fondée sur des ensembles de données « de référence »

Pour vérifier les évaluations dans les études sur le GAIN et la MIDA (Gondara et Wang, 2018; Yoon, Jordon et Schaar, 2018; Lu et coll., 2020), nous avons également comparé les deux modèles d'apprentissage profond avec le CART en fonction des cinq ensembles de données de référence et de la procédure de simulation (différente de notre cadre proposé) utilisés dans ces études. Les précisions relatives à ces ensembles de données et à ces simulations sont présentées dans la documentation supplémentaire. Les tailles d'échantillon de ces données ne sont généralement pas assez grandes pour être considérées comme des données de population à partir desquelles nous pouvons échantillonner à répétition sans remise, de sorte que nous ne sommes pas en mesure de les évaluer de façon considérable à l'aide du biais standardisé absolu, de l'EQMR ou de la couverture. Nous évaluons donc les méthodes principalement en fonction de la mesure du biais absolu pondéré. En résumé, le CART dépasse encore de façon constante et importante la MIDA et le GAIN en ce qui concerne le biais absolu pondéré des variables catégoriques et continues, et ce, pour les cinq ensembles de données de référence. L'écart de rendement est particulièrement prononcé pour les variables continues. Nous avons également calculé l'EQM et l'exactitude globales comme les auteurs de ces études l'ont fait. À l'exception d'un ensemble de données, nous n'avons pas pu reproduire les résultats présentés dans ces études, même en utilisant le code des auteurs. Une raison possible est que la documentation du processus d'ajustement et de sélection des hyperparamètres du modèle serait obscure, ce qui est vrai dans le cas présent. Vous trouverez de plus amples renseignements dans la documentation supplémentaire en ligne.

6. Conclusion

Ces dernières années, de nombreuses méthodes d'imputation de données manquantes fondées sur l'apprentissage automatique ont vu le jour, ce qui a suscité l'espoir qu'il y ait des améliorations par rapport aux méthodes d'imputation plus conventionnelles comme la MICE. Cependant, les efforts pour évaluer ces méthodes dans des situations réelles demeurent rares. Dans la présente étude, nous adoptons un cadre d'évaluation des simulations reposant sur des données réelles. Nous menons de vastes études de simulation qui sont fondées sur l'*American Community Survey* afin de comparer les propriétés d'échantillonnage répété de deux méthodes de la MICE et de deux méthodes d'imputation par apprentissage profond s'appuyant sur le GAN (GAIN) et les autoencodeurs débruiteurs (MIDA).

Nous avons découvert que les modèles d'apprentissage profond ont un grand avantage sur le plan des calculs par rapport aux méthodes de la MICE, en partie parce qu'ils peuvent exploiter la puissance du processeur graphique pour effectuer des calculs de manière très efficace. Cependant, nos simulations ainsi que notre évaluation de plusieurs données de « référence » indiquent que la MICE avec la spécification du CART des modèles conditionnels dépasse constamment, et habituellement de façon substantielle, les modèles d'apprentissage profond en ce qui a trait au biais, à l'erreur quadratique moyenne et à la

couverture dans un large éventail de contextes réalistes. En particulier, le GAIN et la MIDA ont tendance à générer des imputations instables comportant d'énormes variations des échantillons répétés comparativement à la MICE. Une explication possible est que les réseaux neuronaux profonds excellent dans la détection de sous-structures complexes de mégadonnées, mais peuvent ne pas convenir aux données ayant une structure simple, comme les données simulées utilisées dans la présente étude. Une autre possibilité est que la taille des échantillons dans nos simulations ne soit pas suffisante pour former des réseaux neuronaux profonds, lesquels exigent habituellement beaucoup plus de données que les modèles statistiques conventionnels.

Ces résultats contredisent les constatations précédentes fondées sur la seule mesure de rendement de l'erreur quadratique moyenne globale qui proviennent de la littérature sur l'apprentissage automatique (par exemple Gondara et Wang, 2018; Yoon, Jordon et Schaar, 2018; Lu et coll., 2020). Cet écart met en évidence les pièges de la pratique courante de l'évaluation des méthodes d'imputation dans la littérature sur l'apprentissage automatique. Il démontre également l'importance d'évaluer les propriétés de l'échantillonnage répété selon plusieurs paramètres des méthodes d'imputation multiple. Une découverte intéressante est que les ensembles d'arbres (par exemple la RF) ne s'améliorent pas selon un seul arbre (par exemple le CART) dans le contexte de la MICE, ce qui correspond aux résultats d'une autre étude récente (Wongkamthong et Akande, 2021). En combinaison avec le fait que la méthode RF est plus intensive sur le plan des calculs que la méthode CART, nous recommandons d'utiliser en pratique la MICE avec CART plutôt que la RF.

Notre étude comporte quelques limites. Premièrement, il existe de nombreuses méthodes d'apprentissage profond qui peuvent être adaptées à l'imputation des données manquantes et elles peuvent toutes avoir des caractéristiques de fonctionnement différentes. Nous avons choisi le GAIN et la MIDA parce que les réseaux antagonistes génératifs et les autoencodeurs débruiteurs sont des méthodes d'apprentissage profond extrêmement reconnues, et les méthodes d'imputation fondées sur ces méthodes ont été reconnues comme étant supérieures à la MICE. Néanmoins, il serait souhaitable d'examiner d'autres méthodes d'imputation fondées sur l'apprentissage profond dans les travaux de recherche futurs. Deuxièmement, le rendement des méthodes d'apprentissage automatique dépend fortement de la sélection des hyperparamètres. On peut donc soutenir que le rendement inférieur du GAIN et de la MIDA peut être attribuable au moins en partie à la sélection d'hyperparamètres sous-optimaux. Cependant, les spécialistes s'appuieraient probablement sur les valeurs d'hyperparamètres par défaut pour toute méthode d'imputation par apprentissage automatique, ce qui est en fait ce que nous avons adopté dans nos simulations et ce qui représente donc la pratique réelle. Troisièmement, nous n'avons pas tenu compte de la distribution conjointe entre des variables catégoriques et continues, mais nos évaluations au sein de variables catégoriques et continues ont produit des conclusions cohérentes. Enfin, comme pour toute étude de simulation, il faut faire preuve de prudence lors de la généralisation des conclusions. En choisissant

soigneusement les données et les mesures, nous avons tenté d'imiter de près les contextes représentatifs des données d'enquête réelles afin que nos conclusions soient informatives pour les spécialistes qui doivent faire face à des situations semblables. D'autres études d'évaluation fondées sur différentes données sont souhaitées afin de mieux comprendre les caractéristiques opérationnelles et le rendement comparatif des différentes méthodes d'imputation des données manquantes. Les données, les codes et la documentation supplémentaire relatifs à la présente étude sont accessibles à l'adresse suivante : https://github.com/zhenhua-wang/MissingData_DL (en anglais seulement).

Remerciements

Les travaux de recherche de Poulos et Li sont soutenus par la *National Science Foundation* dans le cadre de la subvention DMS-1638521 accordée au *Statistical and Applied Mathematical Sciences Institute*.

Bibliographie

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. et Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems [Logiciel disponible sur tensorflow.org], https://www.tensorflow.org/.

Akande, O., Li, F., et Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*, 71(2), 162-170.

Arnold, B.C., et Press, S.J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association*, 84, 152-156.

Barnard, J., et Meng, X.-L. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8(1), 17-36.

Berthelot, D., Schumm, T. et Metz, L. (2017). *BEGAN: Boundary Equilibrium Generative Adversarial Networks*. CoRR, abs/1703.10717. http://arxiv.org/abs/1703.10717.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Breiman, L., Friedman, J.H., Olshen, R.A. et Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworh, Inc.

Burgette, L., et Reiter, J.P. (2010). Multiple imputation via sequential regression trees. *American Journal of Epidemiology*, 172, 1070-1076.

Cao, W., Wang, D., Li, J., Zhou, H., Li, L. et Li, Y. (2018). BRITS: Bidirectional recurrent imputation for time series. *Advances in Neural Information Processing Systems*, 6775-6785.

Che, Z., Purushotham, S., Cho, K., Sontag, D. et Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1), 1-12.

Chen, S., et Haziza, D. (2019). Recent developments in dealing with item nonresponse in surveys: A critical review. *Revue Internationale de Statistique*, 87, S192-S218.

De Leeuw, E.D., Hox, J. et Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, Stockholm, 19(2), 153-176.

Doove, L., Van Buuren, S. et Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72, 92-104.

Dua, D., et Graff, C. (2017). *UCI Machine Learning Repository*. http://archive.ics.uci.edu/ml.

Fortuin, V., Baranchuk, D., Rätsch, G. et Mandt, S. (2020). GP-VAE: Deep probabilistic time series imputation. *International Conference on Artificial Intelligence and Statistics*, 1651-1661.

Gelman, A., et Speed, T.P. (1993). Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55, 185-188.

Glorot, X., et Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Artificial Intelligence and Statistics*, 9, 249-256.

Gondara, L., et Wang, K. (2018). MIDA: Multiple imputation using denoising autoencoders. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 260-272.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. et Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2672-2680.

Ham, H., Jun, T.J. et Kim, D. (2020). Unbalanced Gans: Pre-Training the Generator of Generative Adversarial Network Using Variational Autoencoder. arXiv preprint arXiv:2002.02112.

Harel, O., et Zhou, X.-H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, 26(16), 3057-3077.

Hastie, T., Tibshirani, R. et Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction (2nd Ed.)*, Springer.

Haziza, D., et Vallée, A.-A. (2020). Variance estimation procedures in the presence of singly imputed survey data: A critical review. *Japanese Journal of Statistics and Data Science*, 3(2), 583-623.

Ho, T.K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1, 278-282.

Honaker, J., King, G. et Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7), 1-47.

Horton, N.J., Lipsitz, S.R. et Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, 57(4), 229-232.

Huque, M.H., Carlin, J.B., Simpson, J.A. et Lee, K.J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology*, 18(1), 1-16.

Kingma, D., et Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. arXiv:1412.6980.

Li, F., Yu, Y. et Rubin, D. (2012). *Imputing Missing Data by Fully Conditional Models: Some Cautionary Examples and Guidelines*. Rapport technique, document de travail du Duke University Department of Statistical Science, 11-24.

Li, F., Baccini, M., Mealli, F., Zell, E.R., Frangakis, C.E. et Rubin, D.B. (2014). Multiple imputation by ordered monotone blocks with application to the anthrax vaccine research program. *Journal of Computational and Graphical Statistics*, 23(3), 877-892.

Lipton, Z.C., Kale, D.C. et Wetzel, R. (2016). Modeling missing data in clinical time series with RNNs. *Machine Learning for Healthcare*, 56.

Little, R.J., et Rubin, D.B. (2014). *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons, Inc.

Little, R.J., et Rubin, D.B. (2019). *Statistical Analysis with Missing Data*, 3rd edition. New York: John Wiley & Sons, Inc.

Lu, H.-M., Perrone, G. et Unpingco, J. (2020). *Multiple Imputation with Denoising Autoencoder Using Metamorphic Truth and Imputation Feedback*. arXiv preprint arXiv:2002.08338.

Maas, A.L., Hannun, A.Y. et Ng, A.Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *Proc. ICML*, (1), 3.

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

Monti, F., Bronstein, M. et Bresson, X. (2017). Geometric matrix completion with recurrent multi-graph neural networks. *Advances in Neural Information Processing Systems*, 3697-3707.

Murray, J.S., et Reiter, J.P. (2016). Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516), 1466-1479.

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

Reiter, J.P., et Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480), 1462-1471.

Royston, P., et White, I.R. (2011). Multiple imputation by chained equations (mice): Implementation in Stata. *Journal of Statistical Software*, 45(4), 1-20.

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

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

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473-489.

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

Shah, A., Bartlett, J., Carpenter, J., Nicholas, O. et Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: A caliber study. *American Journal of Epidemiology*, 179, 764-74.

Stekhoven, D.J., et Bühlmann, P. (2012). Missforest – non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112-118.

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

Tang, L., Song, J., Belin, T.R. et Unützer, J. (2005). A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine*, 24(14), 2111-2128.

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

van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M. et Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049-1064.

van Buuren, S., et Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3), 1-67.

Vincent, P., Larochelle, H., Bengio, Y. et Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning*, 1096-1103.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A. et Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12).

White, I.R., Royston, P. et Wood, A.M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4), 377-399.

Wongkamthong, C., et Akande, O. (2021). A comparative study of imputation methods for multivariate ordinal data. *Journal of Survey Statistics and Methodology*, dans la presse.

Yoon, J., Jordon, J. et Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. *International Conference on Machine Learning*, 5689-5698.

Yoon, J., Zame, W.R. et van der Schaar, M. (2018). Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 66(5), 1477-1490.

Yuan, Y. (2011). Multiple imputation using SAS software. *Journal of Statistical Software*, 45(6), 1-25.