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

Techniques d’enquête

Pondération par probabilité inverse équilibrant les quantiles pour des échantillons non probabilistes

par Maciej Beręsewicz, Marcin Szymkowiak et Piotr Chlebicki

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.

Publication autorisée par la ministre responsable de Statistique Canada

© Sa Majesté le Roi du chef du Canada, représenté par la ministre de l'Industrie, 2025

L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2025
Vol. 51, nº 2, pp. 583-612
Statistique Canada, nº 12-001-X au catalogue

Pondération par probabilité inverse équilibrant les quantiles pour des échantillons non probabilistes

Maciej Beręsewicz, Marcin Szymkowiak et Piotr Chlebicki[1]

Résumé

L'utilisation de sources de données non probabilistes à des fins statistiques et pour des statistiques officielles a gagné en popularité au cours des dernières années. L'inférence statistique fondée sur des échantillons non probabilistes devient toutefois plus difficile en raison de sa nature biaisée et de l'absence de représentativité. Dans le présent article, nous proposons un estimateur de pondération probabiliste inverse équilibrant les quantiles (PPIEQ) pour des échantillons non probabilistes. Nous appliquons ici l'idée de Harms et Duchesne (2006) permettant d'utiliser les renseignements sur les quantiles dans le processus d'estimation afin de reproduire des totaux connus et la répartition de variables auxiliaires. Nous parlons de l'estimation des probabilités de PPIEQ et de sa variance. Notre étude par simulation a montré la robustesse des estimateurs proposés par rapport à une spécification erronée de modèle, qui aident ainsi à réduire les biais et l'erreur quadratique moyenne. En dernier lieu, nous avons appliqué les méthodes proposées pour estimer la proportion d'emplois vacants destinés à des travailleurs ukrainiens en Pologne en utilisant un ensemble intégré de données administratives et d'enquête sur les postes vacants.

Mots-clés : Approche par calage; enquête sur les postes vacants; intégration des données.

1. Introduction

Dans les statistiques officielles, les renseignements concernant la population cible et ses caractéristiques sont principalement recueillis au moyen d'enquêtes probabilistes ou de recensement, ou sont obtenus à partir de registres administratifs, qui couvrent la totalité (ou presque) des unités de la population. Cependant, en raison d'une augmentation des taux de non-réponse, surtout de la non-réponse totale et de l'absence de réponse découlant du fardeau croissant du répondant, en plus de la hausse des coûts des enquêtes menées par les instituts nationaux de statistique, les sources de données non probabilistes gagnent en popularité (Beaumont, 2020; Beręsewicz, 2017). L'utilisation d'enquêtes non probabilistes, comme les panels en ligne à participation volontaire, les médias sociaux, les données de lecteurs optiques, les données de téléphones mobiles ou les données de registres volontaires, fait actuellement l'objet d'un examen pour la production de statistiques officielles (Citro, 2014; Daas, Puts, Buelens et Hurk, 2015). Puisque le mode de sélection de ces sources est inconnu, il est impossible d'appliquer directement des méthodes d'inférence classique fondées sur le plan de sondage.

Afin de corriger ce problème, différentes approches fondées sur la pondération probabiliste inverse (PPI), l'estimateur d'imputation massive (IM) et l'estimateur doublement robuste (DR) ont été proposées pour deux scénarios importants : 1) des données au niveau de la population sont disponibles, sous forme de données au niveau de l'unité (par exemple tirées d'un registre couvrant l'ensemble de la population) ou de

1. Maciej Beręsewicz, Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Statistics, Al. Niepodległości 10, 61-875 Poznań, Pologne et Statistical Office in Poznań, ul. Wojska Polskiego 27/29, 60-624 Poznań, Pologne. Courriel : maciej.beresewicz@ue.poznan.pl; Marcin Szymkowiak, Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Statistics, Al. Niepodległości 10, 61-875 Poznań, Pologne et Statistical Office in Poznań, ul. Wojska Polskiego 27/29, 60-624 Poznań, Pologne; Piotr Chlebicki, Stockholm University, Department of Mathematics, Albano hus 1, 106 91 Stockholm.

totaux ou de moyennes de population connus; 2) seules des données d'enquête sont disponibles comme source de renseignements à propos de la population visée (voir Elliott et Valliant, 2017). Wu (2022) a classé ces approches en trois groupes qui exige un cadre de randomisation conjoint faisant intervenir p (plan d'échantillonnage avec probabilités) et un modèle de régression des résultats ξ ou un modèle de score de propension q. Dans cette approche, l'estimateur de PPI se trouve dans un cadre qp, l'estimateur d'IM, dans un cadre ξp, et l'estimateur DR, dans un cadre qp ou ξp.

La majorité des approches présument que les données sur la population sont utilisées pour réduire le biais de l'échantillonnage non probabiliste, grâce à une repondération adéquate permettant de reproduire des totaux et des moyennes de population connus, en modélisant $E(Y \mid \mathbf{X})$ grâce à diverses techniques, ou en combinant les deux approches (par exemple estimateurs doublement robustes, voir Chen, Li et Wu (2020); régression multiniveau et poststratification (RMP) aussi appelée *Mister-P*, voir Gelman et Little (1997)). La majorité de ces méthodes reposent sur un nombre limité de moments de données continues ou de données de dénombrement, avec quelques exceptions. Par exemple, les approches non paramétriques fondées sur le plus proche voisin (PPV), comme celles exposées par Yang, Kim et Hwang (2021), ou une estimation à noyau de la densité décrite par Chen, Yang et Kim (2022) font également partie des propositions. Habituellement, l'approche classique, comme la PPI, semble limitée, même si les données d'enquête et de registre renferment des données continues ou des données de dénombrement qu'il est possible d'utiliser pour tenir compte des erreurs d'appariement. Cela est attribuable au fait que les estimateurs de PPI classiques ne permettent pas l'appariement de quartiles, de déciles ou de répartition totale entre échantillons.

Dans le présent article, nous portons notre attention sur le cadre qp et étendons les estimateurs de PPI existants pour tenir compte des écarts de répartition au moyen de quantiles. Nous nous appuyons sur l'idée de calage d'enquête pour les quantiles décrite par Harms et Duchesne (2006) et proposons un estimateur de PPI équilibrant les quantiles (ci-après appelé PPIEQ) qui reproduit conjointement des quantiles et des totaux pour un ensemble de variables auxiliaires. Notre contribution peut se résumer ainsi :

- nous étendons l'estimateur de PPI pour les échantillons non probabilistes afin de tenir compte non seulement des totaux et des moyennes, mais aussi pour la répartition des variables auxiliaires sous la forme d'un ensemble de quantiles;
- nous montrons que les méthodes d'inférence existantes s'appliquent aux techniques proposées;
- nous apportons des preuves de l'existence de solutions pour les techniques proposées;
- nous montrons que l'ajout de renseignements sur les quantiles équivaut à utiliser une régression par morceaux (constante), qui peut servir à estimer une relation non linéaire grâce à un modèle linéaire, ce qui rend la PPI plus robuste pour des erreurs de spécification du modèle et, par conséquent, réduit le biais et rehausse l'efficacité de l'estimation.

Dans le présent article, nous respectons en général la notation utilisée par Wu (2022). L'article s'organise autour de la structure suivante : dans la section 2, nous présentons le scénario de base, à savoir la notation

et une description abrégée du calage pour les totaux et les quantiles séparément. Dans la section 3, nous parlons de l'estimateur de PPI pour les échantillons non probabilistes. Dans la section 4, nous présentons l'estimateur proposé. La section 5 comprend les résultats d'une étude par simulation, et finalement à la section 6, nous résumons les résultats d'une étude dans laquelle l'approche proposée a été appliquée aux données combinées provenant d'une enquête sur les postes vacants et des postes vacants de sources administratives (services publics de placement). L'article se termine par les conclusions et cerne les problèmes de recherche qui nécessitent un examen plus approfondi.

2. Scénario de base

2.1 Notation

Supposons que $U = \{1,...,N\}$ correspond à la population cible constituée de N unités étiquetées. Chaque unité k s'accompagne d'un vecteur de variables auxiliaires $\mathbf{x}$ et de la variable cible y, et de leurs valeurs correspondantes $\mathbf{x}_k$ et y_k, respectivement. Supposons que $\{(y_k, \mathbf{x}_k), k \in S_A\}$ est un ensemble de données d'échantillon non probabiliste S_A de taille n_A, et $\{(\mathbf{x}_k, d_k^B), k \in S_B\}$ un ensemble de données d'échantillon probabiliste S_B de taille n_B. Seuls les renseignements sur les variables auxiliaires $\mathbf{x}$ se trouvent dans les deux ensembles de données, et $d_k^B = 1/\pi_k^B$ sont des facteurs de pondération fondés sur le plan attribués à chaque unité de l'échantillon S_B. Le tableau 2.1 résume le scénario, qui ne compte aucun chevauchement.

Tableau 2.1
Scénario selon deux sources de données

Échantillon	ID	Poids de sondage $d = \pi^{-1}$	Covariables x	Variable à l'étude y
Échantillon non probabiliste (S_A)	1	?	√	√
	⋮	?	⋮	⋮
	n_A	?	√	√
Échantillon probabiliste (S_B)	1	√	√	?
	⋮	⋮	⋮	?
	n_B	√	√	?

Supposons que $R_k = I(k \in S_A)$ est la variable indicatrice montrant que l'unité k est comprise dans l'échantillon non probabiliste S_A, qui est défini pour toutes les unités de la population U. Supposons que $\pi_k^A = P(R_k = 1 | \mathbf{x}_k)$ pour $k = 1,\ldots,N$ sont leurs scores de propension. Pour π_k^A, nous pouvons présumer un modèle paramétrique $\pi(\mathbf{x}_k; \boldsymbol{\gamma})$ (par exemple régression logistique) comportant des paramètres inconnus du modèle $\boldsymbol{\gamma}$.

Le but consiste à estimer le total d'une population finie $\tau_y = \sum_{k \in U} y_k$ ou la moyenne $\bar{\tau}_y = \tau_y / N$ de la variable d'intérêt y. Dans des enquêtes de probabilité (présumant que y_k sont connus), l'estimateur de

Horvitz-Thompson est l'estimateur bien connu du total d'une population finie, qui est exprimé comme $\hat{\tau}_{y\pi} = \sum_{k \in S_B} d_k^B y_k$. Cet estimateur est sans biais pour τ_y, c'est-à-dire $E(\hat{\tau}_{y\pi}) = \tau_y$. Il faut toutefois utiliser une approche différente pour les enquêtes de non-probabilité, puisque π_k^A sont inconnus et doivent être estimés (tout en prenant en considération que dans notre scénario, y_k sont inconnus dans S_B et qu'il est impossible d'utiliser $\hat{\tau}_{y\pi}$).

Dans le présent article, nous nous appuyons sur l'idée de Harms et Duchesne (2006) pour les quantiles, ce qui constitue une extension de l'approche de calage proposée par Deville et Särndal (1992) pour les totaux, par conséquent, nous exposons brièvement ces deux méthodes dans la section suivante.

2.2 Estimateur par calage pour un total

Supposons que $\mathbf{x}$ est un vecteur de variables auxiliaires (variables repères) pour lequel $\tau_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$ est présumé connu. L'idée principale derrière le calage d'échantillons probabilistes consiste à trouver de nouveaux poids de calage w_k^B aussi rapprochés que possible des poids initiaux d_k^B et de reproduire avec exactitude des totaux de population connue $\tau_{\mathbf{x}}$. Autrement dit, afin de trouver un nouveau vecteur de poids de calage $\mathbf{w} = \text{argmin}_{\mathbf{v}} D(\mathbf{d}, \mathbf{v})$, nous minimisons une fonction de distance $D(\mathbf{d}, \mathbf{v}) = \sum_{k \in S_B} d_k^B \, G\left(\frac{v_k^B}{d_k^B}\right) \rightarrow \min$ soumise à des équations de calage $\sum_{k \in S_B} v_k^B \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$, où $\mathbf{d} = (d_1^B, \ldots, d_{n_B}^B)^T$, $\mathbf{v} = (v_1^B, \ldots, v_{n_B}^B)^T$ et $G(\cdot)$ est une fonction devant répondre à certaines conditions de régularité. À titre d'exemple, si $G(x) = \frac{(x-1)^2}{2}$, à laquelle est ensuite appliquée la méthode des multiplicateurs de Lagrange, les poids de calage w_k^B peuvent être exprimés comme suit : $w_k^B = d_k^B + d_k^B \left(\tau_{\mathbf{x}} - \hat{\tau}_{\mathbf{x}\pi}\right)^T \left(\sum_{j \in S_B} d_j^B \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \mathbf{x}_k$. L'estimateur de calage définitif d'un total de population τ_y (en présumant que y_k est connu dans un échantillon S_B) peut s'exprimer sous la forme $\hat{\tau}_{y\mathbf{x}} = \sum_{k \in S_B} w_k y_k$.

2.3 Estimateur par calage pour un quantile

Harms et Duchesne (2006) ont examiné l'estimation de quantiles au moyen d'une approche par calage d'une manière semblable à celle qu'ont proposée Deville et Särndal (1992) pour un total de population finie τ_y. Par analogie, leur approche ne nécessite pas de connaître les valeurs de toutes les variables auxiliaires pour toutes les unités de la population. Il suffit de connaître les quantiles correspondants pour les variables repères. Nous abordons brièvement ci-dessous le problème que soulève l'établissement des poids de calage pour un échantillon probabiliste S_B.

Nous cherchons à estimer un quantile $Q_{y,\alpha}$ d'ordre $\alpha \in (0,1)$ de la variable d'intérêt y, qui peut s'exprimer sous la forme $Q_{y,\alpha} = \inf\left\{t \,\middle|\, F_y(t) \geq \alpha\right\}$, où $F_y(t) = N^{-1} \sum_{k \in U} H(t - y_k)$ et la fonction Heaviside est donnée par

$$H(t - y_k) = \begin{cases} 1, & t \geq y_k, \\ 0, & t < y_k. \end{cases} \tag{2.1}$$

Nous présumons que $\mathbf{Q}_{\mathbf{x},\alpha}$ est un vecteur des quantiles d'une population connue d'ordre α pour un vecteur de variables auxiliaires $\mathbf{x}$, où $\alpha \in (0,1)$. Même en utilisant la même notation pour un vecteur $\mathbf{x}$, dans une approche par calage appliquée à des totaux et à des quantiles, ses éléments se révéleront différents pour les uns et les autres. Dans la majorité des cas, les quantiles seront connus pour des variables auxiliaires continues, contrairement aux totaux qui sont habituellement connus pour les variables catégoriques. Il est toutefois possible que pour une variable auxiliaire particulière, son total de population et le quantile correspondant d'ordre α soient connus. Pour trouver un nouveau vecteur de poids de calage $\mathbf{w}$ qui reproduit des quantiles de population connue dans un vecteur $\mathbf{Q}_{\mathbf{x},\alpha}$, un estimateur de la fonction de répartition interpolée de $F_y(t)$ est défini comme

$$\hat{F}_{y,\text{cal}}(t) = \frac{\sum_{k\in S_B} w_k^B H_{y,S_B}(t, y_k)}{\sum_{k\in S_B} w_k^B},$$

où la fonction de Heaviside de la formule (2.1) est remplacée par la fonction modifiée $H_{y,S_B}(t, y_k)$ donnée par

$$H_{y,S_B}(t, y_k) = \begin{cases} 1, & y_k \leqslant L_{y,S_B}(t), \\ \vartheta_{y,S_B}(t), & y_k = U_{y,S_B}(t), \\ 0, & y_k > U_{y,S_B}(t), \end{cases} \tag{2.2}$$

où les paramètres appropriés sont définis comme $L_{y,S_B}(t) = \max\{\{y_k, k \in S_B \mid y_k \leqslant t\} \cup \{-\infty\}\}$, $U_{y,S_B}(t) = \min\{\{y_k, k \in S_B \mid y_k > t\} \cup \{\infty\}\}$ et

$$\vartheta_{y,S_B}(t) = \frac{t - L_{y,S_B}(t)}{U_{y,S_B}(t) - L_{y,S_B}(t)}$$

pour $k = 1,\ldots,n_B$, $t \in \mathbb{R}$. D'un point de vue pratique, il serait possible d'utiliser une légère approximation de la forme « en escalier », qui repose sur la fonction logistique, c'est-à-dire $H(x) \approx \frac{1}{2} + \frac{1}{2}\tanh kx = (1+e^{-2kx})^{-1}$, dans laquelle une valeur plus grande de k correspond à une transition plus nette à $x = 0$.

Un estimateur par calage de quantile $Q_{y,\alpha}$ d'ordre α pour une variable y est défini comme étant $\hat{Q}_{y,\text{cal},\alpha} = \hat{F}^{-1}_{y,\text{cal}}(\alpha)$, où un vecteur $\mathbf{w} = \text{argmin}_{\mathbf{v}}\, D(\mathbf{d}, \mathbf{v})$ constitue une des solutions à un problème d'optimisation $D(\mathbf{d}, \mathbf{v}) = \sum_{k\in S_B} d_k^B\, G\left(\frac{v_k^B}{d_k^B}\right) \to \min$ soumis à des contraintes de calage $\sum_{k\in S_B} v_k^B = N$ et $\hat{\mathbf{Q}}_{\mathbf{x},\text{cal},\alpha} = \mathbf{Q}_{\mathbf{x},\alpha}$ ou, de façon équivalente, $\hat{F}_{x_j,\text{cal}}(Q_{x_j,\alpha}) = \alpha$ pour toutes les variables auxiliaires x_j provenant d'un vecteur $\mathbf{x}$.

Comme dans le cas précédent, si $G(x) = \frac{(x-1)^2}{2}$, alors, en utilisant la méthode des multiplicateurs de Lagrange, les poids définitifs w_k^B peuvent être exprimés comme suit : $w_k^B = d_k^B + d_k^B \left(\mathbf{T}_{\mathbf{a}} - \sum_{j\in S_B} d_j^B \mathbf{a}_j\right)^T \left(\sum_{j\in S_B} d_j^B \mathbf{a}_j \mathbf{a}_j^T\right)^{-1} \mathbf{a}_k$, où $\mathbf{T}_{\mathbf{a}} = (N, \alpha, \ldots, \alpha)^T$ et les éléments de $\mathbf{a}_k$ (dont le premier élément est 1) sont donnés par

$$a_{kj} = \begin{cases} N^{-1}, & x_{kj} \le L_{x_j, S_B}(Q_{x_j,\alpha}), \\ N^{-1}\vartheta_{x_j, S_B}(Q_{x_j,\alpha}), & x_{kj} = U_{x_j, S_B}(Q_{x_j,\alpha}), \\ 0, & x_{kj} > U_{x_j, S_B}(Q_{x_j,\alpha}). \end{cases} \tag{2.3}$$

Dans la méthode décrite ci-dessus, on présume qu'un quantile de population connu est reproduit pour un ensemble de variables auxiliaires, c'est-à-dire que le processus de calage repose sur un quantile particulier (d'ordre α). Il pourrait par exemple s'agir de la médiane $\alpha =$ 0,5. Il s'agit d'un processus simple qui permet de trouver des poids de calage qui révèlent avec précision des quantiles de population (par exemple quartiles) pour un ensemble donné de variables auxiliaires.

3. Estimateur par pondération de probabilité inverse

3.1 Hypothèses

Tout au long de l'article, nous suivons des hypothèses types sur l'inférence qui reposent sur des échantillons non probabilistes compris dans le cadre qp, c'est-à-dire :

(A1) indépendance conditionnelle de R_k et y_k, étant donné $\mathbf{x}_k$;

(A2) toutes les unités dans la population cible ont des scores de propension non nuls $\pi_k^A > 0$;

(A3) $R_1, \ldots, R_N$ sont indépendants compte tenu des variables auxiliaires $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$.

Pour estimer la variance de l'estimateur de PPI, nous appliquons les mêmes conditions de régularité que celles mentionnées dans la documentation supplémentaire de l'article de Chen, Li et Wu (2020) et dans la section 3.2 de Tsiatis (2006). Par souci de simplicité, nous utiliserons $\mathbf{x}_k$ (pour lequel les totaux sont connus ou estimés) et $\mathbf{a}_k$ (pour lequel des α-quantiles précis de variables sélectionnées dans $\mathbf{x}_k$ sont connus ou estimés) dans les définitions des estimateurs de PPI.

3.2 Estimateur de pondération probabiliste inverse

Chen, Li et Wu (2020) ont traité de la pondération par probabilité inverse (PPI) et de ses propriétés asymptotiques. Supposons qu'il soit possible de modéliser des scores de propension de façon paramétrique comme étant $\pi_k^A = P(R_k = 1 \mid \mathbf{x}_k) = \pi(\mathbf{x}_k; \boldsymbol{\gamma}_0)$, où $\boldsymbol{\gamma}_0$ constitue la valeur réelle des paramètres de modèle inconnus. L'estimation du maximum de vraisemblance de π_k^A se calcule comme suit : $\hat{\pi}_k^A = \pi(\mathbf{x}_k; \hat{\boldsymbol{\gamma}})$, où $\hat{\boldsymbol{\gamma}}$ maximise le logarithme du rapport de vraisemblance en vertu de renseignements complets :

$$\ell(\boldsymbol{\gamma}) = \log\left\{\prod_{k=1}^{N} (\pi_k^A)^{R_k} (1-\pi_k^A)^{1-R_k}\right\} = \sum_{k \in S_A} \log\left(\frac{\pi_k^A}{1-\pi_k^A}\right) + \sum_{k \in U} \log(1-\pi_k^A).$$

Cependant, dans la pratique, des variables auxiliaires de référence $\mathbf{x}$ peuvent être fournies par l'échantillon probabiliste S_B et ainsi $\ell(\boldsymbol{\gamma})$ est remplacé par une fonction de pseudo-vraisemblance (PV) :

$$\ell^*(\boldsymbol{\gamma}) = \sum_{k\in S_A} \log\left(\frac{\pi_k^A}{1-\pi_k^A}\right) + \sum_{k\in S_B} d_k^B \log(1-\pi_k^A).$$

L'estimateur de pseudo-maximum de vraisemblance $\hat{\boldsymbol{\gamma}}$ peut être obtenu par la résolution des équations de pseudo-scores données par $\mathbf{U}(\boldsymbol{\gamma}) = \partial\ell^*(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma} = \mathbf{0}$. Si l'hypothèse de la régression logistique est présumée pour π_k^A, $\mathbf{U}(\boldsymbol{\gamma})$ est alors donnée par

$$\mathbf{U}(\boldsymbol{\gamma}) = \sum_{k\in S_A} \mathbf{x}_k - \sum_{k\in S_B} d_k^B \pi\left(\mathbf{x}_k;\boldsymbol{\gamma}\right)\mathbf{x}_k. \tag{3.1}$$

Il convient de mentionner que les poids fondés sur $\pi\left(\mathbf{x}_k;\hat{\boldsymbol{\gamma}}\right)$, pour lesquels l'estimation de $\hat{\boldsymbol{\gamma}}$ repose sur (3.1), ne reproduisent ni la taille ni les totaux de population pour les variables comprises dans un vecteur $\mathbf{x}$.

Une autre solution pourrait être de remplacer des équations de pseudo-scores $\mathbf{U}(\boldsymbol{\gamma}) = \mathbf{0}$ par un système d'équations générales d'estimation (Kim et Riddles, 2012; Wu, 2022). Supposons que $\mathbf{h}(\mathbf{x}_k;\boldsymbol{\gamma})$ est un vecteur personnalisé de fonctions ayant les mêmes dimensions que $\boldsymbol{\gamma}$. Les équations générales d'estimation correspondront à :

$$\mathbf{G}(\boldsymbol{\gamma}) = \sum_{k\in S_A} \mathbf{h}\left(\mathbf{x}_k;\boldsymbol{\gamma}\right) - \sum_{k\in S_B} d_k^B \pi(\mathbf{x}_k;\boldsymbol{\gamma})\,\mathbf{h}(\mathbf{x}_k;\boldsymbol{\gamma}), \tag{3.2}$$

et, quand nous supposons que $\mathbf{h}(\mathbf{x}_k;\boldsymbol{\gamma}) = \mathbf{x}_k/\pi(\mathbf{x}_k;\boldsymbol{\gamma})$, alors (3.2) se réduit à des équations de calage :

$$\mathbf{G}(\boldsymbol{\gamma}) = \sum_{k\in S_A} \frac{\mathbf{x}_k}{\pi\left(\mathbf{x}_k;\boldsymbol{\gamma}\right)} - \sum_{k\in S_B} d_k^B \mathbf{x}_k. \tag{3.3}$$

Dans ce paramètre, les données au niveau de l'unité ne sont pas nécessaires pour obtenir $\sum_{k\in S_B} d_k^B \mathbf{x}_k$. Des totaux connus ou estimés de la population tirés de l'enquête de probabilité de référence peuvent remplacer ces sommes. Kim et Riddles (2012) ont montré que l'estimateur (3.3) permet d'obtenir une estimation optimale lorsqu'un modèle de régression linéaire s'applique à y et à $\mathbf{x}$.

Après avoir estimé $\boldsymbol{\gamma}$, nous obtenons un score de propension $\hat{\pi}_k^A$ et deux versions de l'estimateur de PPI :

$$\hat{\bar{\tau}}_{\text{PPI1}} = \frac{1}{N}\sum_{k\in S_A} \frac{y_k}{\hat{\pi}_k^A} \quad \text{et} \quad \hat{\bar{\tau}}_{\text{PPI2}} = \frac{1}{\hat{N}_A}\sum_{k\in S_A} \frac{y_k}{\hat{\pi}_k^A}, \tag{3.4}$$

où $\hat{N}_A = \sum_{k\in S_A} \hat{\pi}_k^A$.

L'estimateur de PPI classique ne prend en compte qu'une quantité limitée de renseignements et ne conserve pas la répartition des variables auxiliaires. Pour surmonter cette limite, nous avons fait appel à l'approche par calage pour les quantiles décrite à la section 2.3 et modifié l'estimateur de PPI afin d'équilibrer les répartitions à l'aide de quantiles particuliers (par exemple quartiles, déciles).

4. Pondération par probabilité inverse équilibrant les quantiles

4.1 Approche proposée

Dans notre proposition, nous modifions les scores de propension π_k^A afin de tenir compte de $\mathbf{a}_k$ de manière à obtenir la fonction à partir de $\pi_k^A = \pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})$. L'approche proposée se justifie par le fait que la forme paramétrique $\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})$ est une représentation sursaturée qui comprend $\pi(\mathbf{x}_k; \boldsymbol{\gamma})$ en tant que cas spécial. Il convient également de mentionner que nous faisons une distinction entre $\boldsymbol{\gamma}$ et $\boldsymbol{\eta}$, puisque ce dernier présente des paramètres additionnels faisant référence à $\mathbf{a}_k$. L'inclusion de $\mathbf{a}_k$ permet de corriger les écarts entre les quantiles de variables auxiliaires de S_A en faisant référence à S_B.

L'approche proposée est semblable à celles utilisées dans la littérature sur l'échantillonnage d'enquête ou l'inférence causale, qui présentent des arguments en faveur de l'inclusion de moments plus élevés pour $\mathbf{x}$ variables auxiliaires (Ai, Linton et Zhang, 2020; Imai et Ratkovic, 2014). Il s'agit d'autre part d'une version plus simple du score de propension qui équilibre les répartitions, comparativement à des méthodes faisant appel à la densité du noyau (Hazlett, 2020) ou à une intégration numérique (Sant'Anna, Song et Xu, 2022).

Notre approche ne se limite pas à une méthode particulière d'estimation des paramètres du score de propension π_k^A. Ainsi, nous commençons par utiliser la fonction de vraisemblance complète ou de pseudo-vraisemblance comportant la modification proposée, puis la résolvons à l'aide de (3.1), qui prend la forme suivante si $\mathbf{x}_k$ et $\mathbf{a}_k$ sont appliqués

$$\mathbf{U}(\boldsymbol{\gamma}) = \begin{pmatrix} \sum_{k \in S_A} \mathbf{x}_k - \sum_{k \in S_B} d_k^B \pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})\, \mathbf{x}_k \\ \sum_{k \in S_A} \mathbf{a}_k - \sum_{k \in S_B} d_k^B \pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})\, \mathbf{a}_k \end{pmatrix}. \tag{4.1}$$

Il convient de mentionner que cette procédure ne permet pas de reproduire les quantiles pour les variables $\mathbf{x}$ sélectionnées, mais elle réduit les écarts entre les répartitions de S_A et de S_B. Dans l'étude par simulation, nous avons montré que cela se révélait particulièrement important dans les modèles de sélection non linéaires. En revanche, l'ajout d'un plus grand nombre de variables pourrait entraîner une instabilité de la procédure d'estimation et des poids importants.

Pour surmonter les limites d'une estimation du maximum de vraisemblance (EMV), nous pouvons utiliser les équations généralisées d'estimation dans (3.2), qui permettent de reproduire des totaux et des quantiles à partir de l'échantillon probabiliste S_B. Dans ce cas, $\mathbf{G}(\boldsymbol{\eta})$ devient

$$\mathbf{G}(\boldsymbol{\eta}) = \sum_{k \in S_A} \mathbf{h}(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta}) - \sum_{k \in S_B} d_k^B \pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})\, \mathbf{h}(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta}).$$

Si nous supposons une régression logistique pour $\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})$ et $\mathbf{h}(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta}) = \begin{pmatrix} \mathbf{x}_k \\ \mathbf{a}_k \end{pmatrix} \Big/ \pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})$, alors $\mathbf{G}(\boldsymbol{\eta})$ se réduit à

$$\mathbf{G}(\boldsymbol{\gamma}) = \begin{pmatrix} \sum_{k \in S_A} \dfrac{\mathbf{x}_k}{\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})} - \sum_{k \in S_B} d_k^B \mathbf{x}_k \\ \sum_{k \in S_A} \dfrac{\mathbf{a}_k}{\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})} - \sum_{k \in S_B} d_k^B \mathbf{a}_k \end{pmatrix}. \tag{4.2}$$

Cette approche permet d'estimer les paramètres $\boldsymbol{\eta}$, de sorte que la taille de population N, les totaux de population connus (ou estimés) pour certaines variables auxiliaires comprises dans un vecteur $\mathbf{x}$ ou des quantiles de population connus pour d'autres variables issues de ce vecteur soient tous reproduits. Dans la section A de l'annexe, nous exposons les conditions qui garantissent l'existence et l'unicité des solutions pour $\mathbf{U}(\boldsymbol{\eta}) = \mathbf{0}$ et $\mathbf{G}(\boldsymbol{\eta}) = \mathbf{0}$.

Note 1. *Harms et Duchesne (2006, page 41) indiquent que si la relation entre y et une variable auxiliaire scalaire x est exactement linéaire, alors le calage sur un α -quantile de la variable auxiliaire donne le α-quantile de population exact de la variable y cible.*

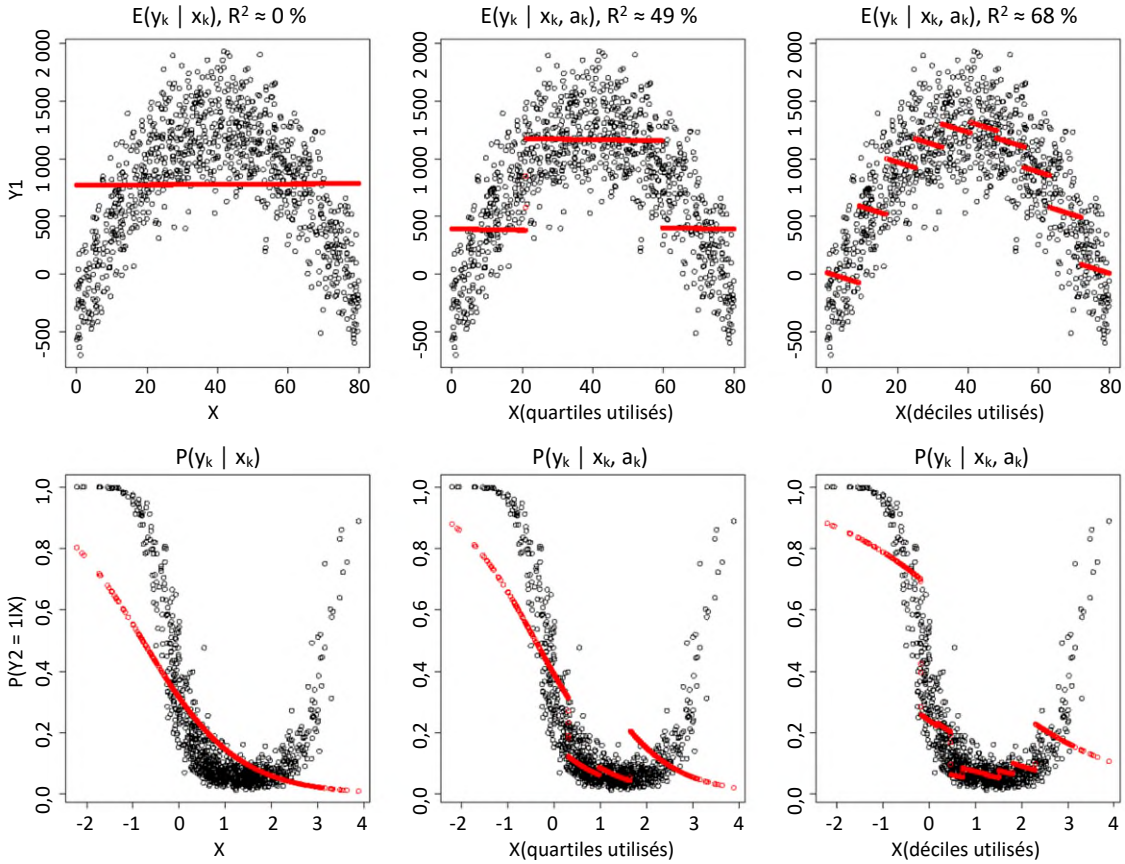
Note 2. *L'utilisation de $\boldsymbol{a}_k$ dans (4.2) est liée à l'utilisation d'une régression par morceaux (constante) où les variables $\boldsymbol{x}_k$ sont fractionnées (par exemple en quartiles ou en déciles) pour ensuite inclure $\boldsymbol{a}_k$ dans une régression linéaire comme celle présentée ci-dessous*

$$\hat{y}_k = \underbrace{\boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}}_{\text{partie linéaire}} + \underbrace{\boldsymbol{a}_k^T \hat{\boldsymbol{\beta}}^*}_{\text{partie fragmentaire}} .$$

Nous faisons donc appel à une méthode par morceaux pour estimer la relation non linéaire qui existe entre y et $\boldsymbol{x}$. Examinons l'exemple suivant : créer $n = 1\,000$ observations de $X \sim$ Uniforme(0, 80), $Y_1 \sim N(1\,300 - (X - 40)^2, 300)$ et $p = P(Y_2 = 1) = \text{logit}(-3 + (X - 1{,}5)^2 + N(0; 0{,}5))$ pour obtenir le triplet (y_k, p_k, x_k) pour $k = 1,\ldots,n$. Puis, supposons que x_k est une variable auxiliaire dans les paramètres suivants : 1) x_k est utilisé tel quel; 2) x_k est utilisé avec a_k précisé pour des quartiles; 3) x_k est utilisé avec a_k précisé pour des déciles. La figure 4.1 expose la manière dont la prédiction de Y_1 (première ligne) et de $P(Y_2 = 1)$ (deuxième ligne) change avec l'utilisation de x_k ou de x_k et de a_k. Nous avons appliqué une régression linéaire pour Y_1 et des probabilités que nous avons modélisées à partir d'une régression logistique pour Y_2. Comme il était prévu, l'insertion de quantiles dans $\boldsymbol{a}_k$ permet d'obtenir de meilleures prédictions pour les deux modèles.

Note 3. *L'estimateur de PPIEQ obtenu en résolvant (4.2) est en réalité doublement robuste, c'est-à-dire qu'il est asymptotiquement sans biais si le modèle de probabilité inverse ou le modèle de résultat est correct. L'utilisation de quantiles dans $\boldsymbol{a}_k$ permet d'estimer les relations non linéaires dans les deux modèles. Dans l'étude par simulation, nous montrerons que l'approche proposée réduit le biais de manière significative lorsque les modèles sont non linéaires.*

Figure 4.1 Prédiction de Y_1 (ligne du haut) et de $\mathrm{P}(Y_2 = 1)$ (ligne du bas) fondée sur x_k ou x_k et a_k (quartiles ou déciles) tirée de l'exemple de la note 2



Note : Les cercles noirs indiquent les y_k observés et les cercles rouges, des prédictions $\hat{m}_k$ fondées sur le modèle linéaire et $\hat{p}_k$ fondées sur la régression logistique.

Après une estimation des paramètres $\boldsymbol{\eta}$, nous sommes en mesure d'estimer la moyenne de la population à partir de $\hat{\bar{\tau}}_{\mathrm{PPI1}}$ ou de $\hat{\bar{\tau}}_{\mathrm{PPI2}}$ obtenus dans l'équation (3.4), que nous désignons $\hat{\bar{\tau}}_{\mathrm{PPIEQ1}}$ ou $\hat{\bar{\tau}}_{\mathrm{PPIEQ2}}$, respectivement. Pour résumer l'approche que nous proposons :

- les scores de propension estimés reproduiront des moments connus ou estimés, de même que des α-quantiles précis (par exemple médiane, quartiles ou déciles). Nous insérons plus de renseignements sur les variables catégoriques ou continues s'ils sont présents dans les deux ensembles de données;
- l'inclusion des α-quantiles dans π_k^A permet d'estimer la relation entre l'inclusion de R et $\mathbf{x}$. Par conséquent, l'inclusion des α-quantiles rend les estimations plus robustes en vue de modéliser les spécifications incorrectes de la propension et du modèle de résultat lorsque l'estimateur de PPI calée est utilisé.

L'inférence fondée sur l'estimateur de PPI ne change pas avec l'ajout de nouvelles variables au modèle de score de propension. Nous ne présumons pas que les quantiles reposent sur, disons, des variables instrumentales ou qu'ils sont recueillis à partir de paradonnées (Park, Kim et Kim, 2019), mais nous les

traitons comme si nous ajoutions des moments plus élevés. L'estimation de la variance des estimateurs de PPIEQ est présentée à la section suivante.

4.2 Estimation de la variance

Une des solutions permettant d'estimer la variance des estimateurs de PPIEQ consiste à modifier l'équation $\boldsymbol{\Phi}_n(\boldsymbol{\eta}) = \mathbf{0}$ (Wu, 2022, équation 6.1) en imposant des contraintes aux quantiles. Dans le cas où $\mathbf{x}_k$ et $\mathbf{a}_k$ sont utilisés et que $\mathbf{h}(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})$ est défini comme précédemment, l'équation se présentera ainsi :

$$\boldsymbol{\Phi}_n(\boldsymbol{\eta}) = N^{-1} \begin{pmatrix} \sum_{k \in U} R_k \dfrac{y_k - \overline{\tau}_y}{\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})} \\ \sum_{k \in U} R_k \dfrac{\mathbf{x}_k}{\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})} - \sum_{k \in S_B} d_k^B \mathbf{x}_k \\ \sum_{k \in U} R_k \dfrac{\mathbf{a}_k}{\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})} - \sum_{k \in S_B} d_k^B \mathbf{a}_k \end{pmatrix}. \tag{4.3}$$

Il est possible d'obtenir la variance asymptotique de l'estimateur de PPIEQ à partir de la forme classique en sandwich donnée par :

$$\mathrm{VA}(\hat{\boldsymbol{\eta}}) = \left[E\left\{ \boldsymbol{\phi}_n(\boldsymbol{\eta}_0) \right\} \right]^{-1} \mathrm{Var}\left\{ \boldsymbol{\Phi}_n(\boldsymbol{\eta}_0) \right\} \left[E\left\{ \boldsymbol{\phi}_n(\boldsymbol{\eta}_0) \right\}' \right]^{-1},$$

où $\boldsymbol{\eta}_0$ correspond à la valeur réelle des paramètres inconnus du modèle et $\boldsymbol{\phi}_n(\boldsymbol{\eta}) = \partial \boldsymbol{\Phi}_n(\boldsymbol{\eta}) / \partial \boldsymbol{\eta}$ dépend de la forme de π_k^A et de $\mathbf{h}$. Chen, Li et Wu (2020) ont fourni la forme analytique de l'estimateur de variance lorsqu'une régression logistique est utilisée pour π_k.

Notre approche, qui repose sur $\mathbf{a}_k$ ou $\mathbf{x}_k$, pourrait mener à une variance plus élevée, puisque le nombre de variables pourrait être considérablement plus grand qu'avec $\mathbf{x}_k$ seul et que l'ajout de nouvelles variables pourrait accroître la variance de $\pi_k^A(1-\pi_k^A)$.

5. Étude par simulation

L'étude par simulation que nous avons menée pour comparer les estimateurs proposés respecte la procédure décrite par Kim et Wang (2019) et par Yang, Kim et Hwang (2021). Nous créons une population finie $\mathcal{F}_N = \{\mathbf{x}_k = (x_{k1}, x_{k2}),\ \mathbf{y}_k = (y_{k1}, y_{k2}) : k = 1, \ldots, N\}$ dont la taille est de $N =$ 100 000, où y_{k1} est le résultat continu, y_{k2} est le résultat binaire, $x_{k1} \sim N(1,1)$ et $x_{k2} \sim \mathrm{Exp}(1)$.

La création des variables cibles de la population finie (Y_1 et Y_2) et du vaste échantillon S_A découle de l'utilisation de variables de résultat (indiquées en tant que modèle de résultat, ou MR) et de la probabilité d'inclusion (indiquée en tant que modèle de probabilité, ou MP) qui sont présentées dans le tableau 5.1, où $\alpha_k \sim N(0,1)$, $\epsilon_k \sim N(0,1)$ et x_{k1}, x_{k2}, α_k et ϵ_k sont mutuellement indépendants. La variable α_k entraîne une dépendance de y_{k1} et de y_{k2}, et ce, même après correction en fonction de x_{k1} et de x_{k2}.

À partir de cette population, nous avons sélectionné un vaste échantillon S_A d'une taille d'environ 70 000 (pour MP1) et de 50 000 (pour MP2), selon le mécanisme de sélection, en présumant que l'indicateur d'inclusion $\delta_{Bk} \sim \text{Bernoulli}(p_k)$ avec p_k désigne la probabilité d'inclusion pour l'unité k. Nous avons ensuite sélectionné un échantillon aléatoire simple S_B de taille $n = 1\,000$, tiré de $\mathcal{F}_N$. Le but consiste à estimer la moyenne de la population $\bar{\tau}_j = N^{-1}\sum_{k=1}^{N} y_{kj},\ j = 1,2.$

Tableau 5.1
Modèles de résultat et probabilité d'inclusion dans un échantillon S_A utilisé dans l'étude par simulation

Type	Forme	Formule
Continue	Linéaire (MR1)	$y_{k1} = 1 + x_{k1} + x_{k2} + \alpha_k + \epsilon_k$
	Non linéaire (MR2)	$y_{k2} = 0{,}5(x_{k1} - 1{,}5)^2 + x_{k2}^2 + \alpha_k + \epsilon_k$
Binaire	Linéaire (MR3)	$P(y_{k1} = 1 \mid x_{k1}, x_{k2}; \alpha_k) = \text{logit}(1 + x_{k1} + x_{k2} + \alpha_k)$
	Non linéaire (MR4)	$P(y_{k2} = 1 \mid x_{k1}, x_{k2}; \alpha_k) = \text{logit}\{0{,}5(x_{k1} - 1{,}5)^2 + x_{k2}^2 + \alpha_k\}$
Sélection (S_A)	Linéaire (MP1)	$\text{logit}(p_k) = x_{k2}$
	Non linéaire (MP2)	$\text{logit}(p_k) = -3 + (x_{k1} - 1{,}5)^2 + (x_{k2} - 2)^2$

Note : MP = modèle de probabilité; MR = modèle de résultat.

Dans les résultats, nous précisons le biais de Monte-Carlo (B), l'erreur-type (ET) et l'erreur quadratique moyenne (EQM) fondés sur $R = 500$ simulations pour chaque variable y :

$$\text{B} = \bar{\hat{\tau}} - \bar{\tau},\ \text{ET} = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\tau}^{(r)} - \bar{\hat{\tau}})^2}{R-1}} \quad \text{et} \quad \text{EQM} = \sqrt{\text{B}^2 + \text{SE}^2}, \quad \text{où} \quad \bar{\hat{\tau}} = \sum_{r=1}^{R}\hat{\tau}^{(r)}/R \quad \text{et} \quad \hat{\tau}^{(r)}$$

est une estimation de la moyenne dans la r^{e} réplique. Les paquets `nonprobsvy` (Chrostowski et Beręsewicz, 2024) et `jointCalib` (Beręsewicz, 2023) ont servi à réaliser la simulation dans R (R Core Team, 2023). Pour trouver la solution aux contraintes de calage relatives à l'estimateur de PPIEQ, nous avons utilisé la méthode de Newton combinée à une stratégie globale de région de confiance de type *double deglog* proposée par Dennis Jr. et Schnabel (1996) et instaurée à l'aide du paquet `nleqslv` (Hasselman, 2023).

Les résultats de la simulation sont présentés dans les tableaux 5.2 et 5.3. Ce dernier montre la couverture empirique (CE) d'intervalles de confiance (IC) de 95 % fondée sur des estimateurs de la variance analytique pour les estimateurs de PPI. La CE est définie comme étant $\frac{1}{R}\sum_{r=1}^{R} I\left(\bar{\tau} \in \text{IC}^{(r)}\right) \times 100,$ où $\text{IC}^{(r)}$ correspond à l'IC calculée à la r^{e} itération, que nous nous attendons à trouver proche du niveau nominal de 95 %. Le tableau B.1 de l'annexe présente également des renseignements sur la qualité des sommes et des quantiles reproduits.

Les estimateurs suivants ont été examinés dans un ensemble d'équations de calage comprenant des totaux ou des quantiles pour des variables auxiliaires estimées à partir de l'échantillon S_B :

- valeurs naïves calculées à partir de l'échantillon S_A seulement;

- imputation massive par le plus proche voisin comprenant 5 voisins, à l'aide de x_1, x_2 seulement (PPV);
- imputation massive fondée sur une régression linéaire, à l'aide de x_1, x_2 seulement (MLG);
- estimateur doublement robuste employant seulement x_1, x_2 avec deux variantes de PPI : EMV (EMV DR; où l'EMV de PPI repose sur (3.1)) et équations d'estimation généralisées (EEG) (calées; EEG DR; où l'EEG de PPI repose sur (3.3));
- EMV de PPI et EEG de PPI, avec les variantes suivantes : totaux estimés pour x_1, x_2 seulement;
- EMV de PPIEQ et EEG de PPIEQ selon (4.1) et (4.2), respectivement, avec les variantes suivantes :
  - quartiles et totaux estimés pour x_1, x_2 (PPIEQ1);
  - déciles et totaux estimés pour x_1, x_2 (PPIEQ2).

Le tableau 5.2 renferme les résultats de quatre scénarios, dans lesquels sont examinées des combinaisons de MR et de MP pour des modèles de résultat continus et binaires, selon la définition du tableau 5.1. Dans le scénario I ayant trait au MR continu, la majorité des estimateurs montrent un rendement satisfaisant qui comporte peu de biais. Les estimateurs proposés, qui font appel à des quantiles et à des totaux, affichent une erreur quadratique moyenne (EQM) comparable à celle des estimateurs du plus proche voisin (PPV), du modèle linéaire généralisé (MLG) ou doublement robuste (DR). Comme prévu, la PPIEQ utilisée avec des EEG se distingue par une erreur-type plus faible que celle de l'EMV. En ce qui concerne les résultats binaires, le rendement des estimateurs de PPIEQ, surtout ceux qui reposent sur des EEG, se révèle encore plus marqué. L'EQM est comparable à celle des estimateurs du MLG et DR. De plus, la PPIEQ proposée à laquelle sont appliquées des EEG démontre une amélioration de l'EQM par rapport à l'estimateur des EEG de PPI.

Tableau 5.2
Résultats pour un Y continu et binaire (biais de Monte Carlo, erreur-type et erreur quadratique moyenne sont multipliés par 100)

	Scénario I			**Scénario II**			**Scénario III**			**Scénario IV**		
MR	**Linéaire (1)**			**Linéaire (1)**			**Non linéaire (2)**			**Non linéaire (2)**		
MP	**Linéaire (1)**			**Non linéaire (2)**			**Linéaire (1)**			**Non linéaire (2)**		
	B	**ET**	**EQM**	**B**	**ET**	**EQM**	**B**	**ET**	**EQM**	**B**	**ET**	**EQM**
						Y continu (MR1 et MR2)						
Naïve	18,71	0,40	18,70	-41,84	0,50	41,80	60,67	0,60	60,70	31,11	0,80	31,10
PPV	0,20	6,10	6,10	0,66	6,30	6,40	0,43	15,20	15,20	0,90	14,90	14,90
MLG	0,43	4,50	4,60	-0,19	4,50	4,50	-19,92	13,60	24,10	110,72	15,00	111,70
DR (EMV)	0,32	4,60	4,60	0,01	4,60	4,60	1,72	5,80	6,00	227,42	77,00	240,10
DR (EEG)	0,33	4,60	4,60	-0,19	4,50	4,50	0,96	7,10	7,10	123,79	18,10	125,10
Pondération par probabilité inverse (EMV)												
PPI	-0,01	5,70	5,70	69,73	27,50	75,00	0,42	9,30	9,40	467,97	175,10	499,60
PPIEQ1	-0,02	5,70	5,70	6,58	11,70	13,50	0,44	9,60	9,60	29,93	22,20	37,30
PPIEQ2	-0,50	5,80	5,80	3,12	11,90	12,30	0,73	12,70	12,70	4,04	21,50	21,80

Note : DR = doublement robuste; EEG = équations d'estimation généralisées; EMV = estimation du maximum de vraisemblance; EQM = erreur quadratique moyenne; ET = erreur-type; MLG = modèle linéaire généralisé; MP = modèle de probabilité; MR = modèle de résultat; PPI = pondération probabiliste inverse; PPIEQ = pondération probabiliste inverse équilibrant les quantiles; PPV = plus proche voisin.

Tableau 5.2 (suite)
Résultats pour un Y continu et binaire (biais de Monte Carlo, erreur-type et erreur quadratique moyenne sont multipliés par 100)

MR / MP	Scénario I Linéaire (1) Linéaire (1)			Scénario II Linéaire (1) Non linéaire (2)			Scénario III Non linéaire (2) Linéaire (1)			Scénario IV Non linéaire (2) Non linéaire (2)		
	B	**ET**	**EQM**	**B**	**ET**	**EQM**	**B**	**ET**	**EQM**	**B**	**ET**	**EQM**
Y continu (MR1 et MR2)												
Pondération par probabilité inverse (calée; EEG)												
PPI	0,33	4,60	4,60	-0,19	4,50	4,50	0,96	7,10	7,10	123,79	18,10	125,10
PPIEQ1	0,33	4,50	4,50	0,38	4,60	4,60	2,61	9,70	10,00	27,84	12,50	30,50
PPIEQ2	0,97	4,20	4,30	0,11	4,50	4,50	4,78	10,70	11,70	8,24	12,50	14,90
Y binaire (MR3 et MR4)												
Naïve	1,04	0,07	1,04	-4,19	0,07	4,19	2,85	0,09	2,85	-1,51	0,10	1,52
PPV	0,11	0,99	0,99	0,10	1,00	1,00	0,03	1,43	1,43	0,02	1,37	1,37
MLG	0,02	0,34	0,34	0,02	0,33	0,33	-0,19	0,50	0,54	2,40	0,45	2,44
DR (EMV)	0,03	0,34	0,34	0,05	0,34	0,34	0,06	0,41	0,41	3,31	0,43	3,33
DR (EEG)	0,03	0,34	0,34	0,03	0,33	0,33	0,04	0,42	0,42	2,78	0,42	2,82
Pondération par probabilité inverse (EMV)												
PPI	-0,02	0,47	0,47	0,37	0,70	0,79	-0,02	0,77	0,77	2,32	1,31	2,66
PPIEQ1	-0,02	0,47	0,47	-0,54	0,60	0,81	-0,02	0,75	0,75	-0,06	0,75	0,76
PPIEQ2	-0,08	0,50	0,51	0,07	0,81	0,81	-0,14	0,79	0,80	-0,09	1,19	1,19
Pondération par probabilité inverse (calée; EEG)												
PPI	0,01	0,39	0,39	-1,86	0,23	1,87	0,03	0,61	0,61	-0,88	0,28	0,92
PPIEQ1	-0,01	0,37	0,37	-0,54	0,25	0,60	-0,02	0,58	0,58	0,09	0,52	0,52
PPIEQ2	0,03	0,35	0,35	-0,11	0,27	0,29	0,05	0,52	0,53	-0,05	0,57	0,57

Note : DR = doublement robuste; EEG = équations d'estimation généralisées; EMV = estimation du maximum de vraisemblance; EQM = erreur quadratique moyenne; ET = erreur-type; MLG = modèle linéaire généralisé; MP = modèle de probabilité; MR = modèle de résultat; PPI = pondération probabiliste inverse; PPIEQ = pondération probabiliste inverse équilibrant les quantiles; PPV = plus proche voisin.

Dans le scénario II, où le MR est linéaire et le MP est non linéaire, il est évident que les estimateurs affichent des schémas de rendement variés. Dans le cas des résultats continus, les estimateurs du PPV, du MLG et DR montrent un rendement constant et fiable, comportant peu de biais et une EQM faible. L'estimateur de l'EMV de PPI présente néanmoins une augmentation prononcée du biais et de l'EQM, ce qui souligne sa sensibilité à l'endroit des modèles de probabilité non linéaires. Les estimateurs de PPIEQ proposés affichent un rendement supérieur à celui de PPI classique, alors que l'estimateur PPIEQ2 (EEG) présente le biais le plus faible et une ET raisonnable parmi les estimateurs de PPI. Dans le cas des résultats binaires, les répercussions du MP non linéaire sont moins importantes. La majorité des estimateurs affichent des biais faibles, et le rendement des estimateurs de PPIEQ sous EEG s'avère particulièrement favorable. Il convient de souligner que PPIEQ2 (EEG) obtient l'EQM la plus faible (0,29) de tous les estimateurs de PPI, surpassant même l'estimateur classique de PPI (EEG).

Le troisième scénario présente un MR non linéaire avec un MP linéaire, ce qui pose des difficultés particulières pour les méthodes paramétriques. Dans le cas de résultats continus, l'estimateur du MLG affiche un niveau élevé de biais et une racine de l'erreur quadratique moyenne, ce qui révèle ses limites inhérentes ayant trait au captage exact des relations non linéaires. L'estimateur du PPV laisse voir une tendance à maintenir un biais faible, et ce, même s'il s'accompagne d'une ET plus marquée. Le rendement

des méthodes DR varie. Dans la méthode DR (EMV), le biais s'avère supérieur (1,72) à celui de la méthode DR (EEG) (0,96). Parmi les estimateurs de PPI, ceux de la catégorie PPIEQ affichent un meilleur rendement, le PPIEQ2 (EEG) étant celui qui obtient l'EQM la plus faible dans ce groupe. Dans le cas des résultats binaires, les répercussions du MR non linéaire sont moins importantes. La majorité des estimateurs affichent peu de biais et, là encore, les estimateurs de PPIEQ sous EEG présentant un rendement solide. Le PPIEQ2 (EEG) obtient l'EQM la plus faible (0,53) de tous les estimateurs de PPI, présentant un niveau de rendement comparable à celui des estimateurs du MLG et DR.

Dans le scénario IV, qui combine un MR et un MP non linéaires, tous les estimateurs se heurtent aux conditions les plus difficiles. Dans le cas des résultats continus, la majorité des estimateurs sont incapables d'aboutir à un rendement satisfaisant. Les estimateurs du MLG et DR montrent un niveau considérable de biais et d'erreurs. L'estimateur du PPV laisse voir une tendance à maintenir un biais faible, et ce, malgré une erreur-type plus marquée. Parmi les estimateurs de PPI, ceux reposant sur une estimation du maximum de vraisemblance (EMV) s'accompagnent de biais considérables et d'une erreur quadratique moyenne. Les estimateurs de PPIEQ proposés, auxquels s'appliquent une EMV et des EEG, procurent un rendement nettement supérieur, le PPIEQ2 affichant le biais et l'EQM les plus faibles de tous les estimateurs. Dans le cas des résultats binaires, même si le rendement général est supérieur à celui observé avec les résultats continus, des difficultés demeurent. Les estimateurs du MLG et DR présentent un biais plus important, les valeurs allant de 2,40 à 3,31. Les estimateurs de PPIEQ proposés, sous EEG, procurent là encore un rendement supérieur à celui d'autres estimateurs de PPI. Le PPIEQ2 obtient l'EQM la plus faible (0,57) de cette catégorie, qui se révèle comparable à celle de l'estimateur du PPV (EQM : 1,37). Ce scénario met en évidence l'efficacité des estimateurs de PPIEQ proposés, surtout selon l'EEG, pour traiter des relations non linéaires complexes dans des modèles de résultat et de probabilité.

Le tableau 5.3 expose les taux de couverture (TC) empirique correspondant aux estimateurs proposés. Dans le scénario I, les estimateurs proposés affichent un TC près du niveau nominal de 95 %. Dans le scénario II, ce phénomène s'avère particulièrement évident dans les EEG de PPI, où le score de propension non linéaire est employé et la spécification du modèle présumé est erronée. Le PPIEQ2 proposé (comportant des déciles et des totaux) présente néanmoins des TC près du niveau nominal, tant pour les résultats continus que pour les résultats binaires. Notamment, l'approche proposée améliore grandement le rendement de PPI classique lorsque l'EMV est utilisé comme technique d'estimation. L'IC le plus court est observé avec les estimateurs DR et des EEG de PPI, qui représente le tiers (pour les continus) ou la moitié (pour les binaires) de la largeur de l'estimateur du PPV.

Tableau 5.3
Taux de couverture empirique et longueur moyenne de l'intervalle de confiance selon l'estimation de la variance analytique

	MR	Scénario I Linéaire (1)		Scénario II Linéaire (1)		Scénario III Non linéaire (2)		Scénario IV Non linéaire (2)	
	MP	Linéaire (1)		Non linéaire (2)		Linéaire (1)		Non linéaire (2)	
Type	**Estimateur**	**Taux de couverture**	**Longueur**	**Taux de couverture**	**Longueur**	**Taux de couverture**	**Longueur**	**Taux de couverture**	**Longueur**
					Continu				
De base	PPV	95,60	24,73	94,80	24,78	94,80	59,82	96,60	59,84
	MLG	93,60	17,63	94,60	17,69	69,00	54,66	0,00	60,57
	DR (EMV)	93,60	17,69	95,40	18,40	95,40	24,00	6,80	363,34
	DR (EEG)	93,60	17,65	94,60	17,61	96,00	28,21	0,00	72,78
EMV de PPI	PPI	95,80	22,36	10,40	122,99	99,20	38,23	2,40	787,33
	PPIEQ1	94,80	21,67	98,20	47,09	100,00	45,92	83,40	92,19
	PPIEQ2	95,60	23,55	97,59	55,69	99,60	56,49	96,98	102,93
EEG de PPI	PPI	93,80	17,77	96,00	18,76	96,40	28,40	0,00	79,91
	PPIEQ1	93,80	17,69	94,80	18,35	95,40	39,57	45,60	51,73
	PPIEQ2	94,80	17,80	95,80	18,43	98,00	46,65	92,20	49,54
					Binaire				
De base	PPV	94,20	3,82	95,00	3,82	93,20	5,34	95,00	5,34
	MLG	94,80	1,36	94,60	1,35	95,00	2,11	0,20	1,91
	DR (EMV)	92,80	1,30	94,00	1,32	95,80	1,72	0,00	1,97
	DR (EEG)	92,60	1,30	94,00	1,31	94,80	1,71	0,00	1,77
EMV de PPI	PPI	94,60	1,81	98,40	3,15	95,20	3,12	84,00	5,91
	PPIEQ1	94,40	1,74	75,60	2,29	93,00	2,84	95,80	3,02
	PPIEQ2	95,20	1,90	96,39	3,57	96,60	3,05	92,77	5,30
EEG de PPI	PPI	93,20	1,46	0,00	0,95	94,80	2,43	18,00	1,22
	PPIEQ1	94,20	1,43	49,60	1,08	96,20	2,37	96,40	2,24
	PPIEQ2	94,20	1,44	93,80	1,07	96,80	2,40	96,40	2,36

Note : DR = doublement robuste; EEG = équations d'estimation généralisées; EMV = estimation du maximum de vraisemblance; MLG = modèle linéaire généralisé; MP = modèle de probabilité; MR = modèle de résultat; PPI = pondération probabiliste inverse; PPIEQ = pondération probabiliste inverse équilibrant les quantiles; PPV = plus proche voisin.

Dans le scénario III, les approches proposées ont donné des résultats qui se rapprochent des taux nominaux qui ont affiché un excellent rendement. Enfin, en ce qui concerne le scénario IV, on peut constater que quand la spécification des deux modèles est erronée, seuls ceux qui comprennent des déciles et des totaux (accompagnés du PPV) dans le cas des données continues et de PPIEQ1 (selon l'EMV et l'EEG) pour les données binaires présentent un taux de couverture près de la valeur nominale de 95 %. Dans la section B de l'annexe, nous présentons une analyse détaillée de la qualité des totaux et des quantiles reproduits pour les méthodes proposées.

L'étude par simulation illustre les avantages considérables de l'intégration de quantiles aux estimateurs de PPI. Cette approche rehausse nettement la robustesse et l'exactitude des estimations, surtout dans les cas où la spécification des modèles de résultat et de probabilité est erronée. Les estimateurs de PPIEQ affichent un rendement supérieur dans des scénarios non linéaires complexes qui comportent des résultats continus et binaires, alors que les versions qui reposent sur des EEG surpassent habituellement leurs EMV équivalentes. L'étendue de l'IC semble comparable aux estimateurs du MLG et DR, ce qui laisse supposer que l'utilisation d'un plus grand nombre de contraintes ne ferait pas augmenter la variance. L'approche proposée procure également un IC plus étroit que l'estimateur du PPV. Par conséquent, les chercheurs procédant à

l'analyse de variables continues ou chiffrées issues de multiples sources devraient songer à employer des méthodes de PPI fondées sur les quantiles afin d'accroître la robustesse de leurs résultats, surtout lorsqu'ils doivent composer avec des spécifications de modèle incertaines ou des relations complexes dans l'analyse de données concrètes.

6. Application sur des données réelles

6.1 Description des données

Dans la présente section, nous exposons une tentative d'intégration de données administratives et d'enquête sur les postes vacants à la fin du deuxième trimestre de 2022 en Pologne. L'objectif consistait à estimer la proportion de postes vacants destinés à des travailleurs ukrainiens. Après l'invasion de l'Ukraine par la Russie, le 24 février 2022, environ 3,5 millions de personnes (principalement des femmes et des enfants) sont arrivées en Pologne entre le 24 février et la mi-mai 2022 (Duszczyk et Kaczmarczyk, 2022). Certaines d'entre elles se sont dirigées vers d'autres pays européens, mais près de 1 million de personnes sont restées en Pologne (en date de 2023, voir Statistics Poland (2023)).

La première source que nous avons utilisée est l'enquête sur les postes vacants (EPV, appelée Labour Demand Survey en Pologne), qui correspond à un échantillon stratifié de 100 000 unités, comportant un taux de réponse d'environ 60 % (S_B). La population observée se compose d'entreprises et de leurs unités locales comprenant 1 employé ou plus. La base de sondage renferme des renseignements selon le code de la Nomenclature statistique des activités économiques dans les communautés européennes (NACE) (19 niveaux), la région (16 niveaux), le secteur (2 niveaux), la taille (3 niveaux) et le nombre d'employés selon les données administratives intégrées par Statistics Poland. L'échantillon de l'EPV est composé de 304 strates créées séparément pour les entreprises qui comptent jusqu'à 9 employés et celles qui en comptent plus de 10 (voir Statistics Poland, 2021).

L'enquête permet de déterminer si une entreprise a des postes vacants à la fin du trimestre (au 30 juin 2022) selon la définition suivante : *les postes vacants sont des postes ou des emplois inoccupés à la suite d'un roulement de personnel ou de postes nouvellement créés qui répondent simultanément aux trois conditions suivantes :* 1) le poste était réellement vacant le jour de l'enquête; 2) l'employeur a fait des efforts pour trouver une personne prête à occuper le poste; 3) si un candidat qualifié a été trouvé pour pourvoir le poste, l'employeur serait disposé à l'embaucher. De plus, l'EPV restreint la définition de postes vacants en excluant les stages, les contrats de mandat, les contrats liés à des travaux particuliers et les contrats interentreprises. Des 60 000 unités déclarantes, environ 7 000 ont déclaré avoir au moins un poste vacant. Notre population cible comprenait des unités ayant au moins un poste vacant, ce qui correspondait à un nombre situé entre 38 000 et 43 000 à la fin du deuxième trimestre de 2022, selon l'enquête.

La deuxième source est la base de données centrale des offres d'emplois (Central Job Offers Database (CBOP)), qui consiste en un registre de tous les postes vacants soumis aux services publics de placement (PEO – S_A). La CBOP est accessible en ligne et à l'aide d'un programme (par l'entremise de l'API). Puisqu'elle renferme tous les types de contrats et d'emplois à l'extérieur de la Pologne, il a fallu nettoyer les données pour qu'elles concordent avec la définition de poste vacant utilisée dans le cadre de l'EPV. Les

données de la CBOP recueillies grâce à l'API comprennent des renseignements indiquant si le poste est toujours vacant (par exemple 17 % des postes n'étaient plus vacants au moment du téléchargement, à la fin du deuxième trimestre de 2022). La CBOP contient également des renseignements sur les identificateurs d'unité (REGON et NIP), ce qui nous a permis de coupler les unités à la base de sondage en vue d'obtenir des variables auxiliaires ayant les mêmes définitions que dans l'enquête (24 % des enregistrements n'avaient pas d'identificateur parce que l'employeur pouvait retenir ce renseignement). L'ensemble de données de la CBOP qui en a résulté contenait environ 8 500 unités comprises dans la base de sondage.

Le chevauchement entre l'EPV et la CBOP était d'environ 2 600 entités (environ 4 % de l'échantillon de l'EPV et 30 % de la CBOP), mais seulement 30 % de ces entités avaient déclaré avoir au moins un poste vacant dans le cadre de l'EPV. Ce résultat laisse entendre une sous-déclaration importante dans l'EPV. Or, ce problème est hors de la portée du présent article. Pour l'étude empirique, nous avons traité les deux sources séparément. Leur traitement approprié avec les méthodes proposées fera l'objet d'une étude ultérieure.

6.2 Analyse des résultats

Nous avons défini notre variable cible comme suit : Y représente *la proportion d'annonces de postes vacants qui ont été traduites en ukrainien* (désigné par UKR), calculée séparément pour chaque unité. Après l'invasion de l'Ukraine par la Russie, de nombreux services de publication d'offres d'emploi en ligne ont ajouté aux annonces des renseignements précisant que l'employeur souhaiterait particulièrement embaucher des Ukrainiens (par exemple une version ukrainienne du site Internet était disponible et les offres d'emploi comprenaient le commentaire suivant : « Nous invitons les Ukrainiens à postuler »).

Le tableau 6.1 montre la répartition des emplois enregistrés (EE) selon l'EPV, pour les entreprises figurant dans la CBOP et les entreprises prêtes à embaucher des Ukrainiens (ayant au moins une annonce de poste vacant traduite en ukrainien, désigné par CBOP-UKR). Les entreprises figurant dans la CBOP sont, en moyenne, de plus grande taille que celles qui répondent à l'EPV. Le nombre médian d'employés pour les entreprises ayant participé à l'EPV était de 8, alors qu'il était de 20 et de 18 pour celles de la CBOP et de la CBOP-UKR, respectivement. Cela laisse supposer que la sélection des entreprises dans la CBOP repose sur leur taille.

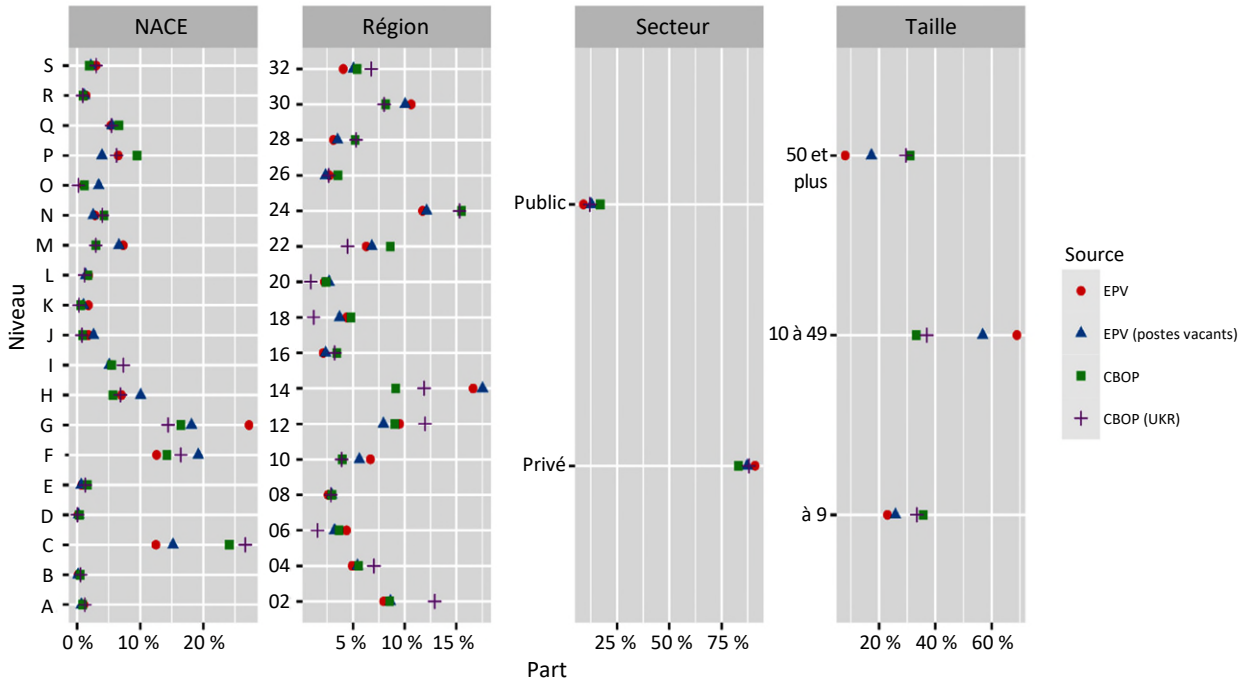
Tableau 6.1
Quantiles des emplois enregistrés selon l'enquête sur les postes vacants, la base de données centrale des offres d'emplois et la variable cible UKR à la fin du deuxième trimestre de 2022

Décile des EE	EPV	CBOP	CBOP-UKR
10 %	3	4	4
20 %	4	6	5
30 %	5	8	8
40 %	6	13	11
50 %	8	20	18
60 %	10	32	28
70 %	17	51	48
80 %	37	90	91
90 %	100	211	215

Note : CBOP = Central Job Offers Database; EE = emplois enregistrés; EPV = Enquête sur les postes vacants; UKR = ukrainien.

La figure 6.1 montre la répartition des quatre variables catégoriques dans les deux sources. Les plus vastes écarts observés entre les sources concernent la taille de l'entreprise, où les parts sont presque égales dans la CBOP. Des différences sont observées entre certaines régions (par exemple 14 Mazowieckie avec Varsovie, la capitale de la Pologne; 02 Dolnośląskie avec Wrocław, et 12 Małopolskie avec Cracovie). Des régions de l'est de la Pologne (06 Lubelskie, 18 Podkarpackie, 20 Podlaskie), mais aussi Pomorskie (22) avec Tricity, se caractérisent par la plus faible proportion de postes vacants destinés à des Ukrainiens. Les plus vastes écarts en ce qui a trait à la NACE se trouvent dans la fabrication (C), le commerce de gros et de détail (G), et les hôtels et restaurants (I).

Figure 6.1 Répartition des quatre variables auxiliaires dans les sources, à la fin du deuxième trimestre de 2022. EPV : échantillon complet; EPV (postes vacants) : entreprises ayant au moins un poste vacant; CBOP : ensemble de données du registre; CBOP-UKR : unités ayant au moins une annonce de poste vacant traduite en ukrainien



Note : EPV : Enquête sur les postes vacants; CBOP = Central Job Offers Database; UKR = Ukrainien.

La figure 6.2 montre la proportion de postes vacants qui visent les Ukrainiens selon quatre variables catégoriques et les déciles des emplois enregistrés (EE) en fonction des déciles estimés à partir de l'EPV. En règle générale, les parts sont de l'ordre de 10 % à 30 % environ, surtout dans les régions. Les parts des déciles des EE varient de 17 % pour les unités les plus importantes à 26 % pour les EE médians (8 employés).

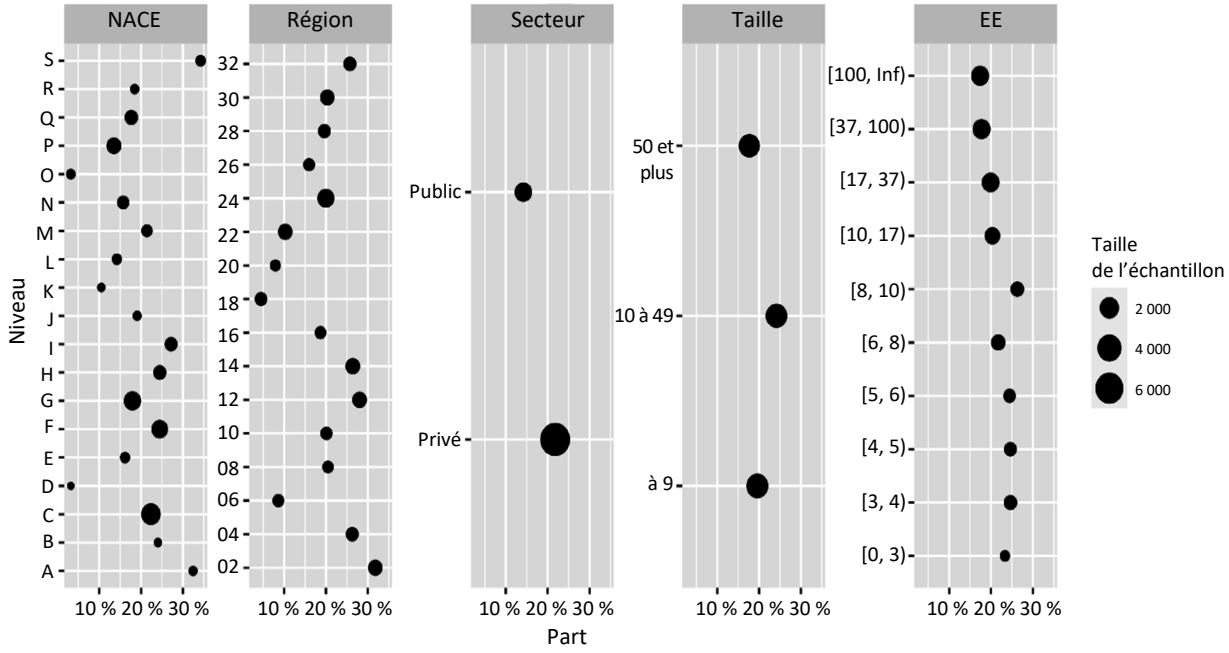
L'examen a porté sur les combinaisons de variables suivantes :

- Ensemble 0 : Région (16 niveaux), NACE (19 niveaux), secteur (2 niveaux), taille (3 niveaux), log(EE), log (nombre de postes vacants), I (nombre de postes vacants = 1) (si l'employeur recherche une seule personne);

- Ensemble 1A : Ensemble 0 (sans log(EE)) + quartiles des EE (estimés à partir de l'EPV);
- Ensemble 1B : Ensemble 0 + quartiles des EE;
- Ensemble 2A : Ensemble 0 (sans log (EE)) + déciles des EE (sauf 10 %);
- Ensemble 2B : Ensemble 0 + déciles des EE (sauf 10 %).

Nous avons décidé de ne pas inclure le nombre de postes vacants pour l'équilibrage des quantiles, puisque près de 45 % de ces postes correspondaient à 1 pour la CBOP et à près de 30 % pour l'EPV. Nous avons examiné les estimateurs suivants, en appliquant une régression linéaire et logistique : MLG avec l'ensemble 0, PPV avec l'ensemble 0 (indiqué en tant que PPV1), PPV avec log (EE) et log (nombre de postes vacants) seulement (indiqué en tant que PPV2), EEG DR avec l'ensemble 0, estimateurs de PPI selon l'EMV et l'EEG joints aux ensembles suivants : PPI avec l'ensemble 0, PPI avec l'ensemble 1A (PPIEQ1A), PPI avec l'ensemble 1B (PPIEQ1B), PPI avec l'ensemble 2A (PPIEQ2A) et PPI avec l'ensemble 2B (PPIEQ2B).

Figure 6.2 Variable cible (proportion d'offres d'emploi destinées aux Ukrainiens), selon cinq variables auxiliaires, à la fin du deuxième trimestre de 2022



La méthode bootstrap suivante a servi à estimer la variance : 1) une approche bootstrap modifiée a permis de rééchantillonner l'échantillon de l'EPV; 2) un échantillonnage aléatoire simple avec remise a servi à rééchantillonner la CBOP. Ce processus a été répété $B = 500$ fois. Le tableau 6.2 montre des estimations ponctuelles (désignées par des points), des erreurs-types bootstrap (désignées par ET), des coefficients de variation (c.v.) et des intervalles de confiance de 95 %.

L'estimateur naïf montre une proportion de 20,51 %, là où d'autres méthodes donnent lieu à des estimations légèrement plus élevées. Les estimateurs du MLG et des EEG DR affichent des résultats semblables (22,68 % et 21,75 %, respectivement), accompagnés de faibles erreurs-types. La variabilité des estimateurs du PPV se révèle plus importante, le PPV1 ayant l'estimation ponctuelle la plus élevée (25,88 %), mais aussi l'erreur-type la plus importante. Les estimateurs de PPI et de PPIEQ, fondés sur l'EMV et l'EEG, donnent des résultats allant de 21,53 % à 22,88 %. Habituellement, les méthodes fondées sur des EEG comportent des erreurs-types plus modestes et des intervalles de confiance plus étroits que leurs EMV équivalentes. Dans l'ensemble, la majorité des méthodes indiquent que la proportion des postes vacants destinés à des Ukrainiens se situe entre 21 % et 23 %, les estimateurs fondés sur des EEG présentant les estimations les plus précises.

Tableau 6.2
Estimations de la proportion de postes vacants destinés aux Ukrainiens à la fin du deuxième trimestre de 2022 en Pologne

Estimateur	Point	ET	c.v.	2,5 %	97,5 %
Naïf					
Naïf	20,51	–	–	–	–
Estimateurs					
MLG	22,68	0,59	2,61	21,52	23,84
PPV1	25,88	6,41	24,78	13,31	38,45
PPV2	23,38	1,91	8,18	19,63	27,13
DR (EEG)	21,75	0,60	2,78	20,57	22,94
Pondération par probabilité inverse (EMV)					
PPI	22,88	0,80	3,52	21,31	24,46
PPIEQ1A	22,24	0,76	3,43	20,74	23,73
PPIEQ1B	22,88	0,80	3,52	21,31	24,46
PPIEQ2A	22,26	0,78	3,48	20,75	23,78
PPIEQ2B	22,88	0,81	3,53	21,30	24,46
Pondération par probabilité inverse (EMV)					
PPI	21,75	0,60	2,78	20,57	22,94
PPIEQ1A	21,53	0,60	2,78	20,36	22,70
PPIEQ1B	21,69	0,60	2,78	20,51	22,87
PPIEQ2A	21,69	0,61	2,83	20,49	22,89
PPIEQ2B	21,74	0,61	2,83	20,53	22,94

Note : c.v. = coefficients de variation; DR = doublement robuste; EEG = équations d'estimation généralisées; EMV = estimation du maximum de vraisemblance; ET = erreur-type; MLG = modèle linéaire généralisé; PPI = pondération probabiliste inverse; PPIEQ = pondération probabiliste inverse équilibrant les quantiles; PPV = plus proche voisin.

7. Résumé

Nous avons assisté, ces dernières années, à une révolution dans le domaine de la recherche par enquête en ce qui a trait à l'utilisation et à l'intégration de nouvelles sources de données pour l'inférence statistique. Les efforts déployés dans cette discipline ont porté sur l'intégration de données d'enquête et sur l'inférence reposant sur des échantillons non probabilistes, avec des applications dans les statistiques officielles (Salvatore, 2023). Cette situation s'explique principalement par le fait que les échantillons probabilistes ont tendance à se révéler très coûteux et sont associés au fardeau croissant du répondant. D'autre part, l'inférence statistique fondée sur des échantillons non probabilistes constitue un défi, principalement en raison des biais inhérents au processus d'estimation. Il est par conséquent nécessaire d'élaborer de nouvelles

méthodes d'inférence statistique fondée sur des échantillons non probabilistes, qui pourraient être utiles dans des applications concrètes (Wu, 2022).

Dans le présent article, nous avons décrit un estimateur de pondération par probabilité inverse équilibrant les quantiles pour les échantillons non probabilistes. D'après notre approche, il est possible, pour certaines des méthodes examinées dans l'article (PPI et DR), de reproduire non seulement des totaux pour un ensemble de variables auxiliaires, mais aussi pour un ensemble de quantiles (ou de quantiles estimés). Ce genre de solution accroît la robustesse par rapport à une spécification erronée du modèle, favorise une réduction des biais et rehausse l'efficacité de l'estimation. Ces gains sont confirmés par notre étude par simulation, grâce à laquelle nous avons montré que l'inclusion de quantiles aux estimateurs de pondération par probabilité inverse procure des estimations de meilleure qualité.

Les estimateurs proposés dans le présent article comportent des possibilités pour des applications concrètes, à condition de disposer de variables continues de qualité suffisante pour les utiliser en tant que variables auxiliaires. L'exemple de notre étude empirique, dans laquelle nous estimons la proportion d'emplois vacants destinés à des travailleurs ukrainiens, montre l'efficacité des méthodes proposées pour générer des estimations de grande précision.

Pour ce qui est des éléments non explorés dans le présent article, une des questions qui mériteraient un examen plus approfondi est le problème de la sous-couverture dans les échantillons non probabilistes, lié à la violation de l'hypothèse de positivité (A2). L'approche proposée dans l'article correspond aux répartitions de variables des échantillons probabilistes (ou la population) et des échantillons non probabilistes, et peut servir de solution de rechange à d'autres approches proposées dans la littérature qui consistent, par exemple, à utiliser un échantillon additionnel de personnes de la population non couvert par l'échantillon non probabiliste (Chen, Li et Wu, 2023).

Remerciements

Le travail des auteurs a été financé par le National Science Centre de la Pologne, OPUS 20, subvention nº 2020/39/B/HS4/00941.

Nous remercions le rédacteur en chef adjoint et les examinateurs pour leurs commentaires qui nous ont permis d'améliorer l'article. Les auteurs tiennent en outre à remercier Łukasz Chrostowski, qui a créé le paquet `nonprobsvy` dans lequel sont intégrés des estimateurs de pointe proposés dans la littérature. L'ensemble est accessible dans `CRAN` et à l'adresse : https://github.com/ncn-foreigners/nonprobsvy. Nous remercions également Natalie Shlomo et Tymon Świtalski pour leurs précieux commentaires sur la version provisoire du présent article.

Les codes permettant de reproduire l'étude par simulation sont offerts gratuitement dans le répertoire GitHub : https://github.com/ncn-foreigners/paper-nonprob-qcal. Un paquet R qui met en œuvre un calage conjoint est accessible à l'adresse : https://github.com/ncn-foreigners/jointCalib. Le paquet repose sur le calage mis en place dans les paquets `survey`, `sampling` ou `laeken`.

Annexe

A. Hypothèses qui garantissent une solution unique aux équations d'estimation

Dans l'annexe, nous faisons appel à la notation $\mathbf{z}_k = \begin{pmatrix} \mathbf{x}_k \\ \mathbf{a}_k \end{pmatrix}$.

(B1) La matrice $\sum_{k \in S_A} \mathbf{z}_k \mathbf{z}_k^T$ est définie positive.

(B2) La matrice $\sum_{k \in S_B} \mathbf{z}_k \mathbf{z}_k^T$ est définie positive.

Dans l'annexe, nous utilisons la notation suivante :

$$\begin{aligned} \pi(\mathbf{z}; \boldsymbol{\eta}) &:= \frac{\exp(\mathbf{z}^T \boldsymbol{\eta})}{1 + \exp(\mathbf{z}^T \boldsymbol{\eta})}, \\ \pi(t) &:= \frac{\exp(t)}{1 + \exp(t)}, \\ \pi'(\mathbf{z}; \boldsymbol{\eta}) &:= \pi(\mathbf{z}; \boldsymbol{\eta}) \big(1 - \pi(\mathbf{z}; \boldsymbol{\eta})\big). \end{aligned}$$

Théorème 1. *L'hypothèse (B1) constitue une condition nécessaire et suffisante de l'existence d'une solution unique à :*

$$\boldsymbol{G}(\boldsymbol{\eta}) = \sum_{k \in S_A} \frac{\boldsymbol{z}_k}{\pi(\boldsymbol{z}_k; \boldsymbol{\eta})} - \sum_{k \in S_B} d_k^B \boldsymbol{z}_k = \boldsymbol{0},$$

et l'hypothèse (B2), une condition nécessaire et suffisante de l'existence d'une solution unique à :

$$\boldsymbol{U}(\boldsymbol{\eta}) = \sum_{k \in S_A} \boldsymbol{z}_k - \sum_{k \in S_B} d_k^B \pi(\boldsymbol{z}_k; \boldsymbol{\eta}) \boldsymbol{z}_k = \boldsymbol{0}.$$

Le processus d'analyse des données permet de vérifier aisément les conditions (B1) et (B2). Il est possible de remédier à la violation de l'une ou de l'autre en réduisant le nombre de quantiles utilisés. Il convient de mentionner que (B1) correspond à $n_A \geq J$ où J est la dimension de $\mathbf{z}$ et qu'il y a au moins J $\mathbf{z}$ linéairement indépendants dans S_A. (B2) correspond à des propriétés analogues de S_B.

Preuve. Nous commencerons par prouver que ℓ^* (l'antidérivée de $\mathbf{U}$) a un maximum global unique. Notons que si (B2) renferme la matrice hessienne de ℓ^* :

$$\frac{\partial \ell^*}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T}(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \mathbf{U}(\boldsymbol{\eta}) = -\sum_{k \in S_B} d_k^B \pi'(\mathbf{z}; \boldsymbol{\eta}) \mathbf{z}_k \mathbf{z}_k^T,$$

est définie positive pour tout $\boldsymbol{\eta} \in \mathbb{R}^J$, de sorte que ℓ^* est strictement concave. Par conséquent, elle a tout au plus un maximum global et, puisque la matrice hessienne est toujours inversible, le second test dérivé nous révèle que $\mathbf{U}(\boldsymbol{\eta})$ a tout au plus une racine.

Il convient de mentionner que (B2) suppose, selon le théorème de nullité des rangs, que nous ayons :

$$\text{Span}(\mathbf{z}_1, \ldots, \mathbf{z}_{n_A})^\perp = (\mathbb{R}^J)^\perp = \{\mathbf{0}\},$$

où $(\cdot)^\perp$ correspond au composant orthogonal qui, en retour, nous donne :

$$\exists j \in S_A : |\mathbf{z}_j^T \boldsymbol{\eta}_m| \to \infty, \tag{A.1}$$

de sorte que π_j^A tend vers 0 ou 1. Cela suppose que la fonction PL :

$$\ell^*(\boldsymbol{\eta}) = \sum_{k \in S_A} \left(\log(\pi_k^A) - \log(1-\pi_k^A) \right) + \sum_{k \in S_B} d_k^B \log(1 - \pi_k^A),$$

qui a une borne supérieure, doit tendre vers $-\infty$.

Il en va de même pour un raisonnement analogique concernant l'antidérivée de $\mathbf{G}$, par exemple :

$$H(\boldsymbol{\eta}) = \sum_{k \in S_A} \int_q^{\boldsymbol{\eta}^T \mathbf{z}_k} \frac{1}{\pi(t)}\, dt - \sum_{k \in S_B} d_k^B \boldsymbol{\eta}^T \mathbf{z}_k,$$

pour certains $q \in \mathbb{R}$, dont la matrice hessienne est :

$$\frac{\partial}{\partial \boldsymbol{\eta}} \mathbf{G}(\boldsymbol{\eta}) = -\sum_{k \in S_A} \frac{\mathbf{z}_k \mathbf{z}_k^T}{\pi(\mathbf{z}_k; \boldsymbol{\eta})^2} \pi'(\mathbf{z}_k; \boldsymbol{\eta}),$$

qui assure que $\mathbf{G}$ a tout au plus une racine. Pour des raisons qui deviendront plus évidentes en fin de compte, nous optons pour $q < 0$, de sorte que :

$$\left\| \sum_{k \in S_B} d_k^B \mathbf{z}_k - \sum_{k \in S_A} \mathbf{z}_k \right\| < n_A (e^{-q} - q),$$

qui est possible puisque $q \mapsto e^{-q} - q$ est une bijection continue allant de $\mathbb{R}$ à $\mathbb{R}$.

La concavité stricte de H suppose une condition suffisante de l'existence d'une racine de $\mathbf{G}$, qui est :

$$\lim_{\|\boldsymbol{\eta}\| \to \infty} H(\boldsymbol{\eta}) = -\infty, \tag{A.2}$$

où $\|\cdot\|$ correspond à la norme euclidienne $\boldsymbol{\zeta} \overset{\|\cdot\|}{\mapsto} \sqrt{|\boldsymbol{\zeta}^T \boldsymbol{\zeta}|}$. Pour prouver (A.2), il suffit de prouver que $\lim_{m \to \infty} |H(\boldsymbol{\eta}_m)| = \infty$ pour des séquences $(\boldsymbol{\eta}_m)_{m \in \mathbb{N}} \subseteq \mathbb{R}^J$ avec $\lim_{m \to \infty} \|\boldsymbol{\eta}_m\| = \infty$, de sorte que la séquence $\operatorname{sgn}(\boldsymbol{\eta}_m) = \frac{\boldsymbol{\eta}_m}{\|\boldsymbol{\eta}_m\|}$ est constante.

Pour confirmer la véracité de la déclaration précédente, observons d'abord que pour tout vecteur $\boldsymbol{\eta}$ et $s \geq 1$:

$$\begin{aligned}
H(s\boldsymbol{\eta}) - sH(\boldsymbol{\eta}) &= \sum_{k \in S_A} \left(\int_q^{s\boldsymbol{\eta}^T \mathbf{z}_k} \frac{1}{\pi(t)}\, dt - \int_q^{\boldsymbol{\eta}^T \mathbf{z}_k} \frac{s}{\pi(t)}\, dt \right) \\
&\leq \sum_{k \in S_A} \left(\int_q^{\boldsymbol{\eta}^T \mathbf{z}_k} \frac{s}{\pi(st)}\, dt + \int_{q/s}^{q} \frac{s}{\pi(st)}\, dt - \int_q^{\boldsymbol{\eta}^T \mathbf{z}_k} \frac{s}{\pi(t)}\, dt \right) \\
&= \sum_{k \in S_A} \int_q^{\boldsymbol{\eta}^T \mathbf{z}_k} \frac{e^{-s} - 1}{e^t}\, dt + n_A \int_{q/s}^{q} \frac{s}{\pi(st)}\, dt \\
&= \sum_{k \in S_A} \left(e^{-s} - 1 \right) \left(-e^{-\boldsymbol{\eta}^T \mathbf{z}_k} + e^{-q} \right) + n_A \int_{q/s}^{q} \frac{s}{\pi(st)}\, dt \\
&= \left(1 - e^{-s} \right) \left(\sum_{k \in S_A} e^{-\boldsymbol{\eta}^T \mathbf{z}_k} - n_A e^{-q} \right) + n_A \left(sq - q + e^{-q} - e^{-qs} \right),
\end{aligned}$$

où la deuxième ligne suit, par substitution de $us = t, sdu = dt$, et le dernier terme a une borne supérieure dans s. Grâce à un bornage local de H, il existe un R constant non négatif, de sorte que :

$$\forall s > 0 \colon H\left(s\boldsymbol{\eta}\right) \leq sH\left(\boldsymbol{\eta}\right) + R.$$

Ainsi, nous avons pour toute séquence de vecteurs normatifs $\left(\boldsymbol{\omega}_m\right)_{m\in\mathbb{N}} \subseteq \mathbb{R}^J$:

$$\begin{aligned} \sup\{t > 0 \colon H(t\boldsymbol{\omega}_m) > -M\} &\leq \sup\{t > 0 \colon tH(\boldsymbol{\omega}_m) > -M - R\} \\ &\leq \frac{-M-R}{H(\boldsymbol{\omega}_m)} \to \infty \Leftrightarrow H(\boldsymbol{\omega}_m) \to 0. \end{aligned}$$

Cependant, par continuité, la dernière suppose que pour chaque $\varepsilon > 0,$ il existe $\mathbf{u}$ avec $\|\mathbf{u}\| = 1$ et $H(\mathbf{u}) > -\varepsilon,$ ce qui est impossible puisque :

$$\begin{aligned} \sum_{k\in S_B} d_k^B \mathbf{u}^T \mathbf{z}_k - \varepsilon &> \sum_{k\in S_A} \int_q^{\mathbf{u}^T\mathbf{z}_k} \frac{1}{\pi(t)}\, dt = n_A(e^{-q} - q) + \sum_{k\in S_A} (\mathbf{u}^T\mathbf{z}_k + e^{-\mathbf{u}^T\mathbf{z}_k}) \\ \Rightarrow \left\| \sum_{k\in S_B} d_k^B \mathbf{z}_k - \sum_{k\in S_A} \mathbf{u}^T\mathbf{z}_k \right\| &\geq \left| \mathbf{u}^T \left(\sum_{k\in S_B} d_k^B \mathbf{z}_k - \sum_{k\in S_A} \mathbf{u}^T\mathbf{z}_k \right) \right| \\ &\geq n_A(e^{-q} - q) + \sum_{k\in S_A} e^{-\mathbf{u}^T\mathbf{z}_k} + \varepsilon, \end{aligned}$$

ce qui est impossible selon la nature du choix constant de $q.$ Par conséquent, nous avons :

$$\forall M > 0 \colon \sup_{\|\boldsymbol{\omega}\|=1} \inf\{t > 0 \colon \forall s > t \colon H(s\boldsymbol{\omega}) < -M\} = \sup_{\|\boldsymbol{\omega}\|=1} \sup\{t > 0 \colon H(t\boldsymbol{\omega}) \geq -M\} < \infty.$$

Supposons maintenant que :

$$\forall M > 0, \|\boldsymbol{\omega}\| = 1 \; \exists N > 0 \; \forall r > N \colon H(r\boldsymbol{\omega}) < -M.$$

Choisissons toute valeur $M > 0$ et supposons que :

$$T := \sup_{\|\boldsymbol{\omega}\|=1} \sup\{t > 0 \colon H(t\boldsymbol{\omega}) \geq -M\} < \infty.$$

Maintenant, pour toute valeur $r > T,$ nous avons :

$$\forall \|z\| = 1 \colon H(rz) < -M.$$

Leur mise en commun nous donne :

$$\forall M > 0 \; \exists T > 0 \; \forall r > T \; \forall \|\boldsymbol{\omega}\| = 1 \colon H(rz) < -M.$$

Cela suffit pour vérifier que :

$$H(\boldsymbol{\eta}_m) \to \infty,$$

pour les séquences de $\boldsymbol{\eta}_m$ accompagnées d'un signe constant et croissant à l'infini dans la norme, comme il était affirmé précédemment.

Nous poursuivons maintenant pour prouver l'affirmation centrale concernant $H.$ Nous avons déjà prouvé que :

$$\forall \boldsymbol{\eta} \in \mathbb{R}^J \; \exists R > 0 \; \forall s > 0 \colon H\left(s\boldsymbol{\eta}\right) \leq sH\left(\boldsymbol{\eta}\right) + R,$$

et choisi q, de sorte que :

$$\forall \|\mathbf{z}\| = 1\colon\ H(\mathbf{z}) < 0.$$

La convergence vers $-\infty$ des séquences $\left(\boldsymbol{\eta}_m\right)_{m\in\mathbb{N}}$ définies plus tôt est obtenue simplement.

Il est facile de constater la nécessité de (B2) pour l'existence d'une solution unique à $\mathbf{G} = \mathbf{0}$. Si on enfreint alors (B2), puisque la matrice dans (B2) est nettement définie non négative, nous pouvons prendre quelques $\boldsymbol{\xi}$, de sorte que :

$$\mathbf{0} \neq \boldsymbol{\xi} \in \text{Span}\left(\mathbf{z}_1,\ldots,\mathbf{z}_{n_A}\right)^{\perp}.$$

Maintenant, si $\mathbf{G} = \mathbf{0}$ a une solution $\boldsymbol{\eta}$, alors :

$$\mathbf{0} = \mathbf{G}(\boldsymbol{\eta}) = \sum_{k\in S_A} \frac{\mathbf{z}_k}{\pi\left(\boldsymbol{\eta}^T\mathbf{z}_k\right)} - \sum_{k\in S_B} d_k^B\mathbf{z}_k = \sum_{k\in S_A} \frac{\mathbf{z}_k}{\pi\left(\boldsymbol{\eta}^T\mathbf{z}_k + \boldsymbol{\xi}^T\mathbf{z}_k\right)} - \sum_{k\in S_B} d_k^B\mathbf{z}_k = \mathbf{G}(\boldsymbol{\eta}+\boldsymbol{\xi}).$$

Ainsi, $\boldsymbol{\xi}+\boldsymbol{\eta}$ constitue aussi une solution à $\mathbf{G} = \mathbf{0}$, ce qui fait en sorte que l'espace de solutions de l'EEG est d'une dimension égale en tous points à $J - \text{Rank}\left(\sum_{k\in S_A}\mathbf{z}_k\mathbf{z}_k^T\right)$ si ce genre de solution existe. Un argument analogique montre que l'hypothèse (B1) n'est pas seulement suffisante, mais qu'elle est également nécessaire pour l'existence d'une solution unique à $\mathbf{U} = \mathbf{0}$.

B. Qualité des totaux et des quantiles reproduits

La présente section fournit les résultats de l'évaluation de la qualité de la reproduction des totaux et des quantiles pour tous les estimateurs de PPI employés dans l'étude par simulation. Pour les quantités suivantes et chaque ronde de simulation $(r = 1,\ldots,500)$, le calcul de la norme euclidienne de la différence entre elles et leurs estimations, respectivement, s'est fait comme suit :

- Taille de population :

$$\nu_{r,\hat{N}} = |\hat{N}_r - N|.$$

- Quantiles (pour x_1 et x_2 ensemble, en examinant les quartiles et les déciles, où $\alpha \in$ {0,25; 0,50; 0,75; 0,1; …, 0,9}) :

$$\nu_{r,\hat{Q}} = \sqrt{\sum_{\alpha}\sum_{p=1}^{2}\left(\hat{Q}_{r,x_p,\alpha} - Q_{r,x_p,\alpha}\right)^2}.$$

- Totaux (pour x_1 et x_2 ensemble) :

$$\nu_{r,\hat{\tau}} = \sqrt{\sum_{p=1}^{2}\left(\hat{\tau}_{r,x_p} - \tau_{r,x_p}\right)^2}.$$

Les valeurs de référence pour $\hat{\tau}_{x_p,r}$ et $Q_{r,x_p,\alpha}$ sont estimées à partir de l'échantillon probabiliste S_B. Elles pourraient varier au cours de la simulation, puisque nous n'avons pas calé l'échantillon S_B.

Le tableau B.1 présente les valeurs moyennes et les valeurs médianes de $\nu_{r,\hat{N}}$, $\nu_{r,\hat{\tau}}$ et $\nu_{r,\hat{Q}}$. L'inclusion de ces deux mesures vise à tenir compte des cas où l'algorithme ne converge pas, ce qui entraîne une reproduction imparfaite des totaux et des quantiles pour l'estimateur des EEG.

Comme le laissait prévoir le plan de sondage, les estimateurs fondés sur des EEG affichent un rendement supérieur à celui de méthodes fondées sur l'EMV, et ce, pour toutes les mesures. Il convient de souligner que les méthodes fondées sur des EEG affichent un rendement supérieur en ce qui a trait à la reproduction des quantiles, la PPIEQ2 ayant obtenu les plus faibles taux d'erreur. Les méthodes PPIEQ1 et PPIEQ2 proposées font preuve d'un rendement exceptionnel, surtout dans le contexte du scénario MP2 qui est le plus exigeant. Cela s'explique par leur plan de sondage, qui tient compte des quantiles et des totaux. La PPIEQ1 tient compte des quartiles, tandis que la PPIEQ utilise des déciles. Notre interprétation porte essentiellement sur les résultats *médians*.

Les résultats montrent un écart marqué entre les scénarios MP1 et MP2 en ce qui a trait au rendement, tous les estimateurs affichant de meilleurs résultats dans MP1. Cet écart est évident non seulement dans la reproduction de la taille de la population et des totaux, mais aussi dans la reproduction des quantiles. Par exemple, dans MP1, la majorité des méthodes présentent des erreurs moyennes des quantiles inférieures à 0,2, tandis que dans MP2, ces erreurs dépassent souvent 1,0 dans les méthodes fondées sur l'EMV. Les méthodes PPIEQ1 et PPIEQ2 proposées, fondées sur des EEG, font néanmoins preuve d'une résilience remarquable en matière d'estimation des quantiles, et ce, même dans MP2 (avec des erreurs moyennes des quantiles d'aussi peu que 0,05 dans PPIEQ2), ce qui illustre bien leur capacité à capter la répartition d'une population à l'aide de divers modèles de probabilité. Le rendement constant en matière de reproduction des quantiles, auquel s'ajoute l'estimation exacte de la taille et des totaux de population, sert à rehausser l'efficacité des méthodes proposées dans l'incorporation de renseignements sur les quantiles et les totaux dans une estimation de PPI. Le rendement supérieur de PPIEQ1 et de PPIEQ2 dans MP2 permet de démontrer leur capacité à gérer des modèles probabilistes complexes en tirant parti d'un ensemble complet de caractéristiques de la population.

Tableau B.1
Qualité de la taille de population, des totaux et des quantiles reproduits

Estimateur	MP1			MP2		
	$\nu_{r,\hat{N}}$	$\nu_{r,\hat{Q}}$	$\nu_{r,\hat{\tau}}$	$\nu_{r,\hat{N}}$	$\nu_{r,\hat{Q}}$	$\nu_{r,\hat{\tau}}$
Moyenne tirée des simulations						
Naïf	–	0,52	–	–	2,28	–
Estimation du maximum de vraisemblance (EMV)						
PPI	658,81	0,16	1 049,47	11 779,12	2,98	78 050,24
PPIEQ1	726,83	0,16	1 094,00	3 403,08	0,90	13 198,45
PPIEQ2	1 334,84	0,15	1 884,13	7 310,56	0,46	20 054,55
Équation d'estimation généralisée (EEG)						
PPI	0,00	0,15	0,00	0,00	1,37	0,00
PPIEQ1	0,11	0,10	2,74	0,00	0,36	0,00
PPIEQ2	133,95	0,05	406,16	41,10	0,05	36,12

Note : EEG = équations d'estimation généralisées; EMV = estimation du maximum de vraisemblance; MP = modèle de probabilité; PPI = pondération probabiliste inverse; PPIEQ = pondération probabiliste inverse équilibrant les quantiles.

Tableau B.1 (suite)
Qualité de la taille de population, des totaux et des quantiles reproduits

Estimateur	MP1			MP2		
	$v_{r,\hat{N}}$	$v_{r,\hat{Q}}$	$v_{r,\hat{\tau}}$	$v_{r,\hat{N}}$	$v_{r,\hat{Q}}$	$v_{r,\hat{\tau}}$
Médianes tirées des simulations						
Naïf	–	0,51	–	–	2,28	–
EMV						
PPI	358,00	0,15	760,42	10 617,02	2,76	70 068,24
PPIEQ1	353,72	0,15	732,87	3 081,35	0,86	10 515,02
PPIEQ2	442,59	0,12	901,13	3 242,52	0,41	8 157,91
EEG						
PPI	0,00	0,14	0,00	0,00	1,37	0,00
PPIEQ1	0,00	0,10	0,00	0,00	0,36	0,00
PPIEQ2	0,00	0,04	0,00	0,00	0,05	0,00

Note : EEG = équations d'estimation généralisées; EMV = estimation du maximum de vraisemblance; MP = modèle de probabilité; PPI = pondération probabiliste inverse; PPIEQ = pondération probabiliste inverse équilibrant les quantiles.

Bibliographie

Ai, C., Linton, O. et Zhang, Z. (2020). A simple and efficient estimation method for models with nonignorable missing data. *Statistica Sinica*, 30, 1949-1970. https://doi.org/10.5705/ss.202018.0107.

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.

Beręsewicz, M. (2017). A two-step procedure to measure representativeness of internet data sources. *Revue Internationale de Statistique*, 85(3), 473-493.

Beręsewicz, M. (2023). *jointCalib: A Joint Calibration of Totals and Quantiles* [R package version 0.1.0]. https://CRAN.R-project.org/package=jointCalib.

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. https://doi.org/10.1093/jssam/smaa036.

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, Y., Li, P. et Wu, C. (2023). Traiter le sous-dénombrement pour les échantillons d'enquête non probabilistes. *Techniques d'enquête*, 49(2), 539-558. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2023002/article/00005-fra.pdf.

Chrostowski, Ł. et Beręsewicz, M. (2024). *nonprobsvy: Package for Inference Based on Non-Probability Samples* [R package version 0.1.0]. https://github.com/ncn-foreigners/nonprobsvy.

Citro, C.F. (2014). Des modes multiples pour les enquêtes à des sources de données multiples pour les estimations. *Techniques d'enquête*, 40(2), 151-181. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2014002/article/14128-fra.pdf.

Daas, P.J., Puts, M.J., Buelens, B. et Hurk, P.A.v.d. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, 31(2), 249-262.

Dennis Jr, J.E. et Schnabel, R.B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM.

Deville, J.-C. et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

Duszczyk, M. et Kaczmarczyk, P. (2022). The war in Ukraine and migration to Poland: Outlook and challenges. *Intereconomics*, 57(3), 164-170.

Elliott, M.R. et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2). https://doi.org/10.1214/16-STS598.

Gelman, A. et Little, T.C. (1997). Stratification a posteriori en un grand nombre de catégories par régression logistique hiérarchique. *Techniques d'enquête*, 23(2), 135-145. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1997002/article/3616-fra.pdf.

Harms, T. et Duchesne, P. (2006). De l'estimation des quantiles par calage. *Techniques d'enquête*, 32(1), 41-57. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006001/article/9255-fra.pdf.

Hasselman, B. (2023). *Nleqslv: Solve Systems of Nonlinear Equations* [R package version 3.3.5]. https://CRAN.R-project.org/package=nleqslv.

Hazlett, C. (2020). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. *SSRN Electronic Journal*, 30(3). https://doi.org/10.2139/ssrn.2746753.

Imai, K. et Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1), 243-263. https://doi.org/10.1111/rssb.12027.

Kim, J.K. et Riddles, M.K. (2012). Théorie concernant les estimateurs ajustés sur le score de propension dans les sondages. *Techniques d'enquête*, 38(2), 171-180. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012002/article/11754-fra.pdf.

Kim, J.-K. et Wang, Z. (2019). Sampling techniques for big data analysis. *Revue Internationale de Statistique*, 87, S177-S191.

Park, S., Kim, J.K. et Kim, K. (2019). Note sur la méthode de pondération des scores de propension par l'utilisation de paradonnées dans l'échantillonnage. *Techniques d'enquête*, 45(3), 483-496. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019003/article/00002-fra.pdf.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienne, Autriche. https://www.R-project.org/.

Salvatore, C. (2023). Inference with non-probability samples and survey data integration: A science mapping study. *Metron*, 1-25.

Sant'Anna, P.H., Song, X. et Xu, Q. (2022). Covariate distribution balance via propensity scores. *Journal of Applied Econometrics*, 37(6), 1093-1120.

Statistics Poland (2021). *Methodological Report the Demand for Labour*. https://stat.gov.pl/obszary-tematyczne/rynek-pracy/popyt-na-prace/zeszyt-metodologiczny-popyt-na-prace,3,1.html.

Statistics Poland (2023). *Residents of Ukraine Under Temporary Protection*. https://stat.gov.pl/en/topics/population/internationa-migration/residents-of-ukraine-under-temporary-protection,9,1.html.

Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data* (Vol. 4). Springer.

Wu, C. (2022). Inférence statistique avec des échantillons d'enquête non probabiliste. *Techniques d'enquête*, 48(2), 307-338. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2022002/article/00002-fra.pdf.

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), 33-64. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021001/article/00004-fra.pdf.