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

Techniques d’enquête

Intégration d’échantillons probabilistes et non probabilistes grâce à l’imputation massive basée sur l’apprentissage profond

par Sixia Chen, Chao Xu et James Cutler

Date de diffusion : le 23 décembre 2025

Statistique Canada
Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Techniques d'enquête, décembre 2025
Vol. 51, nº 2, pp. 537-554
Statistique Canada, nº 12-001-X au catalogue

Intégration d'échantillons probabilistes et non probabilistes grâce à l'imputation massive basée sur l'apprentissage profond

Sixia Chen, Chao Xu et James Cutler[1]

Résumé

Les échantillons probabilistes sont considérés comme la référence pour recueillir des renseignements dans les études basées sur la population, mais l'on utilise fréquemment, dans la pratique, des échantillons non probabilistes en raison de leur faible coût, de leur commodité et de l'absence de base de sondage pour l'enquête. Les estimations naïves fondées sur des échantillons non probabilistes risquent, en l'absence d'ajustements, d'être trompeuses en raison d'un biais de sélection. Une approche valide d'intégration des données comprenant l'imputation massive, la pondération par le score de propension et le calage a récemment été utilisée pour améliorer la représentativité des échantillons non probabilistes. L'efficacité de l'approche d'imputation massive dépend des hypothèses sous-jacentes du modèle. Dans le présent article, nous proposons d'utiliser l'apprentissage profond pour l'imputation massive dans une combinaison d'échantillons probabilistes et non probabilistes et de le comparer à plusieurs approches modernes d'imputation massive basée sur l'apprentissage automatique, y compris la modélisation additive généralisée, l'arbre de régression, la forêt aléatoire et le renforcement extrême du gradient (XGBoosting). Dans l'étude par simulation, les approches basées sur l'apprentissage profond se sont révélées plus robustes et efficaces que d'autres approches d'imputation massive contre l'invalidation des hypothèses sous-jacentes du modèle dans les scénarios de non-linéarité.

Mots-clés : Apprentissage automatique; biais de sélection; échantillon non probabiliste; estimation de la variance; intégration des données.

1. Introduction

Les échantillons probabilistes sont considérés comme la référence pour la recherche basée sur la population, mais l'on utilise fréquemment des échantillons non probabilistes dans les domaines de la santé publique, de l'éducation, de l'économie et de la médecine en raison de leur faible coût, de l'économie de temps qui découle de leur utilisation ou du manque de renseignements sur la base de sondage. En 2015, par exemple, le Pew Research Center (http://www.pewresearch.org [en anglais seulement]) a produit un ensemble de données composé de neuf échantillons non probabilistes et d'un large éventail de mesures, y compris des mesures liées aux caractéristiques démographiques, à l'opinion publique et au comportement en matière de santé. En 2019, le Southern Plains Tribal Health Board a réalisé une enquête afin de recueillir des renseignements sur la santé auprès d'adultes autochtones en combinant les collectes de données effectuées par l'entremise des affaires autochtones et des plateformes de médias sociaux; voir Chen, Campbell, Spain, Milligan et Snider (2022). Même si les échantillons non probabilistes sont actuellement en grande demande pour l'obtention rapide de renseignements, les estimations naïves provenant d'échantillons non probabilistes sans ajustement approprié peuvent être exposées à un biais de sélection; voir Baker, Brick, Bates, Battaglia, Couper, Dever, Gile et Tourangeau (2013).

1. Sixia Chen et Chao Xu, Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences, Oklahoma City, Oklahoma, États-Unis, 73104. Courriel : sixia-chen@ouhsc.edu; James Cutler, Southern Plains Tribal Health Board, Oklahoma City, Oklahoma, 73114.

Les approches d'intégration des données qui combinent les renseignements provenant d'échantillons probabilistes et non probabilistes se sont avérées efficaces pour réduire le biais de sélection résultant de l'utilisation d'échantillons non probabilistes seulement; voir Rivers (2007), Bethlehem (2016), Elliott et Valliant (2017), Lohr et Raghunathan (2017), Beaumont (2020), Rao (2021) et Kim et Tam (2021), qui abordent des méthodes d'intégration des données. En règle générale, les méthodes existantes d'intégration des données se répartissent entre trois grandes catégories : le calage, la pondération par le score de propension et l'imputation massive. L'approche par calage permet de générer des poids calés pour harmoniser la répartition des variables calées d'un échantillon non probabiliste avec celle d'un échantillon probabiliste; voir DiSogra, Cobb, Chan et Dennis (2011) et Elliott et Valliant (2017) pour obtenir plus d'analyses. La validité d'une approche par calage dépend des hypothèses sous-jacentes du modèle touchant les variables d'intérêt de l'étude et les covariables. Dans les approches de pondération par le score de propension (Lee, 2006; Lee et Valliant, 2009; Chen, Li et Wu, 2020), on modélise le score de propension pour la participation à l'échantillon non probabiliste et on l'estime en résolvant les équations d'estimation formées à l'aide des renseignements provenant des échantillons probabilistes et non probabilistes. La validité des approches de pondération par le score de propension dépend des hypothèses sous-jacentes du modèle de score de propension. Pour améliorer la robustesse, dans les approches d'imputation massive, on construit d'abord des modèles de prédiction des variables d'intérêt dépendantes et des covariables en utilisant l'échantillon non probabiliste et on génère des valeurs imputées des variables dépendantes pour toutes les unités de l'échantillon probabiliste, en vue d'une estimation statistique; voir Rivers (2007), Kim, Park, Chen et Wu (2021) et Chen, Yang et Kim (2022) pour obtenir des analyses plus détaillées. La validité des approches d'imputation massive dépend des hypothèses sous-jacentes du modèle d'imputation. Pour améliorer davantage la robustesse des approches d'intégration des données mentionnées ci-dessus, on a proposé l'approche doublement robuste (Chen et coll., 2020) et l'approche robuste multiple (Chen et Haziza, 2022). De plus, les méthodes basées sur l'apprentissage automatique ont été examinées par Buelens, Burger et van den Brakel (2018), Castro-Martín, Rueda, Ferri-García et Hernando-Tamayo (2021), Castro-Martín, Rueda et Ferri-García (2020) et Ferri-García et Rueda (2020). Par contre, aucun d'entre eux n'a étudié les méthodes basées sur l'apprentissage profond pour l'intégration des données.

Les approches d'imputation massive fondées sur des modèles paramétriques sont vulnérables à l'invalidation des hypothèses sous-jacentes de ces modèles (Chen, Yang et Kim, 2022; Chen et Haziza, 2022). Malgré l'élaboration de l'approche d'imputation massive non paramétrique et de l'approche d'imputation massive robuste multiple, composer avec une structure de linéarité plus complexe entre les prédicteurs et une dimensionnalité élevée présente toujours un défi. La littérature montre que les méthodes d'apprentissage automatique telles que la forêt aléatoire (Breiman, 2001), le renforcement extrême du gradient (Chen et Guestrin, 2016) et l'apprentissage profond (LeCun, Bengio et Hinton, 2015) sont très efficaces sur le plan de la prédiction. On a démontré, par ailleurs, que la méthode de renforcement extrême du gradient et la méthode d'apprentissage profond sont très efficaces pour l'imputation de valeurs manquantes lorsque le lien entre les covariables est fortement non linéaire ou lorsque la dimension est élevée; voir, entre autres, Yoon, Jordon et Schaar (2018), Gondara et Wang (2018), Dagdoug, Goga et Haziza (2023a) et Chen et Xu (2022).

De plus, les méthodes d'apprentissage profond commencent à surpasser considérablement le rendement des autres méthodes d'apprentissage statistique dans de nombreux domaines, y compris la reconnaissance de la parole (Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath et Kingsbury, 2012; Graves, Mohamed et Hinton, 2013), l'analyse d'images (Farabet, Couprie, Najman et LeCun, 2012; Krizhevsky, Sutskever et Hinton, 2012; Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Erhan, Vanhoucke et Rabinovich, 2015) et le traitement du langage naturel (Collobert, Weston, Bottou, Karlen, Kavukcuoglu et Kuksa, 2011; Sarikaya, Hinton et Deoras, 2014). Les propriétés asymptotiques des estimateurs statistiques basés sur l'apprentissage profond, y compris la régression non paramétrique, l'analyse de survie et la régression quantile, ont été plus récemment étudiées, notamment par Bauer et Kohler (2019), Schmidt-Hieber (2020), Farrell, Liang et Misra (2021), Zhong, Mueller et Wang (2022) et Padilla, Tansey et Chen (2020). Cependant, l'élaboration d'approches d'imputation massive fondées sur l'apprentissage profond, pour la combinaison d'échantillons probabilistes et non probabilistes, n'a fait l'objet d'aucune étude dans la littérature actuelle. Pour combler cette importante lacune de la recherche, nous avons proposé d'utiliser des approches d'apprentissage profond pour l'imputation massive et de comparer le rendement obtenu à celui d'autres approches populaires existantes basées sur l'apprentissage automatique (par exemple la modélisation additive généralisée, l'arbre de régression, la forêt aléatoire et le renforcement extrême du gradient) à l'aide d'une étude par simulation approfondie et d'une application réelle. Les deux montrent que l'approche d'apprentissage profond surpasse les autres méthodes sur le plan de l'équilibre entre le biais et la variance. Cette approche est la meilleure de toutes.

Dans le présent article, nous procédons de la manière suivante. La section 2 décrit les configurations de base et les notations mathématiques rattachées à notre question de recherche. La section 3 porte sur les méthodes d'imputation massive fondées sur l'apprentissage profond. À la section 4, on effectue une étude par simulation de Monte-Carlo pour comparer différentes méthodes d'imputation massive basées sur l'apprentissage automatique. La section 5 comprend une analyse des résultats et certains domaines de recherche future.

2. Configurations de base

Supposons que nous avons une population finie $\mathcal{F}_N = \{(\mathbf{x}_i, y_i),\ i = 1,2,\ldots,N\}$ où N est sa taille, $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ est le vecteur de covariable ayant une dimension p et y représente la variable d'intérêt scalaire de l'étude. La population finie $\mathcal{F}_N$ représente un échantillon aléatoire d'un modèle de super-population

$$y_i = m(\mathbf{x}_i; \boldsymbol{\beta}^*) + \epsilon_i, \quad i = 1,2,\ldots,N, \tag{2.1}$$

où $m(\mathbf{x}_i; \boldsymbol{\beta}^*)$ est une fonction inconnue ayant un paramètre réel inconnu $\boldsymbol{\beta}^*$ et l'erreur aléatoire satisfait $E(\epsilon_i \mid \mathbf{x}_i) = 0.$ Compte tenu de la population finie $\mathcal{F}_N,$ un échantillon probabiliste S_A est sélectionné à l'aide d'un plan d'échantillonnage probabiliste dont les probabilités d'inclusion de premier et de deuxième ordre sont π_i et π_{ij} pour $i \neq j.$ Le poids de sondage correspondant peut ensuite s'écrire comme suit : $w_i = \pi_i^{-1}.$

Supposons également que le vecteur de covariable $\mathbf{x}$ n'est observable que dans l'échantillon probabiliste S_A. Nous sélectionnons aussi un échantillon non probabiliste S_B à partir de $\mathcal{F}_N$ en utilisant δ_i comme indicateur de sélection, de sorte que $\delta_i = 1$ si l'unité i est sélectionnée et $\delta_i = 0$ autrement. Nous supposons que la probabilité de sélection est inconnue et ne dépend que de $\mathbf{x}$ comme suit :

$$\Pr(\delta = 1 \mid \mathbf{x}, y) = \Pr(\delta = 1 \mid \mathbf{x}). \tag{2.2}$$

Supposons maintenant que nous observons $\mathbf{x}$ et y dans l'échantillon non probabiliste S_B. Cette hypothèse reflète étroitement les scénarios de la vie réelle, car les échantillons non probabilistes sont habituellement conçus pour répondre à des questions d'intérêt particulières pour la recherche, tout en comprenant aussi des renseignements démographiques et socioéconomiques de base. En revanche, les échantillons probabilistes concernent fréquemment des ensembles de données à grande échelle établis tels que ceux du Behavioral Risk Factor Surveillance System, du National Health and Nutrition Examination Survey et de l'American Community Survey. Ces échantillons probabilistes mettent principalement l'accent sur la collecte de données démographiques et socioéconomiques exhaustives de base, en plus d'autres renseignements pertinents liés à la population cible. Supposons que le paramètre d'intérêt de la population finie $\boldsymbol{\theta}_N$ corresponde à la solution unique de l'équation d'estimation :

$$\mathbf{U}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{u}(\mathbf{x}_i, y_i; \boldsymbol{\theta}) = \mathbf{0}, \tag{2.3}$$

où $\mathbf{u}(\mathbf{x}, y; \boldsymbol{\theta})$ est une fonction prédéfinie de $\mathbf{x}$, de y et de $\boldsymbol{\theta}$. Par exemple, $\mathbf{u}(\mathbf{x}, y; \boldsymbol{\theta}) = y - \theta$ correspond à la moyenne de la population finie $\theta_N = N^{-1}\sum_{i=1}^{N} y_i$, tandis que $\mathbf{u}(\mathbf{x}, y; \boldsymbol{\theta}) = I(y \le \theta_N) - \tau$ correspond au $100\tau^{\text{e}}$ centile de la population finie.

3. Méthode proposée

Dans la présente section, nous analysons des méthodes d'imputation massive basées sur l'apprentissage profond pour estimer le paramètre de population finie $\boldsymbol{\theta}_N$ défini à la section précédente. Selon le cadre de Bauer et Kohler (2019) et de Schmidt-Hieber (2020), nous comptons utiliser la structure suivante du réseau neuronal profond comportant L couches pour la modélisation de $m(\mathbf{x}; \boldsymbol{\beta}^*)$. Les nœuds de la première couche masquée sont définis comme suit :

$$g_j^{(1)}(\mathbf{x}) = \sigma\left(\beta_{j0}^{(1)} + \sum_{k=1}^{p} \beta_{jk}^{(1)} x_k\right) \tag{3.1}$$

pour $j = 1, 2, \ldots, d_1$, où $\sigma(t)$ représente la fonction d'activation utilisée pour introduire la non-linéarité. Les nœuds des $L-1$ couches suivantes sont définis récursivement comme suit :

$$g_j^{(l)}(\mathbf{x}) = \sigma\left(\beta_{j0}^{(l)} + \sum_{k=1}^{d_{l-1}} \beta_{jk}^{(l)} g_k^{(l-1)}(\mathbf{x})\right) \tag{3.2}$$

pour $j = 1,2,\ldots,d_l$ et $l = 2,3,\ldots,L,$ où le nombre de couches masquées L représente la profondeur du réseau neuronal profond et le nombre de nœuds d_l pour $l = 1,2,\ldots,L$ représente la largeur de chaque couche. Un nœud d'apprentissage profond est une unité de calcul fondamentale possédant une ou plusieurs connexions d'entrée pondérées, une fonction de transfert qui amalgame ces entrées et une connexion de sortie. Les nœuds sont ensuite organisés en couches pour former un réseau neuronal complet. Une couche désigne un niveau organisationnel au sein d'un réseau neuronal. Les réseaux neuronaux sont composés de couches de nœuds ou de neurones interconnectés. Chaque couche vise un objectif précis dans le traitement et la transformation des données d'entrée. Par la suite, les nœuds de la couche finale $g_1^{(L)}(\mathbf{x}), g_2^{(L)}(\mathbf{x}),\ldots,$ $g_{d_L}^{(L)}(\mathbf{x})$ sont utilisés comme fonctions de base pour la prédiction de y au moyen du modèle de travail suivant :

$$E(y \mid \mathbf{x}) = \beta_0^{(L+1)} + \sum_{j=1}^{d_L} \beta_j^{(L+1)} g_j^{(L)}(\mathbf{x}) \triangleq \tilde{m}(\mathbf{x};\, \tilde{\boldsymbol{\beta}}, L, \mathbf{d}, \sigma), \tag{3.3}$$

où $\tilde{\boldsymbol{\beta}}$ désigne le vecteur contenant tous les coefficients de régression décrits dans (3.1), (3.2) et (3.3) et L, $\mathbf{d} = (d_1, d_2, \ldots, d_L)^\top$ et σ sont les paramètres d'ajustement pouvant être sélectionnés à l'aide de la technique de validation croisée k-fold. L'option la plus répandue pour la fonction d'activation σ comprend l'unité linéaire rectifiée (ULR), soit $\sigma(t) = \max(t, 0),$ et l'unité linéaire exponentielle mise à l'échelle (ULEME), soit $\sigma(t) = \lambda t$ si $t > 0$ ou $\lambda\alpha(e^t - 1)$ si $t \le 0;$ λ et α sont des valeurs approximatives fixes. Un exemple de réseau neuronal profond à quatre couches est présenté à la figure 3.1. Pour consulter un examen des différentes fonctions d'activation dans les réseaux neuronaux profonds, voir Rasamoelina, Adjailia et Sinčák (2020). Pour les paramètres d'ajustement donnés $L^*,$ $\mathbf{d}^*$ et $\sigma^*,$ le modèle peut être adapté en réduisant à son minimum la fonction de coût suivante, à l'aide de l'échantillon non probabiliste :

$$Q(\tilde{\boldsymbol{\beta}}) = \frac{1}{n_B} \sum_{i \in S_B} \left\{ y_i - \tilde{m}(\mathbf{x}_i;\, \tilde{\boldsymbol{\beta}}, L^*, \mathbf{d}^*, \sigma^*) \right\}^2, \tag{3.4}$$

où n_B représente la taille de l'échantillon non probabiliste S_B et σ^* représente la fonction d'activation donnée définie précédemment dans les équations (3.1) et (3.2). La procédure d'optimisation ci-dessus peut être effectuée numériquement au moyen de la descente par gradient stochastique (Bottou, 2010) et de la rétropropagation (LeCun et coll., 2015), et cela garantit un rendement fiable pour des données non vues (ce qui assure la généralisation). Désignons par $\hat{\tilde{\boldsymbol{\beta}}}$ les coefficients de régression estimés à l'aide de la procédure d'optimisation ci-dessus; on peut alors obtenir l'estimateur imputé massivement $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}_N,$ en résolvant l'équation d'estimation pondérée suivante :

$$\hat{\mathbf{U}}(\boldsymbol{\theta}) = \sum_{i \in S_A} w_i \mathbf{u}(\mathbf{x}_i, y_i^*;\, \boldsymbol{\theta}) = \mathbf{0}, \tag{3.5}$$

où y_i^* représente la valeur imputé de l'unité i dans l'échantillon probabiliste $S_A;$ on peut aussi l'écrire comme suit : $y_i^* = \tilde{m}(\mathbf{x}_i;\, \hat{\tilde{\boldsymbol{\beta}}}, L^*, \mathbf{d}^*, \sigma^*) + \epsilon_i^*,$ où ϵ_i^* est sélectionnée à partir du groupe de valeurs résiduelles $\{\hat{\epsilon_i} = y_i - \tilde{m}(\mathbf{x}_i;\, \hat{\tilde{\boldsymbol{\beta}}}, L^*, \mathbf{d}^*, \sigma^*),\, i \in S_A\}$ au moyen du même plan d'échantillonnage que celui utilisé pour

l'échantillon probabiliste initial S_A. C'est ce que l'on appelle communément l'imputation aléatoire, qui permet de préserver la répartition des valeurs imputées; voir Chen et Haziza (2019) pour obtenir des analyses plus précises. Pour estimer la moyenne de la population finie $\theta_N = N^{-1}\sum_{i=1}^{N} y_i$, on peut utiliser l'estimateur imputé de façon déterministe suivant pour réduire la variabilité de l'estimateur imputé :

$$\hat{\theta} = \frac{\sum_{i\in S_A} w_i \tilde{m}(\mathbf{x}_i; \hat{\tilde{\boldsymbol{\beta}}}, L^*, \mathbf{d}^*, \sigma^*)}{\sum_{i\in S_A} w_i}. \tag{3.6}$$

Selon les conditions de régularité précisées dans Farrell et coll. (2021) et leurs arguments et calculs, on peut montrer la cohérence de notre estimateur proposé $\hat{\theta}$. Pour estimer la fonction de répartition de la population finie $\theta_N = N^{-1}\sum_{i=1}^{N} I(y_i \le t)$, l'équation d'estimation correspondante est $\mathbf{u}(\mathbf{x}_i, y_i; \boldsymbol{\theta}) = I(y_i \le t) - \boldsymbol{\theta}$ et l'estimateur imputé massivement peut s'écrire comme suit :

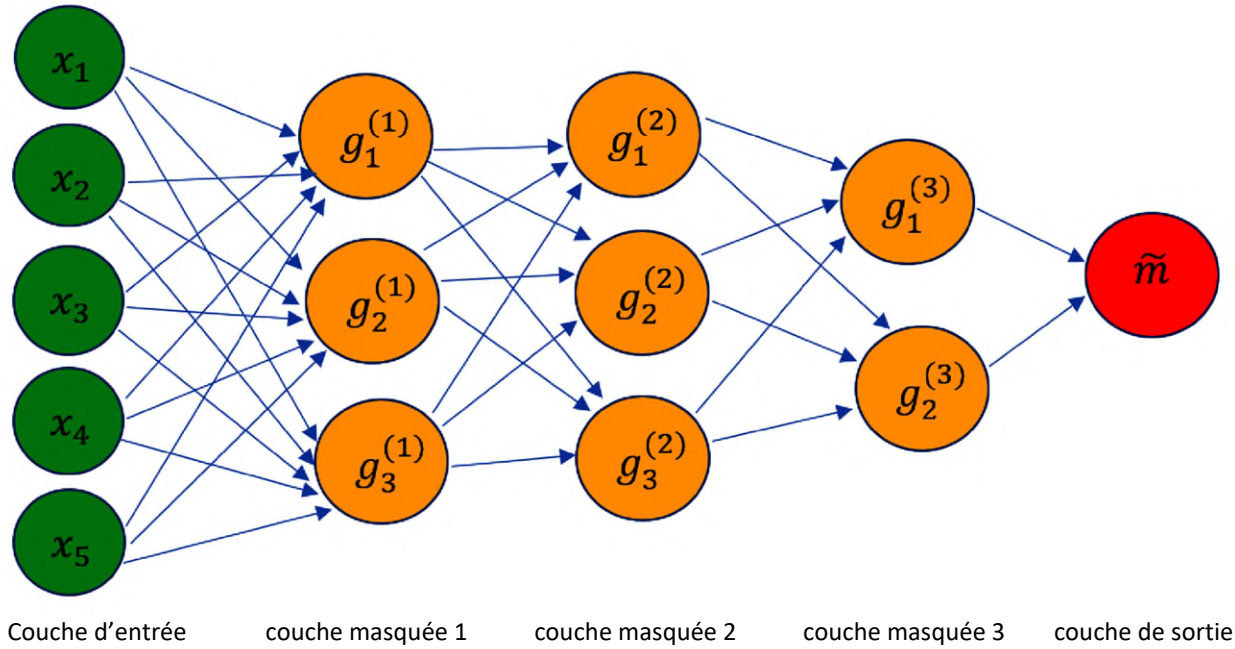
$$\hat{\theta} = \frac{\sum_{i\in S_A} w_i I(y_i^* \le t)}{\sum_{i\in S_A} w_i}. \tag{3.7}$$

Pour l'estimation du $100\tau^{e}$ centile de la population finie, l'équation d'estimation correspondante est $\mathbf{u}(\mathbf{x}_i, y_i; \boldsymbol{\theta}) = I(y_i \le \theta) - \tau$ et l'estimateur imputé massivement peut s'obtenir en résolvant :

$$\hat{\mathbf{U}}(\boldsymbol{\theta}) = \sum_{i\in S_A} w_i \left\{ I(y_i^* \le \theta) - \tau \right\} = \mathbf{0}, \tag{3.8}$$

où y_i^* représente la valeur imputée générée à partir de l'imputation aléatoire.

Figure 3.1 Réseau neuronal profond à quatre couches



Remarque 3.1 *L'inférence statistique fondée sur (3.6) représente un défi important quand la taille de l'échantillon non probabiliste n_B est petite. Par contre, lorsque la taille de l'échantillon n_B est grande par rapport à n_A et, selon certaines conditions de régularité, en utilisant des arguments semblables à ceux de Yang, Kim et Hwang (2021) et d'Angelopoulos, Bates, Fannjiang, Jordan et Zrnic (2023), on peut démontrer que le caractère aléatoire, dans l'estimation des valeurs prédictives basées sur un échantillon non probabiliste, peut être ignoré sans risque; nous avons donc*

$$\begin{aligned}\hat{\theta} &= \frac{\sum_{i\in S_A} w_i \tilde{m}(\mathbf{x}_i)}{\sum_{i\in S_A} w_i} \\ &= \frac{\sum_{i\in S_A} w_i m^*(\mathbf{x}_i)}{\sum_{i\in S_A} w_i} + \frac{\sum_{i\in S_A} w_i \left\{\tilde{m}(\mathbf{x}_i) - m^*(\mathbf{x}_i)\right\}}{\sum_{i\in S_A} w_i} \\ &= \frac{\sum_{i\in S_A} w_i m^*(\mathbf{x}_i)}{\sum_{i\in S_A} w_i} + o_p(n_A^{-1/2}),\end{aligned} \tag{3.9}$$

où $\tilde{m}(\mathbf{x}_i) = \tilde{m}(\mathbf{x}_i; \hat{\tilde{\boldsymbol{\beta}}}, L^*, \mathbf{d}^*, \sigma^*)$ *et* $m^*(\mathbf{x}_i)$ *représente la limite de probabilité de* $\tilde{m}(\mathbf{x}_i)$. *La variance asymptotique de* $\hat{\theta}$ *peut donc s'écrire comme suit :*

$$\begin{aligned}\mathrm{Var}(\hat{\theta}) &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\left\{m^*(\mathbf{x}_i)-\theta^*\right\}\left\{m^*(\mathbf{x}_j)-\theta^*\right\} \\ &\quad + o(n_A^{-1}),\end{aligned} \tag{3.10}$$

où θ^* *représente la limite de probabilité de* $\hat{\theta}$. *C'est la raison pour laquelle l'estimateur de variance peut prend la forme suivante :*

$$\hat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{\hat{N}^2}\sum_{i\in S_A}\sum_{i\in S_A}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\left\{\tilde{m}(\mathbf{x}_i)-\hat{\theta}\right\}\left\{\tilde{m}(\mathbf{x}_j)-\hat{\theta}\right\}, \tag{3.11}$$

où $\hat{N} = \sum_{i\in S_A} w_i$.

4. Études par simulation

Nous générons $B =$ 200 échantillons de Monte-Carlo de populations finies d'une taille de $N =$ 10 000 à partir des trois modèles différents de super-population suivants :

(M1). $y_i = 1 + x_{1,i} + x_{2,i} + x_{3,i} + x_{4,i} + \epsilon_i$ pour $i = 1, 2, \ldots, N$, où $x_{1,i}$, $x_{2,i}$, $x_{3,i}$, $x_{4,i}$ et ϵ_i sont des variables aléatoires mutuellement indépendantes générées à partir d'une distribution normale standard.

(M2). $y_i = 1 + x_{1,i}^2 + x_{1,i}x_{2,i}x_{3,i} + x_{1,i}x_{2,i} + x_{1,i}x_{2,i}x_{3,i}x_{4,i}^2 + \epsilon_i$ pour $i = 1, 2, \ldots, N$, où $x_{1,i}$, $x_{2,i}$, $x_{3,i}$, $x_{4,i}$ et ϵ_i sont des variables aléatoires mutuellement indépendantes générées à partir d'une distribution normale standard.

(M3). La variable d'étude y est générée à partir des modèles multicouches suivants :

$$
\begin{aligned}
a_{k,i}^{(1)} &= \log\left\{1+\exp\left(\alpha_{k0}^{(1)}+\sum_{r=1}^{20}\alpha_{kr}^{(1)}x_{r,i}\right)\right\},\ k=1,2,3,\\
a_{j,i}^{(2)} &= \log\left\{1+\exp\left(\alpha_{j0}^{(2)}+\sum_{k=1}^{3}\alpha_{jk}^{(2)}a_{k,i}^{(1)}\right)\right\},\ j=1,2,3,\\
y_i &= \beta_0+\sum_{j=1}^{3}\beta_j a_{j,i}^{(2)}+\epsilon_i, \qquad\qquad i=1,2,\ldots,N,
\end{aligned}
$$

où ϵ_i représente l'échantillon distribué de manière indépendante et identique à partir de la distribution normale standard. Les coefficients $\alpha_{k0}^{(1)}$, $\alpha_{kr}^{(1)}$, $\alpha_{j0}^{(2)}$, $\alpha_{jk}^{(2)}$, β_0 et β_j sont générés indépendamment une fois à partir d'une distribution uniforme ayant une fourchette de -2 à 2, au début de l'étude par simulation, et restent inchangés dans tous les échantillons de Monte Carlo. $x_{r,i}$ pour $r=1,2,\ldots,20$ sont générées indépendamment à partir d'une distribution uniforme ayant une fourchette de -1 à 1.

Étant donné la population finie, un échantillon probabiliste S_A de taille n_A = 500 est sélectionné à l'aide d'un : a) échantillonnage aléatoire simple (EAS) sans remise; b) échantillonnage systématique randomisé avec probabilité proportionnelle à la taille, dont la variable de taille $z_i = 0{,}5\psi_i + 1$ et ψ_i est générée à partir de la distribution standard du khi carré comportant un degré de liberté. Pour les modèles (M1) et (M2), un échantillon non probabiliste S_B est sélectionné à partir d'une distribution de Bernoulli comportant la probabilité

$$
p(\mathbf{x}_i) = \frac{\exp\left(c+x_{1,i}+x_{2,i}+x_{3,i}+x_{4,i}\right)}{1+\exp\left(c+x_{1,i}+x_{2,i}+x_{3,i}+x_{4,i}\right)},
$$

où $\mathbf{x}_i = (x_{1,i}, x_{2,i}, x_{3,i}, x_{4,i})$ et c est choisie pour que la taille de l'échantillon n_B avoisine 500. Pour le modèle (M3), un échantillon non probabiliste S_B de taille n_B avoisinant 500 est sélectionné à partir d'une distribution de Bernoulli comportant la probabilité

$$
p(\mathbf{x}_i) = \frac{\exp\left(c+\sum_{r=1}^{20}x_{r,i}\right)}{1+\exp\left(c+\sum_{r=1}^{20}x_{r,i}\right)},
$$

où $\mathbf{x}_i = (x_{1,i}, x_{2,i},\ldots, x_{20,i})$ et c est choisie pour que la taille de l'échantillon n_B avoisine 500. Tous les paramètres sont choisis avant l'étude par simulation et restent inchangés pour tous les échantillons de Monte Carlo. Supposons que nous nous intéressions à l'estimation de la moyenne de super-population (θ_0) de y. Nous avons comparé les méthodes suivantes : 1) Modèle de régression linéaire avec sélection du modèle par étapes basée sur le critère d'Akaike; 2) Modélisation additive généralisée; 3) Arbre de régression; 4) Forêt aléatoire; 5) Renforcement extrême du gradient; 6) Nos méthodes d'apprentissage profond proposées comportant deux fonctions d'activation (par exemple ULR et ULEME) et trois options pour le nombre de

couches (par exemple 3, 4 et 5 couches). Chaque couche contient 50 neurones ou nœuds. L'activation linéaire a été utilisée dans la couche de résultats pour les données continues. L'estimation correspondante pour chaque méthode a été obtenue au moyen de

$$\hat{\theta} = \frac{\sum_{i \in S_A} w_i \hat{m}(\mathbf{x}_i)}{\sum_{i \in S_A} w_i}, \tag{4.1}$$

où $\hat{m}(\mathbf{x}_i)$ est la valeur prédictive du sujet $i \in S_A$ selon la méthode correspondante.

Les méthodes d'apprentissage automatique (arbre de régression, forêt aléatoire et renforcement extrême du gradient) ont été ajustées à l'aide des paquets `parsnip`, `workflows` et `recipes` de R. Les hyperparamètres clés que nous avons ajustés pour ces méthodes sont indiqués dans le tableau 4.1. Les hyperparamètres sont des paramètres d'apprentissage automatique dont la valeur permet de contrôler le processus d'apprentissage et de déterminer le rendement du modèle. Pour en savoir plus à leur sujet, voir Hastie, Tibshirani, Friedman et Friedman (2009). Pour chacune de ces méthodes, 30 ensembles de combinaison d'hyperparamètres d'ajustement ont été étudiés, sur la base d'une validation croisée à 10 reprises, pour sélectionner l'ensemble optimal en réduisant au minimum la racine de l'erreur quadratique moyenne (REQM) prédictive. Comme nous n'avions aucune idée de la fourchette de recherche de chaque hyperparamètre, nous avons d'abord fait une tentative en utilisant la fourchette de recherche par défaut des paquets de R sur un ensemble de données simulé. Une fourchette de recherche affinée peut être établie en vérifiant visuellement les tracés de la REQM par rapport à chaque hyperparamètre. Nous avons enfin appliqué la fourchette de recherche affinée à tous les ensembles de données de simulation afin d'obtenir les hyperparamètres d'ajustement finaux. En tenant compte de ceux-ci, le modèle d'estimation des paramètres reposant sur ces modèles a été établi et ajusté selon le cas.

Dans le contexte de l'apprentissage profond, le taux d'apprentissage est un hyperparamètre qui détermine la taille des étapes accomplies pendant le processus d'optimisation. Il représente la proportion selon laquelle les paramètres du modèle sont mis à jour à chaque itération du processus. Pour les méthodes d'apprentissage profond, nous avons ajusté le taux d'apprentissage de (0,1; 0,01; 0,001; 0,0001; 0,00001) au moyen d'une procédure similaire. Sur la base de 200 ensembles de données simulés, nous avions la possibilité d'établir un taux d'apprentissage optimal pour tous les ensembles de données en comparant la fonction cible REQM. Les paquets `TensorFlow` et `Keras` de Python ont été utilisés pour la mise en œuvre. Le fractionnement de validation s'établissait à 0,25. Les époques étaient fixées à 200. Les époques sont des itérations sur l'ensemble de données complet pendant la phase d'entraînement, au cours de laquelle le modèle fait des apprentissages à partir des données et met à jour ses paramètres pour réduire au minimum la fonction de perte. Les données ont été mélangées avant le début de l'entraînement. Pour éviter davantage les ajustements excessifs, nous avons appliqué la règle d'arrêt précoce si le rendement du modèle, à l'étape de la validation, ne s'améliorait plus au bout de 15 époques.

Tableau 4.1
Liste des paramètres d'ajustement utilisés dans différentes approches

Méthode	Paramètres	Définition
Arbre de régression	tree_depth	Profondeur maximale de l'arbre
	min_n	Nombre minimal de points de données dans un nœud
	cost_complexity	Paramètre de coût ou de complexité
Forêt aléatoire	mtry	Nombre de prédicteurs sélectionnés
	trees	Nombre d'arbres contenus dans l'ensemble
	min_n	Nombre minimal de points de données dans un nœud
Renforcement extrême du gradient	tree_depth	Profondeur maximale de l'arbre
	trees	Nombre d'arbres contenus dans l'ensemble
	learn_rate	Taux d'apprentissage
	mtry	Nombre de prédicteurs sélectionnés
	min_n	Nombre minimal de points de données dans un nœud

Nous comparons le rendement de différents estimateurs en utilisant le biais relatif (BR) de Monte Carlo, l'erreur-type relative (ETR) de Monte Carlo et la racine de l'erreur quadratique moyenne relative (REQMR) de Monte Carlo, puisqu'ils ont été utilisés dans la littérature existante (Chen et Xu, 2023; Dagdoug, Goga et Haziza, 2023b; Chen et Haziza, 2023) comme mesures efficaces pour comparer le rendement des estimateurs dans une étude par simulation de Monte Carlo. Une valeur plus faible de ces mesures indique une meilleure précision et exactitude. Le calcul se fait de la manière suivante :

$$\text{BR} = \frac{B^{-1}\sum_{k=1}^{B}\hat{\theta}_k - \theta_0}{|\theta_0|}, \tag{4.2}$$

$$\text{ETR} = \frac{\left\{(B-1)^{-1}\sum_{k=1}^{B}(\hat{\theta}_k - \bar{\hat{\theta}})^2\right\}^{1/2}}{|\theta_0|}, \tag{4.3}$$

et

$$\text{REQMR} = (\text{BR}^2 + \text{ETR}^2)^{1/2}, \tag{4.4}$$

où $\hat{\theta}_k$ représente l'estimateur basé sur le k^{e} échantillon de Monte Carlo, B est le nombre d'échantillons de Monte Carlo décrits précédemment et $\bar{\hat{\theta}} = B^{-1}\sum_{k=1}^{B}\hat{\theta}_k$. Il convient de prendre note que $\hat{\theta}_k$ désigne l'un des six estimateurs pris en compte dans l'étude par simulation.

Les résultats sont présentés dans les tableaux 4.2 (EAS) et 4.3 (échantillonnage avec probabilité proportionnelle à la taille). Dans le modèle de régression linéaire (M1), les méthodes de régression linéaire, de modélisation additive généralisée et d'apprentissage profond affichent un BR, un ETR et une REQMR inférieurs par rapport aux autres, puisque celles qui sont basées sur un arbre présentent un problème d'extrapolation, au point qu'elles ne peuvent prédire les valeurs hors de la fourchette dans les données d'entraînement; voir Hengl, Nussbaum, Wright, Heuvelink et Gräler (2018). Dans nos données simulées, les fourchettes de la covariable x de S_A et des données d'entraînement S_B sont très différentes. Cela est attribuable à l'ampleur du biais de sélection introduit par le mécanisme de probabilité de l'échantillon non

probabiliste. Une étude plus récente a permis de prouver qu'en présence de données d'entraînement appropriées, les réseaux neuronaux profonds peuvent extrapoler des fonctions cibles linéaires (Xu, Zhang, Li, Du, Kawarabayashi et Jegelka, 2020); cela explique notre configuration. Dans le modèle (M1), les méthodes de régression linéaire et de modélisation additive généralisée affichent les valeurs de BR, d'ETR et de REQMR les moins élevées, puisqu'elles reposent sur les bonnes hypothèses du modèle. La méthode d'apprentissage profond qui repose sur l'utilisation de trois couches et d'une fonction d'activation (ULR ou ULEME) procure des résultats comparables à ceux des méthodes de régression linéaire et de modélisation additive généralisée sur le plan du BR. En ce qui concerne les modèles non linéaires (M2) et (M3), les méthodes d'apprentissage profond sont considérablement meilleures que les autres méthodes sur le plan de la production d'un biais relatif plus faible et de la racine de l'erreur quadratique moyenne relative. Les méthodes basées sur un arbre ont toujours un problème d'extrapolation. La méthode de régression linéaire n'a pas bien fonctionné en raison de l'erreur de spécification du modèle. La modélisation additive généralisée affichait des biais importants, car elle ne tenait pas compte des interactions entre les variables auxiliaires. En ce qui concerne le modèle (M1), la méthode d'apprentissage profond reposant sur l'utilisation de trois couches et de la fonction d'activation ULEME a produit le meilleur résultat selon l'EAS. La méthode d'apprentissage profond reposant sur l'utilisation de trois couches et de la fonction d'activation ULR a donné le meilleur résultat selon l'échantillonnage avec probabilité proportionnelle à la taille. En ce qui concerne le modèle (M2), la méthode d'apprentissage profond reposant sur l'utilisation de cinq couches et des fonctions d'activation ULR et ULEME a produit le meilleur résultat selon l'EAS. La méthode d'apprentissage profond reposant sur l'utilisation de trois couches et de la fonction d'activation ULR et celle reposant sur l'utilisation de cinq couches et de la fonction d'activation ULEME ont produit le meilleur résultat selon l'échantillonnage avec probabilité proportionnelle à la taille. Pour le modèle (M3), toutes les méthodes d'apprentissage profond ont donné des résultats comparables selon l'EAS ou l'échantillonnage avec probabilité proportionnelle à la taille. En résumé, les méthodes d'apprentissage profond que nous proposons montrent des résultats qui sont meilleurs à ceux de toutes les autres méthodes ou s'y comparent bien, et le rendement, dans tous les scénarios, est stable. Nous avons constaté quelques variations de rendement pour les méthodes d'apprentissage profond ayant des fonctions d'activation et un nombre de couches différents. Lorsque le modèle sous-jacent est linéaire, la régression linéaire et la modélisation additive généralisée ont obtenu le meilleur rendement. Lorsque le modèle sous-jacent n'est pas linéaire et comporte des termes d'interaction ou une structure de modèle multicouche, la régression linéaire et la modélisation additive généralisée risquent d'engendrer un biais considérable et les méthodes d'apprentissage profond affichent le meilleur rendement. Les méthodes basées sur un arbre, y compris la forêt aléatoire, l'arbre de régression et le renforcement extrême du gradient, risquent de ne pas offrir un bon rendement lorsqu'il y a un grand écart entre les fourchettes des covariables dans les deux types d'échantillons. Le code de calcul de l'étude par simulation est accessible au lien suivant : https://github.com/xu1912/imputationDL.

Tableau 4.2
Biais relatif (BR) [en %], erreur-type relative (ETR) [en %] et racine de l'erreur quadratique moyenne relative (REQMR) [en %] de différentes méthodes selon les modèles (M1) à (M3) et l'échantillonnage aléatoire simple sans remise

Modèle	Méthode	BR	ETR	REQMR
	Régression linéaire	0,18	11,94	11,94
	Modélisation additive généralisée	0,31	12,63	12,64
	Forêt aléatoire	86,93	12,15	87,77
	Arbre de régression	95,44	18,56	97,23
	Renforcement extrême du gradient	34,06	16,49	37,84
M1	Apprentissage profond, 3 couches, ULR	-0,89	18,97	18,99
	Apprentissage profond, 3 couches, ULEME	0,57	14,51	14,52
	Apprentissage profond, 4 couches, ULR	-3,45	17,59	17,92
	Apprentissage profond, 4 couches, ULEME	4,14	15,81	16,35
	Apprentissage profond, 5 couches, ULR	4,89	22,61	23,13
	Apprentissage profond, 5 couches, ULEME	8,02	13,87	16,02
	Régression linéaire	-133,36	28,23	136,31
	Modélisation additive généralisée	-135,63	41,61	141,87
	Forêt aléatoire	-25,29	20,64	32,65
	Arbre de régression	-24,85	24,75	35,07
	Renforcement extrême du gradient	-15,42	19,40	24,78
M2	Apprentissage profond, 3 couches, ULR	-1,19	13,37	13,42
	Apprentissage profond, 3 couches, ULEME	3,07	13,60	13,94
	Apprentissage profond, 4 couches, ULR	2,84	13,99	14,27
	Apprentissage profond, 4 couches, ULEME	-3,74	24,12	24,41
	Apprentissage profond, 5 couches, ULR	0,69	12,91	12,93
	Apprentissage profond, 5 couches, ULEME	-0,40	12,68	12,69
	Régression linéaire	15,88	6,54	17,18
	Modélisation additive généralisée	15,25	6,41	16,55
	Forêt aléatoire	11,72	4,57	12,58
	Arbre de régression	10,49	6,81	12,51
	Renforcement extrême du gradient	13,17	5,52	14,28
M3	Apprentissage profond, 3 couches, ULR	1,06	4,98	5,09
	Apprentissage profond, 3 couches, ULEME	5,42	4,91	7,31
	Apprentissage profond, 4 couches, ULR	2,10	5,45	5,85
	Apprentissage profond, 4 couches, ULEME	1,43	4,71	4,92
	Apprentissage profond, 5 couches, ULR	1,39	4,99	5,18
	Apprentissage profond, 5 couches, ULEME	1,32	4,90	5,08

Note : ULEME = unité linéaire exponentielle mise à l'échelle; ULR = unité linéaire rectifiée.

Tableau 4.3
Biais relatif (BR) [en %], erreur-type relative (ETR) [en %] et racine de l'erreur quadratique moyenne relative (REQMR) [en %] de différentes méthodes selon les modèles (M1) à (M3) et l'échantillonnage systématique randomisé avec probabilité proportionnelle à la taille

Modèle	Méthode	BR	ETR	REQMR
	Régression linéaire	-0,46	12,37	12,38
	Modélisation additive généralisée	-0,35	12,46	12,47
	Forêt aléatoire	85,47	11,56	86,25
	Arbre de régression	93,69	18,06	95,41
	Renforcement extrême du gradient	33,66	17,18	37,79
M1	Apprentissage profond, 3 couches, ULR	-0,44	17,74	17,75
	Apprentissage profond, 3 couches, ULEME	-3,76	18,13	18,51
	Apprentissage profond, 4 couches, ULR	1,52	17,63	17,69
	Apprentissage profond, 4 couches, ULEME	-2,88	17,80	18,03
	Apprentissage profond, 5 couches, ULR	-2,16	18,82	18,94
	Apprentissage profond, 5 couches, ULEME	-1,86	17,81	17,91

Note : ULEME = unité linéaire exponentielle mise à l'échelle; ULR = unité linéaire rectifiée.

Tableau 4.3 (suite)
Biais relatif (BR) [en %], erreur-type relative (ETR) [en %] et racine de l'erreur quadratique moyenne relative (REQMR) [en %] de différentes méthodes selon les modèles (M1) à (M3) et l'échantillonnage systématique randomisé avec probabilité proportionnelle à la taille

Modèle	Méthode	BR	ETR	REQMR
	Régression linéaire	-132,00	25,46	134,43
	Modélisation additive généralisée	-132,82	42,09	139,33
	Forêt aléatoire	-23,53	19,90	30,82
	Arbre de régression	-24,18	30,11	38,62
	Renforcement extrême du gradient	-12,35	18,17	21,97
M2	Apprentissage profond, 3 couches, ULR	0,14	11,09	11,09
	Apprentissage profond, 3 couches, ULEME	-4,26	11,34	12,11
	Apprentissage profond, 4 couches, ULR	-3,41	12,94	13,38
	Apprentissage profond, 4 couches, ULEME	3,30	10,82	11,31
	Apprentissage profond, 5 couches, ULR	-5,23	11,84	12,94
	Apprentissage profond, 5 couches, ULEME	-0,38	11,23	11,24
	Régression linéaire	16,35	8,40	18,38
	Modélisation additive généralisée	14,52	8,29	16,72
	Forêt aléatoire	45,94	6,26	46,36
	Arbre de régression	50,08	10,36	51,14
	Renforcement extrême du gradient	27,46	7,71	28,52
M3	Apprentissage profond, 3 couches, ULR	6,92	8,94	11,30
	Apprentissage profond, 3 couches, ULEME	4,35	10,61	11,46
	Apprentissage profond, 4 couches, ULR	10,66	10,14	14,71
	Apprentissage profond, 4 couches, ULEME	8,90	12,14	15,05
	Apprentissage profond, 5 couches, ULR	5,73	10,39	11,87
	Apprentissage profond, 5 couches, ULEME	3,03	10,40	10,83

Note : ULEME = unité linéaire exponentielle mise à l'échelle; ULR = unité linéaire rectifiée.

5. Discussion

Dans le présent article, nous avons comparé les méthodes d'imputation massive basées sur l'apprentissage profond à certaines méthodes existantes d'imputation massive basées sur l'apprentissage automatique, aux fins d'intégration des données. L'étude par simulation a montré les avantages de nos méthodes proposées basées sur l'apprentissage profond pour la réduction du biais de Monte Carlo, de l'erreur-type de Monte Carlo et de l'erreur quadratique moyenne de Monte Carlo. Plus précisément, le rendement des méthodes basées sur l'apprentissage profond a considérablement dépassé celui des autres méthodes basées sur l'apprentissage automatique au moyen d'une structure de modèle hiérarchique fortement non linéaire. Sur le plan des éléments à considérer liés aux calculs, les méthodes basées sur l'apprentissage profond nécessitent plus de temps pour l'ajustement, mais ce problème devrait se régler facilement grâce à la puissance de calcul moderne. Dans notre étude par simulation, il a suffi de 5 minutes pour effectuer 200 simulations pour le modèle 3, le plus complexe, selon un taux d'apprentissage de 0,1. Il a fallu 27 minutes pour effectuer 200 simulations pour ce même modèle quand le taux d'apprentissage était de 0,00001. Nous avons utilisé une unité centrale graphique NVIDIA Quadro P600 pour l'apprentissage profond. Ce modèle est apparu dans le marché il y a 7 ans. Les nouvelles unités centrales graphiques offrant une vitesse de calcul beaucoup plus rapide, le temps de calcul pour exécuter les modèles d'apprentissage profond devrait être négligeable. Après l'optimisation d'un modèle de ce genre, son application à de nouvelles données exigerait peu de temps. Le principal élément à considérer, en ce qui concerne l'apprentissage profond, demeurerait

l'évitement des ajustements excessifs. Les arrêts précoces, l'abandon aléatoire et un fractionnement de validation élevé pourraient aider à les éviter. L'obtention de la sélection optimale des paramètres d'ajustement, notamment le nombre de couches et de nœuds, reste une question exigeant d'autres recherches. L'inférence statistique à l'aide de l'imputation massive basée sur l'apprentissage profond représente un défi très complexe en présence d'un échantillon non probabiliste de petite taille, mais nous prévoyons nous y attaquer dans nos recherches futures. De plus, davantage de données empiriques fondées sur des exemples réels sont nécessaires pour la recherche future.

Remerciements

Le professeur Chen a reçu le soutient de ASA/NSF/Census Research Fellowship du Bureau du recensement des États-Unis (contrat n° 1333LB24P00000032), du National Institute on Minority Health and Health Disparities des National Institutes of Health (1R21MD014658-01A1) et de l'Oklahoma Shared Clinical and Translational Resources (U54GM104938) avec une bourse de développement institutionnel du National Institute of General Medical Sciences. La teneur du présent article relève de la seule responsabilité des auteurs et ne représente pas nécessairement l'opinion officielle des National Institutes of Health et du Bureau du recensement des États-Unis. Une partie des calculs rattachés au projet a été réalisée au Supercomputing Center for Education & Research de l'Université de l'Oklahoma.

Bibliographie

Angelopoulos, A.N., Bates, S., Fannjiang, C., Jordan, M.I. et Zrnic, T. (2023). Prediction-powered inference. *Science*, 382(6671), 669-674.

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. et Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1(2), 90-143.

Bauer, B. et Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4), 2261-2285.

Beaumont, J.-F. (2020). Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles? *Techniques d'enquête*, 46(1), 1-30. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2020001/article/00001-fra.pdf.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching? *Social Science Computer Review*, 34(1), 59-77.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT'2010*, Springer, 177-186.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Buelens, B., Burger, J. et van den Brakel, J.A. (2018). Comparing inference methods for non-probability samples. *Revue Internationale de Statistique*, 86(2), 322-343.

Castro-Martín, L., Rueda, M.d.M. et Ferri-García, R. (2020). Inference from non-probability surveys with statistical matching and propensity score adjustment using modern prediction techniques. *Mathematics*, 8(6), 879.

Castro-Martín, L., Rueda, M.d.M., Ferri-García, R. et Hernando-Tamayo, C. (2021). On the use of gradient boosting methods to improve the estimation with data obtained with self-selection procedures. *Mathematics*, 9(23), 2991.

Chen, S., Campbell, J., Spain, E., Milligan, A. et Snider, C. (2022). Improving the representativeness of the tribal behavioral risk factor surveillance system through data integration, soumis.

Chen, T. et Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

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

Chen, S. et Haziza, D. (2022). A general multiply robust framework for combining probability and non-probability samples in surveys. *Scandinavian Journal of Statistics*.

Chen, S. et Haziza, D. (2023). General purpose multiply robust data integration procedures for handling nonprobability samples. *Scandinavian Journal of Statistics*, 50(2), 697-724.

Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011-2021.

Chen, S. et Xu, C. (2022). Handling high-dimensional data with missing values by modern machine learning techniques. *Journal of Applied Statistics*, 1-19.

Chen, S. et Xu, C. (2023). Handling high-dimensional data with missing values by modern machine learning techniques. *Journal of Applied Statistics*, 50(3), 786-804.

Chen, S., Yang, S. et Kim, J.K. (2022). Nonparametric mass imputation for data integration. *Journal of Survey Statistics and Methodology*, 10(1), 1-24.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. et Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 2493-2537.

Dagdoug, M., Goga, C. et Haziza, D. (2023a). Imputation procedures in surveys using nonparametric and machine learning methods: An empirical comparison. *Journal of Survey Statistics and Methodology*, 11(1), 141-188.

Dagdoug, M., Goga, C. et Haziza, D. (2023b). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 118(542), 1234-1251.

DiSogra, C., Cobb, C., Chan, E. et Dennis, J.M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. *Joint Statistical Meetings (JSM), Survey Research Methods*, 4501-4515.

Elliott, M.R. et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Farabet, C., Couprie, C., Najman, L. et LeCun, Y. (2012). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1915-1929.

Farrell, M.H., Liang, T. et Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1), 181-213.

Ferri-García, R. et Rueda, M.d.M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PloS One*, 15(4), e0231500.

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

Graves, A., Mohamed, A.-r. et Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 6645-6649.

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

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B. et Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6, e5518.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N. et Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6), 82-97.

Kim, J.K., Park, S., Chen, Y. et Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(3), 941-963.

Kim, J.-K. et Tam, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *Revue Internationale de Statistique*, 89(2), 382-401.

Krizhevsky, A., Sutskever, I. et Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.

LeCun, Y., Bengio, Y. et Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22(2), 329.

Lee, S. et Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3), 319-343.

Lohr, S.L. et Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science*, 32(2), 293-312.

Padilla, O., Tansey, W. et Chen, Y. (2020). Quantile regression with relu networks: Estimators and minimax rates. *arXiv Preprint arXiv:2010.08236*, 13.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83(1), 242-272.

Rasamoelina, A.D., Adjailia, F. et Sinčák, P. (2020). A review of activation function for artificial neural network. *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, IEEE, 281-286.

Rivers, D. (2007). Sampling for web surveys. *Joint Statistical Meetings*, 4.

Sarikaya, R., Hinton, G.E. et Deoras, A. (2014). Application of deep belief networks for natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4), 778-784.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4), 1875-1897.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. et Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1-9.

Xu, K., Zhang, M., Li, J., Du, S.S., Kawarabayashi, K.-i. et Jegelka, S. (2020). How neural networks extrapolate: From feedforward to graph neural networks. *arXiv Preprint arXiv:2009.11848*.

Yang, S., Kim, J.K. et Hwang, Y. (2021). Intégration de données d'enquêtes probabilistes et de mégadonnées aux fins d'inférence de population finie au moyen d'une imputation massive. *Techniques d'enquête*, 47(1), 29-58. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021001/article/00004-fra.pdf.

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

Zhong, Q., Mueller, J. et Wang, J.-L. (2022). Deep learning for the partially linear cox model. *The Annals of Statistics*, 50(3), 1348-1375.