Une approche par prédiction pour l'intégration de données

Estelle Medous[1,2], Anne Ruiz-Gazen[1], Camelia Goga[3], Jean-François Beaumont[4], Alain Dessertaine[2] et Pauline Puech[2].

[1] *Toulouse School of Economics, Université Toulouse 1 Capitole*
1, Esplanade de l'Université, 31000 Toulouse
E-mail: estelle.medous@laposte.fr, anne.ruiz-gazen@tse-fr.eu

[2] *La Poste, 3 rue Jean Richepin, 93192 Noisy le Grand cedex.*
Email: alain.dessertaine@laposte.fr, pauline.puech@laposte.fr

[3] *Laboratoire de Mathématiques de Besançon, Université de Bourgogne Franche-Comté*
Email: camelia.goga@univ-fcomte.fr

[4] *Statistique Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada*
Email: jean-francois.beaumont@canada.ca

Résumé. De nos jours, l'accès à des sources de données non probabilistes est fréquent et de nombreuses méthodes existent pour combiner ce type de données à des données d'enquête probabiliste (voir Beaumont, 2019, Kim et al., 2021 et Rao, 2021 pour des survols récents concernant ces méthodes). Dans le cadre de l'étude du trafic postal en France, nous proposons deux nouveaux estimateurs d'un total qui permettent de prendre en compte un échantillon non probabiliste dans une situation où la variable d'intérêt n'est pas connue au niveau de cet échantillon. L'objet de la présentation est de préciser le contexte de La Poste, de donner un survol des méthodes d'intégration de données existantes et d'introduire les estimateurs que nous proposons.

Mots-clés. Base de données massives, Biais de sélection, Enquête non-probabiliste, Enquête probabiliste, Inférence statistique.

Abstract. Nowadays, access to non-probabilistic data sources is common and many methods exist to combine this type of data with probabilistic survey data (see Beaumont, 2019, Kim et al., 2021 and Rao, 2021 for recent overviews regarding these methods). In the context of the study of postal traffic in France, we propose two new estimators of a total. They allow to take into account a non-probabilistic sample in a situation where the variable of interest is not known at the level of that sample. The purpose of the presentation is to present the context of La Poste, to give an overview of existing data integration methods and to introduce our proposed estimators.

Keywords. Big data, Non probability survey, Probability survey, Selection bias, Statistical inference.

1 Introduction

La Poste française donne des estimations de trafic et de revenu mensuel postal par type de pli. La population d'intérêt est celle des plis, observée indirectement grâce à la population des tournées. Les plis sont échantillonnés par tirage probabiliste et differentes caractéristiques, comme le type de pli (lettre, recommandé, colis...) ou le coût de l'affranchissement, sont observées. Dans les années à venir, à cause de contrainte de coût, la taille de l'échantillon probabiliste va diminuer et La Poste cherche des solutions pour ne pas perdre trop de précision dans ses estimations. L'intégration de sources de données non probabilistes est une solution envisagée. En effet, La Poste dispose d'une base de données massives régulièrement mise à jour, contenant des informations sur chaque lettre traitée en machine. Cette base de données, appelée TAE (Traitement Automatique de l'Enveloppe), ne constitue pas toutefois l'ensemble de la population des lettres. En effet certaines lettres ne passent pas en machine de tri, soit parce que leur dimension ne le permet pas (lettre trop grande, lettre bulle avec un objet ou colis, ...), soit parce qu'elles passent par un autre circuit (certaines lettres recommandées,...). Pour faire l'estimation du trafic mensuel à partir de TAE, il faut donc corriger le biais de sélection et/ou la sous-couverture de TAE. Des méthodes existent pour corriger ce biais, mais elles requièrent la connaissance des valeurs prises par la variable d'intérêt Y (ici le type de lettre ou le coût du timbre) sur la base non probabiliste. Les variables d'intérêt dans TAE souffrent en général d'erreur de mesure et ne sont pas observables. TAE permet toutefois d'accèder à de l'information auxiliaire. L'idée proposée dans cette présentation est d'utiliser un échantillon probabiliste pour imputer les valeurs des variables d'intérêt dans TAE. Dans un premier temps nous rappelons brièvement les méthodes existantes d'intégration de bases de données, puis nous présentons les deux estimateurs proposés.

2 Méthodes existantes

Soit U la population d'intérêt. On cherche à estimer un paramètre θ_Y sur U, où Y est la variable d'intérêt. Un échantillon probabiliste s_P est tiré selon un plan de sondage dans U. On suppose qu'il existe un échantillon non probabiliste s_{NP} de U pour lequel les valeurs de certaines variables sont connues. On note δ l'indicatrice d'appartenance à s_{NP}, où $\delta_k = 1$ si k appartient à s_{NP}, 0 sinon. L'idée des méthodes présentées dans cette section est d'utiliser les données connues dans s_{NP} afin d'améliorer l'estimation de θ_Y. Il existe plusieurs façons de faire selon les données disponibles dans s_{NP} :

- Estimation à partir de l'échantillon non probabiliste s_{NP}.
- Amélioration grâce à s_{NP} de l'estimateur calculé sur l'échantillon probabiliste s_P.

La première catégorie de méthodes nécessite de connaître les valeurs de la variable Y pour tous les individus de s_{NP}. On cherche à estimer le paramètre θ_Y sur l'échantillon s_{NP}. Le méchanisme d'échantillonnage pour s_{NP} étant inconnu, calculer directement le paramètre θ_Y sur s_{NP} pose des problèmes de biais de sélection. L'échantillon probabiliste permet de corriger ce biais. Cela peut se faire en modifiant le poids de sondage pour s_{NP} (calcul du score

de propensité, calibration weighting, voir Yang and Kim; 2019, Chen, Valliant and Elliott, 2018; Chen and al., 2019; McConville and al., 2017; Beaumont, 2019) ou en rajoutant une estimation sur s_P de la zone non couverte par s_{NP} (classification semi-supervisée, estimation par le ratio, voir Kim and Tam, 2020). Il faut noter que l'estimation sur s_P nécessite souvent de connaître les valeurs prises par la variable δ sur s_P. Ces méthodes utilisent majoritairement les données de s_{NP}. L'échantillon probabiliste permet d'apporter des corrections et peut donc être de taille faible, et tel que la variable Y n'est pas nécessairement mesurée. Cela permet de réduire les coûts liés à l'observation des variables.

La deuxième catégorie regroupe les méthodes où l'estimation se fait sur s_P, tandis que s_{NP} sert soit à corriger les problèmes de s_P, comme la non-récupération de la variable Y sur s_P, soit à améliorer l'estimation avec un calage sur s_{NP}. Si Y n'est pas observée pour les individus de s_P, mais connue sur s_{NP}, on peut utiliser les valeurs de Y sur s_{NP} pour imputer les valeurs de s_P. Ces valeurs permettent de calculer l'estimateur de θ_Y sur s_P. Ce type d'imputation de masse (voir Yang and Kim, 2020; Kim et al., 2021; Beaumont, 2020) répond à des problèmes de coûts et/ou de taux de réponse faible qui rendent l'obtention des valeurs de Y sur s_P difficile. L'amélioration de l'estimation sur s_P par calage sur s_{NP} nécessite la connaissance de la variable Y sur s_P mais pas forcément sur s_{NP} (voir Beaumont, 2020). La variable δ est supposée connue sur s_P. On peut estimer θ_Y uniquement grâce à s_P, puis améliorer la précision de l'estimateur obtenu en calant sur des totaux connus pour s_{NP}.

Dans la section suivante, nous proposons une méthode alternative dans le cas où la variable Y est connue sur s_P mais pas forcément sur s_{NP}.

3 Méthode proposée

Soit U la population d'intérêt. On cherche à estimer le total T_Y de la variable Y sur U. Pour cela on dispose d'un échantillon probabiliste s_P tiré selon un plan de sondage $p(s \mid \boldsymbol{Z})$ où $\boldsymbol{Z}$ est la matrice contenant les informations nécessaires au tirage. Soit I_k, $k \in U$, l'indicateur d'appartenance à l'échantillon, qui vaut 1 si l'individu k est échantillonné, 0 sinon. Le vecteur contenant les valeurs de I_k pour tous les individus de la population est noté $\boldsymbol{I}$. La probabilité que l'individu k soit échantillonné est $\pi_k = E(I_k \mid \boldsymbol{Z})$. On observe les valeurs y_k et $\boldsymbol{x}_k$ pour les individus k de s_P, où $\boldsymbol{x}$ est un vecteur de variables auxiliaires. On note $\boldsymbol{X}$ la matrice contenant les valeurs de $\boldsymbol{x}$ pour tous les individus de U. On suppose disposer des valeurs $\boldsymbol{x}_k$ pour tous les individus d'un échantillon non probabiliste $s_{NP} \subset U$. On rappelle que l'indicateur d'appartenance à s_{NP} pour $k \in U$ est noté δ_k. Le vecteur contenant les valeurs de δ_k, $k \in U$, est noté $\boldsymbol{\delta}$. On suppose δ_k connu pour tous les individus de s_P et par construction, tous les individus de s_{NP}. On note $\boldsymbol{Y}$ le vecteur contenant les valeurs de y_k, $k \in U$.

Les vecteurs utilisés pour l'inférence sont $\boldsymbol{Y}$, $\boldsymbol{I}$, $\boldsymbol{\delta}$, $\boldsymbol{Z}$ and $\boldsymbol{X}$. On suppose qu'ils respectent les hypothèses suivantes:

Hypothèse 1: $F(\boldsymbol{I} \mid \boldsymbol{Y}, \boldsymbol{\delta}, \boldsymbol{X}, \boldsymbol{Z}) = F(\boldsymbol{I} \mid \boldsymbol{Z})$
Hypothèse 2: $F(\boldsymbol{Y} \mid \boldsymbol{I}, \boldsymbol{\delta}, \boldsymbol{X}, \boldsymbol{Z}) = F(\boldsymbol{Y} \mid \boldsymbol{X})$

où $F(x)$ désigne la distribution de la variable x. La distribution $F(\boldsymbol{Y} \mid \boldsymbol{X})$ est inconnue en

pratique et nécessite l'élaboration d'un modèle.

Nous faisons les hypothèses de modèle suivantes :

1. les variables y_k, $k \in U$, sont indépendantes conditionnellement à $\boldsymbol{X}$;
2. $\mathrm{E}(y_k \mid \boldsymbol{X}) = \mu_k \equiv \mu(\boldsymbol{x}_k)$, pour une fonction $\mu(.)$;
3. $\mathrm{Var}(y_k \mid \boldsymbol{X}) = \sigma_k^2 \equiv \nu(\boldsymbol{x}_k)$, pour une fonction $\nu(.)$.

Soient $\boldsymbol{Y}_{s_{NP}}$ et $\boldsymbol{Y}_{U-s_{NP}}$, les vecteurs contenant les valeurs y_k pour les individus $k \in s_{NP}$ et $k \in U - s_{NP}$ respectivement. On considère $\boldsymbol{Y}_{s_{NP}}$ et $\boldsymbol{I}$ aléatoires pour notre inférence et on conditionne par rapport à $\boldsymbol{\Omega} = \{\boldsymbol{Y}_{U-s_{NP}}, \boldsymbol{\delta}, \boldsymbol{X}, \boldsymbol{Z}\}$. $\boldsymbol{Y}_{U-s_{NP}}$ est donc considéré comme non aléatoire. En effet, comme le vecteur $\boldsymbol{x}$ n'est pas connu pour les individus de $s_P - (s_P \cap s_{NP})$, il n'est pas possible d'estimer les paramètres μ_k et ν_k pour ces individus et par extension il n'est pas possible d'estimer la variance de nos estimateurs.

Considérons des poids de sondages w_k, $k \in U$, tels que $\sum_{k \in s_P} w_k y_k$ est un estimateur sans biais de θ_Y sous le plan de sondage. Soit $\hat{y}_k$ un prédicteur sans biais sous le modèle de y_k pour $k \in s_{NP}$. Les estimateurs de T_Y que nous proposons sont définis par :

$$\hat{T}_Y^{*} = \sum_{k \in s_{NP}} \hat{y}_k + \sum_{k \in s_P \cap s_{NP}} (y_k - \hat{y}_k) + \sum_{k \in s_P - (s_P \cap s_{NP})} w_k y_k, \tag{1}$$

$$\hat{T}_Y^{**} = \sum_{k \in s_{NP}} \hat{y}_k + \sum_{k \in s_P \cap s_{NP}} w_k (y_k - \hat{y}_k) + \sum_{k \in s_P - (s_P \cap s_{NP})} w_k y_k.$$

On peut par exemple considérer pour $\hat{y}_k$ une imputation par régression linéaire telle que $\mu_k = \boldsymbol{x}_k^{'} \boldsymbol{\beta}$, où $\boldsymbol{\beta}$ est un vecteur de paramètres inconnus et

$$\hat{y}_k = \sum_{i \in s_P \cap s_{NP}} \phi_{ki} y_i \tag{2}$$

avec ϕ_{ki} donné par

$$\phi_{ki} = \boldsymbol{x}_k^{'} \left(\sum_{i \in s_P \cap s_{NP}} \omega_i \boldsymbol{x}_i \boldsymbol{x}_i^{'} \right)^{-1} \boldsymbol{x}_i \omega_i,$$

où ω_i permet de prendre en compte une éventuelle hétéroscedasticité.

L'objectif de cette présentation et de proposer une première étude théorique de ces deux estimateurs en termes de biais et de variance dans quelques cas simples.

Bibliographie

[1] Beaumont, J.-F. (2020). Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles?

[2] Chen, J. K. T., Valliant, R. and Elliott, M. R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO, *Survey Methodology* 44: 117–144.

[3] Chen, J. K. T., Valliant, R. L. and Elliott, M. R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68: 657–681.

[4] Kim, J. K., Tam, S. M. (2020). Data integration by combining big data and survey sample data for finite population inference, *International Statistical Review*.

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

[6] McConville, K. S., Breidt, F. J., Lee, T. C. and Moisen, G. G. (2017). Model-assisted survey regression estimation with the LASSO, *Journal of Survey Statistics and Methodology* 5: 131–158.

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

[8] Yang, s. et Kim, J.K. (2020). Statistical Data Integration in Survey Sampling: A review, *Japanese Journal of Statistics and Data Science*, 1-26.