

## Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

### Stratégies d'imputation axées sur les données et leurs indicateurs de qualité connexes dans les enquêtes économiques

par Matei Mireuta, Ahalya Sivathayalan et Stephen Styles

Date de diffusion : le 8 septembre 2025



Statistique  
Canada

Statistics  
Canada

Canada

## Stratégies d'imputation axées sur les données et leurs indicateurs de qualité connexes dans les enquêtes économiques

Matei Mireuta, Ahalya Sivathayalan et Stephen Styles<sup>1</sup>

### Résumé

Des méthodes modernes d'imputation axées sur les « données » pour traiter la non-réponse dans le contexte d'enquêtes relevant du Programme intégré de la statistique des entreprises de Statistique Canada ont été évaluées dans le passé. Nous avons observé que ces méthodes permettent d'observer qu'elles peuvent mener à une imputation de grande qualité, en plus de pouvoir entraîner des gains d'efficacité globaux lors de l'établissement d'une stratégie de vérification et d'imputation (V&I) d'une enquête donnée. Cependant, il reste ardu d'évaluer la variance totale connexe, plus précisément la composante attribuable à l'imputation. Dans le présent article, deux méthodes pour évaluer la variance totale sont proposées, et on montre les résultats préliminaires qui ont amené à poursuivre les recherches dans ce domaine.

Mots clés : Imputation cubiste; forêt aléatoire; variance attribuable à l'imputation; PISE.

### 1. Introduction

Déployé pour la première fois en 2013 pour une quarantaine d'enquêtes, le Programme intégré de la statistique des entreprises (PISE) (Statistique Canada, 2015) propose désormais un cadre normalisé pour environ 130 enquêtes économiques sur des sujets aussi diversifiés que l'agriculture, les industries de la fabrication et les services. Les enquêtes que traite désormais le PISE couvrent également un vaste éventail de fréquences de traitement, allant de la fréquence mensuelle à la fréquence quinquennale, ainsi qu'un vaste éventail de tailles d'échantillons, allant de quelques unités à des millions d'unités.

Le système du PISE repose sur le concept du modèle de contenu harmonisé. Ainsi, les enquêtes doivent mettre en application des classifications statistiques normalisées pour recueillir les données d'entrée et de sortie sur les entreprises et les classer. En outre, le traitement dans le cadre du PISE est réalisé, de manière prédominante, à l'aide de l'ensemble de systèmes généralisés de Statistique Canada, comme G-Sam pour l'échantillonnage (Statistique Canada, 2013), Banff pour le contrôle et l'imputation (Statistique Canada, 2017) et G-Est pour l'estimation (Statistique Canada, 2019). La méthodologie du PISE est adaptée au modèle de traitement des estimations en continu, un traitement automatisé et itératif qui est exécuté régulièrement pendant le cycle d'une enquête jusqu'à l'atteinte d'un niveau acceptable de qualité (Godbout et coll., 2011).

Parmi les objectifs importants du système du PISE, il y avait la mise en œuvre d'une méthodologie pour évaluer la variance totale des estimations, c'est-à-dire la variance attribuable à l'échantillonnage et celle attribuable à l'imputation (VI). La variance attribuable à l'imputation est évaluée à l'aide du Système d'estimation de la variance due à la non-réponse et à l'imputation (SEVANI) de Statistique Canada (Beaumont et Bissonnette, 2011). Des indicateurs de la qualité se fondant sur la variance totale sont désormais produits pour de nombreuses enquêtes économiques.

---

<sup>1</sup>Matei Mireuta, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6 ([matei.mireuta@statcan.gc.ca](mailto:matei.mireuta@statcan.gc.ca)); Ahalya Sivathayalan, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6 ([ahalya.sivathayalan@statcan.gc.ca](mailto:ahalya.sivathayalan@statcan.gc.ca)); Stephen Styles, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6 ([stephen.styles@statcan.gc.ca](mailto:stephen.styles@statcan.gc.ca))

## 2. Contexte

Les stratégies de vérification et d'imputation (V&I) sont devenues beaucoup trop complexes pour la plupart des enquêtes économiques traitées par le PISE. En raison de la différence au niveau de la disponibilité des données auxiliaires entre les unités, l'imputation composite a toujours été une méthode de choix pour les enquêtes du PISE. Cependant, cela mène à des dizaines (et souvent plus) de modèles d'imputation possibles pour chaque variable, ce qui rend très difficiles et coûteux la mise en œuvre, le soutien et l'analyse d'une stratégie d'imputation globale d'une enquête.

Dans le cadre d'un projet de recherche antérieur, notre équipe a évalué trois solutions de rechange qui sont axées sur les « données », c'est-à-dire que nous avons pris en compte tous les renseignements auxiliaires et que nous ne supposons aucune connaissance antérieure de relations sous-jacentes entre les variables. La première méthode prenait la forme d'une méthode de régression linéaire paramétrique, la deuxième prenait la forme d'un algorithme de forêt aléatoire, et la troisième prenait la forme d'un algorithme d'imputation cubiste (Quinlan, 1992, et Quinlan, 1993). Dans l'ensemble, nous avons constaté que ces méthodes mènent à une imputation de qualité élevée qui va de pair avec les résultats signalés par d'autres auteurs avec des données synthétiques (Dagdoug et coll., 2023). Nous avons également observé que, dans l'ensemble, les estimations étaient très semblables aux estimations publiées dans le passé, lorsque des méthodes d'imputation standard étaient utilisées. Parmi les avantages évidents de ces méthodes « axées sur les données », il y a leur capacité d'automatisation, leur coût relativement bas en ce qui concerne la mise en œuvre, et leur potentiel de réduire et de simplifier considérablement les stratégies de V&I actuelles du PISE. Cependant, l'estimation de la variance attribuable à l'imputation de ces méthodes, surtout les modèles par forêt aléatoire et cubiste, représente le principal défi méthodologique auquel nous avons dû faire face. Puisque SEVANI a été créé comme solution analytique pour l'imputation composite linéaire, il ne peut pas, dans sa forme actuelle, servir à estimer la VI dans le contexte de méthodes d'apprentissage statistique. Pour mesurer la VI de la méthode par forêt aléatoire et de la méthode cubiste (et d'autres méthodes d'apprentissage statistique en général), les techniques basées sur les répliques représentent la première option. Il en existe plusieurs dans la littérature (voir, par exemple, Davison et coll., 2007). Toutefois, il s'agit tout de même d'un problème non résolu en grande partie qui nécessiterait des recherches méthodologiques et des investissements supplémentaires importants pour le PISE. Cela pourrait être nécessaire en fin de compte. Cependant, avant d'aller de l'avant, nous voulions évaluer l'exactitude et la faisabilité de deux solutions approximatives qui pourraient être mises en œuvre dans l'infrastructure actuelle. Nous décrivons de manière détaillée ces deux méthodes dans les sections 3 et 4. Nous fournissons les résultats préliminaires de simulations dans la section 5.

## 3. Imputation à l'aide de prédictions *a priori*

### 3.1 Description de la méthode

La méthode générale prend la forme suivante. Supposons un échantillon de taille  $n$  et posons  $\mathbf{y}_t$  le vecteur de toutes les  $n$  valeurs d'échantillon pour la variable  $y$  au moment  $t$ ; certaines pouvant provenir de répondants et d'autres pouvant être imputées. Posons  $\mathbf{y}_t^R$  et  $\mathbf{y}_t^I$  les vecteurs des valeurs répondues et des valeurs imputées pour la variable  $y$  au moment  $t$ . La première étape de la méthode consiste à ajuster un modèle  $M$  au moment  $t - 1$ , comme suit :

$$\mathbf{y}_{t-1}^R \sim M(\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \mathbf{X}_{t-3}, \mathbf{H}_{t-2}, \mathbf{H}_{t-3}) \quad (1)$$

où  $\mathbf{X}$  et  $\mathbf{H}$  désignent la matrice des renseignements auxiliaires observés (ou accessibles) et des renseignements historiques respectivement, à différents points dans le temps. Les résultats présentés à la section 3.3 ont été obtenus avec la configuration ci-dessus (c'est-à-dire deux périodes pour les données historiques et trois points dans le temps pour les données auxiliaires). Cependant, en général, l'utilisateur peut souhaiter ajouter plus ou moins de renseignements auxiliaires et historiques s'il le juge approprié pour un objectif particulier. Ensuite, la deuxième étape de la méthode permet d'obtenir une prédiction pour chaque unité d'échantillonnage au moment  $t$ , comme suit :

$$\mathbf{y}_t^P = M(\mathbf{X}_t, \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \mathbf{H}_{t-1}, \mathbf{H}_{t-2})$$

Le vecteur des prédictions  $\mathbf{y}_t^P$  dépend des données auxiliaires et des données historiques, mais pas des données actuelles  $\mathbf{y}_t$ . Il est donc accessible avant le début de la collecte. L'étape finale de la méthode consiste à ajuster et à appliquer, au moment  $t$  et pour chaque itération d'estimation en continu, un modèle d'imputation  $F_I$  en utilisant uniquement  $\mathbf{y}_t^P$  comme données auxiliaires :

$$\begin{aligned} \mathbf{y}_t^R &= F_I(\mathbf{y}_t^P) + \varepsilon \\ \mathbf{y}_t^I &= \hat{F}_I(\mathbf{y}_t^P) \end{aligned} \quad (2)$$

Le modèle  $M$  de (1) peut, de manière générale, être plutôt complexe. Cependant, nous avons constaté lors des études préliminaires que, souvent, une méthode linéaire simple pour  $F_I$  peut procurer des résultats d'imputation d'un degré d'exactitude élevé. Nous avons étudié la régression linéaire et la méthode du plus proche voisin pour  $F_I$ , dont des résultats sélectionnés sont présentés à la section 3.3. Parmi les principaux avantages associés à la méthode d'imputation à l'aide de prédictions *a priori*, il y a, dans le cas des fonctions linéaires  $F_I$ , le fait que la variance conditionnelle des estimations touchant  $\mathbf{y}_t$  peut être obtenue au moyen de l'infrastructure actuelle du PISE. En ce qui concerne les estimations des totaux, par exemple, il est simple d'estimer la variance totale  $Var(\hat{T}_y - T_y | \mathbf{y}_t^P)$  à l'aide de SEVANI, dans sa forme actuelle. Une hypothèse cruciale de cette méthode qui peut être remise en question est de savoir si l'inférence peut raisonnablement être conditionnée sur  $\mathbf{y}_t^P$ .

### 3.2 Modèle M et agrégation

Le modèle  $M$  de (1) peut être un modèle simple ou une agrégation de modèles distincts. Il peut être paramétrique ou non. En fait, il peut s'agir de n'importe quel modèle ou de n'importe quelle méthode qui produit des prédictions qui sont raisonnablement corrélées avec les valeurs des répondants. En général, nous prévoyons que le choix de  $M$  sera principalement orienté de manière empirique.

Les résultats figurant dans la section 3.3 ont été obtenus au moyen d'une agrégation linéaire convexe des prédictions produites par trois modèles différents, c'est-à-dire une méthode de régression linéaire par étapes, une méthode d'imputation par forêt aléatoire et une méthode d'imputation cubiste, comme le montre la figure 3.3-1. Dans ce cas,  $\mathbf{y}_t^P = \alpha_1 \mathbf{y}_t^{PLR} + \alpha_2 \mathbf{y}_t^{PRF} + \alpha_3 \mathbf{y}_t^{PCub}$ , où  $\alpha_1, \alpha_2$  and  $\alpha_3$  sont estimés à l'aide de techniques d'optimisation pour un sous-ensemble indépendant d'unités, avec la contrainte additionnelle que  $\alpha_1 + \alpha_2 + \alpha_3 = 1$ . De manière générale, de nombreuses autres méthodes d'agrégation pourraient être pertinentes, en fonction de plusieurs mesures de la qualité possibles, ou aucune si un modèle dépasse grandement tous les autres.

### 3.3 Résultats préliminaires sélectionnés

Nous avons étudié la méthode d'imputation à l'aide de prédictions *a priori* pour 10 variables de l'Enquête annuelle sur les industries manufacturières et de l'exploitation forestière (EAMEF) de Statistique Canada pour la période de référence 2022.

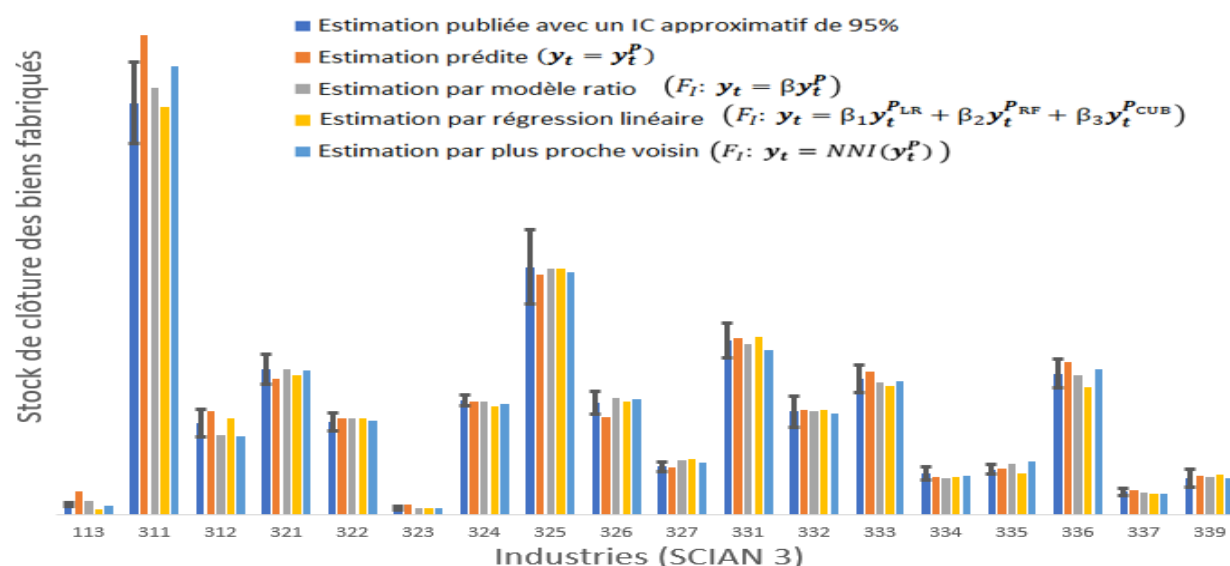
Comme mentionné dans la section 3.2, une prédiction agrégée  $\mathbf{y}_t^P = \alpha_1 \mathbf{y}_t^{PLR} + \alpha_2 \mathbf{y}_t^{PRF} + \alpha_3 \mathbf{y}_t^{PCub}$  a été utilisée dans les études préliminaires. Les résultats présentés dans cette section se fondent sur cette méthode d'agrégation, ainsi que plusieurs choix pour  $F_I$ . La figure 3.3-1 montre des résultats au niveau des macroestimations pour une variable représentative des 10 variables examinées. Dans ce cas-ci, il s'agit des stocks de clôture des biens fabriqués. Dans le cadre des études préliminaires, nous avons étudié plusieurs choix pour  $F_I$  et avons constaté que, souvent, de nombreux modèles mènent à des estimations très semblables. En raison de la mise en œuvre simple, on recommanderait d'utiliser un modèle par ratio simple appliqué à des prédictions agrégées (troisième barre à partir de la gauche dans la figure 3.3-1) ou une méthode du plus proche voisin (dernière barre).

La quatrième barre à la figure 5 montre une option de régression linéaire pour  $F_I$  appliquée aux prédictions de modèle individuelles. La méthode du plus proche voisin (dernière barre) peut être particulièrement utile dans le cas des variables qui affichent des relations non linéaires significatives avec des prédictions *a priori*. Deux autres arguments peuvent être fournis en faveur de l'utilisation de la méthode du plus proche voisin. Tout d'abord, de manière générale, le modèle utilisé pour obtenir des prédictions *a priori* n'est pas paramétrique. C'est pourquoi

l'utilisation d'une méthode non paramétrique pour  $F_I$  semble aussi très conservatrice. Ensuite, si une variable ne montre aucune corrélation avec des prédictions *a priori*, l'utilisation de la méthode du plus proche voisin comme modèle pour  $F_I$  fait de la procédure d'imputation une imputation hot deck aléatoire, ce qui peut être un scénario acceptable.

**Figure 3.3-1**

**Estimations des stocks de clôture totaux des biens fabriqués par industrie pour l'EAMEF 2022. La barre en bleu foncé correspond aux estimations actuelles de production publiées avec des barres d'erreur représentant un intervalle de confiance d'environ 95 %.**



## 4. Imputation par linéarisation cubiste

### 4.1 Description du modèle cubiste

L'approche par imputation cubiste est un modèle fondé sur des règles qui combine plusieurs méthodologies (Quinlan, 1992, et Quinlan, 1993) et produit des modèles de régression linéaire à chaque nœud d'extrémité d'un arbre fondé sur des règles (voir la figure 4.1-1 comme exemple). La procédure de lissage à chaque nœud tient compte de la variance et de la covariance des modèles aux nœuds enfants et parents, et combine ces modèles linéaires en un seul modèle lissé par nœud. En outre, le surajustement est atténué en intégrant un taux d'erreur ajusté comme critère d'élagage ou de combinaison de règles. La méthode cubiste peut être utilisée comme un arbre simple, qui sera désigné par  $M_{cub(1,0)}$  dans cette section, ou comme une méthode d'ensemble. Comme le montre la figure 4.1-1, il est simple de faire correspondre automatiquement l'arborescence produite par un modèle  $M_{cub(1,0)}$  à une stratégie de V&I linéaire compatible avec le PISE et d'utiliser toutes les fonctionnalités du PISE (y compris SEVANI) sans faire des investissements additionnels dans l'infrastructure du système. La stratégie de V&I obtenue ne compte pas obligatoirement moins d'étapes que la stratégie de V&I classique qu'utilise actuellement le PISE. En fait, elle est souvent beaucoup plus longue. Toutefois, cette stratégie est plus simple, car elle est produite automatiquement sans intervention manuelle.

### 4.2 Description de la méthode

Nous proposons d'utiliser  $M_{cub(1,0)}$  comme approximation linéaire de toute autre méthode d'apprentissage automatique ou méthode d'agrégation de modèles proposée dans la section 3.2. Ainsi,  $M$  pourrait correspondre à n'importe quelle méthode, comme le montre l'équation (1), utilisée pour obtenir des prédictions  $y_t^P$  de la même

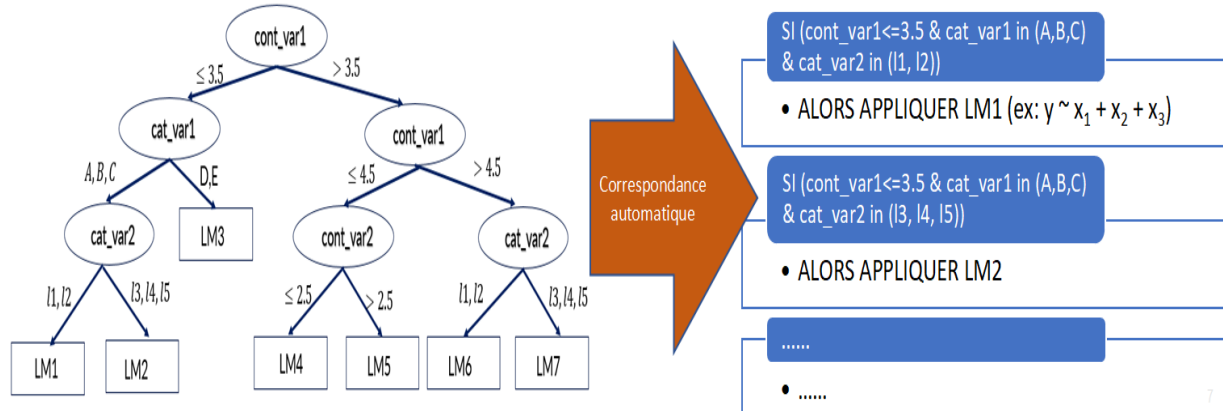
manière qu'illustré dans la section 3.1. Puis, dans le cadre d'une deuxième étape, un modèle  $M_{cub(1,0)}$  peut être ajusté en fonction de  $y_t^P$  et son résultat peut correspondre à  $F_I$  :

$$y_t^P \sim M_{cub(1,0)}(X_t, X_{t-1}, X_{t-2}, H_{t-1}, H_{t-2})$$

$$F_I \cong M_{cub(1,0)}(X_t, X_{t-1}, X_{t-2}, H_{t-1}, H_{t-2})$$

**Figure 4.1-1**

**Représentation schématique de la correspondance de  $M_{cub(1,0)}$  sous forme de stratégie de V&I du PISE.**



## 5. Variance inconditionnelle – résultats de la simulation

Parmi les inquiétudes que posent les deux méthodes proposées, il y a le fait que l'estimation de la variance est conditionnelle, que ce soit au niveau des prédictions, ou de l'arbre et des modèles de nœuds associés, qui dépendent tous deux des données historiques et auxiliaires observées. Afin d'évaluer la variance non conditionnelle et le biais de l'estimateur imputé et de l'estimateur de la variance, nous avons réalisé une brève étude par simulation. Nous avons évalué la méthode des prédictions *a priori* décrite dans la section 3, même si nous avons eu recours à la méthode d'imputation cubiste. Les idées décrites dans la section 4 seront évaluées dans le cadre d'études futures.

Une pseudo-population de 540 unités a été créée comprenant des répondants de l'Enquête annuelle sur les industries manufacturières et de l'exploitation forestière pour les années 2021 et 2022. Pour chaque itération, un échantillon de 108 unités a été choisi à l'aide de l'EAS pour l'année 2021. La non-réponse a été produite avec des valeurs manquantes entièrement au hasard. Nous avons mis à l'essai plusieurs choix pour la probabilité de non-réponse, sans que cela ait des répercussions importantes sur les résultats (données non illustrées). Nous avons ensuite utilisé une méthode cubiste pour obtenir des prédictions pour l'année 2022 à l'aide des valeurs des répondants de 2021, ainsi qu'un ensemble de données auxiliaires. Le même échantillon a été retenu pour l'année 2022, et la non-réponse a été produite encore une fois (valeurs manquantes entièrement au hasard). Enfin, des estimations de la population totale pour l'année 2022 ont été obtenues avec les estimations de la variance correspondantes, conformément à la méthodologie décrite dans le SEVANI (Beaumont et Bissonnette, 2011) et à l'aide d'un modèle d'imputation par ratio ( $F_I: y_t = \beta y_t^{P_{cub}}$ ).

La figure 5-1 montre le biais relatif moyen Monte-Carlo (MC) et les estimations de la variance MC et la moyenne des estimations de la variance MC pour une variable ayant de bonnes prédictions (revenus totaux, figure 5-1 A) et une variable difficile à prédire (total des employés dans la main-d'œuvre indirecte, figure 5-1 B). Dans les deux cas, l'estimateur imputé demeure raisonnablement non biaisé, sauf si les conditions sont extrêmes (taux de non-réponse >80 %). Dans le même ordre d'idées, l'estimateur de la variance conditionnelle demeure raisonnablement non biaisé pour la variance non conditionnelle, sauf si, encore une fois, les conditions sont extrêmes. Dans les deux cas, les propriétés MC de l'estimateur repondéré (si l'on suppose des valeurs manquantes entièrement au hasard) sont affichées à titre comparatif (en jaune). Dans les études futures, nous aimerions évaluer le comportement de ces

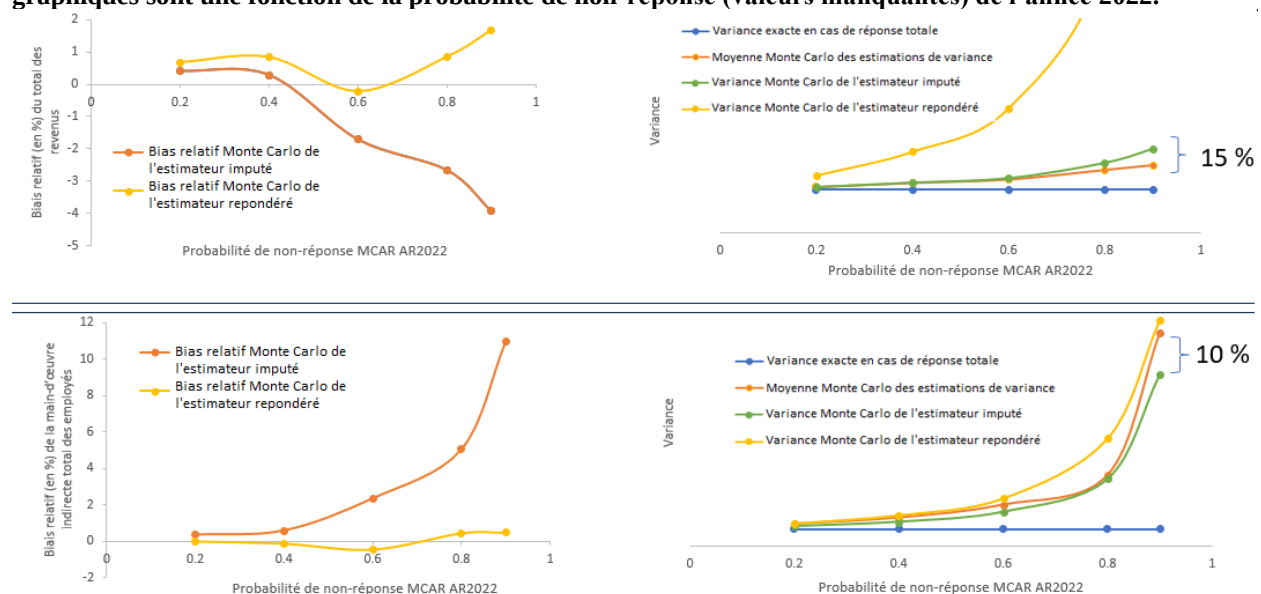
estimations dans des modèles de non-réponse plus réalistes et pour d'autres méthodes de prédiction et méthodes d'imputation.

## 6. Conclusion

Dans le présent article, nous avons décrit des idées préliminaires pour deux approches générales qui pourraient servir à ajouter des méthodes d'imputation modernes au cadre actuel du PISE d'une manière évolutive et automatisée. Nous avons observé des résultats prometteurs dans le cadre d'études préliminaires, ainsi que de nombreux autres avantages de ces méthodes, notamment pour la simplification des stratégies de V&I pour la production, ainsi que leur utilité dans le cadre des méthodes de collecte du PISE. Cependant, nous reconnaissons que les deux approches reposent sur des hypothèses et que, de manière générale, d'autres travaux seront nécessaires pour évaluer de manière exhaustive ces méthodes et les préciser.

**Figure 5-1**

**Biais relatif MC, estimations de la variance MC et moyenne des estimations de la variance MC pour de nombreux estimateurs obtenus au moyen d'une simulation pour deux variables de l'EAMEF de 2022. Les graphiques sont une fonction de la probabilité de non-réponse (valeurs manquantes) de l'année 2022.**



## Remerciements

Les auteurs souhaitent remercier les collègues suivants pour leurs conseils et leur discussion : Steve Matthews, Marie-Claude Duval, Jean-François Beaumont, Keven Bosa et David Haziza, ainsi que Pierre Daoust et Darren Gray pour leur examen grandement apprécié de cet article.

## Bibliographie

- Beaumont, J-F., et Bissonnette, J. (2011), "Variance estimation under composite imputation: The methodology behind SEVANI", *Survey Methodology*, Statistics Canada, 37(2), p. 171-179.
- Dagdoug, M., Goga, C., et Haziza, D. (2023), "Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison", *Journal of Survey Statistics and Methodology*, 11, 141-188.

- Davison, A. C., et Sardy, S. (2007), “Resampling variance estimation in surveys with missing data”, *Journal of Official Statistics*, 23 (2007), p. 371-386.
- Godbout, S., Beaucage, Y., et Turmelle, C. (2011), “Achieving Quality and Efficiency Using a Top-Down Approach in the Canadian Integrated Business Statistics Program”, *Proceedings of the Conference of European Statisticians*. Ljubljana, Slovenia.
- Quinlan, J. (1993), “Combining instance-based and model-based learning”, *Proceedings of the tenth International Conference on machine learning*, p. 236–243.
- Quinlan, J. (1992), “Learning with continuous classes”, *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, 92, p. 343–348. World Scientific.
- Statistique Canada (2013), G-Sam Methodology Guide, Rapport technique (Ébauche).
- Statistique Canada (2015), Integrated Business Statistics, Program Overview. Catalogue no. 68-515-X, ISBN 978-0-660- 02486-8.
- Statistique Canada (2017), Functional Description of the Banff System for Edit and Imputation, Rapport technique.
- Statistique Canada (2019), G-Est Methodology Guide, Rapport technique.