Estimation du trafic postal en France : perte de précision due à l'utilisation d'un sondage indirect double.

Estelle Medous[1], Camelia Goga[2], Anne Ruiz-Gazen[3], Jean-François Beaumont[4], Alain Dessertaine[5] et Pauline Puech[5].

[1] *Laboratoire de Mathématiques Jean Leray, Université de Nantes*
Faculté des Sciences et des Techniques, 2 Chem. de la Houssinière, 44322 Nantes
E-mail : emedous@hotmail.fr
[2] *Laboratoire de Mathématiques de Besançon, Université de Bourgogne Franche-Comté*
Email : camelia.goga@univ-fcomte.fr
[3] *Toulouse School of Economics, Université Toulouse 1 Capitole*
1, Esplanade de l'Université, 31000 Toulouse
E-mail : anne.ruiz-gazen@tse-fr.eu
[4] *Statistique Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada*
Email : jean-francois.beaumont@canada.ca
[5] *La Poste, 3 rue Jean Richepin, 93192 Noisy le Grand cedex.*
Email : alain.dessertaine@laposte.fr, pauline.puech@laposte.fr

Résumé. Dans les enquêtes probabilistes, lorsqu'il n'y a pas de base de sondage pour la population cible, une solution consiste à trouver une base de sondage liée à la population cible et à utiliser un échantillonnage indirect. Les poids d'échantillonnage peuvent être déterminés à l'aide de la méthode généralisée du partage des poids (MGPP). De plus, on montre l'existence de poids optimaux, qui minimisent la variance des estimateurs obtenus pour toute variable d'intérêt. Toutefois, cette méthode ne peut pas être appliquée lorsque certains des liens entre la base de sondage et l'échantillon de la population cible sont manquants ou difficiles à récupérer de manière exhaustive. Une solution pour éviter ce problème est de considérer une population intermédiaire liée à la fois à la base de sondage et à la population cible et d'utiliser un double échantillonnage indirect. La MGPP peut alors être utilisée deux fois, d'abord entre la population de base et la population intermédiaire, puis entre la population intermédiaire et la population cible. Comme l'illustre l'enquête française sur le trafic postal, ce double échantillonnage indirect peut détériorer la précision des estimateurs dans certaines situations. Mathématiquement, il est possible de mettre en évidence l'ampleur de la perte de précision dans des situations pratiques proches du contexte de La Poste. Les résultats sont illustrés par des simulations Monte-Carlo et par une application à l'estimation du trafic postal français.

Mots-clés. Enquêtes, Estimation de variance, Méthode généralisée du partage des poids, Plan de sondage complexe, Population finie.

Abstract. In probabilistic surveys, when there is no sampling frame for the target population, a solution is to find a frame population linked in some way to the target population and use indirect sampling. The sampling weights can be determined using the generalized weight share method (GWSM). Moreover, we show the existence of optimal weights that

minimize the variance of the resulting estimators regardless of the variable of interest. However, this method cannot be applied when some of the links between the frame population and the sample in the target population are missing or difficult to retrieve exhaustively. A solution to avoid this issue is to consider an intermediate population linked in some way to both the frame and target populations and use a double indirect sampling. Then the GWSM can be used twice, first between the frame and intermediate populations and then between the intermediate and target populations. As illustrated with the French postal traffic survey, this double indirect sampling appears to be deteriorating the precision of estimators in some situations. Using mathematical derivations, it is possible to highlight the magnitude of the loss of precision in practical situations similar to the French postal context. Results are illustrated through Monte Carlo simulations and with an application to the French postal traffic estimation.

Keywords. Complex sampling design, Finite population, Generalized Weight Share Method, Surveys, Variance estimation.

1 Introduction

En France, l'estimation du trafic postal mensuel par "La Poste" est basée sur un tirage d'échantillon probabiliste. Jusqu'au début des années 2010, les échantillons étaient tirés directement dans la population des tournées de facteurs, qui constitue la population cible ou d'intérêt. Récemment, l'organisation des tournées a évolué de telle façon que cette population n'est plus stable dans le temps. Il n'est plus possible d'échantillonner directement les tournées, et le plan de sondage a été modifié en un tirage dans la population des adresses, qui constitue la base de sondage. Chaque tournée de facteur étant constituée d'adresses, il est possible de relier la population cible à la base de sondage et d'utiliser un plan de sondage indirect pour récupérer un échantillon de tournées.

L'échantillonnage indirect a été étudié de manière intensive dans la littérature (voir par exemple Deville and Lavallée (2006), Lavallée (2009) et Deville and Maumy-Bertrand (2006)). La méthode d'estimation privilégiée dans ce contexte est la méthode dite méthode généralisée de partage des poids (MGPP). Elle consiste à utiliser les liens qui existent entre la base de sondage et la population cible pour exprimer un total d'intérêt sur la population cible comme un total sur la base de sondage. Les méthodes d'estimation classiques comme l'estimateur d'Horvitz-Thompson peuvent alors être utilisées. La MGPP offre l'avantage de pouvoir déterminer des poids de liens optimaux, qui minimisent la variance des estimateurs obtenus quelle que soit la variable d'intérêt (voir Deville and Lavallée (2006)). Cependant les conditions d'existence de tels poids sont peu étudiées dans la littérature. La MGPP est une méthode simple, mais elle nécessite que les liens entre la base de sondage et la population cible soient connus. Pour l'exemple de La Poste, il s'agit de connaître toutes les adresses dont le courrier est délivré par un facteur lors d'une tournée échantillonnée. On a une moyenne d'environ 500 adresses par tournée et il n'est pas possible de collecter toute l'information avant le départ du facteur. Pour contourner le problème, La Poste a mis en place un sondage doublement indirect en utilisant les casiers de tri du courrier comme une population intermédiaire entre la population des adresses et celle des tournées. Pour ce plan de sondage, il suffit de connaître les casiers des tournées échantillonnées (50 en moyenne) et les adresses

du casier associé à l'adresse échantillonnée (10 en moyenne). Avec 60 éléments d'information à collecter en moyenne par tournée échantillonnée pour ce sondage indirect double, au lieu de 500 pour le sondage indirect simple, il devient possible de mettre en œuvre la méthode d'échantillonnage et ainsi maîtriser les biais d'estimation. Toutefois, La Poste a observé une détérioration importante de la précision des estimateurs de trafic postal après avoir mis en place ce double sondage indirect.

L'objectif du présent article est dans un premier temps de déterminer des conditions suffisantes à l'existence de poids optimaux pour les estimateurs MGPP obtenus par sondage indirect simple puis d'évaluer la détérioration de la précision en comparant les variances des estimateurs MGPP obtenus avec des poids optimaux et non optimaux. Ces calculs mathématiques aboutissent à une évaluation précise de la différence de variances dans un contexte proche de celui de La Poste, et permettent d'expliquer la perte de précision observée en pratique. Dans un deuxième temps, cet article introduit la notion de sondage indirect double et de MGPP double comme cas particulier de MGPP. Des simulations de type Monte-Carlo valident les résultats et permettent de distinguer des situations où la perte de précision liée à l'utilisation d'un sondage indirect double est faible, voire nulle, de situations où la perte peut-être très forte. Au-delà de la compréhension de la perte de précision pour l'estimation du trafic postal, les résultats obtenus permettent de donner des recommandations pour une mise en œuvre efficace d'un plan de sondage indirect double.

Dans la section 2, nous rappelons les notations et les définitions des estimateurs MGPP pour un sondage indirect simple et déterminons des conditions d'existence de poids optimaux, minimisant la variance des estimateurs pour toute variable d'intérêt, ainsi que la perte de précision associée avec l'utilisation de poids non optimaux. Dans la section 3, nous présentons les définitions et propriétés des estimateurs MGPP pour un sondage indirect double.

2 Sondage indirect simple

2.1 Principe

Soit y la variable d'intérêt et y_k la valeur de y pour l'individu k dans la population cible U_T. L'objectif est d'estimer le total $t_y = \sum_{k \in U_T} y_k$ de la variable y sur U_T. On suppose que la liste exhaustive des unités de U_T n'est pas disponible, mais qu'il existe une base de sondage U_F reliée à U_T de telle façon que chaque unité de U_T soit liée à au moins une unité de U_F. Dans ce cas, l'échantillonnage indirect, tel que détaillé par Deville and Lavallée (2006), permet de sélectionner un échantillon s_F de U_F par un plan de sondage classique, noté p, et d'utiliser des méthodes standards d'estimation de paramètres sur U_T. Dans l'exemple de La Poste, la population cible est constituée des tournées de facteurs, la base de sondages est composée d'adresses et chaque tournée contient au moins une adresse.

Dans un plan indirect, la base de sondage U_F et la population cible U_T peuvent être reliées de diverses manières (voir Deville and Lavallée (2006) pour plus de détails). Dans le cas de La Poste, pour un jour donné, une adresse n'est délivrée que par une seule tournée (hors organisations dédiées à la distribution des colis) et on parle de liens de type "tous pour un", puisqu'une unité de U_F n'est liée qu'à une seule unité de U_T, mais qu'une unité de U_T peut être reliée à plusieurs unités de U_F. Dans la suite de cet article, nous étudierons

en particulier ce type de liens. À chaque paire (i,k) de $U_F \times U_T$, est associé un indicateur pondéré de lien (ou plus simplement poids de lien), noté θ_{ik}. On a $\theta_{ik} = 0$, si les unités i et k ne sont pas liées, et un poids strictement positif $\theta_{ik} > 0$ sinon. Pour pouvoir exprimer un total sur la population U_T comme un total pondéré sur U_F, il est nécessaire de normaliser les θ_{ik}. Dans la suite, on note $\tilde{\theta}_{ik} = \theta_{ik} / \sum_{i' \in U_F} \theta_{i'k}$ les poids de liens normalisés et on a $\sum_{i \in U_F} \tilde{\theta}_{ik} = 1$ pour tout k de U_T. On peut alors écrire le total t_y de y sur U_T, comme un total sur U_F pour une variable artificielle $\tilde{y}_i = \sum_{k \in U_T} \tilde{\theta}_{ik} y_k$:

$$t_y = \sum_{k \in U_T} y_k = \sum_{k \in U_T} \left(\sum_{i \in U_F} \tilde{\theta}_{ik} \right) y_k = \sum_{i \in U_F} \sum_{k \in U_T} \tilde{\theta}_{ik} y_k = \sum_{i \in U_F} \tilde{y}_i.$$

Pour estimer t_y, on tire un échantillon s_F dans U_F avec le plan de sondage p et on note $\pi_i = p(i \in s_F)$, $i \in U_F$, les probabilités d'inclusion d'ordre un, supposées strictement positives pour tout $i \in U_F$. L'estimateur Horvitz-Thompson (HT) de t_y est donné par :

$$\hat{t}_{y1} = \sum_{i \in s_F} \frac{\tilde{y}_i}{\pi_i} = \sum_{i \in s_F} \frac{1}{\pi_i} \left(\sum_{k \in U_T} \tilde{\theta}_{ik} y_k \right) = \sum_{k \in U_T} \left(\sum_{i \in s_F} \frac{\tilde{\theta}_{ik}}{\pi_i} \right) y_k. \tag{1}$$

Cet estimateur est appelé dans la suite estimateur MGPP simple et il est sans biais pour t_y si et seulement si les poids de liens sont normalisés.

2.2 Poids optimaux

L'estimateur $\hat{t}_{y1}$ dépend de la valeur choisie pour les poids de liens $\tilde{\theta}_{ik}$, $i \in U_F$, $k \in U_T$. Dans cette section, nous étudions l'existence de poids de liens minimisant la variance de l'estimateur $\hat{t}_{y1}$ pour toute variable d'intérêt.

Soit N_F la taille de la base de sondage U_F et, pour $k \in U$, U_{Fk} la sous-population de U_F de taille N_{Fk} composée des individu reliés à k. On note

$$\Delta_{ii'} = \frac{\pi_{ii'} - \pi_i \pi_{i'}}{\pi_i \pi_{i'}}, \quad i, i' \in U_F,$$

$\mathbf{\Delta} = (\Delta_{ii'})_{i,i' \in U_F}$ la matrice de taille $N_F \times N_F$ et $\mathbf{\Delta}_{kk'} = (\Delta_{ii'})_{i \in U_{Fk}, i' \in U_{Fk'}}$ la sous-matrice de $\mathbf{\Delta}$ de taille $N_{Fk} \times N_{Fk'}$ composée des éléments en position i, i' si i (resp. i') est lié à k (resp k') et $\mathbf{\Delta}_k = \mathbf{\Delta}_{kk}$. Dans la figure 1, l'individu k de la population cible U_T est lié aux individus notés 1, 2 et 3 dans la population U_F. La sous-matrice Δ_k est donc composée des 9 valeurs $\Delta_{ii'}$, $i, i' = 1 \ldots 3$, situées en haut à gauche de la matrice (en rouge). On peut similairement obtenir les trois autres matrices de la figure 1.

Quand les liens sont de type TpU, les sous-populations U_{Fk} forment une partition de U_F et on suppose les individus de U_F ordonnés selon U_{Fk}, $k \in U$, (voir le graphique de gauche de la figure 1). On a alors (voir le graphique de droite de la figure 1)

$$\mathbf{\Delta} = (\mathbf{\Delta}_{kk'})_{k,k' \in U}.$$

Soit $\mathbb{1}_k$ le vecteur de taille N_{Fk} dont les composantes sont égales à 1. Si les liens sont de type TpU, alors un plan de sondage satisfait la $\mathbf{\Delta}$-propriété si

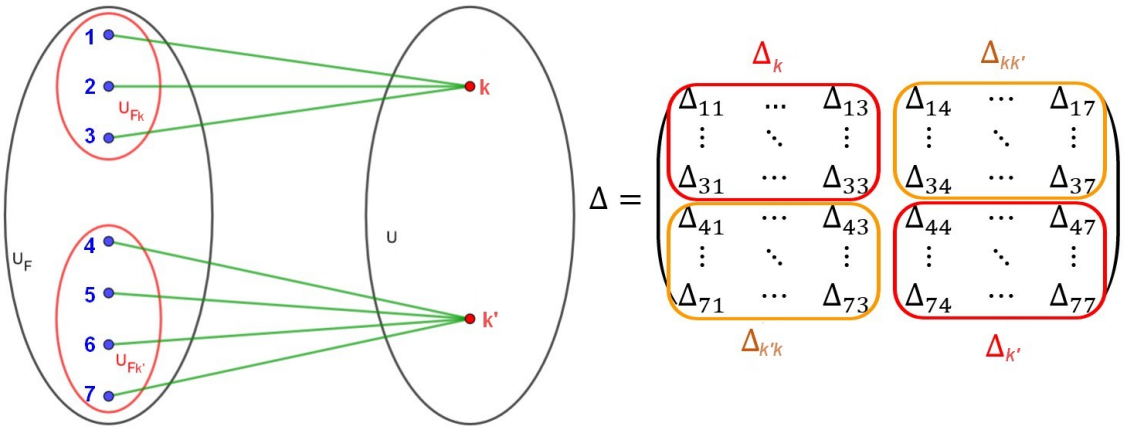

FIGURE 1 – Décomposition de la matrice $\boldsymbol{\Delta}$ dans un cas TpU.

— pour tout $k \in U$, $\boldsymbol{\Delta}_k$ est inversible,
— pour $k \neq k' \in U$,

$$\boldsymbol{\Delta}_{kk'} = c_{kk'} \mathbb{1}_k \mathbb{1}_{k'}^t \quad \text{avec} \quad c_{kk'} \quad \text{ne dépendant ni de } i \text{ ni de } i'. \tag{2}$$

Les plans de Poisson, Sondage Aléatoire Simple Sans Remise (SASSR) et, sous certaines conditions, SASSR stratifié (STASSR) satisfont la $\boldsymbol{\Delta}$-propriété (voir Medous et al., 2023). Medous et al. (2023) montrent que, si les liens sont de type TpU et que le plan de sondage dans U_F satisfait la Δ-propriété, il existe alors des poids de liens standardisés $\tilde{\theta}_{ik}^{opt}$, $i \in U_F$, $k \in U$, donnés par :

$$(\tilde{\theta}_{ik}^{opt})_{i \in U_{Fk}} = \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k \left(\mathbb{1}_k^t \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k \right)^{-1}, \text{ pour tout } k \in U_T, \tag{3}$$

qui minimisent la variance de l'estimateur MGPP simple quelle que soit la variable d'intérêt.

La Δ-propriété permet de simplifier l'expression de la variance de l'estimateur par MGPP simple et on montre dans ce cas que la différence entre la variance de l'estimateur MGPP simple $\hat{t}_{y1}$ utilisant des poids quelconques $\tilde{\theta}_{ik}$, $i \in U_F$, $k \in U$, et la variance de l'estimateur MGPP simple optimal $\hat{t}_{y1}^{opt}$ utilisant les poids optimaux $\tilde{\theta}_{ik}^{opt}$ est donnée par :

$$\mathrm{Var}(\hat{t}_{y1}) - \mathrm{Var}(\hat{t}_{y1}^{opt}) = \sum_{k \in U_T} y_k^2 \mathrm{Var}(\hat{t}_{\tilde{\theta}_k} - \hat{t}_{\tilde{\theta}_k^{opt}}) = \sum_{k \in U_T} y_k^2 \mathrm{Var}(\hat{t}_{\tilde{\theta}_k - \tilde{\theta}_k^{opt}}), \tag{4}$$

avec $\hat{t}_{\tilde{\theta}_k^{opt}} = \sum_{i \in s_F} d_i \tilde{\theta}_{ik}^{opt}$ l'estimateur de HT du total $t_{\tilde{\theta}_k^{opt}} = \sum_{i \in U_F} \tilde{\theta}_{ik}^{opt} = 1$, et $\hat{t}_{\tilde{\theta}_k - \tilde{\theta}_k^{opt}} = \sum_{i \in s_F} d_i (\tilde{\theta}_{ik} - \tilde{\theta}_{ik}^{opt})$ l'estimateur de HT du total $t_{\tilde{\theta}_k - \tilde{\theta}_k^{opt}} = 0$, pour tout $k \in U_T$. L'équation (4) montre que la différence entre la variance de l'estimateur MGPP simple quelconque et la variance de l'estimateur MGPP simple optimal du total T_y dépend de la variance de la différence entre les poids standardisés $\tilde{\theta}_k - \tilde{\theta}_k^{opt}$, $k \in U$. La standardisation dépend de la répartition des liens entre les populations et influence aussi l'augmentation de variance liée à l'utilisation de poids quelconques.

Le premier plan indirect prévu par La Poste prévoyait l'utilisation d'une MGPP simple optimale. La Poste souhaitait connaître la performance, en terme de variance, de l'estimateur

MGPP simple optimal, comparé à l'estimateur de HT direct de T_y. Il n'est en général pas possible de déterminer lequel des deux estimateurs sera le plus précis en terme de variance (Kiesl, 2016). Dans l'article, on se penche sur un cas particulier où un plan de Poisson est utilisé pour tirer l'échantillon direct s dans U_T et l'échantillon s_F dans U_F, avec π_i (resp. π_k) la probabilité d'inclusion d'ordre un de l'individu $i \in U_F$ (resp. $k \in U$) utilisée pour tirer l'échantillon s_F (resp. l'échantillon direct s). La probabilité π_k, $k \in U$ est obtenue comme suit :

$$\pi_k = P(\text{au moins un } i \in U_F \text{ lié à } k \text{ est échantillonné dans } s_F) = 1 - \Pi_{i \in U_F}(1-\pi_i)^{l_{ik}}.$$

Dans ce cas, on montre que la variance de l'estimateur de HT direct de T_y est plus faible que la variance de l'estimateur MGPP simple optimal de T_y.

3 Sondage indirect double

Pour La Poste, U_F est la population des adresses et U_T celle des tournées. Ces deux populations sont reliées par des liens de type TpU et La Poste utilise un STASSR tel que toutes les adresses desservies par une même tournée appartiennent à la même strate. Le plan de sondage utilisé à La Poste satisfait la $\mathbf{\Delta}$-propriété et il existe donc, en théorie, des poids de liens qui permettent de minimiser la variance de l'estimateur MGPP simple initialement prévu par La Poste quelle que soit la variable d'intérêt. Ces poids sont égaux à l'inverse du nombre d'adresses par tournées, qui vaut 500 en moyenne. L'utilisation de ces poids optimaux ne peut pas être effectuée car il n'est pas possible d'identifier, ni de compter, toutes les adresses d'une tournée.

La Poste a donc été contrainte de réduire le nombre de liens à observer. Pour ce faire, elle a utilisé la MGPP double.

Soit une population intermédiaire U_M, reliée à U_F et à U_T de telle manière que tout individu $k \in U$ soit relié à au moins un individu de U_M et que tout individu de U_M soit relié à au moins un individu de U_F. Dans le cas de La Poste, U_M est la population des cases de tri. On attribue à chaque lien entre $i \in U_F$ et $j \in U_M$ (resp. entre $j \in U_M$ et $k \in U$) un poids de liens θ_{ij}^{FM} (resp. θ_{jk}^{MT}) tel que $\theta_{ij}^{FM} > 0$ si i et j sont liés (resp. $\theta_{jk}^{MT} > 0$ si j et k sont liés), 0 sinon. Le lien entre i et k est pondéré par $\theta_{ik} = \sum_{j \in U_M} \theta_{ij}^{FM}\theta_{jk}^{MT}$et la MGPP double est donc un cas particulier de MGPP simple. Soit $\tilde{\theta}_{ij}^{FM}$, $\tilde{\theta}_{jk}^{MT}$ et $\tilde{\theta}_{ik}$ les poids standardisés associés à θ_{ij}^{FM}, θ_{jk}^{MT} et $\theta_{ik} = \sum_{j \in U_M} \theta_{ij}^{FM}\theta_{jk}^{MT}$. Ces poids sont calculés de manière à avoir

$$\sum_{i \in U_F} \tilde{\theta}_{ik} = \sum_{i \in U_F} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM}\tilde{\theta}_{jk}^{MT} = 1.$$

On parle alors de standardisation globale. L'estimateur MGPP double est obtenu en remplaçant $\tilde{\theta}_{ik}$ par $\sum_{j \in U_M} \tilde{\theta}_{ij}^{FM}\tilde{\theta}_{jk}^{MT}$ dans l'expression de l'estimateur MGPP simple donnée par l'équation (1) :

$$\hat{t}_{y2} = \sum_{k \in s} \left(\sum_{i \in s_F} \frac{1}{\pi_i} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM}\tilde{\theta}_{jk}^{MT} \right) y_k. \tag{5}$$

On considère le cas où les liens entre les trois populations sont de type TpU-TpU, c'est à dire TpU entre U_F et U_M puis entre U_M et U_T. C'est le cas de La Poste, comme illustré par

la figure 2. La somme $\sum_{j\in U_M} \tilde{\theta}_{ij}^{FM}\tilde{\theta}_{jk}^{MT}$ ne contient dans ce cas qu'un seul élément non nul pour un certain $i \in U_F$ et $k \in U$.

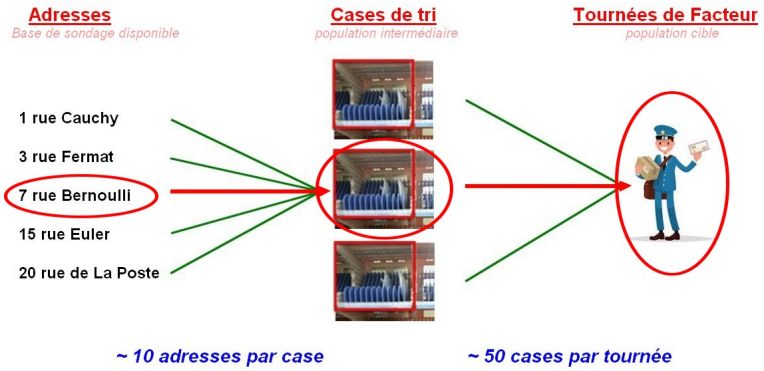

FIGURE 2 – TpU-TpU observé par La Poste.

Dans l'exemple de la figure 2, cet élément non nul est le produit des poids associés aux deux liens rouges. Les poids $\tilde{\theta}_{ij}^{FM}$ et $\tilde{\theta}_{jk}^{MT}$ utilisés à La Poste sont respectivement l'inverse du nombre d'adresses par case et l'inverse du nombre de cases par tournée. Le poids utilisé par La Poste dans l'exemple de la figure 2 est donc $1/5 * 1/3 = 1/15$. Dans le cas de La Poste, ce poids est généralement différent du poids optimal et on a donc une perte de précision de l'estimateur MGPP double par rapport à l'estimateur MGPP simple optimal.

L'intérêt de la MGPP double, comme utilisée par La Poste, est de diminuer le nombre de liens à observer pour calculer les poids standardisés. Cependant, cet avantage se limite aux liens de type TpU-TpU. Pour ce type de liens, on s'intéresse à deux méthodes de standardisation :

— La standardisation double, $\sum_{i\in U_F} \tilde{\theta}_{ij}^{FM} = 1$ et $\sum_{j\in U_M} \tilde{\theta}_{jk}^{MT} = 1$.

— D'autres types de standardisation globale, $\sum_{i\in U_F}\sum_{j\in U_M} \tilde{\theta}_{ij}^{FM}\tilde{\theta}_{jk}^{MT} = 1$, moins contraignants.

Dans l'article Medous et al. (2023), plusieurs méthodes de standardisation sont comparées : la standardisation de la MGPP simple, la standardisation double, et un cas particulier de standardisation globale, où $\theta_{ik} = \sum_{j\in U_M} \theta_{ij}^{FM}\theta_{jk}^{MT}$, $i \in U_F$, $k \in U$ est divisé par le total $\sum_{i\in U_F} \theta_{ik}$. Soit N_{Fk} le nombre de liens entre U_F et $k \in U$, N_{Mk} le nombre de liens entre U_M et k et N_{Fj} le nombre de liens entre U_F et $j \in U_M$. Dans le cas de La Poste, N_{Fk} est le nombre d'adresses dans la tournée k, N_{Mk} le nombre de cases dans la tournée k et N_{Fj} le nombre d'adresses dans la case j. Si les liens sont de type TpU-TpU, $N_{Fk} = \sum_{\substack{j\in U_M \\ j \text{ lié à k}}} N_{Fj}$, c'est-à-dire que le nombre d'adresses dans une tournée est égal à la somme des adresses contenues dans toutes les cases de la tournée. Il faut observer N_{Fk} liens pour la standardisation de la MGPP simple, $N_{Fk} + N_{Mk}$ liens pour la standardisation globale considérée et $N_{Fj} + N_{Mk}$ pour la standardisation double (voir l'exemple de la figure 3). En pratique, si $N_{Fj} + N_{Mk} < N_{Fk}$, alors la standardisation double permet un gain en terme de liens à observer. En particulier, si $N_{Fk} = N_{Fj}N_{Mk}$ et $N_{Fj} > 2$, $N_{Mk} > 2$, alors $N_{Fj} + N_{Mk} < N_{Fk}$ et la MGPP double permet de réduire le nombre de liens à observer par rapport à la MGPP simple. Dans le cas de La Poste, la condition $N_{Fk} = N_{Fj}N_{Mk}$ est satisfaite si toutes les cases contiennent le

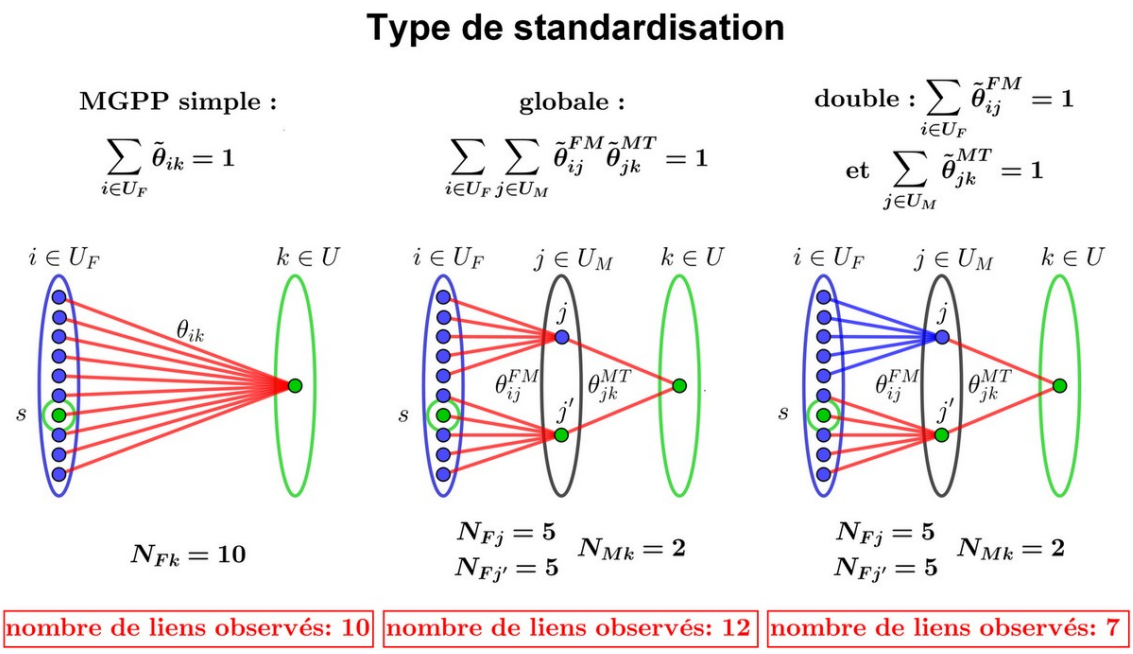

FIGURE 3 – Comparaison des méthodes de standardisation.

même nombre d'adresses, ce qui n'est pas réaliste. Par contre, le nombre d'adresses par case plus le nombre de cases par tournée vaut en moyenne 60 alors que le nombre d'adresses par tournée vaut en moyenne 500. La condition $N_{Fj} + N_{Mk} < N_{Fk}$ est donc satisfaite pour un large nombre de tournées et de cases, ce qui permet un gain important en terme de nombre de liens à observer.

On a vu précédemment que la différence entre la variance de l'estimateur MGPP simple quelconque et la variance de l'estimateur MGPP simple optimal du total dépend, via la standardisation, de la répartition des liens entre les populations. Pour mieux comprendre l'impact de la répartition des liens sur la précision de la MGPP utilisée à La Poste, deux études par simulation ont été faites. Dans chaque simulation, la population cible est la population des tournées, la population intermédiaire celle des cases et la population Frame celle des adresses.

La première étude montre que l'augmentation de variance de la MGPP double comparée à la MGPP simple optimale dans le cas de La Poste est négligeable quand le nombre d'adresses par case est constant, indépendamment du nombre de cases par tournée. Si le nombre d'adresses par case varie de façon importante selon les cases, la perte de précision de la MGPP double dépend du nombre de cases par tournée. La perte de précision est alors plus importante si le nombre de cases varie selon les tournées.

La deuxième étude cherche à reproduire la situation de La Poste pour évaluer la perte de précision de la MGPP double par rapport à la MGPP simple optimale, mais aussi de la MGPP simple optimale par rapport à l'estimateur direct. Les populations sont générées de manière à reproduire la situation observée sur les historiques de La Poste. Cette simulation montre que la MGPP optimale est un peu moins précise que l'estimateur direct et que la MGPP double est bien moins précise que la MGPP optimale. L'utilisation d'une MGPP double non optimale par La Poste est donc responsable de la détérioration de la variance

des estimateurs du trafic postal. L'utilisation de la MGPP simple optimale dans le cas de La Poste pourrait permettre de conserver une variance proche de celle de l'estimateur direct.

Références

Deville, J. and Maumy-Bertrand, M. (2006). Extension of the indirect sampling method and its application to tourism. *Survey Methodology*, 32(2) :177.

Deville, J.-C. and Lavallée, P. (2006). Indirect sampling : The foundations of the generalized weight share method. *Survey Methodology*, Vol. 32(2) :165–176.

Kiesl, H. (2016). Indirect sampling : a review of theory and recent applications. *AStA Wirtschafts-und Sozialstatistisches Archiv*, 10(4) :289–303.

Lavallée, P. (2009). *Indirect sampling*. Springer Science & Business Media.

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.-F., Dessertaine, A., and Puech, P. (2023). Many-to-one indirect sampling with application to the french postal traffic estimation. *The Annals of Applied Statistics*, 17(1) :838–859.