Nº 11-522-x au catalogue
ISSN : 1709822-X

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Une approche bayésienne approximative pour améliorer les estimateurs d'un échantillon probabiliste à l'aide d'un échantillon non probabiliste supplémentaire

par Yong You, Abel DaSylva et Jean-François Beaumont

Date de diffusion : le 29 octobre 2021

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Une approche bayésienne approximative pour améliorer les estimateurs d'un échantillon probabiliste à l'aide d'un échantillon non probabiliste supplémentaire

Yong You, Abel DaSylva et Jean-François Beaumont[1]

Résumé

Les organismes nationaux de statistique étudient de plus en plus la possibilité d'utiliser des échantillons non probabilistes en complément des échantillons probabilistes. Nous examinons le scénario où la variable d'intérêt et les variables auxiliaires sont observées à la fois dans un échantillon probabiliste et un échantillon non probabiliste. Nous cherchons à utiliser les données de l'échantillon non probabiliste pour améliorer l'efficacité des estimations pondérées par les poids d'enquête obtenues à partir de l'échantillon probabiliste. Récemment, Sakshaug, Wiśniowski, Ruiz et Blom (2019) et Wiśniowski, Sakshaug, Ruiz et Blom (2020) ont proposé une approche bayésienne visant à intégrer les données des deux échantillons aux fins de l'estimation des paramètres du modèle. Dans leur méthode, on utilise les données de l'échantillon non probabiliste pour déterminer la distribution a priori des paramètres du modèle et on obtient la distribution a posteriori en supposant que le plan de sondage probabiliste est ignorable (ou non informatif). Nous étendons cette approche bayésienne à la prédiction de paramètres d'une population finie dans le cadre d'un échantillonnage non ignorable (ou informatif) en nous appuyant sur des statistiques pondérées par des poids d'enquête appropriées. Nous illustrons les propriétés de notre prédicteur au moyen d'une étude par simulations.

Mots clés : prédiction bayésienne; échantillonnage de Gibbs; échantillonnage non ignorable; intégration des données statistiques.

1. Introduction

Les organismes gouvernementaux considèrent de plus en plus les échantillons non probabilistes comme un moyen de réduire les coûts d'enquête et d'obtenir des estimations plus actuelles que la plupart des estimations d'enquêtes probabilistes. En particulier, à la suite de la pandémie de COVID-19, Statistique Canada a lancé une série d'enquêtes en ligne auprès de volontaires, appelés enquêtes par approche participative, afin d'obtenir de l'information sur différents aspects de la vie de la population canadienne. Cependant, on sait depuis des décennies que l'utilisation d'échantillons non probabilistes seulement, comme les échantillons par approche participative, peut mener à des estimations présentant un biais de sélection important. Ces dernières années, des recherches en nombre croissant s'intéressent aux méthodes qui utilisent les données auxiliaires d'un échantillon probabiliste pour réduire le biais de sélection des estimateurs d'échantillons non probabilistes. Ces méthodes s'appliquent dans un scénario où les variables d'intérêt sont observées seulement dans l'échantillon non probabiliste, mais les variables auxiliaires courantes s'observent dans les deux échantillons. La pondération par le score de propension (p. ex. Chen, Li et Wu, 2020) fournit une méthode d'intégration des données des deux échantillons. Elle consiste à pondérer les participants de l'échantillon non probabiliste par l'inverse de leur probabilité de participation estimée. L'appariement statistique ou l'appariement d'échantillons (p. ex. Yang, Kim et Hwang 2021; ou Rivers, 2007) est une autre solution. Il consiste à imputer les variables d'intérêt manquantes dans l'échantillon probabiliste au moyen de données de l'échantillon non probabiliste. À propos de l'intégration des données statistiques dans ce scénario, citons les études récentes dans Beaumont (2020), Elliot et Valliant (2017), Rao (2021) et Valliant (2020).

[1]Yong You, Abel Dasylva et Jean-François Beaumont, Centre de collaboration internationale et d'innovation en méthodologie (CCIIM), Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), K1A 0T6. Adresse courriel de contact : yong.you@statcan.gc.ca, abel.dasylva@statcan.gc.ca et jean-francois.beaumont@statcan.gc.ca.
Avertissement : Cet article expose les opinions des auteurs qui ne sont pas nécessairement celles de Statistique Canada. Il décrit des méthodes théoriques qui pourraient ne pas correspondre à celles qu'emploie l'organisme.

Nous envisageons un scénario d'intégration des données différent, dans lequel les variables d'intérêt et les variables auxiliaires sont observées dans les deux échantillons. Nous ne supposons pas que les totaux de population des variables auxiliaires sont connus ni que l'indicateur de participation dans l'échantillon non probabiliste est observé dans l'échantillon probabiliste. La littérature concernant ce scénario est rare. Elliott et Haviland (2007) ont proposé un estimateur composite, qui est tout simplement une moyenne pondérée des estimateurs de l'échantillon non probabiliste et de l'échantillon probabiliste. Ils indiquent que l'estimateur proposé nécessite un échantillon probabiliste relativement grand pour que le biais de l'estimateur de l'échantillon non probabiliste puisse être estimé avec une précision suffisante. Plus récemment, des estimateurs bayésiens qui intègrent les deux échantillons ont été proposés par Sakshaug, Wiśniowski, Ruiz et Blom (2019), Wiśniowski, Sakshaug, Ruiz et Blom (2020) et Nandram et Rao (2021). En supposant un plan de sondage probabiliste ignorable, Sakshaug et coll. (2019) et Wiśniowski et coll. (2020) ont proposé et évalué au moyen d'études par simulations une approche bayésienne intéressante qui intègre les données des deux échantillons pour estimer les paramètres du modèle. Leur méthode comporte des distributions a priori intrigantes pour les paramètres du modèle : les données de l'échantillon non probabiliste servent à déterminer la moyenne a priori, mais les données des deux échantillons sont utilisées pour la détermination de la variance a priori. En effet, le biais estimé de l'estimateur de l'échantillon non probabiliste sert à gonfler la variance a priori. Dans le présent article, nous étendons leur approche bayésienne à l'estimation des moyennes de population finie selon un plan de sondage probabiliste non ignorable et l'évaluons au moyen d'une étude par simulations.

2. Problème d'estimation

Supposons que nous voulons estimer la moyenne de population $\theta = N^{-1}\sum_{i\in U} y_i$, où y_i est la valeur de la variable d'intérêt *y* pour l'unité *i* de la population finie *U* de taille *N*. Un échantillon probabiliste s de taille *n* est tiré de *U* au moyen d'un plan de sondage probabiliste, et la variable d'intérêt y est observée pour toutes les unités $i \in s$. L'estimateur pondéré par les poids d'enquête de la moyenne de la population est $\hat{\theta}_w = \hat{N}^{-1}\sum_{i\in s} w_i y_i$, où $\hat{N} = \sum_{i\in s} w_i$ et w_i est un poids d'enquête pour l'unité d'échantillon *i*. Le poids d'enquête peut être le poids de sondage de base $w_i = 1/\pi_i$, où π_i est la probabilité dans laquelle l'unité *i* est sélectionnée dans s, ou il peut être un poids de calage (p. ex. Deville et Särndal, 1992). La variable d'intérêt *y* est également observée pour toutes les unités d'un échantillon non probabiliste s_{NP} de taille n_{NP}. De plus, nous observons un vecteur de *q* variables auxiliaires, $\boldsymbol{x}_i$, pour toutes les unités dans s et s_{NP}. Nous supposons que $\boldsymbol{x}_i$ comprend une ordonnée à l'origine.

Nous postulons le modèle linéaire

$$y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 \rightarrow N(\boldsymbol{x'}_i\boldsymbol{\beta}, \sigma^2),\ i \in U, \tag{1}$$

où $\boldsymbol{\beta}$ et σ^2 sont des paramètres inconnus de modèle. L'idée de Sakshaug et coll. (2019) et Wiśniowski et coll. (2020) est d'utiliser l'échantillon non probabiliste pour obtenir la moyenne a priori de $\boldsymbol{\beta}$. Comme ces auteurs, nous considérons la distribution a priori suivante pour $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \rightarrow N\big(\widehat{\boldsymbol{\beta}}_{NP}, \boldsymbol{\Phi}_0\big), \tag{2}$$

où $\widehat{\boldsymbol{\beta}}_{NP} = \big(\sum_{i\in s_{NP}} \boldsymbol{x}_i\boldsymbol{x'}_i\big)^{-1}\sum_{i\in s_{NP}} \boldsymbol{x}_i y_i$ est l'estimateur par le maximum de vraisemblance de $\boldsymbol{\beta}$, en supposant que la sélection de l'échantillon non probabiliste est ignorable (voir Rubin, 1976) par rapport au modèle (1), et $\boldsymbol{\Phi}_0$ est une matrice connue de variance-covariance spécifiée. De plus amples détails sur le choix de $\boldsymbol{\Phi}_0$ sont donnés à la section 3.3. Nous supposons une probabilité a priori non informative pour σ^2, avec la fonction de densité de probabilité proportionnelle à σ^{-2}.

En supposant que le plan de sondage probabiliste est ignorable pour les inférences fondées sur un modèle (p. ex. l'échantillon probabiliste est sélectionné au moyen d'un échantillonnage aléatoire simple), Sakshaug et coll. (2019) et Wiśniowski et coll. (2020) estiment $\boldsymbol{\beta}$ à partir de sa moyenne a posteriori en prenant les données observées, $\boldsymbol{Y}_s = \{y_i, i \in s\}$. Notre objectif est différent. Nous voulons prédire la moyenne de population finie $\theta = N^{-1}\sum_{i\in U} y_i$. En supposant que $\boldsymbol{x}_i$ est disponible pour toute la population *U* et un échantillonnage ignorable, nous pouvons calculer la moyenne a posteriori de θcomme étant

$$\tilde{\theta} = E(\theta|\boldsymbol{Y}_s) = N^{-1}\big(\sum_{i\in s} y_i + \sum_{i\in U-s} \boldsymbol{x'}_i\widehat{\boldsymbol{\beta}}\big) = N^{-1}\big(\sum_{i\in U} \boldsymbol{x'}_i\widehat{\boldsymbol{\beta}} + \sum_{i\in s}\big(y_i - \boldsymbol{x'}_i\widehat{\boldsymbol{\beta}}\big)\big), \tag{3}$$

où $\widehat{\boldsymbol{\beta}} = E(\boldsymbol{\beta}|\boldsymbol{Y}_s)$ est la moyenne a posteriori de $\boldsymbol{\beta}$. La moyenne a posteriori de θ, donnée dans (3), est calculable à condition que N et $\sum_{i\in U} \boldsymbol{x}_i$ soient connus. Dans bien des scénarios, ce n'est pas le cas. On peut remplacer les quantités de population inconnues dans (3) par des estimateurs pondérés par les poids d'enquête pour obtenir

$$\hat{\theta} = \hat{N}^{-1}\left(\sum_{i\in s} w_i \boldsymbol{x'}_i \widehat{\boldsymbol{\beta}} + \sum_{i\in s}\left(y_i - \boldsymbol{x'}_i \widehat{\boldsymbol{\beta}}\right)\right). \tag{4}$$

Si la fraction d'échantillonnage $f = n/N$ est petite, le premier membre de (4) peut être estimé approximativement comme suit :

$$\hat{\theta}_a(\widehat{\boldsymbol{\beta}}) = \hat{N}^{-1} \sum_{i\in s} w_i \boldsymbol{x'}_i \widehat{\boldsymbol{\beta}}. \tag{5}$$

L'estimateur par le pseudo maximum de vraisemblance de $\boldsymbol{\beta}$ basé sur les données de l'échantillon probabiliste est

$$\widehat{\boldsymbol{\beta}}_w = \left(\sum_{i\in s} w_i \boldsymbol{x}_i \boldsymbol{x'}_i\right)^{-1} \sum_{i\in s} w_i \boldsymbol{x}_i y_i.$$

Dans des conditions de régularité, $\widehat{\boldsymbol{\beta}}_w$ est convergent pour $\boldsymbol{\beta}$ selon le modèle (1) et le plan de sondage. Il est facile de montrer que $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}_w)$ se réduit à l'estimateur pondéré par les poids d'enquête $\hat{\theta}_w = \hat{N}^{-1} \sum_{i\in s} w_i y_i$, à condition qu'une ordonnée à l'origine soit incluse dans $\boldsymbol{x}_i$. Par conséquent, on s'attend à ce que l'estimateur (5) réalise des gains d'efficacité non négligeables par rapport à l'estimateur pondéré par les poids d'enquête $\hat{\theta}_w$ seulement quand les données des deux échantillons sont combinées pour l'estimation de $\boldsymbol{\beta}$. L'approche bayésienne que nous examinons réalise cette intégration des données en utilisant des données de l'échantillon non probabiliste pour déterminer la moyenne a priori de $\boldsymbol{\beta}$. À partir de (4) et (5), nous observons également que les valeurs attendues de $\hat{\theta}$ et $\hat{\theta}_a$ sont proches de $\hat{\theta}_w$ quand le modèle linéaire (1) a une forte puissance prédictive (petit σ^2) de sorte que $y_i \approx \boldsymbol{x}'_i \widehat{\boldsymbol{\beta}}$. Cela soulève la question suivante : devrait-on utiliser une variable auxiliaire afin d'estimer des paramètres de population finie?

3. Inférence bayésienne

3.1 Plan de sondage probabiliste ignorable

Dans certaines conditions (voir Rubin, 1976), on peut ignorer le plan de sondage probabiliste quand on fait des inférences fondées sur un modèle (conditionnellement à s). Dans ce cas, les résultats types sur la régression linéaire bayésienne peuvent servir à obtenir les distributions a posteriori conditionnelles de $\boldsymbol{\beta}$ et σ^2. La distribution a posteriori conditionnelle de $\boldsymbol{\beta}$ est donnée par

$$\boldsymbol{\beta}|\boldsymbol{Y}_s, \sigma^2 \to N\left(\widehat{\boldsymbol{\beta}}_c, \sigma^2 \boldsymbol{\Phi}_c^{-1}\right), \tag{6}$$

où $\boldsymbol{\Phi}_c = \sum_{i\in s} \boldsymbol{x}_i \boldsymbol{x}'_i + \sigma^2 \boldsymbol{\Phi}_0^{-1}$ et $\widehat{\boldsymbol{\beta}}_c = \boldsymbol{\Phi}_c^{-1}\left(\sum_{i\in s} \boldsymbol{x}_i y_i + \sigma^2 \boldsymbol{\Phi}_0^{-1} \widehat{\boldsymbol{\beta}}_{NP}\right)$.

L'estimateur $\widehat{\boldsymbol{\beta}}_c$ se réduit à l'estimateur non pondéré $\widehat{\boldsymbol{\beta}}_{uw} = \left(\sum_{i\in s} \boldsymbol{x}_i \boldsymbol{x}'_i\right)^{-1} \sum_{i\in s} \boldsymbol{x}_i y_i$ quand une probabilité a priori non informative pour $\boldsymbol{\beta}$ est utilisée (p. ex. quand $\boldsymbol{\Phi}_0^{-1}$ est spécifié proche de $\mathbf{0}$). L'estimateur $\widehat{\boldsymbol{\beta}}_{uw}$ est l'estimateur par le maximum de vraisemblance de $\boldsymbol{\beta}$ selon l'hypothèse que le plan d'échantillonnage est ignorable.

La distribution a posteriori conditionnelle de σ^2 est donnée par

$$\sigma^2|\boldsymbol{Y}_s, \boldsymbol{\beta} \to IG\left(\frac{n}{2}, \frac{n}{2} s_{uw}^2(\boldsymbol{\beta})\right), \tag{7}$$

où $IG(a, b)$ correspond à la distribution gamma inverse avec le paramètre de forme a et le paramètre d'échelle b, et

$$s_{uw}^2(\boldsymbol{\beta}) = n^{-1}\sum_{i\in s}(y_i - \boldsymbol{x}'_i\boldsymbol{\beta})^2.$$

En utilisant (6) et (7), on peut appliquer la procédure d'échantillonnage itératif de Gibbs pour l'inférence bayésienne pour générer un grand nombre de valeurs à partir de la distribution a posteriori $\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}_s$. Ces valeurs générées permettent de calculer approximativement la distribution a posteriori de $\boldsymbol{\beta}$ et σ^2. En particulier, on obtient l'estimation bayésienne de $\boldsymbol{\beta}$ en calculant approximativement la moyenne a posteriori $\widehat{\boldsymbol{\beta}} = E(\boldsymbol{\beta}|\boldsymbol{Y}_s)$ par la moyenne des valeurs générées de $\boldsymbol{\beta}$, qu'on note $\widehat{\boldsymbol{\beta}}^*$. La moyenne de population θ est ensuite estimée par $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*)$.

Selon le modèle (1), on sait bien que

$$\widehat{\boldsymbol{\beta}}_{uw}\big|\boldsymbol{\beta}, \sigma^2 \to N(\boldsymbol{\beta}, \sigma^2(\textstyle\sum_{i\in s}\boldsymbol{x}_i\boldsymbol{x}'_i)^{-1}) \tag{8}$$

et

$$s_{uw}^2(\boldsymbol{\beta})|\boldsymbol{\beta}, \sigma^2 \to G\left(\frac{n}{2}, 2\frac{\sigma^2}{n}\right), \tag{9}$$

où $G(a, b)$ correspond à la distribution gamma avec le paramètre de forme a et le paramètre d'échelle b. En utilisant (8) et (9), on montre facilement que la distribution $\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_{uw}, \sigma^2$ est identique à la distribution a posteriori conditionnelle donnée dans (6), et que la distribution $\sigma^2|s_{uw}^2(\boldsymbol{\beta}), \boldsymbol{\beta}$ est identique à la distribution a posteriori conditionnelle donnée dans (7). Cette observation est essentielle pour comprendre la principale idée motivant l'extension de cette approche au cas d'un plan d'échantillonnage non ignorable. Dans la prochaine section, nous élaborerons une procédure d'échantillonnage de Gibbs pour des plans d'échantillonnage non ignorables au moyen d'analogues pondérés par les poids d'enquête de (8) et (9).

3.2 Plan de sondage probabiliste non ignorable

En pratique, dans une perspective fondée sur un modèle et un plan fréquentiste, l'estimateur non pondéré $\widehat{\boldsymbol{\beta}}_{uw}$ est utilisé seulement quand le poids d'enquête w_i est identique pour tous les $i \in s$. Autrement, il est plus courant d'utiliser l'estimateur pondéré par les poids d'enquête $\widehat{\boldsymbol{\beta}}_w$, car il reste convergent dans un plan de sondage probabiliste non ignorable. Nous faisons l'hypothèse habituelle

$$\widehat{\boldsymbol{\beta}}_w\big|\boldsymbol{\beta}, \sigma^2 \to N(\boldsymbol{\beta}, \sigma^2\boldsymbol{\Psi}), \tag{10}$$

où $\boldsymbol{\Psi} = \sigma^{-2}\, var_{mp}(\widehat{\boldsymbol{\beta}}_w)$. L'indice m fait référence au modèle (1), tandis que l'indice p fait référence au plan de sondage probabiliste. Les conditions de régularité de la validité de (10) sont données dans Fuller (2009). Au moyen d'arguments types (p. ex. Binder et Roberts, 2003), nous obtenons

$$\boldsymbol{\Psi} \approx (\textstyle\sum_{i\in U}\boldsymbol{x}_i\boldsymbol{x}'_i)^{-1} + \sigma^{-2}E_m\left[var_p(\widehat{\boldsymbol{\beta}}_w)\right]. \tag{11}$$

Notons que le premier terme du deuxième membre de (11) est négligeable quand la fraction de sondage f est négligeable. Dans un plan d'échantillonnage ignorable, il est facile de montrer que

$$\boldsymbol{\Psi} = \sigma^{-2}E_p\left[var_m(\widehat{\boldsymbol{\beta}}_w)\right] = E_p[(\textstyle\sum_{i\in s}w_i\boldsymbol{x}_i\boldsymbol{x}'_i)^{-1}(\sum_{i\in s}w_i^2\boldsymbol{x}_i\boldsymbol{x}'_i)(\sum_{i\in s}w_i\boldsymbol{x}_i\boldsymbol{x}'_i)^{-1}],$$

qui ne dépend ni de $\boldsymbol{\beta}$ ni de σ^2.

Par analogie avec le cas de l'échantillonnage ignorable, nous considérons la distribution a posteriori conditionnelle $\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_w, \sigma^2$. L'idée à la fois simple et brillante de traiter l'échantillonnage non ignorable, en conditionnant sur un estimateur convergent, a été proposée par Wang, Kim et Yang (2018). À partir de l'hypothèse (10) et de la distribution a priori (2), nous obtenons la distribution a posteriori conditionnelle

$$\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_w, \sigma^2 \to N\left(\widehat{\boldsymbol{\beta}}_{wc}, \sigma^2\boldsymbol{\Phi}_{wc}^{-1}\right), \tag{12}$$

où $\boldsymbol{\Phi}_{wc} = \boldsymbol{\Psi}^{-1} + \sigma^2\boldsymbol{\Phi}_0^{-1}$ et $\widehat{\boldsymbol{\beta}}_{wc} = \boldsymbol{\Phi}_{wc}^{-1}\left(\boldsymbol{\Psi}^{-1}\widehat{\boldsymbol{\beta}}_w + \sigma^2\boldsymbol{\Phi}_0^{-1}\widehat{\boldsymbol{\beta}}_{NP}\right)$.

Dans une perspective fondée sur un plan et un modèle fréquentiste, $s_{uw}^2(\boldsymbol{\beta})$ n'est pas convergent pour σ^2. Une version pondérée par les poids d'enquête de $s_{uw}^2(\boldsymbol{\beta})$ est $s_w^2(\boldsymbol{\beta}) = \widehat{N}^{-1}\sum_{i\in s} w_i(y_i - \boldsymbol{x}'_i\boldsymbol{\beta})^2$. Dans des conditions de régularité, $s_w^2(\boldsymbol{\beta})$ est asymptomatiquement sans biais et convergent pour σ^2 selon le modèle (1) et le plan de sondage**.** Nous supposons que

$$s_w^2(\boldsymbol{\beta})|\boldsymbol{\beta},\sigma^2 \to G\left(\frac{\lambda}{2}, 2\frac{\sigma^2}{\lambda}\right), \tag{13}$$

où $2\lambda^{-1} = \sigma^{-4}\, var_{mp}[s_w^2(\boldsymbol{\beta})]$. La quantité λ peut être écrite comme étant $\lambda = \frac{n}{D}$, où $D = \left(\frac{2\sigma^4}{n}\right)^{-1} var_{mp}[s_w^2(\boldsymbol{\beta})]$ peut être interprété comme un effet de plan. Dans des conditions de régularité, nous obtenons

$$2\lambda^{-1} \approx \frac{2}{N} + \sigma^{-4}E_m\left\{var_p[s_w^2(\boldsymbol{\beta})]\right\}. \tag{14}$$

Encore une fois, le premier terme du deuxième membre de (14) est négligeable quand la fraction de sondage *f* est négligeable. Selon un plan d'échantillonnage ignorable, nous obtenons

$$2\lambda^{-1} = \sigma^{-4}E_p\{var_m[s_w^2(\boldsymbol{\beta})]\} = 2E_p\left(\frac{\sum_{i\in s} w_i^2}{\widehat{N}^2}\right),$$

qui ne dépend ni de $\boldsymbol{\beta}$ ni de σ^2.

Comme dans le cas de l'échantillonnage ignorable, nous considérons la distribution a posteriori conditionnelle $\sigma^2|s_w^2(\boldsymbol{\beta}), \boldsymbol{\beta}$. En utilisant (13) et une fonction de densité de probabilité a priori pour σ^2 proportionnelle à σ^{-2}, nous obtenons

$$\sigma^2|s_w^2(\boldsymbol{\beta}), \boldsymbol{\beta} \to IG\left(\frac{\lambda}{2}, \frac{\lambda}{2}s_w^2(\boldsymbol{\beta})\right). \tag{15}$$

En utilisant (12) et (15), on peut appliquer la procédure d'échantillonnage de Gibbs pour générer un grand nombre de valeurs a posteriori pour $\boldsymbol{\beta}$ et σ^2. La moyenne de population θ est estimée par $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}_{NI}^*)$, où $\widehat{\boldsymbol{\beta}}_{NI}^*$ est la moyenne des valeurs générées de $\boldsymbol{\beta}$.

En pratique, $\boldsymbol{\Psi}$ et λ sont inconnus et il faut les remplacer par des estimateurs convergents avant de générer des valeurs a posteriori pour $\boldsymbol{\beta}$ et σ^2 au moyen de (12) et (15). À partir de (11), un estimateur convergent de $\boldsymbol{\Psi}$ est

$$\widehat{\boldsymbol{\Psi}} = \left(\sum_{i\in s} w_i\boldsymbol{x}_i\boldsymbol{x}'_i\right)^{-1} + \frac{\boldsymbol{v}_p(\widehat{\boldsymbol{\beta}}_w)}{\tilde{s}_w^2(\widehat{\boldsymbol{\beta}}_w)}\ , \tag{16}$$

où $\boldsymbol{v}_p\left(\widehat{\boldsymbol{\beta}}_w\right)$ est un estimateur convergent par rapport au plan de $var_p\left(\widehat{\boldsymbol{\beta}}_w\right)$, obtenu au moyen de méthodes de linéarisation ou d'estimation de la variance par répliques, et

$$\tilde{s}_w^2(\widehat{\boldsymbol{\beta}}_w) = \frac{n}{n-q}s_w^2(\widehat{\boldsymbol{\beta}}_w)\ .$$

À partir de (14), un estimateur convergent de $2\lambda^{-1}$ est

$$2\hat{\lambda}^{-1} = \frac{2}{\widehat{N}} + \frac{v_p[s_w^2(\boldsymbol{\beta})]|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_w}}{\left(\tilde{s}_w^2(\widehat{\boldsymbol{\beta}}_w)\right)^2}, \tag{17}$$

où $v_p[s_w^2(\boldsymbol{\beta})]$ est un estimateur convergent par rapport au plan de $var_p[s_w^2(\boldsymbol{\beta})]$.

3.3 Choix de probabilité a priori

Quand une probabilité a priori non informative pour $\boldsymbol{\beta}$ est utilisée (p. ex. quand $\boldsymbol{\Phi}_0^{-1}$ est spécifié près de $\mathbf{0}$), la distribution a posteriori conditionnelle (12) est réduite à $\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_w, \sigma^2 \rightarrow N\left(\widehat{\boldsymbol{\beta}}_w, \sigma^2\boldsymbol{\Psi}\right)$. Par conséquent, la moyenne a posteriori $\widehat{\boldsymbol{\beta}} = E\left(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_w\right) = \widehat{\boldsymbol{\beta}}_w$ et l'estimateur (5) devient $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}_w) = \hat{N}^{-1}\sum_{i\in s} w_i \boldsymbol{x}'_i\widehat{\boldsymbol{\beta}}_w = \hat{N}^{-1}\sum_{i\in s} w_i y_i = \hat{\theta}_w$. L'approche bayésienne n'apporte donc aucun avantage par rapport à l'approche fréquentiste fondée sur le plan quand on utilise une probabilité a priori non informative pour $\boldsymbol{\beta}$.

Dans le contexte d'un plan d'échantillonnage ignorable, Sakshaug et coll. (2019) ont proposé de définir $\boldsymbol{\Phi}_0$ comme étant une matrice diagonale avec le j^e élément sur la diagonale égal à $\Phi_{0j} = \left(\hat{\beta}_{NP,j} - \hat{\beta}_{w,j}\right)^2$, où $\hat{\beta}_{NP,j}$ et $\hat{\beta}_{w,j}$ sont les j^e éléments de $\widehat{\boldsymbol{\beta}}_{NP}$ et $\widehat{\boldsymbol{\beta}}_w$, respectivement. L'idée est de tenir compte du biais de $\widehat{\boldsymbol{\beta}}_{NP}$ par la spécification de la matrice de variance-covariance a priori pour $\boldsymbol{\beta}$. Nous appelons cette spécification a priori de $\boldsymbol{\Phi}_0$ une spécification de « distance ». Sakshaug et coll. (2019) ont indiqué la mise en garde suivante :

« *En utilisant l'estimateur fondé sur la probabilité pour construire la distribution a priori, la question de l'utilisation des données deux fois se pose. Nous abordons cette question en signalant que l'estimateur par le maximum de vraisemblance (MV) de l'échantillon probabiliste (une mesure de tendance centrale) sert à informer la variance, plutôt que la moyenne. De plus, nous utilisons l'information des données probabilistes seulement en comparaison relative à l'échantillon non probabiliste. Par conséquent, tout rétrécissement potentiel de la variance a posteriori dépend de la combinaison des deux ensembles de données, plutôt que des données probabilistes seulement.* »

Dans le contexte d'un plan d'échantillonnage non ignorable, l'argument reste valable, mais « l'estimateur fondé sur la probabilité » et « l'estimateur par le MV de l'échantillon probabiliste » sont remplacés par l'estimateur de pseudo maximum de vraisemblance $\widehat{\boldsymbol{\beta}}_w$.

Pour poursuivre cette idée, nous suggérons de définir la matrice de variance-covariance a priori $\boldsymbol{\Phi}_0$ comme étant une matrice diagonale avec le j^e élément sur la diagonale égal à

$$\Phi_{0j} = max\left[s_{NP}^2\Psi_{NP,j}, \left(\hat{\beta}_{NP,j} - \hat{\beta}_{w,j}\right)^2 - \tilde{s}_w^2(\widehat{\boldsymbol{\beta}}_w)\widehat{\Psi}_j\right], \tag{18}$$

où $s_{NP}^2 = (n_{NP} - q)^{-1}\sum_{i\in s_{NP}}\left(y_i - \boldsymbol{x}'_i\widehat{\boldsymbol{\beta}}_{NP}\right)^2$, et $\Psi_{NP,j}$ et $\widehat{\Psi}_j$ sont les j^e éléments diagonaux des matrices $\left(\sum_{i\in s_{NP}}\boldsymbol{x}_i\boldsymbol{x}'_i\right)^{-1}$ et $\widehat{\boldsymbol{\Psi}}$, respectivement. Le premier élément de la fonction de maximum dans (18) est un estimateur de la variance de $\hat{\beta}_{NP,j}$ selon le modèle (1), en supposant que le mécanisme de sélection non probabiliste est ignorable. Le deuxième élément est un estimateur fondé sur le plan du carré du biais $\left(\hat{\beta}_{NP,j} - \beta_j\right)^2$, où β_j est le j^e élément de $\boldsymbol{\beta}$. Nous désignons cette spécification a priori de $\boldsymbol{\Phi}_0$ comme étant la spécification de « biais-variance ».

Wiśniowski et coll. (2020) ont proposé quelques matrices de variance-covariance a priori ayant la forme $\boldsymbol{\Phi}_0 = \sigma^2 k_0\mathbf{V}$, où k_0 est une constante spécifiée et $\mathbf{V}$ est une matrice d'hyperparamètres. Une spécification qui a semblé donner de bons résultats dans leur étude par simulations est $k_0 = \frac{1}{log(n_{NP})}$ et de définir $\mathbf{V}$ comme étant une matrice diagonale avec le j^e élément sur la diagonale égal à $V_j = max\left[s_{NP}^2\Psi_{NP,j}, \left(\hat{\beta}_{NP,j} - \hat{\beta}_{w,j}\right)^2\right]$. Wiśniowski et coll. (2020) font référence à cette spécification a priori de $\boldsymbol{\Phi}_0$ comme étant la spécification de « distance-conjuguée ». Elle a une forme semblable à la spécification de biais-variance, mais avec un facteur d'échelle $\sigma^2 k_0$.

4. Étude par simulations

4.1 Configuration de la simulation

Nous avons réalisé une étude par simulations fondée sur le plan, en nous inspirant de Hidiroglou et You (2016), afin d'évaluer l'estimateur bayésien proposé de la moyenne de population et de le comparer à l'estimateur de Horvitz-Thompson. Nous avons créé une population composée de N = 1 000 unités de population comme suit :

i) x_i a été généré à partir d'une distribution gamma avec une moyenne μ_x et une variance σ_x^2, où μ_x et σ_x^2 sont des constantes prédéterminées.

ii) z_i a été généré à partir d'une distribution gamma avec une moyenne μ_z et une variance σ_z^2, où μ_z et σ_z^2 sont des constantes prédéterminées.

iii) $y_i = \alpha_1 x_i + \alpha_2 z_i + \delta_i$, où $\delta_i \to N(0, \sigma_\delta^2)$ et σ_δ^2 est une constante prédéterminée. Le choix de α_1 et α_2 est abordé ci-dessous. Notons que $E(y_i|x_i) = \alpha_0 + \alpha_1 x_i$, où $\alpha_0 = \alpha_2 \mu_z$.

Ensuite, à partir de la population, on a sélectionné l'échantillon non probabiliste en choisissant 300 unités ayant les plus petites valeurs de y_i. Des échantillons probabilistes de taille 20 et 50 ont été tirés avec une probabilité proportionnelle à la taille sans remise (PPTSR), avec z_i comme mesure de la taille, comme dans You, Rao et Kovacevic (2003) et Hidiroglou et You (2016). Pour l'estimation, le modèle $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ a été pris en compte, où $\varepsilon_i \to N(0, \sigma_\varepsilon^2)$.

Pour déterminer un choix approprié de α_1 et α_2, définissons d'abord $e_i = y_i - E(y_i|x_i) = -\alpha_0 + \alpha_2 z_i + \delta_i$. Le degré d'informativité de l'échantillonnage dépend de la corrélation entre e_i et z_i. Définissons le carré du coefficient de corrélation $\rho_{ez}^2 = \frac{[cov(e_i, z_i)]^2}{var(e_i)\, var(z_i)}$. On peut aisément montrer que $\rho_{ez}^2 = \frac{\alpha_2^2 \sigma_z^2}{\alpha_2^2 \sigma_z^2 + \sigma_\delta^2}$. La résolution de α_2 donne :

$$\alpha_2 = \sqrt{\frac{\sigma_\delta^2 \rho_{ez}^2}{\sigma_z^2 (1 - \rho_{ez}^2)}}. \tag{19}$$

La valeur de ρ_{ez}^2 est définie comme étant une constante prédéterminée. De même, nous pouvons également définir le carré du coefficient de corrélation entre x_i et y_i comme $\rho_{xy}^2 = \frac{[cov(x_i, y_i)]^2}{var(x_i)\, var(y_i)}$. Encore une fois, nous pouvons montrer que $\rho_{xy}^2 = \frac{\alpha_1^2 \sigma_x^2}{\alpha_1^2 \sigma_x^2 + \alpha_2^2 \sigma_z^2 + \sigma_\delta^2}$. En constatant que $\alpha_2^2 \sigma_z^2 + \sigma_\delta^2 = \frac{\sigma_\delta^2}{1 - \rho_{ez}^2}$ et en résolvant α_1, on obtient :

$$\alpha_1 = \sqrt{\frac{\sigma_\delta^2 \rho_{xy}^2}{\sigma_x^2 (1 - \rho_{xy}^2)(1 - \rho_{ez}^2)}}. \tag{20}$$

La valeur de ρ_{xy}^2 est également définie comme étant une constante prédéterminée.

Les quantités suivantes sont définies comme étant des constantes prédéterminées : μ_x, σ_x^2, μ_z, σ_z^2, σ_δ^2, ρ_{ez} et ρ_{xy}. Ensuite, α_1 et α_2 sont déterminés comme dans (20) et (19). Dans notre étude par simulations, nous posons μ_x=μ_z=4, σ_x^2=σ_z^2=8, σ_δ^2=36, ρ_{ez}=0,8, et pour ρ_{xy}, nous considérons deux cas, ρ_{xy}=0,8 et ρ_{xy}=0,2. La taille de l'échantillon non probabiliste est de 300 et celle de l'échantillon probabiliste est n = 20 ou n = 50. Pour l'estimation bayésienne de β, la moyenne antérieure de β est estimée en fonction de l'échantillon non probabiliste, tandis que la variance a priori est la spécification de biais-variance proposée donnée dans (18).

4.2 Résultats

Nous comparons les prédicteurs bayésiens de la moyenne de population $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*)$ et $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*_{NI})$, décrits à la section 2, avec l'estimateur de Horvitz-Thompson (HT) $\hat{\theta}_w$ (avec $w_i = \pi_i^{-1}$) en calculant le biais relatif absolu (BRA) de Monte Carlo et la racine de l'erreur quadratique moyenne relative (REQMR) de ces estimateurs. Le BRA est défini comme étant

$$BRA = \left|\frac{1}{R}\sum_{r=1}^{R}\frac{(\hat{\theta}^{(r)}-\theta)}{\theta}\right|,$$

où $\hat{\theta}^{(r)}$ est $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*)$, $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*_{NI})$ ou $\hat{\theta}_w$ basé sur la r^{e} simulation, et R = 5000. La REQMR se définit comme étant

$$REQMR = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^{(r)}-\theta)^2}}{\theta}.$$

Gardons à l'esprit que $\widehat{\boldsymbol{\beta}}^*$ est obtenu sous l'hypothèse que le plan de sondage probabiliste est ignorable, alors que $\widehat{\boldsymbol{\beta}}^*_{NI}$ est obtenu sous l'hypothèse d'un plan de sondage probabiliste non ignorable. Le tableau 3.2.1 présente la comparaison de l'estimateur de HT et des prédicteurs bayésiens, pour ce qui est du BRA et de la REQMR, quand ρ_{xy}=0.8, ce qui indique une association forte entre x_i et y_i.

Tableau 4.2.1
Comparaison du BRA et de la REQMR, $\boldsymbol{\rho_{xy} = 0.8}$

Estimateur	Taille de l'échantillon n = 20		Taille de l'échantillon n = 50	
	BRA	REQMR	BRA	REQMR
$\hat{\theta}_w$	0,88 %	21,7 %	0,91 %	13,9 %
$\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*)$	5,3 %	19,5 %	5,7 %	12,7 %
$\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*_{NI})$	0,95 %	16,8 %	0,98 %	10,5 %

Le tableau 4.2.1 montre clairement que le biais de $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*_{NI})$ est négligeable, car son BRA Monte Carlo est légèrement plus grand que celui de l'estimateur de HT sans biais. Les BRA de $\hat{\theta}_w$ et $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*_{NI})$ sont inférieurs à 1 % pour les deux tailles d'échantillon, tandis que $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*)$ a un biais modéré légèrement supérieur à 5 %. L'estimateur de HT a la REQMR la plus grande et $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*_{NI})$ a la REQMR la plus petite. Par conséquent, il est clair que la prise en compte appropriée du plan de sondage probabiliste dans l'estimation de $\boldsymbol{\beta}$ entraîne la diminution du biais et de la REQMR. Les résultats du tableau 4.2.1 montrent que le prédicteur bayésien proposé a de très bonnes performances quand le plan de sondage probabiliste n'est pas ignorable.

Tableau 4.2.2
Comparaison du BRA et de la REQMR, $\boldsymbol{\rho_{yx} = 0.2}$

Estimateur	Taille de l'échantillon n = 20		Taille de l'échantillon n = 50	
	BRA	REQMR	BRA	REQMR
$\hat{\theta}_w$	0,98 %	15,1 %	0,93 %	9,72 %
$\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*)$	15,6 %	26,5 %	19,5 %	22,9 %
$\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*_{NI})$	1,86 %	20,5 %	1,48 %	12,6 %

Le tableau 4.2.2 présente la comparaison de l'estimateur de HT et des prédicteurs bayésiens quand ρ_{xy}=0.2, ce qui indique une association faible entre x_i et y_i. L'estimateur de HT donne les meilleurs résultats dans ce scénario présentant le plus petit BRA et la plus petite REQMR. Le prédicteur $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*)$ a les moins bonnes performances avec un biais et une REQMR très grands, tandis que $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*_{NI})$ a un petit BRA, inférieur à 2 %, mais légèrement plus grand que le BRA de l'estimateur de HT, et une REQMR plus petite que $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*)$. Ainsi, la prise en compte appropriée du plan de sondage probabiliste dans l'estimation de $\boldsymbol{\beta}$ peut réduire considérablement le biais, y compris en présence

d'une association faible entre x_i et y_i. Cependant, le prédicteur bayésien $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*_{NI})$ n'est pas nécessairement plus efficace que l'estimateur de HT dans cette situation.

5. Conclusion

Nous avons proposé un prédicteur d'une moyenne de population en utilisant un modèle linéaire bayésien qui nous permet de combiner les données d'un échantillon probabiliste et d'un échantillon non probabiliste. Selon une distribution a priori informative appropriée pour les paramètres du modèle, nous avons démontré en quoi l'utilisation d'un échantillon non probabiliste peut améliorer des estimations réalisées à partir d'un échantillon probabiliste, particulièrement quand l'association entre les variables auxiliaires et la variable d'intérêt n'est pas faible. Toutefois, si l'association est parfaite, aucun gain d'efficience ne peut être réalisé par l'utilisation d'un échantillon non probabiliste, car notre prédicteur se réduit à l'estimateur standard pondéré par les poids d'enquête. Nous avons également montré qu'il est important de tenir compte du plan de sondage probabiliste quand il est non ignorable. Le prédicteur de la moyenne de population que nous proposons a donné de bons résultats dans notre étude par simulations.

Dans de futures recherches, nous étudierons l'estimation de l'erreur quadratique moyenne et comparerons notre prédicteur bayésien à celui proposé par Nandram et Rao (2021). Dans le prolongement de cette recherche, un autre sujet intéressant, surtout dans les programmes de statistiques sociales, est l'extension de l'estimation d'une proportion (c.-à-d. le cas d'une variable d'intérêt binaire) au moyen d'un modèle logistique bayésien.

Bibliographie

Beaumont, J.-F. (2020). Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles? *Techniques d'enquête*, 46, 1-28.

Binder, D.A., et G.A. Roberts, (2003). Design-based and model-based methods for estimating model parameters. Dans *Analysis of Survey Data*, Chambers, R.L., et C.J. Skinner, (éd.), Wiley, New York.

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

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

Elliott, M. et A. Haviland (2007). Utilisation d'un échantillon de convenance électronique comme complément à un échantillon probabiliste. *Techniques d'enquête*, 33, 211-215.

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

Fuller, W.A. (2009). *Sampling Statistics*, Wiley, New York.

Hidiroglou, M. et Y. You (2016). Comparaison d'estimateurs sur petits domaines au niveau de l'unité et au niveau du domaine. *Techniques d'enquête*, 42, 41-46.

Nandram, B. et J.N.K. Rao (2021). A Bayesian approach for integrating a small probability sample with a non-probability sample. Dans *Proceedings of the Survey Research Methods Section*, American Statistical Association.

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

Rivers, D. (2007). Sampling from web surveys. Dans *Proceedings of the Survey Research Methods Section*, American Statistical Association.

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

Sakshaug, J.W., A. Wiśniowski, D.A.P. Ruiz, et A.G. Blom (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics*, 35, 653-681.

Valliant (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8, 231-263.

Wiśniowski, A., J.W. Sakshaug, D.A.P. Ruiz, et A.G. Blom (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8, 120-147.

Wang, Z., J.K. Kim, et S. Yang (2018). Approximate Bayesian inference under informative sampling. *Biometrika*, 105, 91-102.

Yang, S., J.K. Kim, et Y. Hwang (2021). Intégration des données des enquêtes probabilistes et des mégadonnées trouvées aux fins d'inférence de population finie au moyen d'une imputation massive *Techniques d'enquête*, 47, 29-58.

You, Y., J.N.K. Rao, et M. Hidiroglou (2003). Estimation des effets fixes et des composantes de la variance par un modèle à valeur aléatoire à l'origine en utilisant des données d'enquête. Dans *La série des symposiums internationaux de Statistique Canada : recueil de 2003*, Statistique Canada.