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

Techniques d’enquête

Prédiction QR pour l’intégration de données statistiques

par Estelle Medous, Camelia Goga, Anne Ruiz-Gazen,
Jean-François Beaumont, Alain Dessertaine et Pauline Puech

Date de diffusion : le 3 janvier 2024

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Techniques d'enquête, décembre 2023
Vol. 49, nº 2, pp. 417-443
Statistique Canada, nº 12-001-X au catalogue

Prédiction QR pour l'intégration de données statistiques

Estelle Medous, Camelia Goga, Anne Ruiz-Gazen, Jean-François Beaumont, Alain Dessertaine et Pauline Puech[1]

Résumé

Dans le présent article, nous examinons la façon dont une grande base de données non probabiliste peut servir à améliorer des estimations de totaux de population finie d'un petit échantillon probabiliste grâce aux techniques d'intégration de données. Dans le cas où la variable d'intérêt est observée dans les deux sources de données, Kim et Tam (2021) ont proposé deux estimateurs convergents par rapport au plan de sondage qui peuvent être justifiés par la théorie des enquêtes à double base de sondage. D'abord, nous posons des conditions garantissant que les estimateurs en question seront plus efficaces que l'estimateur de Horvitz-Thompson lorsque l'échantillon probabiliste est sélectionné par échantillonnage de Poisson ou par échantillonnage aléatoire simple sans remise. Ensuite, nous étudions la famille des prédicteurs QR proposée par Särndal et Wright (1984) pour le cas moins courant où la base de données non probabiliste ne contient pas la variable d'intérêt, mais des variables auxiliaires. Une autre exigence est que la base non probabiliste soit vaste et puisse être couplée avec l'échantillon probabiliste. Les conditions que nous posons font que le prédicteur QR est asymptotiquement sans biais par rapport au plan de sondage. Nous calculons sa variance asymptotique sous le plan de sondage et présentons un estimateur de variance convergent par rapport au plan de sondage. Nous comparons les propriétés par rapport au plan de sondage de différents prédicteurs de la famille des prédicteurs QR dans une étude par simulation. La famille comprend un prédicteur fondé sur un modèle, un estimateur assisté par un modèle et un estimateur cosmétique. Dans nos scénarios de simulation, l'estimateur cosmétique a donné des résultats légèrement supérieurs à ceux de l'estimateur assisté par un modèle. Nos constatations sont confirmées par une application aux données de La Poste, laquelle illustre par ailleurs que les propriétés de l'estimateur cosmétique sont conservées indépendamment de l'échantillon non probabiliste observé.

Mots-clés : Estimateur cosmétique; double base de sondage; estimateur par régression; échantillon non probabiliste; échantillon probabiliste; estimateur de variance.

1. Introduction

Dans le domaine des sciences économiques et sociales, les enquêtes reposent habituellement sur des méthodes d'échantillonnage probabiliste. Dans le service postal français (La Poste) par exemple, le trafic postal s'estime au moyen d'enquêtes probabilistes trimestrielles. En contrôlant le plan de sondage, on peut faire des inférences fondées sur le plan hors de toute modélisation des variables d'intérêt; cette caractéristique est attrayante pour nombre de statisticiens d'enquête. On considère généralement que Neyman (1934) a été l'étude fondatrice de la théorie de l'échantillonnage probabiliste. Depuis lors, la documentation consacrée à ce thème a été en croissance rapide avec une interaction entre la théorie et la pratique (voir Rao (2005) pour connaître les apports les plus importants).

Ces derniers temps, les statisticiens d'enquête ont observé une diminution des taux de réponse, de même qu'une augmentation des coûts d'enquête, ce qui rend l'échantillonnage probabiliste plus difficile. Une autre constatation est que de grands échantillons non probabilistes sous forme de données administratives ou

1. Estelle Medous, Toulouse School of Economics, Université Toulouse Capitole, 1 Esplanade de l'Université, 31000 Toulouse, Université de Franche-Comté, Laboratoire de Mathématiques de Besançon and La Poste, 3 rue Jean Richepin, 93192 Noisy le Grand cedex. Courriel : estelle.medous@tse-fr.eu; Camelia Goga, Université de Franche-Comté, Laboratoire de Mathématiques de Besançon. Courriel : camelia.goga@univ-fcomte.fr; Anne Ruiz-Gazen, Toulouse School of Economics, Université Toulouse Capitole, 1 Esplanade de l'Université, 31000 Toulouse. Courriel : anne.ruiz-gazen@tse-fr.eu; Jean-François Beaumont, Statistique Canada, 100 promenade du pré Tunney, Ottawa, Ontario, Canada. Courriel : jean-francois.beaumont@statcan.gc.ca; Alain Dessertaine et Pauline Puech, La Poste, 3 rue Jean Richepin, 93192 Noisy le Grand cedex. Courriel : alain.dessertaine@laposte.fr et pauline.puech@laposte.fr.

d'enquêtes en ligne sont souvent disponibles à faible coût (pour obtenir plus de détails, voir Beaumont (2020) et Rao (2021)). Ces observations valent aussi pour La Poste où, en raison des coûts, la taille des échantillons probabilistes ne peut que décroître alors qu'une grande base de données sur le trafic postal en traitement automatique est disponible. Bien que le terme « échantillon non probabiliste » soit synonyme de « mécanisme de sélection inconnu » et que ces échantillons puissent être entachés de biais de sélection et d'erreurs de mesure, ils livrent des renseignements à jour sur la population d'intérêt. Ce contexte est ce qui amène les statisticiens d'enquête à étudier ce que peut être l'intégration ou la combinaison de données d'échantillons probabilistes et non probabilistes.

La recherche spécialisée sur l'intégration des données pour les enquêtes par échantillonnage a récemment connu des progrès rapides, et le lecteur peut consulter plusieurs études sur ce thème (voir Beaumont (2020), Yang et Kim (2020), Rao (2021), Kim (2022) et Wu (2022)). Si nous considérons le problème de la combinaison d'échantillons probabilistes et non probabilistes, nous pouvons voir que les méthodes d'intégration de données se rangent dans trois catégories selon que la variable d'intérêt est observée dans le seul échantillon probabiliste, dans le seul échantillon non probabiliste ou dans les deux (voir, par exemple, Rao (2021)). Dans la plupart des méthodes, on considère le problème d'une variable d'intérêt observée dans le seul échantillon non probabiliste, notamment dans Kim (2022). Dans ce contexte, le but est de corriger le biais de sélection en combinant les données de l'échantillon non probabiliste et des données auxiliaires disponibles dans un échantillon probabiliste.

À La Poste, le problème est plutôt que les variables d'intérêt (il peut s'agir, par exemple, des différents types d'envois postaux) sont seulement présentes dans l'échantillon probabiliste alors que l'information auxiliaire est seulement accessible dans l'échantillon non probabiliste. Un tel contexte est assez rare dans la pratique et, par conséquent, n'a pas encore été étudié en détail. Nous nous proposons d'étudier en profondeur ce contexte particulier dans le présent document. La méthode que nous recommandons et étudions en détail est applicable là où : i) le recouvrement entre l'échantillon probabiliste et l'échantillon non probabiliste est idéalement important et à tout le moins non vide; ii) il est possible de mettre en correspondance les observations des deux échantillons. À La Poste, l'échantillon non probabiliste représente plus de 80 % de la population, aussi est-il largement en intersection avec l'échantillon probabiliste. L'appariement demeure toutefois une tâche difficile que La Poste continue à examiner.

Dans le cas où les variables d'intérêt se mesurent dans les deux échantillons, Kim et Tam (2021) proposent une approche par échantillonnage avec double base de sondage pour accroître l'efficacité de l'estimateur de Horvitz-Thompson (Horvitz et Thompson, 1952), qui exploite seulement les échantillons probabilistes. Un estimateur du total de la variable d'intérêt sur toute la population est la somme du vrai total sur l'échantillon non probabiliste et du total estimé sur le complémentaire de l'échantillon non probabiliste dans la population. Dans ce contexte, Kim et Tam (2021) proposent plusieurs estimateurs par calage.

À la section 2, nous réexaminons l'approche de Kim et Tam (2021) et dégageons des résultats généraux quant à l'efficacité des estimateurs à double base de sondage qu'ils proposent. Si la variable d'intérêt n'est pas mesurée dans l'échantillon non probabiliste, nous proposons de remplacer le total réel inconnu de

l'échantillon non probabiliste par une certaine valeur de prédiction. À la section 3, nous adaptons la famille générale des prédicteurs QR proposée par Wright (1983) à l'intégration de données. Cette famille d'estimateurs comprend deux estimateurs bien connus, l'un assisté par un modèle (estimateur RG) et l'autre fondé sur un modèle, auxquels s'ajoute l'estimateur cosmétique (Särndal et Wright, 1984). Nous posons d'abord une condition pour que les prédicteurs QR puissent se formuler comme projection. Nous en déduisons une condition QR telle que les prédicteurs en question soient identiques aux prédicteurs assistés par un modèle. À la section 4, nous considérons les propriétés asymptotiques des estimateurs QR. Nous montrons que ceux-ci sont asymptotiquement sans biais sous le modèle et le plan de sondage. Nous démontrons également que, avec la condition QR, les prédicteurs sont asymptotiquement sans biais par rapport au plan de sondage. Nous calculons leur variance asymptotique par rapport au plan de sondage et présentons un estimateur de variance convergent par rapport au plan de sondage. À la section 5, nous procédons par simulation de Monte Carlo pour comparer un certain nombre de prédicteurs QR et montrons que l'estimateur cosmétique représente un bon compromis dans plusieurs scénarios. À la section 6, nous examinons une application aux données de La Poste et illustrons l'incidence de l'échantillon non probabiliste sur les estimateurs. Nous concluons enfin et exposons les perspectives à la section 7.

2. Variable d'intérêt observée dans les deux échantillons

Nous désirons estimer le total de population $T = \sum_{k \in U} y_k$, où y_k est la variable d'intérêt Y pour l'unité k de la population U. Nous tirons un échantillon probabiliste s_P de U selon le plan de sondage $p(s_P \mid \mathbf{Z})$, où la matrice de population $\mathbf{Z}$ contient l'information concernant le plan de sondage telle que l'appartenance aux strates. L'indicateur d'inclusion dans l'échantillon, I_k, $k \in U$, prend la valeur 1 si l'unité k est sélectionnée dans s_P et la valeur 0 autrement. La probabilité qu'une unité donnée de population k soit sélectionnée dans s_P est $\pi_k = \mathrm{E}_p(I_k \mid \mathbf{Z})$. Nous supposons dans la présente section que la variable d'intérêt Y est observée pour chaque unité de l'échantillon probabiliste, mais aussi pour l'échantillon non probabiliste $s_{\mathrm{NP}} \subset U$. L'indicateur d'inclusion dans s_{NP} pour l'unité de population $k \in U$ est désigné par δ_k (soit $\delta_k = 1$, si $k \in s_{\mathrm{NP}}$, et $\delta_k = 0$ autrement). Nous supposons que δ_k est disponible pour chaque unité de l'échantillon probabiliste s_P. Désignons par N (N_{NP} respectivement) la taille de U (de s_{NP} respectivement) et par n l'espérance de la taille de s_P. Supposons que $\hat{T}_{\mathrm{HT}} = \sum_{k \in s_P} d_k y_k$ est l'estimateur par dilatation ou estimateur d'Horvitz-Thompson bien connu de T avec les poids d'échantillonnage $d_k = 1/\pi_k$. Si $\pi_k > 0$ pour tous les $k \in U$, $\hat{T}_{\mathrm{HT}}$ est un estimateur de T sans biais par rapport au plan.

L'échantillon non probabiliste s_{NP} est habituellement une source de données vaste et peu coûteuse. Son mécanisme de sélection est inconnu et son biais de sélection ne peut être négligé dans la formulation d'inférences. Par ailleurs, l'échantillon probabiliste s_P est jugé représentatif (exempt de biais de sélection), bien qu'étant souvent cher et d'une taille relativement modeste. En réunissant des informations provenant des deux échantillons, nous espérons dégager un estimateur plus précis que l'estimateur par extension issu de s_P.

Kim et Tam (2021) proposent deux estimateurs combinant des données de s_P et s_{NP}, et nous voulons réexaminer les propriétés de ces estimateurs. Le total peut se décomposer ainsi :

$$T = T_{\mathrm{NP}} + T_C,$$

avec $T_{\mathrm{NP}} = \sum_{k \in s_{\mathrm{NP}}} y_k = \sum_{k \in U} \delta_k y_k$ et $T_C = \sum_{k \in U - s_{\mathrm{NP}}} y_k = \sum_{k \in U} (1 - \delta_k) y_k$. Comme y_k est mesuré pour toutes les unités de s_{NP}, T_{NP} est connu, et nous avons seulement à estimer T_C. Kim et Tam (2021) proposent l'estimateur suivant :

$$\hat{T}_{\mathrm{DI}} = T_{\mathrm{NP}} + \sum_{k \in s_P} d_k (1 - \delta_k) y_k, \tag{2.1}$$

où T_C s'estime à l'aide de l'estimateur par dilatation. Comme l'indique Beaumont (2020), on peut y voir un problème à double base de sondage, avec les bases U et s_{NP}, où l'échantillon s_P est prélevé aléatoirement sur U et où s_{NP} est intégralement recensé. Dans ce contexte de deux bases de sondage, l'estimateur $\hat{T}_{\mathrm{DI}}$ s'obtient en appliquant directement la méthode proposée par Bankier (1986). On pourrait penser que $\hat{T}_{\mathrm{DI}}$ est plus efficace que $\hat{T}_{\mathrm{HT}}$, surtout si la taille de l'échantillon non probabiliste est importante, mais il n'en est pas ainsi en général. La proposition qui suit indique que pour l'échantillonnage de Poisson, la variance de T_{DI} est toujours inférieure à celle de T_{HT}, alors que pour l'échantillonnage aléatoire simple sans remise (EASSR), cette propriété n'est vérifiée que si on suppose une condition sur la variable d'intérêt.

Proposition 2.1

(i) Pour l'échantillonnage de Poisson, la variance de $\hat{T}_{\mathrm{DI}}$ est inférieure ou égale à la variance de $\hat{T}_{\mathrm{HT}}$.

(ii) Pour l'échantillonnage aléatoire simple sans remise, la variance de $\hat{T}_{\mathrm{DI}}$ est inférieure ou égale à la variance de $\hat{T}_{\mathrm{HT}}$ si et seulement si

$$\mathrm{CV}_{\mathrm{NP}}^2 \geq -\frac{N_{\mathrm{NP}}}{N_{\mathrm{NP}} - 1}\left(1 + \frac{N_{\mathrm{NP}}}{N} - 2\frac{\overline{Y}_U}{\overline{Y}_{\mathrm{NP}}}\right),$$

où $\overline{Y}_U = \frac{1}{N}\sum_{k \in U} y_k$ est la moyenne de Y sur U, où $\overline{Y}_{\mathrm{NP}} = \frac{1}{N_{\mathrm{NP}}}\sum_{k \in U} \delta_k y_k$ est la moyenne de Y sur s_{NP} et où $\mathrm{CV}_{\mathrm{NP}} = \sqrt{S_{Y,\mathrm{NP}}^2}\big/\overline{Y}_{\mathrm{NP}}$ est le coefficient de variation de Y dans s_{NP}, avec $S_{Y,\mathrm{NP}}^2 = \frac{1}{N_{\mathrm{NP}} - 1}\sum_{k \in U} \delta_k (y_k - \overline{Y}_{\mathrm{NP}})^2$.

La preuve de la proposition 2.1 figure en annexe. Intuitivement, on peut expliquer le résultat (ii) de la proposition 2.1 par le fait que la taille de s_P est fixe pour un échantillonnage aléatoire simple sans remise dans l'expression de $\hat{T}_{\mathrm{HT}}$, alors que la taille de $s_P \cap U - s_{\mathrm{NP}}$ est aléatoire pour $\hat{T}_{\mathrm{DI}}$. En d'autres termes, l'estimateur $\hat{T}_{\mathrm{DI}}$ est calé sur N_{NP} et T_{NP}, mais non sur N, alors que $\hat{T}_{\mathrm{HT}}$ est calé sur N.

Dans le cas où la taille de la population U est connue, Kim et Tam (2021) proposent d'améliorer $\hat{T}_{\mathrm{DI}}$ à l'aide de l'estimateur suivant :

$$\hat{T}_{\mathrm{PDI}} = T_{\mathrm{NP}} + \hat{T}_C^{(\mathrm{Ha})},$$

où

$$\hat{T}_C^{(\mathrm{Ha})} = (N - N_{\mathrm{NP}}) \frac{\sum_{k \in s_P} d_k (1-\delta_k)\, y_k}{\sum_{k \in s_P} d_k (1-\delta_k)}$$

est un estimateur du type Hájek du total T_C. Kim et Tam (2021) ont prouvé que $\hat{T}_{\mathrm{PDI}}$ est un estimateur par la régression généralisée (estimateur RG) calé sur N, N_{NP} et T_{NP}. Cette expression peut encore se généraliser en utilisant des variables de calage supplémentaires.

À la suite de Kim et Tam (2021), il est possible de procéder par linéarisation et de calculer la variance approchée de $\hat{T}_{\mathrm{PDI}}$ notée $\mathrm{Avar}(\hat{T}_{\mathrm{PDI}})$. Dans le cas d'un échantillonnage de Poisson, l'indépendance des indicateurs d'inclusion ramène la comparaison entre $\hat{T}_{\mathrm{PDI}}$ et $\hat{T}_{\mathrm{DI}}$ à une comparaison entre les estimateurs de Horvitz-Thompson et de Hájek pour le total $T_C = \sum_{k \in U} (1-\delta_k)\, y_k$. L'estimateur de Hájek peut se révéler considérablement plus efficace que l'estimateur de Horvitz-Thompson, mais ce n'est pas le cas en général (voir, par exemple, Särndal, Swensson et Wretman (1992)).

Dans le cas de l'échantillonnage de Poisson, l'estimateur $\hat{T}_{\mathrm{PDI}}$ peut être largement plus efficace que l'estimateur de Horvitz-Thompson $\hat{T}_{\mathrm{HT}}$, ainsi que l'illustre notre étude par simulation à la section 5. Pour un échantillonnage aléatoire simple sans remise, la variance approchée de $\hat{T}_{\mathrm{PDI}}$ peut se comparer à la variance de $\hat{T}_{\mathrm{HT}}$ dans un contexte plus général que celui proposé par Kim et Tam (2021). La proposition 2.2 qui suit indique que la variance approchée de $\hat{T}_{\mathrm{PDI}}$ est inférieure à la variance de $\hat{T}_{\mathrm{HT}}$ en échantillonnage aléatoire simple sans remise et donne l'expression de la différence entre les variances.

Proposition 2.2 *Pour un échantillonnage aléatoire simple sans remise,*

$$\mathrm{Var}(\hat{T}_{\mathrm{HT}}) - \mathrm{Avar}(\hat{T}_{\mathrm{PDI}}) = \frac{N^2 (1-f)}{(N-1)n} \left(\sum_{k \in U} \delta_k (y_k - \bar{Y}_U)^2 + \sum_{k \in U} (1-\delta_k)(\bar{Y}_C - \bar{Y}_U)^2 \right),$$

où $\bar{Y}_U = \frac{1}{N} \sum_{k \in U} y_k$ *est la moyenne de* Y *sur* U *et où* $\bar{Y}_C = \frac{1}{N - N_{NP}} \sum_{k \in U} (1-\delta_k)\, y_k$ *est la moyenne de* Y *sur* $U - s_{\mathrm{NP}}$.

Dans la présente section, nous supposons que la variable d'intérêt Y est mesurée dans les deux échantillons s_P et s_{NP}. À la prochaine section, nous allégerons cette hypothèse en considérant que la variable d'intérêt n'est pas connue dans l'échantillon non probabiliste. C'est la situation que connaît La Poste puisque toutes les variables d'intérêt ne sont pas mesurées lors du traitement automatique des objets postaux. La grande base de données non probabiliste provient d'un processus de reconnaissance d'images qui traite environ 80 % du trafic postal. Cette base de données renferme de l'information auxiliaire utile comme par exemple les dates de départ du bureau de poste expéditeur. Toutefois, ces données peuvent être entachées d'un biais de sélection (puisque les objets postaux qui n'ont pas un format standard ne sont pas mis en traitement automatique, par exemple) ainsi que d'erreurs de mesure (erreurs de lecture de code à barres dans le processus de reconnaissance des images, par exemple). Dans un tel cas, nous proposons d'utiliser l'intersection entre la grande base de données et l'échantillon probabiliste, où les variables

auxiliaires et la variable d'intérêt sont disponibles, afin de prédire la variable d'intérêt inconnue y_k pour $k \in s_{\text{NP}} - s_P$.

3. Estimateurs de prédiction de la variable d'intérêt inobservée dans l'échantillon non probabiliste

Rappelons que le total de population finie de Y peut se décomposer en $T = T_{\text{NP}} + T_C$. Le total T_C est estimé comme à la section 2 par l'estimateur de type Hájek $\hat{T}_C^{(\text{Ha})}$. Dans la présente section, y_k est inconnu pour $k \in s_{\text{NP}}$ et, à la différence de ce qui est présenté à la section 2, le total T_{NP} doit être estimé. Pour ce faire, nous introduisons un modèle pour Y et la famille des prédicteurs QR de T_{NP} pour laquelle y_k n'a pas à être connue pour les unités de s_{NP}. Nous supposons qu'un vecteur de variables auxiliaires $\mathbf{x}_k = (X_{k1}, \ldots, X_{kp})^\top$ est disponible pour chaque unité k d'un échantillon non probabiliste $s_{\text{NP}} \subset U$. Nous supposons également que δ_k et $\delta_k \mathbf{x}_k$ sont disponibles pour chaque unité k de l'échantillon probabiliste s_P. Le tableau 3.1 présente un résumé des caractéristiques des données examinées dans le reste du présent document.

Tableau 3.1
Caractéristiques des données dans le contexte d'intégration de données de la section 3.

Échantillon	y_k mesuré	δ_k disponible	Mécanisme de sélection connu	Variables auxiliaires disponibles
s_P	oui	oui	oui	non
s_{NP}	non	oui	non	oui

La variable Y n'est pas disponible dans s_{NP}, et nous ne pouvons plus employer $\hat{T}_{\text{PDI}}$, puisque le total $T_{\text{NP}} = \sum_{k \in U} \delta_k y_k$ est inconnu. L'idée à la base de la famille d'estimateurs que nous présentons dans la présente section est de prédire y_k pour $k \in s_{\text{NP}}$ grâce à une modélisation par régression entre Y et les variables auxiliaires, puis de prédire T_{NP}. Nous adoptons le modèle suivant entre la variable d'intérêt Y et le vecteur de variables auxiliaires $\mathbf{x}_k$:

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \quad k \in s_{\text{NP}}, \tag{3.1}$$

où les erreurs ε_k sont indépendantes d'espérance $\text{E}_m(\varepsilon_k) = 0$ et où la variance $\text{Var}_m(\varepsilon_k)$ est proportionnelle à $v(\mathbf{x}_k) = v_k$ pour certaines constantes positives v_k connues. L'indice m indique que l'espérance et la variance sont prises par rapport au modèle (3.1) et en conditionnant par les variables auxiliaires observées $\mathbf{x}_k$, $k \in s_{\text{NP}}$. Il convient de mentionner que le modèle (3.1) ne doit être valide que pour les unités de l'échantillon non probabiliste. Un modèle pour Y n'a pas à être explicité pour les unités $k \in U - s_{\text{NP}}$, car les inférences que nous effectuons sont toujours conditionnelles à y_k, $k \in U - s_{\text{NP}}$.

Nous définissons un prédicteur $\hat{y}_k$ de y_k pour $k \in s_{\text{NP}}$ par $\hat{y}_k = \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}$ avec

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k \in s_P} q_k \delta_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \left(\sum_{k \in s_P} q_k \delta_k \mathbf{x}_k y_k \right), \tag{3.2}$$

où q_k sont des constantes positives connues pour $k \in s_{\text{NP}}$. Nous supposons que la matrice de dimensions $p \times p$, $\sum_{k \in s_P} q_k \delta_k \mathbf{x}_k \mathbf{x}_k^\top$, est non singulière pour tous les échantillons s_P possibles.

Nous proposons d'estimer $T_{\text{NP}} = \sum_{k \in U} \delta_k y_k$ par un *prédicteur QR* comme le propose Wright (1983) :

$$\begin{aligned} \hat{T}_{\text{NP}}^{(\text{QR})} &= \sum_{k \in U} \delta_k \hat{y}_k + \sum_{k \in s_P} r_k \delta_k (y_k - \hat{y}_k) \\ &= \sum_{k \in U} \delta_k \mathbf{x}_k^\top \hat{\boldsymbol{\beta}} + \sum_{k \in s_P} r_k \delta_k (y_k - \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}), \end{aligned} \tag{3.3}$$

où $r_k \geq 0$ sont des constantes préétablies. Les initiales Q et R renvoient aux constantes q_k et r_k. L'estimateur final de T est alors donné par

$$\hat{T}^{(\text{QR})} = \hat{T}_{\text{NP}}^{(\text{QR})} + \hat{T}_C^{(\text{Ha})}. \tag{3.4}$$

Divers choix de q_k et r_k donnent des prédicteurs $\hat{T}_{\text{NP}}^{(\text{QR})}$ avec des formes familières décrites ci-après.

1. Pour $q_k = d_k v_k^{-1}$ et $r_k = d_k$, nous obtenons l'estimateur assisté par un modèle ou estimateur RG :

   $$\hat{T}_{\text{NP}}^{(\text{MA})} = \sum_{k \in U} \delta_k \hat{y}_k^{(\text{MA})} + \sum_{k \in s_P} \delta_k d_k (y_k - \hat{y}_k^{(\text{MA})}),$$

   où $\hat{y}_k^{(\text{MA})} = \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}^{(\text{MA})}$ avec $\hat{\boldsymbol{\beta}}^{(\text{MA})} = \left(\sum_{k \in s_P} d_k v_k^{-1} \delta_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \left(\sum_{k \in s_P} d_k v_k^{-1} \delta_k \mathbf{x}_k y_k \right)$.

2. Pour $q_k = v_k^{-1}$ et $r_k = 1$, nous obtenons l'estimateur fondé sur un modèle :

   $$\hat{T}_{\text{NP}}^{(\text{MB})} = \sum_{k \in U} \delta_k \hat{y}_k^{(\text{MB})} + \sum_{k \in s_P} \delta_k (y_k - \hat{y}_k^{(\text{MB})}),$$

   où $\hat{y}_k^{(\text{MB})} = \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}^{(\text{MB})}$ avec $\hat{\boldsymbol{\beta}}^{(\text{MB})} = \left(\sum_{k \in s_P} \delta_k v_k^{-1} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \left(\sum_{k \in s_P} \delta_k v_k^{-1} \mathbf{x}_k y_k \right)$.

3. Pour $q_k = (d_k - 1) v_k^{-1}$ et $r_k = 1$, nous obtenons l'estimateur cosmétique (Särndal et Wright, 1984; Brewer, 1999) :

   $$\hat{T}_{\text{NP}}^{(\text{Cos})} = \sum_{k \in U} \delta_k \hat{y}_k^{(\text{Cos})} + \sum_{k \in s_P} \delta_k (y_k - \hat{y}_k^{(\text{Cos})}),$$

   où $\hat{y}_k^{(\text{Cos})} = \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}^{(\text{Cos})}$ avec

   $$\hat{\boldsymbol{\beta}}^{(\text{Cos})} = \left(\sum_{k \in s_P} (d_k - 1) v_k^{-1} \delta_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \left(\sum_{k \in s_P} (d_k - 1) v_k^{-1} \delta_k \mathbf{x}_k y_k \right).$$

Énonçons un certain nombre de propriétés de cette famille de prédicteurs QR. La proposition 3.1 pose une condition générale pour les constantes q_k et r_k, de sorte que le prédicteur QR puisse se définir comme

une somme de prédictions sur la population. La proposition 3.2 pose une autre condition générale sur les constantes q_k et r_k de sorte que le prédicteur QR soit un estimateur de type assisté par un modèle. Les preuves sont fournies en annexe.

Proposition 3.1 *(forme de projection). Considérons le prédicteur QR* $\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})}$ *donné par (3.3). Si nous posons comme condition qu'il existe un vecteur* $\boldsymbol{\mu} \in \mathbf{R}^p$ *tel que*

$$\textit{(Proj) :}\ \boldsymbol{\mu}^{\top}\mathbf{x}_k q_k = r_k \quad \textit{pour tout} \quad k \in s_{\mathrm{NP}}, \tag{3.5}$$

nous obtenons $\sum_{k\in s_P} r_k \delta_k (y_k - \hat{y}_k) = 0$. *Dans ce cas,* $\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})}$ *peut se formuler comme projection :*

$$\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})} = \sum_{k\in U} \delta_k \hat{y}_k.$$

L'estimateur assisté par un modèle $\hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}$ et l'estimateur fondé sur un modèle $\hat{T}_{\mathrm{NP}}^{(\mathrm{MB})}$ satisfont la condition (Proj) s'il existe un vecteur $\boldsymbol{\mu} \in \mathbf{R}^p$ tel que $\boldsymbol{\mu}^{\top}\mathbf{x}_k = v_k$ pour tout $k \in s_{\mathrm{NP}}$. Cette condition est remplie lorsque v_k est une des variables auxiliaires du modèle. Si $v_k = 1$, la condition est respectée pourvu que la variable constante égale à 1 soit incluse dans le modèle. La condition (Proj) se vérifie pour $\hat{T}_{\mathrm{NP}}^{(\mathrm{Cos})}$ si $\boldsymbol{\mu}^{\top}\mathbf{x}_k = v_k (d_k - 1)^{-1}$ pour tout $k \in s_{\mathrm{NP}}$. Une conséquence de la proposition 3.1 est que, avec un plan de sondage équiprobable comme l'échantillonnage aléatoire simple sans remise, les estimateurs assisté par un modèle, fondé sur un modèle et cosmétique sont tous égaux.

À partir du théorème 2 de Wright (1983), nous déduisons la proposition qui suit. Pour r_k satisfaisant à la condition (QR) ci-après et pour tout q_k donné, le prédicteur QR de T_{NP} est identique au prédicteur assisté par un modèle de T_{NP} avec les mêmes q_k.

Proposition 3.2 *Supposons que les constantes* r_k *et* q_k *sont telles qu'il existe un certain vecteur* $\boldsymbol{\lambda} \in \mathbf{R}^p$ *de sorte que*

$$\textit{(QR) :}\ 1 - \pi_k r_k = \pi_k q_k \mathbf{x}_k^{\top}\boldsymbol{\lambda} \ \textit{pour tout}\ k \in s_{\mathrm{NP}}. \tag{3.6}$$

Dans ce cas,

$$\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})} = \hat{T}_{\mathrm{NP}}^{(\mathrm{Q}\pi)},$$

où

$$\hat{T}_{\mathrm{NP}}^{(\mathrm{Q}\pi)} = \sum_{k\in U} \delta_k \mathbf{x}_k^{\top}\hat{\boldsymbol{\beta}} + \sum_{k\in s_P} d_k \delta_k (y_k - \mathbf{x}_k^{\top}\hat{\boldsymbol{\beta}}) \tag{3.7}$$

est le prédicteur assisté par un modèle de T_{NP} *avec* $\hat{\boldsymbol{\beta}}$ *donné par (3.2).*

À la suite de Wright (1983), nous notons que la condition (QR) se vérifie toujours pour $\hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}$. Elle vaut aussi pour l'estimateur fondé sur un modèle $\hat{T}_{\mathrm{NP}}^{(\mathrm{MB})}$ si et seulement s'il existe un vecteur $\boldsymbol{\lambda} \in \mathbf{R}^p$ de sorte

que $v_k(d_k - 1) = \mathbf{x}_k^\top \boldsymbol{\lambda}$ pour tout $k \in s_{\mathrm{NP}}$. Cette condition se vérifie si nous prenons $v_k(d_k - 1)$ parmi les variables auxiliaires $\mathbf{x}_k$. La condition (QR) vaut pour l'estimateur cosmétique $\hat{T}_{\mathrm{NP}}^{(\mathrm{Cos})}$ si et seulement s'il existe un vecteur $\boldsymbol{\lambda} \in \mathbf{R}^p$ de sorte que $v_k = \mathbf{x}_k^\top \boldsymbol{\lambda}$ pour tout $k \in s_{\mathrm{NP}}$. Cette condition est vraie si v_k est inclus dans le vecteur des variables auxiliaires.

4. Propriétés asymptotiques et estimation de variance des prédicteurs QR

Considérons la famille de prédicteurs QR $\hat{T}^{(\mathrm{QR})}$ donnée en (3.4) et étudions d'abord l'erreur de prédiction $\hat{T}^{(\mathrm{QR})} - T$ sous le modèle (3.1) et le plan de sondage $p(\cdot)$. Nous avons

$$\hat{T}^{(\mathrm{QR})} - T = \left(\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})} - T_{\mathrm{NP}}\right) + \left(\hat{T}_C^{(\mathrm{Ha})} - T_C\right).$$

Le premier terme du côté droit dépend du modèle et du plan de sondage, tandis que le second terme dépend seulement du plan de sondage. Si nous supposons que le plan de sondage n'est pas informatif par rapport au modèle (3.1), nous pouvons démontrer que l'espérance de $\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})} - T_{\mathrm{NP}}$ calculée par rapport au modèle est égale à 0. En fait, le biais sous le modèle de $\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})}$ est donné par :

$$\mathrm{E}_m(\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})} - T_{\mathrm{NP}}) = \sum_{k \in U} \delta_k \mathrm{E}_m(\mathbf{x}_k^\top \hat{\boldsymbol{\beta}} - y_k) + \sum_{k \in s_P} r_k \delta_k \mathrm{E}_m(y_k - \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}),$$

avec $\hat{\boldsymbol{\beta}} = \left(\sum_{k \in s_P} q_k \delta_k \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \left(\sum_{k \in s_P} q_k \delta_k \mathbf{x}_k y_k\right)$. Dans le modèle (3.1), $\mathrm{E}_m(y_k) = \mathbf{x}_k^\top \boldsymbol{\beta}$ pour tout $k \in s_{\mathrm{NP}}$, $\mathrm{E}_m(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ et $\mathrm{E}_m(\mathbf{x}_k^\top \hat{\boldsymbol{\beta}} - y_k) = 0$, ce qui implique que

$$\mathrm{E}_m(\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})} - T_{\mathrm{NP}}) = 0. \tag{4.1}$$

L'estimateur $\hat{T}_C^{(\mathrm{Ha})}$ est un estimateur du type Hájek pour T_C et il n'est pas sans biais par rapport au plan pour T_C. À la suite de Särndal (1980), nous nous intéressons plutôt à son absence asymptotique de biais par rapport au plan. Un estimateur de $\hat{T}$ est dit asymptotiquement sans biais par rapport au plan pour le total de population finie T si $\lim_{N \to \infty} N^{-1}[\mathrm{E}_p(\hat{T}) - T] = 0$, où $\mathrm{E}_p(\cdot)$ est l'espérance calculée par rapport au plan de sondage. Le cadre asymptotique d'Isaki et Fuller (1982) peut être considéré pour une croissance à l'infini des tailles de population et d'échantillon. Si nous supposons que la probabilité d'observer un ensemble à intersection vide $s_P \cap s_{\mathrm{NP}}$ est négligeable, alors $\hat{T}_C^{(\mathrm{Ha})}$ sera asymptotiquement sans biais par rapport au plan pour T_C. Si nous combinons cette propriété à la relation (4.1) et considérons un plan de sondage non informatif, nous obtenons un $\hat{T}^{(\mathrm{QR})}$ asymptotiquement sans biais sous le modèle (3.1) et sous le plan.

4.1 Propriétés de biais de $\hat{T}^{(\mathrm{Q}\pi)}$

Étudions maintenant la famille de prédicteurs QR qui satisfait à la condition (QR) posée en (3.6). Pour cette famille de prédicteurs, l'estimateur final de T est

$$\hat{T}^{(\mathrm{Q}\pi)} = \hat{T}_{\mathrm{NP}}^{(\mathrm{Q}\pi)} + \hat{T}_C^{(\mathrm{Ha})}$$

et l'erreur totale normalisée est donnée par :

$$\frac{1}{N}\left(\hat{T}^{(\mathrm{Q}\pi)}-T\right) = \frac{1}{N}\left(\hat{T}_{\mathrm{NP}}^{(\mathrm{Q}\pi)}-T_{\mathrm{NP}}\right)+\frac{1}{N}\left(\hat{T}_{C}^{(\mathrm{Ha})}-T_{C}\right).$$

L'estimateur $\hat{T}^{(\mathrm{Q}\pi)}$ n'est pas exactement sans biais par rapport au plan en raison de la non-linéarité de $\hat{\boldsymbol{\beta}}$ et de l'estimateur de Hájek $\hat{T}_{C}^{(\mathrm{Ha})}$. Wright (1983) a prouvé que la condition (QR) de la proposition 3.2 est suffisante pour que $\hat{T}_{\mathrm{NP}}^{(\mathrm{Q}\pi)}$ soit asymptotiquement sans biais par rapport au plan pour T_{NP}, pourvu que

$$\lim_{N\to\infty}\frac{1}{N}\mathrm{E}_{p}\left[\left(\sum_{k\in U}\delta_{k}\mathbf{x}_{k}-\sum_{k\in s_{P}}d_{k}\delta_{k}\mathbf{x}_{k}\right)^{\top}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})\right] = 0, \tag{4.2}$$

où $\tilde{\boldsymbol{\beta}}=\left(\sum_{k\in U}\pi_{k}q_{k}\delta_{k}\mathbf{x}_{k}\mathbf{x}_{k}^{\top}\right)^{-1}\sum_{k\in U}\pi_{k}q_{k}\delta_{k}\mathbf{x}_{k}y_{k}$ et en supposant que $\sum_{k\in U}\pi_{k}q_{k}\delta_{k}\mathbf{x}_{k}\mathbf{x}_{k}^{\top}$ soit non singulière. Comme dans Breidt et Opsomer (2000), si la fraction d'échantillonnage n/N converge vers une constante différente de 0 et que nous supposons des conditions faibles sur les probabilités d'inclusion du premier et du second ordre du plan de sondage ainsi que sur les vecteurs d'information auxiliaire $\mathbf{x}_{k}$ pour tout $k\in s_{\mathrm{NP}}$, il peut être démontré que

$$\lim_{N\to\infty}\mathrm{E}_{p}\left\|N^{-1}\left(\sum_{k\in U}\delta_{k}\mathbf{x}_{k}-\sum_{k\in s_{P}}d_{k}\delta_{k}\mathbf{x}_{k}\right)\right\|^{2} = 0,$$

où $\|\cdot\|$ est la norme euclidienne habituelle. L'équation (4.2) s'ensuit si nous supposons que l'estimateur de coefficient de régression satisfait à $\lim_{N\to\infty}\mathrm{E}_{p}\left\|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\right\|^{2}=0$ (voir Cardot, Goga et Lardin (2013) pour obtenir plus de détails). L'estimateur $\hat{T}_{C}^{(\mathrm{Ha})}$ est un estimateur du type Hájek dont on peut démontrer qu'il est asymptotiquement sans biais par rapport au plan pour T_{C} si la probabilité d'observer l'ensemble vide pour $s_{P}\cap s_{\mathrm{NP}}$ est négligeable. Nous en concluons que le prédicteur QR $\hat{T}^{(\mathrm{Q}\pi)}$ est asymptotiquement sans biais par rapport au plan pour T.

4.2 Variance asymptotique et estimation de variance de $\hat{T}^{(\mathrm{Q}\pi)}$

Comme l'estimateur QR $\hat{T}^{(\mathrm{Q}\pi)}$ est asymptotiquement sans biais par rapport au plan, nous estimons sa variance asymptotique sous le plan plutôt que son erreur quadratique moyenne sous le plan. Nous pouvons formuler l'erreur totale normalisée comme :

$$\frac{1}{N}(\hat{T}^{(\mathrm{Q}\pi)}-T) = \frac{1}{N}\left(\sum_{k\in s_{P}}d_{k}(E_{k}+e_{k})-\sum_{k\in U}(E_{k}+e_{k})\right)+R_{1}+R_{2},$$

où

$$E_{k} = \delta_{k}(y_{k}-\mathbf{x}_{k}^{\top}\tilde{\boldsymbol{\beta}}), \quad e_{k} = (1-\delta_{k})\left(y_{k}-\frac{\sum_{k'\in U}(1-\delta_{k'})\,y_{k'}}{N-N_{\mathrm{NP}}}\right),$$

$$R_1 = -\frac{1}{N}\left(\sum_{k\in s_P} d_k\delta_k\mathbf{x}_k - \sum_{k\in U}\delta_k\mathbf{x}_k\right)^\top (\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})$$

et

$$R_2 = \left(\frac{1}{N}\sum_{k\in s_P} d_k(1-\delta_k)\right)^{-1}\frac{1}{N}\left(N-N_{\mathrm{NP}}-\sum_{k\in s_P} d_k(1-\delta_k)\right)\frac{1}{N}\left(\sum_{k\in s_P} d_k e_k - \sum_{k\in U} e_k\right).$$

Comme à la section 4.1, nous supposons les conditions habituelles pour la fraction d'échantillonnage n/N, les probabilités d'inclusion de premier et de second ordre, la variable d'intérêt y_k et le vecteur d'information auxiliaire $\mathbf{x}_k$. Nous utilisons les notations de Landau avec grand O_p et petit o_p. Si $\|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\|^2 = o_p(1)$ et $\left(\sum_{k\in s_P} d_k(1-\delta_k)/N\right)^{-1} = O_p(1)$, alors $R_1 = o_p(n^{-1/2})$, $R_2 = O_p(n^{-1})$ et l'erreur totale normalisée peut être estimée par

$$\frac{1}{N}(\hat{T}^{(\mathrm{Q}\pi)}-T) \approx \frac{1}{N}\left(\sum_{k\in s_P} d_k(E_k+e_k) - \sum_{k\in U}(E_k+e_k)\right),$$

où le côté droit de l'expression qui précède est de l'ordre $O_p(n^{-1/2})$. Comme $\sum_{k\in s_P} d_k(E_k+e_k)$ est l'estimateur de Horvitz-Thompson du total $\sum_{k\in U}(E_k+e_k)$, la variance asymptotique de l'estimateur QR $\hat{T}^{(\mathrm{Q}\pi)}$ est donnée par

$$\mathrm{Avar}(\hat{T}^{(\mathrm{Q}\pi)}) = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}d_k d_l(E_k+e_k)(E_l+e_l).$$

Si nous supposons que $\pi_{kl}>0$ pour tout $k,l\in U$, un estimateur de la variance asymptotique est donné par

$$\hat{V}(\hat{T}^{(\mathrm{Q}\pi)}) = \sum_{k\in s_P}\sum_{l\in s_P}\frac{\Delta_{kl}}{\pi_{kl}}d_k d_l(\hat{E}_k+\hat{e}_k)(\hat{E}_l+\hat{e}_l),$$

où $\hat{E}_k = \delta_k(y_k - \mathbf{x}_k^\top\hat{\boldsymbol{\beta}})$ et $\hat{e}_k = (1-\delta_k)(y_k - \sum_{k'\in s_P} d_{k'}(1-\delta_{k'})y_{k'}/(N-N_{\mathrm{NP}}))$. Nous pouvons démontrer, avec les hypothèses décrites dans Breidt et Opsomer (2000) et Goga, Deville et Ruiz-Gazen (2009), que cet estimateur est convergent par rapport au plan de sondage pour $\mathrm{Avar}(\hat{T}^{(\mathrm{Q}\pi)})$, au sens que $N^{-2}n(\hat{V}(\hat{T}^{(\mathrm{Q}\pi)}) - \mathrm{Avar}(\hat{T}^{(\mathrm{Q}\pi)})) = o_p(1)$.

5. Simulations

Dans la présente section, nous présentons les résultats d'une étude de Monte Carlo où nous comparons l'efficacité de trois prédicteurs QR $\hat{T}^{\mathrm{QR}} = \hat{T}_{\mathrm{NP}}^{\mathrm{QR}} + \hat{T}_C^{(\mathrm{HA})}$ particuliers de la section 3, à savoir l'estimateur assisté par un modèle, l'estimateur fondé sur un modèle et l'estimateur cosmétique, si nous supposons aussi que $v_k = 1$ dans le modèle (3.1). Nous comparons également ces estimateurs à l'estimateur par dilatation et à l'estimateur PDI défini à la section 2. Pour illustrer que la supériorité relative des estimateurs dépend de la structure des données, nous définissons trois scénarios différents selon des populations artificielles

différentes. Comme mentionné à la section 3, si on prélève des échantillons probabilistes par échantillonnage aléatoire simple sans remise, les trois estimateurs QR sont tous égaux. Ainsi, nous nous concentrons sur un échantillonnage de Poisson où les probabilités d'inclusion sont proportionnelles à une variable auxiliaire.

5.1 Populations et scénarios

Nous générons les variables à l'aide de distributions Gamma pour nous assurer qu'elles sont positives. Des résultats de simulation semblables ont été obtenus avec des distributions gaussiennes, mais nous ne les présentons pas ci-dessous. Toutes les populations sont de taille $N = 1\,000$. Nous générons deux variables auxiliaires X_1 et X_2 où X_1 (X_2 respectivement) suit une distribution Gamma avec une moyenne égale à 20 (30 respectivement) et un écart-type de 15 (20 respectivement). Nous employons différents modèles pour définir la variable Y pour toutes les unités de la population. Dans chaque modèle, $Y \mid X_1, X_2$ suit une distribution Gamma avec une variance $\sigma^2_{Y|X_1, X_2}$ constante et une moyenne $\mu_{Y|X_1, X_2}$ qui dépend du modèle.

1. Pour le modèle 1, $\mu_{Y|X_1, X_2}$ est une fonction linéaire de X_1 et X_2 :

$$\mu_{Y|X_1, X_2} = a_0 + a_1 X_1 + a_2 X_2.$$

2. Pour le modèle 2, $\mu_{Y|X_1, X_2}$ est une fonction quadratique de X_1 et une fonction linéaire de X_2 :

$$\mu_{Y|X_1, X_2} = b_0 + b_1 (X_1 - \bar{X}_1)^2 + b_2 X_2 \text{ avec } \bar{X}_1 \text{ la moyenne de } X_1 \text{ sur } U.$$

3. Pour le modèle 3, $\mu_{Y|X_1, X_2}$ est une fonction linéaire de X_2 :

$$\mu_{Y|X_1, X_2} = c_0 + c_2 X_2.$$

Pour rendre les résultats comparables entre les trois modèles, nous déterminons les constantes a_0, a_1, a_2, b_0, b_1, b_2, c_0, c_2 et $\sigma^2_{Y|X_1, X_2}$ de sorte que les caractéristiques suivantes sont les mêmes :

- moyenne μ et variance σ^2 inconditionnelles de la variable Y,
- coefficient de détermination du modèle désigné par R^2,
- rapport des variances pour les variables explicatives :

$$\gamma = \text{Var}(a_1 X_1) / \text{Var}(a_2 X_2) = \text{Var}(b_1 (X_1 - \bar{X}_1)^2) / \text{Var}(b_2 X_2).$$

Ce rapport est applicable seulement pour les modèles 1 et 2, car X_1 n'est pas inclus dans le modèle 3.

Dans ce qui suit, nous posons $\mu = 100$, $\sigma^2 = 100$ et $\gamma = 0{,}5$. À la section 5.2, la valeur R^2 est fixe à 0,8 ou varie de 0,1 à 0,96. Les grandes caractéristiques des trois modèles de population sont résumées dans le tableau 5.1.

Tableau 5.1
Modèles de population avec $\mu = 100$, $\sigma^2 = 100$ et $\gamma = 0{,}5$.

Modèle	Moyenne de (X_1, X_2)	Écart-type de (X_1, X_2)	Moyenne de $Y \mid X_1, X_2$	R^2
1	(20, 30)	(15, 20)	$\mu_Y = a_0 + a_1 X_1 + a_2 X_2$	égalité
2			$\mu_Y = b_0 + b_1 (X_1 - \bar{X}_1)^2 + b_2 X_2$	entre
3			$\mu_Y = c_0 + c_2 X_2$	populations

Nous tirons un échantillon non probabiliste de taille 900 par échantillonnage aléatoire simple sans remise. Il sera le même pour toutes les populations. Les échantillons probabilistes sont tirés selon un échantillonnage de Poisson avec une taille d'espérance de 200 ou 50 et des probabilités proportionnelles à X_1. Nous considérons trois scénarios et, dans chacun, nous générons $Y \mid X_1, X_2$ à l'aide d'un des trois modèles de population. Nous calculons $\hat{y}_k, k \in s_{\text{NP}}$ pour différents prédicteurs QR. Les variables servant de variables explicatives dans les modèles de prédiction diffèrent selon les scénarios, comme suit :

1. Scénario 1 : cas informatif. Le modèle de population 1 sert à générer les valeurs de population Y et seul X_2 sert de variable explicative dans le modèle de prédiction avec une ordonnée à l'origine.
2. Scénario 2 : cas quadratique. Le modèle de population 2 sert à générer les valeurs de population Y et les deux variables auxiliaires X_1 et X_2 servent de variables explicatives dans le modèle de prédiction avec l'ordonnée à l'origine.
3. Scénario 3 : cas non informatif. Le modèle de population 3 sert à générer les valeurs de la population Y et seul X_2 sert de variable explicative dans le modèle de prédiction avec l'ordonnée à l'origine.

Pour les scénarios informatif et quadratique, le modèle de prédiction diffère du modèle de population pour Y, et tandis que le bon modèle est employé dans le scénario non informatif. Le tableau 5.2 présente un résumé des trois scénarios.

Tableau 5.2
Scénarios étudiés.

Scénario	Population	Variables servant à la prédiction	Bonne spécification du modèle
Informatif	$\mu_Y = a_0 + a_1 X_1 + a_2 X_2$	$\mathbf{x}_k^\top = (1, x_{2k})$	non
Quadratique	$\mu_Y = b_0 + b_1 (X_1 - \bar{X}_1)^2 + b_2 X_2$	$\mathbf{x}_k^\top = (1, x_{1k}, x_{2k})$	non
Non informatif	$\mu_Y = c_0 + c_2 X_2$	$\mathbf{x}_k^\top = (1, x_{2k})$	oui

5.2 Résultats

Considérons les trois scénarios qui précèdent et comparons les estimateurs suivants :

- $\hat{T}_{\text{HT}} = \sum_{k \in s_P} d_k y_k$,

- $\hat{T}_{\mathrm{PDI}} = T_{\mathrm{NP}} + (N - N_{\mathrm{NP}}) \dfrac{\sum_{k \in s_P} d_k (1-\delta_k)\, y_k}{\sum_{k \in s_P} d_k (1-\delta_k)}$,

- $\hat{T}^{(\mathrm{MB})} = \sum_{k \in U} \delta_k \hat{y}_k^{(\mathrm{MB})} + \sum_{k \in s_P} \delta_k (y_k - \hat{y}_k^{(\mathrm{MB})}) + (N - N_{\mathrm{NP}}) \dfrac{\sum_{k \in s_P} d_k (1-\delta_k)\, y_k}{\sum_{k \in s_P} d_k (1-\delta_k)}$,

- $\hat{T}^{(\mathrm{MA})} = \sum_{k \in U} \delta_k \hat{y}_k^{(\mathrm{MA})} + \sum_{k \in s_P} \delta_k d_k (y_k - \hat{y}_k^{(\mathrm{MA})}) + (N - N_{\mathrm{NP}}) \dfrac{\sum_{k \in s_P} d_k (1-\delta_k)\, y_k}{\sum_{k \in s_P} d_k (1-\delta_k)}$,

- $\hat{T}^{(\mathrm{Cos})} = \sum_{k \in U} \delta_k \hat{y}_k^{(\mathrm{Cos})} + \sum_{k \in s_P} \delta_k (y_k - \hat{y}_k^{(\mathrm{Cos})}) + (N - N_{\mathrm{NP}}) \dfrac{\sum_{k \in s_P} d_k (1-\delta_k)\, y_k}{\sum_{k \in s_P} d_k (1-\delta_k)}$.

Pour chaque scénario, $L =$ 10 000 échantillons probabilistes s_P sont prélevés conformément au plan d'échantillonnage de Poisson comme nous l'avons décrit plus haut et plusieurs mesures de Monte Carlo sont calculées. Nous calculons le biais relatif de Monte Carlo d'un estimateur donné $\hat{T}$ (soit $\hat{T}_{\mathrm{HT}}$, $\hat{T}^{(\mathrm{MB})}$, $\hat{T}^{(\mathrm{MA})}$, $\hat{T}^{(\mathrm{Cos})}$ ou $\hat{T}_{\mathrm{PDI}}$) comme

$$\mathrm{BR}_{\mathrm{MC}}(\hat{T}) = 100 \times L^{-1} \sum_{l=1}^{L} \frac{\hat{T}^{(l)} - T}{T},$$

où $\hat{T}^{(l)}$ est une estimation de T calculée pour le l^{e} échantillon, $l = 1, \ldots, L$.

Comme mesure d'efficacité, nous calculons l'erreur quadratique moyenne relative (EQMR) de Monte Carlo pour un estimateur $\hat{T}$ (relativement à $\hat{T}^{(\mathrm{Cos})}$) :

$$\mathrm{EQMR}_{\mathrm{MC}}(\hat{T}) = 100 \times \frac{\mathrm{EQM}_{\mathrm{MC}}(\hat{T})}{\mathrm{EQM}_{\mathrm{MC}}(\hat{T}^{(\mathrm{Cos})})},$$

où

$$\mathrm{EQM}_{\mathrm{MC}}(\hat{T}) = L^{-1} \sum_{l=1}^{L} \left(\hat{T}^{(l)} - T \right)^2.$$

Nous calculons aussi la variance relative (VarR) de Monte Carlo d'un estimateur $\hat{T}$ (relativement à $\hat{T}^{(\mathrm{Cos})}$) :

$$\mathrm{VarR}_{\mathrm{MC}}(\hat{T}) = 100 \times \frac{\mathrm{Var}_{\mathrm{MC}}(\hat{T})}{\mathrm{Var}_{\mathrm{MC}}(\hat{T}^{(\mathrm{Cos})})},$$

où

$$\mathrm{Var}_{\mathrm{MC}}(\hat{T}) = L^{-1} \sum_{l=1}^{L} \left(\hat{T}^{(l)} \right)^2 - \left(L^{-1} \sum_{l=1}^{L} \hat{T}^{(l)} \right)^2.$$

Le tableau 5.3 présente les résultats de simulation pour les trois scénarios lorsque $R^2 = 0{,}8$. Dans tous les scénarios, nous confirmons que $\hat{T}_{\text{PDI}}$ et $\hat{T}_{\text{HT}}$ ont un biais Monte Carlo négligeable. Pour ce qui est de l'erreur quadratique moyenne, $\hat{T}_{\text{PDI}}$ est l'estimateur le plus précis, alors que $\hat{T}_{\text{HT}}$ est l'estimateur le moins précis de tous. Ce résultat était à prévoir, car l'estimateur par dilatation n'exploite pas d'information auxiliaire, alors que $\hat{T}_{\text{PDI}}$ tient compte des valeurs réelles de la variable d'intérêt y_k pour $k \in s_{\text{NP}}$. $\hat{T}_{\text{PDI}}$ prend donc en compte les valeurs réelles de Y pour 900 unités sur les 1 000 de la population. Dans notre contexte où la variable d'intérêt n'est pas observée dans s_{NP}, l'estimateur $\hat{T}_{\text{PDI}}$ n'est pas calculable et sert plutôt de référence.

Tableau 5.3
Biais relatif (en % de la valeur réelle), variance relative (en % à $\hat{T}^{(\text{Cos})}$) et EQM des différents estimateurs pour les trois différents scénarios; l'espérance de la taille de l'échantillon probabiliste est de 200 et la taille de l'échantillon non probabiliste, est de 900.

Paramètres de population	Scénario	Mesures de Monte Carlo	$\hat{T}_{\text{HT}}$	$\hat{T}^{(\text{MB})}$	$\hat{T}^{(\text{MA})}$	$\hat{T}^{(\text{Cos})}$	$\hat{T}_{\text{PDI}}$
$\mu = 100$ $\sigma^2 = 100$ $R^2 = 0{,}8$ $\gamma = 0{,}5$	Scénario 1	BR_{MC}	-0,13	3,34	0,11	0,11	0,03
		VarR_{MC}	23 566,93	55,62	114,06	100,00	20,97
		EQMR_{MC}	22 897,58	2 715,21	113,91	100,00	20,65
	Scénario 2	BR_{MC}	-0,07	-1,65	-0,06	-0,05	0,02
		VarR_{MC}	36 947,99	84,94	118,21	100,00	23,17
		EQMR_{MC}	36 638,27	1 056,44	118,42	100,00	23,15
	Scénario 3	BR_{MC}	0,03	-0,01	0,01	0,01	0,01
		VarR_{MC}	41 088,93	58,38	100,49	100,00	33,47
		EQMR_{MC}	41 080,51	58,39	100,48	100,00	33,51

Notes : BR = biais relatif; VarR = variance relative; EQMR = erreur quadratique moyenne relative; MC = Monte Carlo.

Le biais de Monte Carlo de $\hat{T}^{(\text{MA})}$ et $\hat{T}^{(\text{Cos})}$ est négligeable dans les trois scénarios, tandis que $\hat{T}^{(\text{MB})}$ accuse un biais dans les scénarios informatif et quadratique. Dans ces deux scénarios, le modèle de prédiction diffère du modèle de population servant à générer les valeurs de Y. Dans le scénario non informatif où le modèle de prédiction est bien spécifié, le biais de $\hat{T}^{(\text{MB})}$ est également négligeable. L'estimateur $\hat{T}^{(\text{MA})}$ présente la plus grande variance des prédicteurs QR dans les scénarios informatif et quadratique, tandis que $\hat{T}^{(\text{MB})}$ présente la variance la plus petite dans les trois scénarios. Dans le scénario quadratique, la variance de $\hat{T}^{(\text{MB})}$ est semblable à la variance de $\hat{T}^{(\text{Cos})}$, mais $\hat{T}^{(\text{MB})}$ a la plus grande EQM parmi les prédicteurs QR dans les scénarios informatif et quadratique. Cela signifie que le biais de $\hat{T}^{(\text{MB})}$ voit son EQM se dégrader largement malgré sa petite variance. Dans le scénario non informatif, $\hat{T}^{(\text{MB})}$ a la plus faible EQM parmi les prédicteurs QR. Nous pouvons voir dans le tableau 5.3 que cela tient à l'absence de biais pour $\hat{T}^{(\text{MB})}$ dans ce scénario ainsi qu'à sa petite variance.

Dans les scénarios informatif et quadratique, $\hat{T}^{(\text{Cos})}$ est plus précis avec une variance plus faible que $\hat{T}^{(\text{MA})}$. Les estimateurs $\hat{T}^{(\text{MA})}$ et $\hat{T}^{(\text{Cos})}$ sont semblables dans le scénario non informatif. Les deux estimateurs font appel à une régression pondérée avec des poids légèrement différents (d_k pour $\hat{T}^{(\text{MA})}$ et $d_k - 1$ pour $\hat{T}^{(\text{Cos})}$).

En résumé, quand le modèle de prédiction est mal spécifié comme dans les scénarios informatif et quadratique, $\hat{T}^{(\mathrm{MA})}$ et $\hat{T}^{(\mathrm{Cos})}$ sont nettement plus efficaces que $\hat{T}^{(\mathrm{MB})}$, en raison du biais de $\hat{T}^{(\mathrm{MB})}$, et même si ce biais n'est pas grand. À l'opposé, si le modèle est bien spécifié mais que les poids de sondage et Y sont sans corrélation, comme dans le scénario non informatif, $\hat{T}^{(\mathrm{MB})}$ est mieux que $\hat{T}^{(\mathrm{MA})}$ et $\hat{T}^{(\mathrm{Cos})}$ pour l'EQM. Dans tous les scénarios, $\hat{T}^{(\mathrm{Cos})}$ est plus efficace que $\hat{T}^{(\mathrm{MA})}$ ou d'une efficacité semblable.

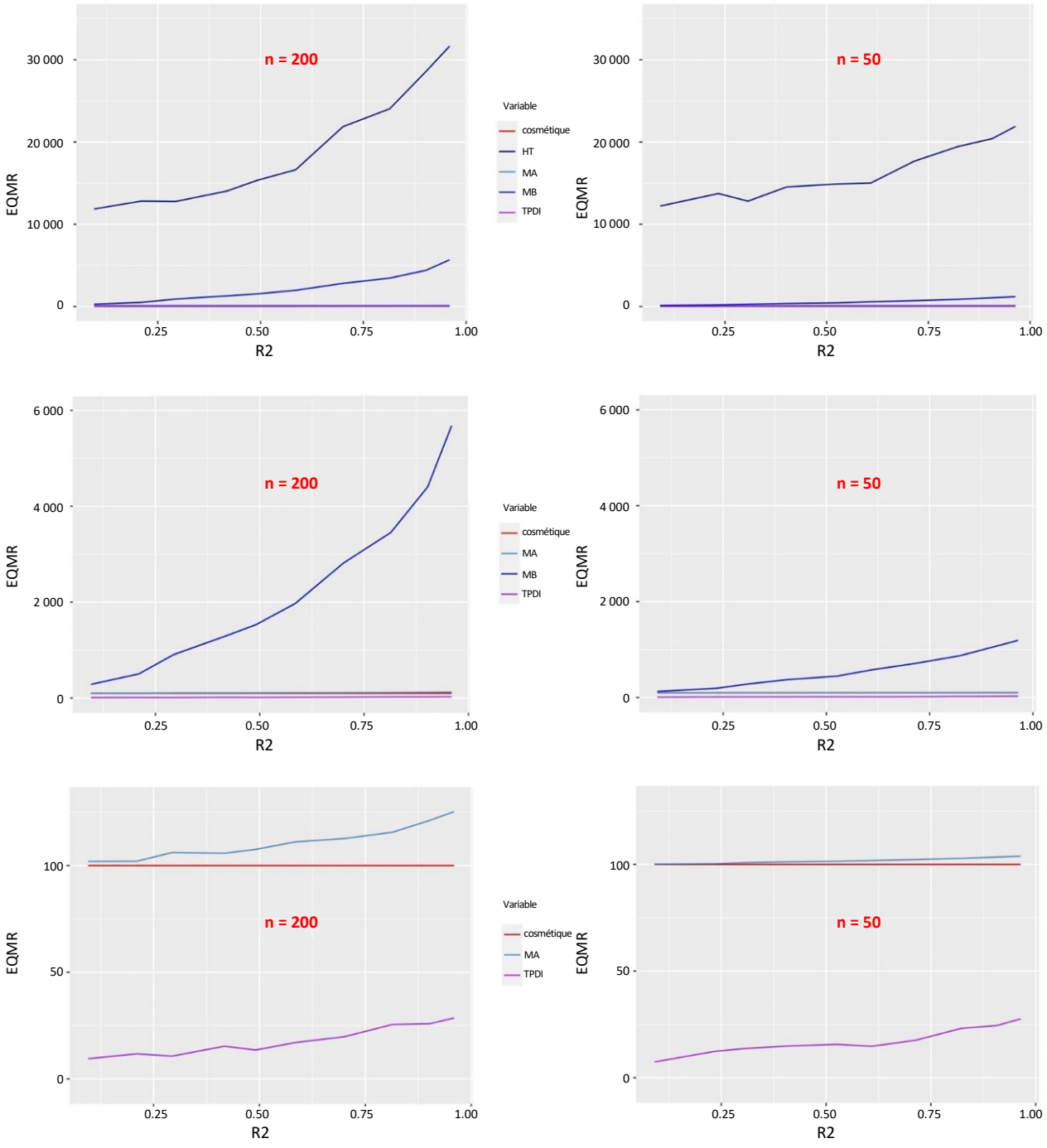
Pour mieux comprendre l'incidence de R^2 sur les résultats, nous représentons graphiquement, sur l'axe des y des figures 5.1, 5.2 et 5.3, l'$\mathrm{EQMR}_{\mathrm{MC}}$ pour 10 valeurs différentes de R^2 représentées sur l'axe des x: 0,1; 0,2; …, 0,9; 0,96. Pour ce faire, nous générons 10 populations pour chaque scénario, une pour chaque valeur de R^2. La figure 5.1 (figure 5.2 et 5.3 respectivement) présente les résultats pour le scénario 1 (2 et 3 respectivement) pour une taille d'échantillon égale à 200 (à 50 respectivement) dans la colonne de gauche (de droite respectivement) des graphiques. Dans tous les graphiques, les courbes correspondent aux différents estimateurs avec la courbe rouge à 100 pour $\hat{T}^{(\mathrm{Cos})}$ (l'EQMR étant relative à $\hat{T}^{(\mathrm{Cos})}$) et d'autres couleurs pour $\hat{T}_{\mathrm{HT}}$, $\hat{T}^{(\mathrm{MA})}$, $\hat{T}^{(\mathrm{MB})}$ et $\hat{T}_{\mathrm{PDI}}$. Les graphiques dans le haut des figures comprennent tous les estimateurs, tandis que pour la deuxième rangée (en troisième rangée pour les figures 5.1 et 5.2), $\hat{T}_{\mathrm{HT}}$ (et $\hat{T}^{(\mathrm{MB})}$ pour les figures 5.1 et 5.2) est retiré de manière à mettre en évidence et à faciliter la comparaison entre $\hat{T}^{(\mathrm{Cos})}$, $\hat{T}^{(\mathrm{MA})}$, $\hat{T}^{(\mathrm{MB})}$ et $\hat{T}_{\mathrm{PDI}}$. L'échelle sur l'axe des y est maintenue fixe pour les deux colonnes (tailles d'échantillon).

Comme on pouvait s'y attendre, $\hat{T}_{\mathrm{PDI}}$ est de loin le meilleur estimateur avec l'EQM la plus faible dans tous les scénarios. Quelle que soit la figure, $\hat{T}_{\mathrm{HT}}$ a une EQM relative très médiocre, particulièrement là où R^2 est élevé. Fait à noter, en réalité, l'EQM absolue de $\hat{T}_{\mathrm{HT}}$ demeure stable lorsque R^2 augmente (résultats non présentés), alors que l'EQM des autres estimateurs s'améliore. Ce résultat est attendu parce que $\hat{T}_{\mathrm{HT}}$ ne dépend pas de la distribution de $Y \mid X_1, X_2,$ mais de μ et σ^2 qui sont constants pour toutes les populations.

La figure 5.1 (2 respectivement) décrit l'évolution de l'$\mathrm{EQMR}_{\mathrm{MC}}$ par rapport à R^2 dans le scénario informatif (scénario quadratique respectivement) pour l'échantillon s_p avec une espérance de taille de 200 (colonne de gauche) et de 50 (colonne de droite). Dans ces deux scénarios, non seulement $\hat{T}^{(\mathrm{Cos})}$ un estimateur plus performant que $\hat{T}^{(\mathrm{MB})}$ ou $\hat{T}^{(\mathrm{MA})}$ (voir le tableau 5.3), mais cet avantage est celui qui s'accroît le plus avec R^2. $\hat{T}^{(\mathrm{MA})}$ gagne aussi en précision, mais quelque peu plus lentement. L'EQM de $\hat{T}^{(\mathrm{MB})}$ se dégrade avec R^2, parce que le modèle de prédiction diffère trop du modèle de population dans ces scénarios, ce qui implique un plus grand biais de $\hat{T}^{(\mathrm{MB})}$ quand R^2 augmente. Dans les scénarios informatif et quadratique, une moindre taille veut dire moins de différence entre les $\mathrm{EQMR}_{\mathrm{MC}}(\hat{T})$ des prédicteurs QR.

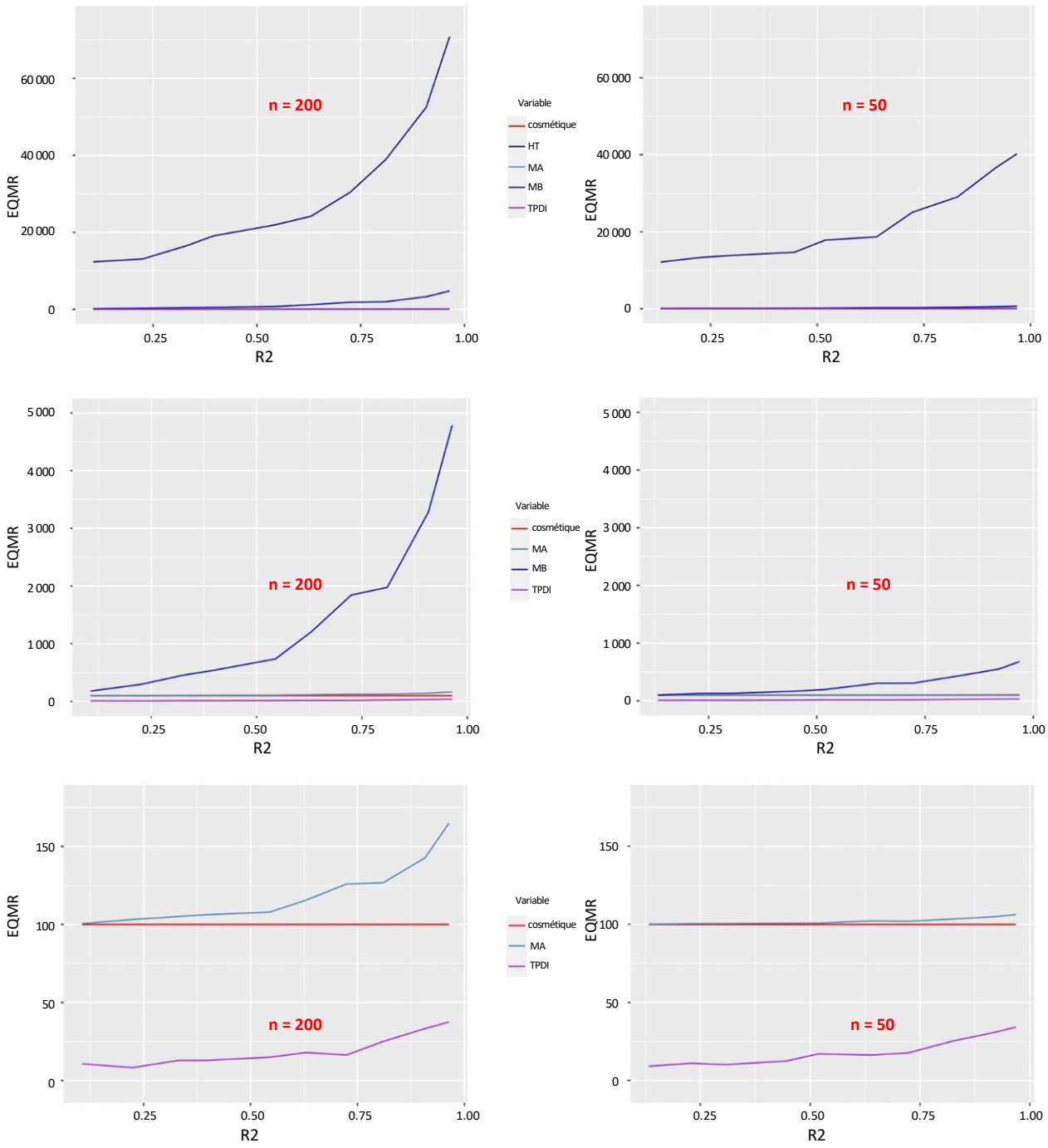
La figure 5.3 décrit l'évolution de l'$\mathrm{EQMR}_{\mathrm{MC}}$ en fonction de R^2 dans le scénario non informatif. Cette fois, $\hat{T}^{(\mathrm{MB})}$ ne perd pas de précision quand R^2 augmente, puisque le modèle de prédiction est identique au modèle de population. Tous les prédicteurs QR gagnent en précision avec R^2; $\hat{T}^{(\mathrm{Cos})}$ et $\hat{T}^{(\mathrm{MA})}$ ayant une précision similaire pour toutes les valeurs de R^2. Dans ce scénario, les graphiques sont comparables pour les deux tailles d'échantillon, car le modèle est correctement spécifié pour tous les modèles de prédiction.

Figure 5.1 EQM relative (en %) avec l'EQM de l'estimateur cosmétique comme référence, par rapport à R^2 dans le scénario informatif.



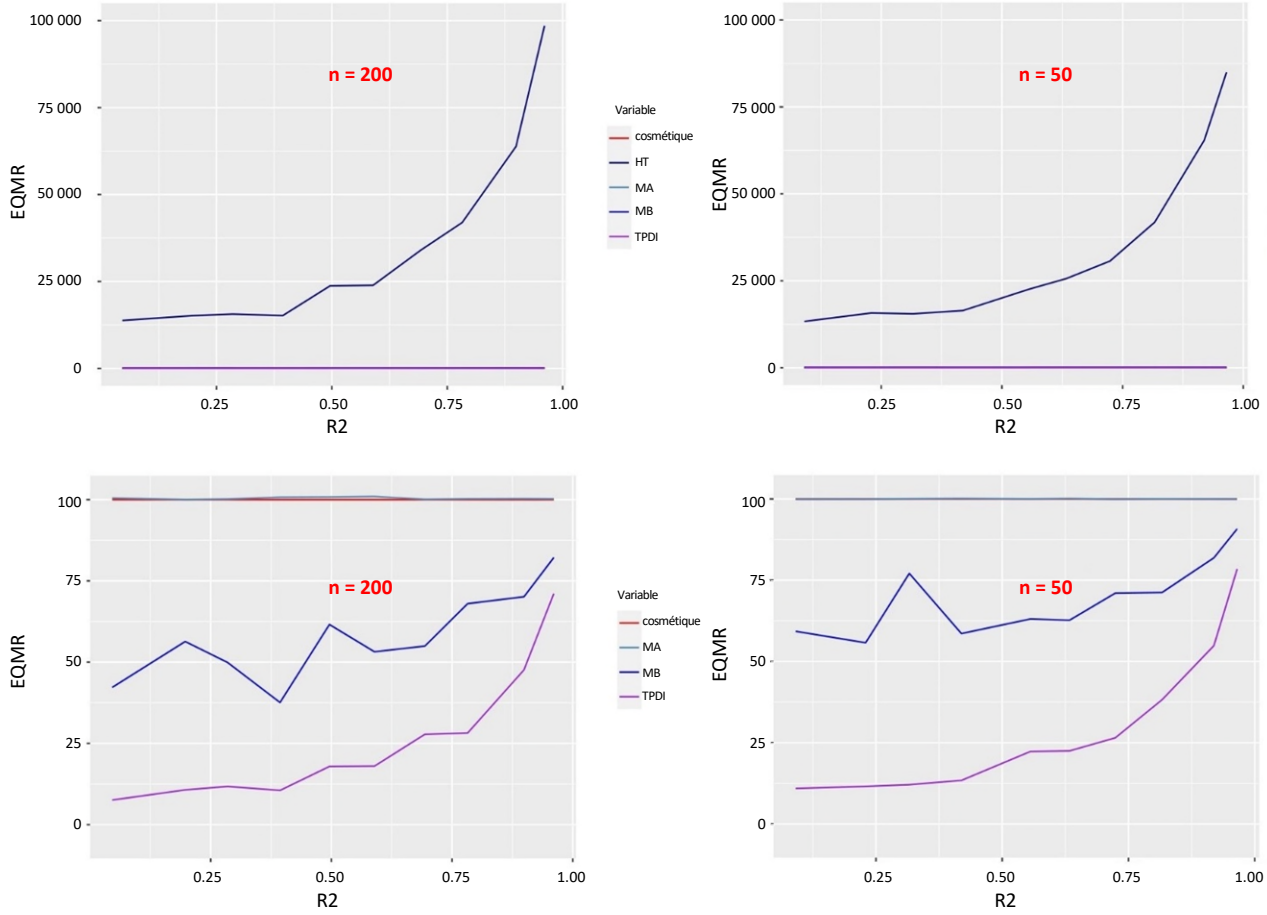
Note : EQMR = erreur quadratique moyenne relative.

Figure 5.2 EQM relative (en %) avec l'EQM de l'estimateur cosmétique comme référence, par rapport à R^2 dans le scénario quadratique.



Note : EQMR = erreur quadratique moyenne relative.

Figure 5.3 EQM relative (en %) avec l'EQM de l'estimateur cosmétique comme référence, par rapport à R^2 dans le scénario non informatif.



Note : EQMR = erreur quadratique moyenne relative.

Pour résumer, si le modèle de prédiction est mal spécifié, l'estimateur cosmétique est le meilleur choix dans tous nos scénarios. De tous les prédicteurs QR, il présente l'EQM la plus faible et il gagne plus rapidement en précision avec R^2 que les autres estimateurs. L'avantage qu'offre $\hat{T}^{(\mathrm{Cos})}$ sur $\hat{T}^{(\mathrm{MA})}$ pourrait toutefois disparaître dans un scénario où la taille d'échantillon probabiliste serait une fraction moindre de la taille de la population. L'estimateur $\hat{T}^{(\mathrm{MB})}$ accuse un biais et présente l'EQM la plus importante même pour des valeurs moindres de R^2. Si le modèle est correctement spécifié et que Y n'est pas corrélé avec X_1 alors que les probabilités d'inclusion du premier ordre sont proportionnelles à X_1, $\hat{T}^{(\mathrm{MB})}$ est le meilleur choix pour l'EQM. Cependant, le gain d'efficacité qu'apporte $\hat{T}^{(\mathrm{MB})}$ comparativement à $\hat{T}^{(\mathrm{Cos})}$ dans ce troisième scénario est significativement inférieur à la perte d'efficacité observée si on choisit $\hat{T}^{(\mathrm{MB})}$ plutôt que $\hat{T}^{(\mathrm{Cos})}$ dans les deux premiers scénarios. Nous recommandons donc de choisir l'estimateur cosmétique qui est un bon compromis dans tous les scénarios, suivi de près par l'estimateur assisté par un modèle. Des observations semblables peuvent se faire à partir de données réelles, comme nous le verrons à la prochaine section.

6. Application aux données de La Poste

6.1 Présentation des données

En France, plus de 90 % des lettres acheminées par la poste sont triées mécaniquement. Les informations recueillies par les machines de tri reposent sur des photographies des lettres et constituent notre base de données non probabiliste. La Poste a également accès à un échantillon probabiliste et souhaite utiliser les deux bases de données pour estimer les totaux mensuels des types de lettres en employant les méthodes d'intégration de données que nous avons proposées précédemment. Certaines lettres, comme les « *lettres vertes* » (lettres qui portent un timbre vert et font l'objet d'un transport écologique) ne sont pas reconnues par les machines de tri qui saisissent seulement des images en noir et blanc, avec pour conséquence que la variable d'intérêt n'est pas présente dans la base non probabiliste. L'information auxiliaire dans le cas des lettres triées automatiquement n'est pas facile à coupler avec l'échantillon probabiliste. C'est une question qui est actuellement étudiée par La Poste. Ainsi, l'exemple qui suit repose sur des données recueillies lors d'enquêtes antérieures réalisées au fil des ans à La Poste.

Les données de ces enquêtes sont recueillies par l'entremise des tournées de facteurs et portent en particulier sur le nombre de *lettres vertes* et de « *produits 1b* » (qui comprennent plusieurs types de lettres, dont les *lettres vertes*) ainsi que le nombre total de lettres de la tournée. L'idée est d'imiter la situation à La Poste où le nombre de *lettres vertes* figure dans l'échantillon probabiliste mais pas dans la base de données non probabiliste, alors que le nombre de *produits 1b* et le nombre total de lettres figurent dans la base de données non probabiliste mais pas dans l'échantillon probabiliste.

L'objectif de la présente section est de voir, d'une part, si la conclusion qui se dégage des populations simulées se vérifie pour les données réelles et, d'autre part, si la méthode de sélection de l'échantillon non probabiliste influe sur les estimateurs. La population d'intérêt est constituée de 11 906 tournées issues des données historiques. Dans ce qui suit, nous supposons que la variable d'intérêt est le nombre de *lettres vertes* et que la variable explicative X_1 (X_2 respectivement) est le nombre total de lettres (nombre de *produits 1b* respectivement).

Nous considérons trois scénarios. Pour chaque scénario, un échantillon non probabiliste d'une taille de 9 524 est tiré de la population; la fraction d'échantillonnage non probabiliste est donc de 80 %. Dans le premier scénario, l'échantillon non probabiliste est sélectionné par échantillonnage aléatoire simple sans remise (EASSR). Dans le deuxième scénario (le troisième respectivement), l'échantillon non probabiliste contient les 9 524 tournées ayant les valeurs de Y les plus élevées (les plus faibles respectivement). Seul X_2 sert de variable explicative dans le modèle de prédiction avec l'ordonnée à l'origine.

Nous comparons les mêmes estimateurs qu'à la section 5.2. Pour chacun des trois échantillons non probabilistes, nous tirons $L = 1\,000$ échantillons probabilistes s_P d'une taille attendue de 2 000 par échantillonnage de Poisson et avec des probabilités proportionnelles à X_1. Nous calculons les mêmes mesures de Monte Carlo qu'à la section 5.2.

6.2 Résultats

Le tableau 6.1 présente les résultats de simulation des données de La Poste pour les trois échantillons non probabilistes. Dans tous les scénarios, le modèle de prédiction semble être significativement mal spécifié, ce qui rend $\hat{T}^{(\mathrm{MB})}$ moins efficace, tant pour la variance que pour l'EQM, que dans les simulations à la section 5.2. Nous pourrions calculer différentes statistiques de diagnostic pour repérer les défauts de spécification du modèle et nous pourrions proposer des modèles prédictifs alternatifs. Mais l'évaluation de tels modèles dépasse la portée du présent document.

Tableau 6.1
Biais relatif (en % de la valeur réelle), variance relative (en %) à $\hat{T}^{(\mathrm{Cos})}$ et EQM des différents estimateurs pour les trois échantillons non probabilistes; l'espérance de la taille de l'échantillon probabiliste est de 2 000 et la taille de l'échantillon non probabiliste, est de 9 524.

S_{NP}	Mesures de Monte Carlo	$\hat{T}_{\mathrm{HT}}$	$\hat{T}^{(\mathrm{MB})}$	$\hat{T}^{(\mathrm{MA})}$	$\hat{T}^{(\mathrm{Cos})}$	$\hat{T}_{\mathrm{PDI}}$
EASSR	$\mathrm{BR}_{\mathrm{MC}}$	0,08	8,30	0,08	0,08	0,06
	$\mathrm{VarR}_{\mathrm{MC}}$	253,67	122,66	102,21	100,00	87,41
	$\mathrm{EQMR}_{\mathrm{MC}}$	252,95	5 489,24	102,18	100,00	87,31
Valeurs les plus élevées de Y	$\mathrm{BR}_{\mathrm{MC}}$	0,03	5,51	0,03	0,03	0,01
	$\mathrm{VarR}_{\mathrm{MC}}$	1 528,57	197,54	101,82	100,00	3,78
	$\mathrm{EQMR}_{\mathrm{MC}}$	1 528,57	13 676,65	101,87	100,00	3,82
Valeurs les plus faibles de Y	$\mathrm{BR}_{\mathrm{MC}}$	-0,02	1,23	0,13	0,13	0,41
	$\mathrm{VarR}_{\mathrm{MC}}$	197,09	100,67	100,26	100,00	93,17
	$\mathrm{EQMR}_{\mathrm{MC}}$	195,36	184,83	100,24	100,00	92,95

Notes : BR = biais relatif; VarR = variance relative; EQMR = erreur quadratique moyenne relative; MC = Monte Carlo; EASSR = échantillonnage aléatoire simple sans remise.

Pour tous les scénarios, les résultats sont semblables à ceux que nous avons obtenus dans les scénarios informatif et quadratique à la section 5.2, $\hat{T}^{(\mathrm{MA})}$ et $\hat{T}^{(\mathrm{Cos})}$ ayant une efficacité similaire et étant l'un et l'autre plus efficaces que $\hat{T}^{(\mathrm{MB})}$, qui est le seul estimateur entaché d'un biais. Il convient de mentionner que, bien que le choix de s_{NP} n'influe pas sur l'efficacité relative de $\hat{T}^{(\mathrm{MA})}$, il agit sur la précision relative des estimateurs de Horvitz-Thompson et PDI.

Dans tous les scénarios, s_P est prélevé par échantillonnage de Poisson et la variance des estimateurs par intégration de données peut se simplifier de la manière suivante :

$$\mathrm{Var}(\hat{T}) = \mathrm{Var}(\hat{T}_{\mathrm{NP}}) + \mathrm{Var}(\hat{T}_C^{(\mathrm{Ha})}),$$

avec $\hat{T}_{\mathrm{NP}}$ le prédicteur du total T_{NP} et $\hat{T}_C^{(\mathrm{Ha})}$ l'estimateur de Hájek du total T_C. Pour mieux comprendre l'effet de la sélection de s_{NP} sur les estimateurs, nous étudions la variance de $\hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}$, de $\hat{T}_{\mathrm{NP}}^{(\mathrm{Cos})}$ et de $\hat{T}_C^{(\mathrm{Ha})}$, ainsi que la variance de l'estimateur de Hájek $\hat{T}_{\mathrm{NP}}^{(\mathrm{Ha})} = N_{\mathrm{NP}} \sum_{k \in s_P} d_k \delta_k y_k \Big/ \sum_{k \in s_P} d_k \delta_k$ du total T_{NP}.

Le tableau 6.2 donne la variance relative de $\hat{T}_{\mathrm{HT}}$, $\hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}$, $\hat{T}_{\mathrm{NP}}^{(\mathrm{Cos})}$, $\hat{T}_C^{(\mathrm{Ha})}$ et $\hat{T}_{\mathrm{NP}}^{(\mathrm{Ha})}$ pour les deuxième et troisième scénarios, où s_{NP} contient les valeurs les plus élevées ou les plus faibles de Y relativement à leur

variance dans la première configuration (EASSR). Comme prévu, la précision de l'estimateur de HT ne dépend pas de s_{NP} et est fixe pour les trois scénarios. Dans le deuxième scénario, $\hat{T}_C^{(\mathrm{Ha})}$ est d'une variance bien moindre que dans les autres scénarios, puisque seules les valeurs les plus faibles de Y restent dans $U - s_{\mathrm{NP}}$. De même, sa variance est plus grande dans le troisième scénario où seules les valeurs les plus élevées de Y restent dans $U - s_{\mathrm{NP}}$. Le même raisonnement explique le rapport des variances entre les scénarios pour $\hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}$, $\hat{T}_{\mathrm{NP}}^{(\mathrm{Cos})}$ et $\hat{T}_{\mathrm{NP}}^{(\mathrm{Ha})}$. Le point important est que tant $\hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}$ que $\hat{T}_{\mathrm{NP}}^{(\mathrm{Cos})}$ ne sont pas très sensibles à s_{NP} pour ce qui est de la variance.

Tableau 6.2
Variance relative en pourcentage des estimateurs lorsque s_{NP} contient les valeurs les plus élevées ou les plus faibles de Y, relativement à leur variance lorsque s_{NP} est sélectionné par EASSR.

S_{NP}	$\hat{T}_{\mathrm{HT}}$	$\hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}$	$\hat{T}_{\mathrm{NP}}^{(\mathrm{Cos})}$	$\hat{T}_{\mathrm{NP}}^{(\mathrm{Ha})}$	$\hat{T}_C^{(\mathrm{Ha})}$
Valeurs les plus élevées de Y	101,97	111,39	110,38	101,61	0,92
Valeurs les plus faibles de Y	104,61	73,07	69,03	15,43	181,29

Note : EASSR = échantillonnage aléatoire simple sans remise.

Il est possible de faire valoir que les échantillons non probabilistes sélectionnés dans les deuxième et troisième scénarios ne sont pas réalistes. Pourtant, les résultats observés pour des échantillons non probabilistes plus réalistes, sélectionnés en utilisant l'échantillonnage de Poisson proportionnel à X_1 ou Y, ne sont pas significativement différents des résultats observés avec s_{NP} sélectionné par EASSR, et ne sont pas présentés. De même, l'utilisation des deux variables auxiliaires dans le modèle de prédiction influe seulement sur le biais de $\hat{T}^{(\mathrm{MB})}$, et une fois de plus, les résultats ne sont pas présentés.

Afin d'améliorer la précision, nous pourrions employer d'autres estimateurs de T_C. Comme nous l'avons mentionné précédemment, les méthodes d'intégration de données que nous proposons dans le présent document peuvent être utiles lorsque le recouvrement entre échantillon probabiliste et échantillon non probabiliste n'est pas nul et, idéalement, important. Dans ce contexte et en fonction de nos résultats empiriques, nous recommandons l'estimateur cosmétique ou l'estimateur assisté par un modèle.

7. Conclusion

La plupart des études consacrées à l'intégration de données dans le cas d'une population finie traitent du problème de la variable d'intérêt qui n'est pas observée dans l'échantillon probabiliste. Dans le présent article, nous avons étudié le problème de la variable d'intérêt inobservée dans l'échantillon non probabiliste en supposant qu'elle était observée dans l'échantillon probabiliste et qu'une information auxiliaire était disponible dans les deux échantillons. Nous avons défini une famille générale d'estimateurs de prédiction, à partir de la famille QR déjà connue, qui comprend l'estimateur assisté par un modèle, l'estimateur fondé sur un modèle et l'estimateur cosmétique. Nous avons étudié d'un point de vue théorique leurs propriétés

de biais et de variance. Nous avons aussi proposé un estimateur de variance et comparé les trois types d'estimateurs à l'estimateur habituel de Horvitz-Thompson selon différents scénarios de simulation, et ce, tant pour le biais que pour l'EQM. Nous avons conclu que l'estimateur cosmétique représente en général un bon compromis.

La principale conclusion que nous tirons de nos expériences est que des gains significatifs d'efficacité sont possibles si nous faisons appel à une grande base de données non probabiliste renfermant de l'information auxiliaire liée aux principales variables d'intérêt. Dans le cas des grands domaines, les gains d'efficacité obtenus grâce à l'utilisation de l'estimateur assisté par un modèle, qui inclut l'estimateur cosmétique, peuvent suffire à l'obtention d'estimations de bonne qualité des paramètres de population d'intérêt. Dans le cas des petits domaines, ces estimateurs pourraient ne pas donner la précision recherchée. Ils peuvent néanmoins faire fonction d'estimateurs directs dans un modèle d'estimation sur petits domaines comme dans le modèle bien connu de Fay-Herriot. Ce modèle exige de l'information auxiliaire au niveau du domaine. La grande base de données non probabiliste se prêterait naturellement à l'obtention de l'information auxiliaire nécessaire à la production d'estimations sur petits domaines. Les méthodes d'estimation sur petits domaines offrent souvent une hausse significative de précision par rapport aux estimateurs directs au prix de l'introduction d'hypothèses de modélisation.

Remerciements

Nous aimerions remercier le rédacteur associé et les deux examinateurs de leurs suggestions et commentaires constructifs qui nous ont permis d'améliorer considérablement notre article. Le présent travail a été en partie soutenu par l'Agence nationale de la recherche en France (contrat CIFRE 2019/1966) et par le programme d'investissements d'avenir (subvention ANR-17-EURE-0010).

Annexe

Preuve de la proposition 2.1

Rappelons que $\hat{T}_{\mathrm{DI}} = \sum_{k\in U} \delta_k y_k + \sum_{k\in s_P} (1-\delta_k)\, d_k y_k$ et $\hat{T}_{\mathrm{HT}} = \sum_{k\in s_P} d_k y_k = \sum_{k\in s_P} \delta_k d_k y_k + \sum_{k\in s_P} (1-\delta_k)\, d_k y_k$. Ainsi, nous avons :

$$\mathrm{Var}(\hat{T}_{\mathrm{HT}}) - \mathrm{Var}(\hat{T}_{\mathrm{DI}}) = \mathrm{Var}\left(\sum_{k\in s_P} \delta_k d_k y_k\right) + 2\,\mathrm{Cov}\left(\sum_{k\in s_P} \delta_k d_k y_k, \sum_{k\in s_P} (1-\delta_k)\, d_k y_k\right).$$

(i) En échantillonnage de Poisson, nous avons :

$$\mathrm{Cov}\left(\sum_{k\in s_P} \delta_k d_k y_k, \sum_{k\in s_P} (1-\delta_k)\, d_k y_k\right) = \sum_{k\in U} \delta_k (1-\delta_k)\,(d_k - 1)\, y_k^2 = 0$$

et

$$\mathrm{Var}(\hat{T}_{\mathrm{HT}}) - \mathrm{Var}(\hat{T}_{\mathrm{DI}}) \;=\; \mathrm{Var}\left(\sum_{k\in s_P}\delta_k d_k y_k\right) = \sum_{k\in U}\delta_k\,(d_k-1)\,y_k^2 \geq 0,$$

ce qui prouve la première partie de la proposition.

(ii) En échantillonnage aléatoire simple sans remise, supposons que $\overline{Y}_U = \sum_{k\in U} y_k \big/ N$, $\overline{Y}_{\mathrm{NP}} = \sum_{k\in U}\delta_k y_k \big/ N_{\mathrm{NP}}$, $S^2_{Y,\mathrm{NP}} = \sum_{k\in U}\delta_k (y_k - \overline{Y}_{\mathrm{NP}})^2 \big/ (N_{\mathrm{NP}}-1)$ et $\mathrm{CV}^2_{\mathrm{NP}} = S^2_{Y,\mathrm{NP}} \big/ \overline{Y}^2_{\mathrm{NP}}$. Un simple calcul nous donne :

$$\mathrm{Var}\left(\sum_{k\in s_P}\delta_k d_k y_k\right) \;=\; \frac{N}{n}\,\frac{N-n}{N(N-1)}\left(N\,(N_{\mathrm{NP}}-1)\,S^2_{Y,\mathrm{NP}} + N_{\mathrm{NP}}\overline{Y}^2_{\mathrm{NP}}\left(N-N_{\mathrm{NP}}\right)\right),$$

$$\mathrm{Cov}\left(\sum_{k\in s_P}\delta_k d_k y_k, \sum_{k\in s_P}(1-\delta_k)\,d_k y_k\right) \;=\; -\frac{N}{n}\,\frac{N-n}{N(N-1)}\,N_{\mathrm{NP}}\overline{Y}_{\mathrm{NP}}\left(N\overline{Y}_U - N_{\mathrm{NP}}\overline{Y}_{\mathrm{NP}}\right),$$

et, par conséquent,

$$\mathrm{Var}(\hat{T}_{\mathrm{HT}}) - \mathrm{Var}(\hat{T}_{\mathrm{DI}}) \;=\; \frac{N}{n}\,\frac{N-n}{N(N-1)}\left(N\,(N_{\mathrm{NP}}-1)\,S^2_{Y,\mathrm{NP}} + N_{\mathrm{NP}}\overline{Y}_{\mathrm{NP}}\left((N+N_{\mathrm{NP}})\,\overline{Y}_{\mathrm{NP}} - 2N\overline{Y}_U\right)\right).$$

Nous concluons que $\mathrm{Var}(\hat{T}_{\mathrm{HT}})$ est supérieur ou égal à $\mathrm{Var}(\hat{T}_{\mathrm{DI}})$ si et seulement si

$$N\,(N_{\mathrm{NP}}-1)\,S^2_{Y,\mathrm{NP}} + N_{\mathrm{NP}}\overline{Y}_{\mathrm{NP}}\left((N+N_{\mathrm{NP}})\,\overline{Y}_{\mathrm{NP}} - 2N\overline{Y}_U\right) \;\geq\; 0,$$

ce qui équivaut à :

$$\mathrm{CV}^2_{\mathrm{NP}} \;\geq\; -\frac{N_{\mathrm{NP}}}{N_{\mathrm{NP}}-1}\left(1 + \frac{N_{\mathrm{NP}}}{N} - 2\,\frac{\overline{Y}_U}{\overline{Y}_{\mathrm{NP}}}\right),$$

et ce qui prouve la seconde partie de la proposition.

Preuve de la proposition 2.2

Nous avons :

$$\mathrm{Var}(\hat{T}_{\mathrm{HT}}) \;=\; \mathrm{Var}\left(\sum_{k\in s_P} d_k y_k\right) \;=\; N^2(1-f)\,\frac{S^2_{Y,U}}{n},$$

$$\mathrm{Avar}(\hat{T}_{\mathrm{PDI}}) \;=\; \mathrm{Var}\left(\sum_{k\in s_P}(1-\delta_k)\,d_k\,(y_k-\overline{Y}_C)\right) \;=\; \mathrm{Var}\left(\sum_{k\in s_P} d_k\tilde{y}_k\right) \;=\; N^2\,(1-f)\,\frac{S^2_{\tilde{Y},U}}{n}$$

où

$$\begin{aligned} S^2_{Y,U} \;&=\; \frac{1}{N-1}\sum_{k\in U}(y_k-\overline{Y}_U)^2, \\ \tilde{y}_k \;&=\; (1-\delta_k)\,(y_k-\overline{Y}_C),\, k\in U \\ S^2_{\tilde{Y},U} \;&=\; \frac{1}{N-1}\sum_{k\in U}(\tilde{y}_k-\overline{\tilde{Y}}_C)^2 \;=\; \frac{1}{N-1}\sum_{k\in U}\tilde{y}_k^2. \end{aligned}$$

Un peu de calcul, élémentaire mais fastidieux, nous donne :

$$\begin{aligned}\mathrm{Var}(\hat{T}_{\mathrm{HT}}) - \mathrm{Avar}(\hat{T}_{\mathrm{PDI}}) &= N^2(1-f)\frac{S^2_{Y,U} - S^2_{\tilde{Y},U}}{n} \\ &= N^2(1-f)\frac{1}{n}\frac{1}{N-1}\left(\sum_{k\in U}\delta_k (y_k - \bar{Y}_U)^2 + (N - N_{\mathrm{NP}})(\bar{Y}_C - \bar{Y}_U)^2\right) \\ &= N^2(1-f)\frac{1}{n}\frac{1}{N-1}\left(S^2_{Y,\mathrm{NP}}(N_{\mathrm{NP}}-1) + N_{\mathrm{NP}}\frac{N}{N-N_{\mathrm{NP}}}(\bar{Y}_{\mathrm{NP}} - \bar{Y}_U)^2\right).\end{aligned}$$

Preuve de la proposition 3.1

Soit $\mathbf{R}_{s_P} = \mathrm{diag}(r_k\delta_k)_{k\in s_P}$, $\mathbf{X}_{s_P} = (\mathbf{x}_k^\top)_{k\in s_P}$, $\mathbf{y}_{s_P} = (y_k)_{k\in s_P}$ et $\mathbf{Q}^\top_{xs_P} = \mathbf{X}^\top_{s_P}\mathrm{diag}(q_k\delta_k)_{k\in s_P}$. Dans ce cas, $\hat{\boldsymbol{\beta}} = (\mathbf{Q}^\top_{xs_P}\mathbf{X}_{s_P})^{-1}\mathbf{Q}^\top_{xs_P}\mathbf{y}_{s_P}$. Nous pouvons écrire la somme $\sum_{k\in s_P} r_k\delta_k(y_k - \hat{y}_k)$ sous la forme matricielle suivante :

$$\sum_{k\in s_P} r_k\delta_k(y_k - \hat{y}_k) = \mathbf{1}^\top_{s_P}\mathbf{R}_{s_P}(\mathbf{y}_{s_P} - \mathbf{X}_{s_P}\hat{\boldsymbol{\beta}}),$$

où $\mathbf{1}_{s_P}$ est un vecteur des valeurs 1 ayant comme dimension la taille de s_P. Si la condition $\boldsymbol{\mu}^\top\mathbf{x}_k q_k - r_k = 0$ est remplie pour tout $k \in s_{\mathrm{NP}}$, alors $\delta_k(\boldsymbol{\mu}^\top\mathbf{x}_k q_k - r_k) = 0$ pour tout $k \in s_P$ et, par conséquent, $\boldsymbol{\mu}^\top\mathbf{Q}^\top_{xs_P} = \mathbf{1}^\top_{s_P}\mathbf{R}_{s_P}$. Nous obtenons $\mathbf{1}^\top_{s_P}\mathbf{R}_{s_P}(\mathbf{y}_{s_P} - \mathbf{X}_{s_P}\hat{\boldsymbol{\beta}}) = 0$.

Preuve de la proposition 3.2

Nous avons :

$$\begin{aligned}\hat{T}^{(\mathrm{QR})}_{\mathrm{NP}} - \hat{T}^{(\mathrm{Q}\pi)}_{\mathrm{NP}} &= \sum_{k\in s_P}(r_k - d_k)\,\delta_k(y_k - \mathbf{x}_k^\top\hat{\boldsymbol{\beta}}) \\ &= -\boldsymbol{\lambda}^\top\sum_{k\in s_P} q_k\delta_k\mathbf{x}_k(y_k - \mathbf{x}_k^\top\hat{\boldsymbol{\beta}}) = 0.\end{aligned}$$

Bibliographie

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81(396), 1074-1079.

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.

Breidt, F.J., et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4), 1026-1053.

Brewer, K.R.W. (1999). Le calage esthétique dans le cas de l'échantillonnage avec probabilités inégales. *Techniques d'enquête*, 25, 2, 231-239. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1999002/article/4883-fra.pdf.

Cardot, H., Goga, C. et Lardin, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic Journal of Statistics*, 7, 562-596.

Goga, C., Deville, J.-C. et Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample data. *Biometrika*, 96(3), 691-709.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663-685.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal American Statistical Association*, 77(377), 89-96.

Kim, J.-K. (2022). A gentle introduction to data integration in survey sampling. *The Survey Statistician*, 85, 19-29.

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

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4), 558-625.

Rao, J.N.K. (2005). Évaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage. *Techniques d'enquête*, 31, 2, 127-151. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2005002/article/9040-fra.pdf.

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

Särndal, C.-E. (1980). On π -inverse weighting best linear unbiased weighting in probability sampling. *Biometrika*, 67(3), 639-650.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics. New York: Springer-Verlag.

Särndal, C.-E., et Wright, R.L. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11(3), 146-156.

Wright, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78(384), 879-884.

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., et Kim, J.-K. (2020). Integration of survey data and big observational data for finite population inference using mass imputation. *Japanese Journal of Statistics and Data Science*, 3, 625-650.