AVERTISSEMENT

LIENS

Code la Propriété Intellectuelle – Articles L. 122-4 et L. 335-1 à L. 335-10
Loi n° 92-597 du 1er juillet 1992, publiée au *Journal Officiel* du 2 juillet 1992

http://www.cfcopies.com/V2/leg/leg-droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

L’université n’entend ni approuver ni désapprouver les opinions particulières du candidat.

Université de Toulouse

THÈSE

LA POSTE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse Capitole (UT Capitole)*

Présentée et soutenue le *14/12/2023* **par :**

Estelle Medous

Méthode généralisée de partage des poids et intégration de données pour l'amélioration de la précision des estimateurs de trafic postal en France

JURY

Patrice BERTAIL	PR, Univ. Paris Nanterre	Rapporteur
Yves TILLÉ	PR, Univ. de Neuchâtel	Rapporteur
Pierre-Andre CORNILLON	MCF, Univ. Rennes 2	Examinateur
Eric GAUTIER	PR, Univ. Toulouse Capitole	Président du jury
Anne RUIZ-GAZEN	PR, Univ. Toulouse Capitole	Directrice
Camelia GOGA	PR, Univ. de Franche-Comté	Co-Directrice
Jean-François BEAUMONT	Chef Recherche, Statistique Canada	Invité
Alain DESSERTAINE	Expert Stratégie DATA, La Poste	Invité
Pauline PUECH	Resp. Statistique, La Poste	Invitée

École doctorale et spécialité :

MITT : Domaine Mathématiques : Mathématiques appliquées

Unité de Recherche :

Toulouse School of Economics (TSE-R)

Directrices de Thèse :

Anne RUIZ-GAZEN et *Camelia GOGA*

Rapporteurs :

Patrice BERTAIL et *Yves TILLÉ*

Remerciements

Je tiens en premier lieu à remercier mes directrices de thèse, Anne Ruiz-Gazen et Camelia Goga, pour m'avoir accompagnée tout au long de ces quatre ans et m'avoir transmis une partie de leur connaissance en théorie des sondages ainsi qu'en rédaction, bien que je sois encore loin de leur niveau dans ces deux domaines. Je voudrais aussi remercier mes encadrants à la Poste, Alain Dessertaine et Pauline Puech, qui m'ont formée et encouragée tout au long de mon stage de Master puis pendant la thèse. Grace à eux j'ai pu découvrir le monde des sondages en entreprise et mettre en pratique la théorie.

Je souhaite aussi remercier mes rapporteurs de thèse, Partice Bertail et Yves Tillé, pour avoir pris le temps de lire mon manuscrit et pour les précieux retours qu'ils m'ont apportés. Je tiens aussi à remercier Pierre-Andre Cornillon et Eric Gautier pour avoir accepté de faire partie de mon jury de thèse.

Je remercie aussi Jean-François Beaumont pour m'avoir fait découvrir Ottawa et pour ses remarques et encouragements.

Je remercie aussi La Poste, qui a financé cette thèse et le stage de Master la précédant. Je remercie particulièrement le service Statistique de La Poste pour m'avoir accueillie et encadrée dans le monde entreprenarial malgré les difficultés rencontrées, et les membres de l'UMR TSE-R pour leur accueil et leur échanges.

Je voudrais remercier ma famille pour leur soutien. Je remercie particulièrement mon frère Charles et un ami de longue date, Guillaume, qui ont fait les mêmes choix d'orientation que moi, pour leurs discussions édifiantes. Je remercie aussi Pepper pour avoir daigné m'accompagner un peu plus longtemps et je remercie Amalric pour avoir été là dans les moments les plus difficiles.

FIGURE 1 – Pepper.

Table des matières

Introduction

La Poste possède un observatoire du trafic postal français qui lui permet, entre autres, de s'adapter à la réalité du terrain et de développer de nouveaux produits. Cet observatoire utilise une étude par sondage nommée SYCI 2 afin de recueillir des données comme le volume et la nature du trafic postal, les délais d'acheminement, ou encore les flux géographiques et économiques.

Actuellement, La Poste cherche à réduire la taille des échantillons prélevés pour SYCI 2, ce qui compromet la précision des estimateurs du trafic total. Mon objectif de thèse consiste à comprendre et améliorer la méthodologie de SYCI 2 pour maintenir la précision des estimateurs de La Poste malgré la réduction des tailles d'échantillons. Mes travaux de recherche s'organisent en deux axes.

Le premier axe consiste à améliorer la méthodologie actuelle, basée sur l'utilisation d'une Méthode Généralisée de Partage des Poids. Cette méthode, introduite par Deville and Lavallée (2006) puis développée dans Lavallée (2007), permet, sous certaines contraintes, d'obtenir un estimateur du total, dit optimal, qui est le plus précis possible en terme de variance, quelle que soit la variable d'intérêt. Bien que de nombreux travaux existent sur ce sujet, tels que Deville and Maumy-Bertrand (2006), De Vitiis et al. (2014), Kiesl (2016) et Haziza and Beaumont (2017), la question de la précision de la MGPP n'est abordée que dans Deville and Lavallée (2006) et Lavallée (2007), où les auteurs montrent l'existence d'estimateurs MGPP optimaux dans quelques cas particuliers. Mes recherches visent à identifier des situations pour lesquelles un estimateur MGPP optimal existe et à proposer des alternatives applicables par La Poste en utilisant les données auxiliaires récoltées dans le cadre de l'étude SYCI 2.

Le deuxième axe porte sur l'utilisation des bases de données massives dont dispose La Poste. Certaines informations sur les objets postaux (lettres, colis ...) triés par machines de tri sont automatiquement récupérées. La Poste souhaite développer une nouvelle méthodologie pour SYCI 2 qui prenne en compte ces données par Intégration Statistique de Données (ISD). L' ISD est un sujet récent mais bien étudié dans la littérature, comme en témoignent des études telles que celles menées par Kim and Tam (2021), Kim et al. (2021) ainsi que les revues

exhaustives des différentes méthodes d' ISD réalisées par Beaumont (2020), Rao (2021), Kim (2022), Wu (2022) et Yang and Kim (2020). L' ISD consiste à utiliser, d'une part, une base de données non probabiliste et d'autre part un échantillon probabiliste pour obtenir des estimations de paramètres sur la population d'intérêt, ici les objets postaux. Cependant, les méthodes étudiées dans la littérature requièrent que les valeurs des variables d'intérêt soient connues pour tout individu contenu dans la base non probabiliste. Les données collectées par machine de tri ne remplissent pas cette condition, car certaines variables d'intérêt, comme le poids et les dimensions d'un objet, ne sont pas récoltées par les machines. La Poste ne peut donc pas utiliser ces méthodes. Le but de mes recherches, dans ce second axe, est de proposer de nouvelles méthodes d' ISD adaptées aux données de La Poste.

Le chapitre 1 présente en détail mon sujet de thèse et la problématique de La Poste. Je commence par rappeler les bases de la théorie des sondages, nécessaires à la compréhension de mes travaux. Une présentation détaillée de l'étude SYCI 2 et des données de La Poste permet de mieux saisir les enjeux pour La Poste. Finalement, je résume les principales contributions apportées dans ma thèse.

Les chapitres 2 et 3 détaillent mes travaux sur le premier axe de recherche. Dans le chapitre 2, on trouve l'article Medous et al. (2023a), publié dans *The Annals of Applied Statistics*. On définit des conditions d'existence pour l'estimateur MGPP optimal puis on introduit le concept de MGPP double, qui est le cas particulier de MGPP (aussi appelée MGPP simple) utilisé par La Poste. Ce cas particulier offre des avantages en termes de faisabilité qui sont étudiés dans le chapitre 2.

La MGPP double mène cependant à une perte de précision comparée à la MGPP simple, comme l'a expérimenté La Poste. L'existence d'estimateurs optimaux pour la MGPP double est discutée dans le chapitre 3, qui fera l'objet d'un article que je compte soumettre prochainement en tant que seule autrice. Je propose aussi dans ce travail des estimateurs alternatifs pour les cas où l'estimateur optimal ne peut être utilisé.

Le chapitre 4 résume mes travaux sur le deuxième axe de recherche. Il contient l'article Medous et al. (2023b), accepté dans *Survey Methodology/Techniques d'Enquête* et qui porte sur l'intégration statistique de données. On commence par étudier certaines méthodes de la littérature pour ensuite proposer des méthodes adaptées aux données observées à La Poste.

Les différents travaux présentés dans ce mémoire ont donné lieu à des publications :

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

- E. Medous, C. Goga, A. Ruiz-Gazen, J.F. Beaumont, A. Dessertaine and P.Puech (2023), QR Prediction for Statistical Data Integration,
  Survey Methodology/Techniques d'Enquête, Statistique Canada, Numéro consacré au 11e Colloque International Francophone sur les Sondages, à paraître,

et je les ai présentés lors des conférences suivantes :

- Comparaison des sondages indirects simple et double. Application à l'estimation du trafic postal en France.
  - Computational and Methodological Statistics (CMStat), en ligne, 19-21 Decembre 2020,
  - Journées de Statistique, Nice (en ligne), 7-11 Juin 2021 .
- Pros and Cons of the Double Indirect Sampling for "Many to One" links.
  Conference in Honour of Fred Smith & Chris Skinner, Southampton, Angleterre (en ligne), 7-9 Juillet 2021 .
- Estimation du trafic postal en France : perte de précision dûe à l'utilisation d'un sondage indirect double.
  Colloque francophone sur les sondages, Bruxelles, Belgique, 6-8 Octobre 2021.
- Une approche par prédiction pour l'intégration de données.
  Forum des Jeunes Mathématicien.nes, Besançon, 8-10 Décembre 2021.
- Introduction à l'intégration de données en sondages.
  Rencontre des Jeunes Statisticiens, Porquerolles, 3-7 Avril 2022.
- Improving finite population inference by data integration & Statistical data integration using a prediction approach.
  5th International Conference on Econometrics and Statistics, Kyoto, Japon (en ligne), 4-6 Juin 2022 (avec A. Ruiz-Gazen).
- Optimality of the double Generalized Weight Share Method and alternatives.
  - Colloque francophone sur les sondages, Aubervilliers, 21-24 mars 2023,
  - Journées de Statistiques, Bruxelles, Belgique, 3-7 Juillet 2023.

ainsi qu'aux séminaires suivants :

- Comparaison des sondages indirects simple et double. Application à l'estimation du trafic postal en France.
  Doctoral Workshop on Decision Mathematics and Statistics, Toulouse, 6 Juillet 2021.
- Pros and Cons of the Double Indirect Sampling for "Many to One" links.
  Seminar of the University of Jyväskylä, Jyväskylä, Finlande (en ligne), 5 Novembre 2021.
- Many-to-One indirect sampling with application to the French postal traffic estimation.

Séminaire MAD-Stat, Toulouse, 10 Mars 2022, (avec A. Ruiz-Gazen).

- A prediction approach for statistical data integration in survey sampling.
  Doctoral Workshop on Decision Mathematics and Statistics, Toulouse, 16 Juin 2022.
- Optimality of the double Generalized Weight Share Method and alternatives.
  Séminaire du Laboratoire de mathématiques de Besançon, 31 Janvier 2023.

Mes travaux portant sur le sondage indirect ont été récompensés par un prix lors de la *Student competition in association with the Conference in Honour of Fred Smith & Chris Skinner*, du 7 au 9 Juillet 2021.

Acronymes

DGWSM Double Generalized Weight Share Method.

EQM Erreur Quadratique Moyenne.

GWSM Generalized Weight Share Method.

HT Horvitz-Thompson.

ISD Intégration Statistique de Données.

MGPP Méthode Généralisée de Partage des Poids.

MtO Many-to-One.

MtO-MtO Many-to-One-Many-to-One.

RAO Référentiel des Adresses Organisées.

RRMSE Relative Root Mean Square Error.

SASSR Sondage Aléatoire Simple Sans Remise.

SDI Statistical Data Integration.

SOJ SOrties-Jours.

SRSWOR Simple Random Sampling Without Replacement.

SSRSWOR Stratified Simple Random Sampling Without Replacement.

STASSR sondage STratifié Aléatoire Simple Sans Remise.

SYCI SYstème de Collecte de l'Information.

TAE Traitement Automatisé de l'Enveloppe.

TpU Tous-pour-Un.

TpU-TpU Tous-pour-Un-Tous-pour-Un.

UP Unités Primaires.

US Unités Secondaires.

Chapitre 1

Rappels de théorie des sondages, contexte et contributions

1.1 Théorie des sondages

Dans cette section, on introduit les notions de théorie des sondages nécessaires à la compréhension de mes travaux de thèse. On rappelle les méthodes classiques d'échantillonnage et d'estimation, ainsi que le concept d'asymptotique en théorie des sondage. On présente ensuite le sondage indirect et la méthode généralisée de partage des poids (MGPP) sur lesquels porte une grande partie de mes travaux. Pour plus de détails sur cette section, on peut se référer aux ouvrages suivants : Hájek (1981), Särndal et al. (1989), Ardilly (2006), Lavallée (2007) et Tillé (2019).

1.1.1 Principes

On appelle population finie de taille N, notée U, une collection de N individus possédant un ensemble de caractéristiques appelées variables. On peut considérer par exemple la population des chats français comme dans la figure 1.1 ou, plus communément, la population des personnes habitant en France. Ces personnes forment une population finie de taille 68 042 591 (début 2023) et sont caractérisés par leur sexe, taille, âge, adresse...

On dit qu'une population U est cible si on souhaite connaître certains paramètres pour cette population, comme le total ou la moyenne d'une certaine variable. On peut aussi s'intéresser à des paramètres plus complexes, comme la médiane d'une certaine variable ou les covariances entre différentes variables. Le paramètre étudié et les variables qui permettent de le calculer sont dits d'intérêt. Par exemple, on peut s'intéresser au pourcentage de femmes vivant en France, à leur revenu médian ou au nombre total de chats de compagnie en France.

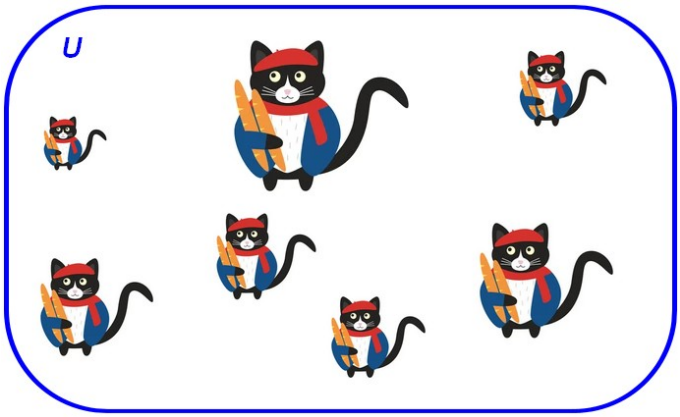

FIGURE 1.1 – Population des (chats) français.

La variable d'intérêt est respectivement le sexe d'un individu, son revenu et le nombre de chats résidant chez lui. Les variables qui ne sont pas d'intérêt, comme le niveau d'étude ou le type de logement, sont dites auxiliaires.

Soit θ_y un paramètre d'intérêt qui est une fonction d'une variable d'intérêt y sur U. On peut calculer la valeur de ce paramètre si on observe tous les individus de la population. Cette méthode, appelée recensement, est coûteuse et complexe à mettre en place. Par exemple, pour recenser les personnes habitant en France, il faut observer chacune des 68 042 591 personnes composant la population et leur poser des questions, soit par téléphone, par mail, en face à face... Cela nécessite donc une infrastructure conséquente.

Une autre méthode, appelée sondage, consiste à sélectionner, le plus souvent de manière aléatoire, un sous-groupe d'individus s, appelé échantillon, comme illustré dans la figure 1.2. Les paramètres d'intérêt sont ensuite estimés à partir des valeurs des variables d'intérêt observées pour les individus échantillonnés. On appelle méthode d'échantillonnage la façon dont est tiré l'échantillon. Différentes méthodes d'échantillonnage seront présentées en section 1.1.2.

Mettre en place un sondage comporte de nombreuses étapes. Il faut d'abord identifier les objectifs de l'étude ainsi que la population et les paramètres d'intérêt. Ensuite, il faut identifier les informations auxiliaires disponibles au niveau de la population d'intérêt. Ces données peuvent servir à décider la façon dont les individus seront sélectionnés dans l'échantillon.

Les données récoltées au niveau de l'échantillon, parfois couplées avec les informations auxiliaires disponibles au niveau de la population, vont permettre d'estimer les paramètres d'intérêt. L'étape qui consiste à calculer ces estimateurs est appelée estimation et sera détaillée en section 1.1.3.

Il existe d'autres étapes dans la mise en pratique d'un sondage, comme la rédaction du questionnaire ou l'encodage des résultats. Cependant, mes travaux portant uniquement sur

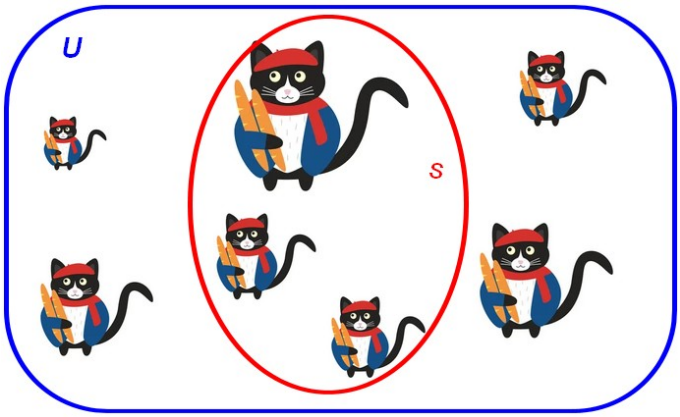

FIGURE 1.2 – Échantillon de (chats) français.

les étapes d'échantillonnage et d'estimation, je ne m'attarderai pas sur les autres étapes dans ce chapitre.

1.1.2 Échantillonnage

Dans cette section, ainsi que dans le reste du chapitre, on considère les variables d'intérêt comme non aléatoires.

Tirer un échantillon de manière aléatoire nécessite d'identifier de manière unique tous les individus de la population. On appelle base de sondage la liste contenant tous les individus de U identifiés de manière unique, comme illustré dans la figure 1.3. La base de sondage contient généralement des informations auxiliaires sur les individus de la population, comme le lieu de naissance, le nom...

Soit S l'ensemble des échantillons possibles et $\mathscr{B}(S)$ la tribu borélienne engendrée par S. La taille d'un échantillon $s \in S$, notée n_s, peut être fixe ou variable. On appelle plan de sondage $p(\cdot)$ une loi de probabilités sur $(S, \mathscr{B}(S))$ telle que $p(s)$ est la probabilité d'échantillonner $s \in S$. Le plan de sondage consiste donc à attribuer une probabilité de tirage à chaque échantillon $s \in S$. Un plan de sondage est déterminé avant le tirage en utilisant les données accessibles dans la base de sondage. Dans cette section, je présente les plans de sondage classiques utilisés dans mes travaux.

Pour un individu $k \in U$, on note I_k l'indicatrice d'inclusion dans un échantillon de l'individu $k \in U$, où $I_k = 1$ si k est sélectionné et 0 sinon. Les indicatrices d'inclusion sont aléatoires et un échantillon est composé de tous les individus $k \in U$ dont la valeur de I_k vaut 1.

On appelle probabilité d'inclusion de premier ordre de l'individu $k \in U$, notée π_k, la probabilité que l'individu k soit sélectionné dans un échantillon et probabilité d'inclusion

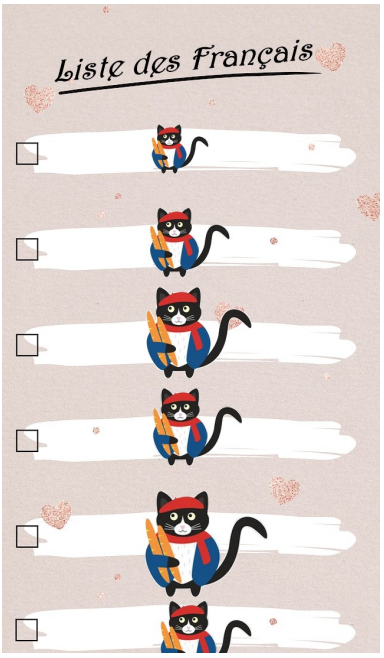

FIGURE 1.3 – Base de sondage des (chats) français.

de second ordre du couple $k, k' \in U$, notée $\pi_{kk'}$, la probabilité que k et k' soit sélectionnés dans un échantillon. Soit $\mathrm{E}[\cdot]$ l'espérance sous le plan de sondage $p(\cdot)$. On a $\pi_k = \mathrm{E}[I_k]$ et $\pi_{kk'} = \mathrm{E}[I_k I_{k'}]$, ce qui implique que $\pi_{kk} = \mathrm{E}[I_k I_k] = \mathrm{E}[I_k] = \pi_k$. La covariance entre I_k et $I_{k'}$ est donnée par :

$$\mathrm{Cov}(I_k, I_{k'}) = \mathrm{E}(I_k I_{k'}) - \mathrm{E}(I_k)\mathrm{E}(I_{k'}) = \pi_{kk'} - \pi_k \pi_{k'}, \ k, k' \in U.$$

Les probabilités d'inclusion de premier et second ordre permettent de construire un estimateur de θ_y et de déterminer sa précision, ce qui sera détaillé dans la section 1.1.3. Les valeurs de π_k et $\pi_{kk'}$, $k, k' \in U$, seront données pour chaque plan présenté dans cette section.

Sondage Aléatoire Simple Sans Remise

Le Sondage Aléatoire Simple Sans Remise (SASSR) est un plan très simple à mettre en oeuvre qui permet de tirer des échantillons sans remise d'une taille donnée n. Le SASSR consiste à attribuer la même probabilité d'être tiré à chaque échantillon sans remise s de taille n et une probabilité nulle aux autres échantillons. Comme il existe $\binom{N}{n}$ échantillons sans remise de taille n dans U, la probabilité de tirage d'un échantillon s est $p(s) = 1/\binom{N}{n}$. On peut alors calculer la probabilité que l'individu k soit sélectionné dans un échantillon, qui est le ratio entre le nombre d'échantillons contenant k et le nombre total d'échantillons

possibles :

$$\pi_k = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}, k \in U.$$

On peut dériver de manière similaire la probabilité d'inclusion de second ordre :

$$\pi_{kk'} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}, k \neq k' \in U.$$

Avec un SASSR, les probabilités d'inclusion de premier ordre sont identiques pour tous les individus $k \in U$. On dit que c'est un plan de sondage à probabilités égales.

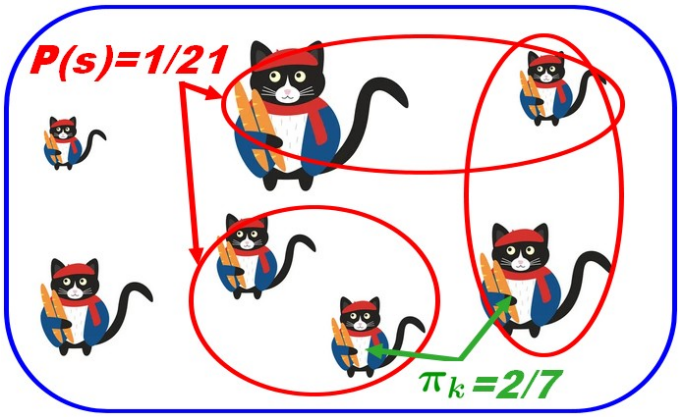

FIGURE 1.4 – Echantillons tirés via SASSR.

Par exemple, dans la figure 1.4, on a une population composée de 7 chats français et on veut tirer un échantillon sans remise de taille 2 dans cette population. Il existe 21 échantillons de taille 2, donc la probabilité de tirer un de ces échantillons est 1/21. De là, on calcule la probabilité d'observer un chat français dans cette population, qui est de 2/7 pour tous les chats.

Plan de Poisson

Le plan de Poisson permet de tirer des échantillons sans remise tels que les indicatrices d'inclusion soient indépendantes.

On fixe la probabilité d'inclusion du premier ordre π_k pour chaque individu $k \in U$ grâce aux informations disponibles dans la base de sondage et on sélectionne de façon indépendante et sans remise les individus avec les probabilités prédéfinies $\pi_1, \ldots, \pi_N$. Si les probabilités π_k, $k \in U$ sont choisies fixes égales à π, le plan est dit de Bernoulli.

Pour un plan de Poisson, les indicatrices d'inclusion I_k pour $k \in U$ suivent une loi de Bernoulli de paramètre π_k et elles sont indépendantes. La taille n_s de l'échantillon est

donc aléatoire. Si les probabilités π_k sont égales à π, alors la taille de l'échantillon suit une binomiale de paramètres N et π.

L'indépendance des indicatrices d'inclusion implique que pour tout $k, k' \neq k \in U$, $\pi_{kk'} = \pi_k \pi_{k'}$ et que la covariance entre deux indicatrices I_k et $I_{k'}$ avec $k \neq k'$ est nulle.

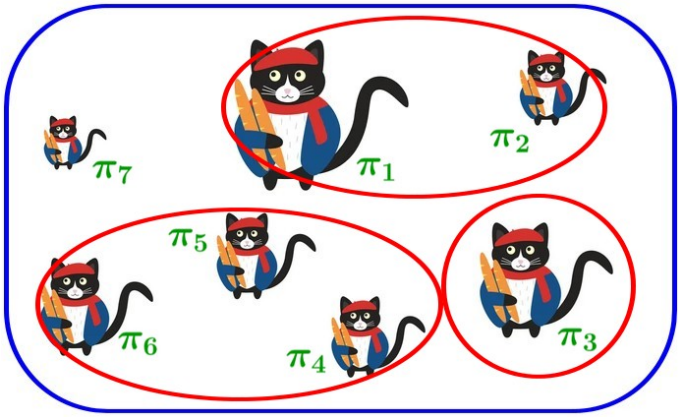

FIGURE 1.5 – Échantillons tirés via plan de Poisson.

Dans la figure 1.5, on fixe les probabilités π_k pour chaque chat français. Trois échantillons possibles avec un plan de Poisson sont présentés en rouge. L'un est de taille 1, le deuxième de taille 2 et le troisième de taille 3.

Plan stratifié

Si la base de sondage contient suffisamment d'information sur les individus de U, il est possible de partitionner U dans des sous-populations disjointes U_h, $h = 1, \ldots, H$ avec $U = \cup_{h=1}^{H} U_h$. Ces sous-populations s'appellent des strates. Un plan de sondage stratifié consiste à tirer de manière indépendante un échantillon s_h dans chaque strate U_h, $h = 1, \ldots, H$. L'échantillon final est donné par $s = \cup_{h=1}^{H} s_h$. Si le tirage au sein de chaque strate s'effectue avec un SASSR, alors le plan est appelé sondage STratifié Aléatoire Simple Sans Remise (STASSR).

Ce type d'échantillonnage est utile si on souhaite par exemple effectuer des estimations pour certaines sous-populations correspondant aux strates. Par exemple, si on souhaite connaître le pourcentage de propriétaires de chat dans chaque région française. Il permet aussi d'améliorer dans certaines situations la précision des estimateurs de θ_y, ce qui sera détaillé en section 1.1.3.

Soit l'ensemble de strates U_h, $h = 1, \ldots, H$ de taille respective N_h correspondant à la partition de U. Pour un STASSR où l'échantillon s_h dans la strate U_h, $h = 1 \ldots H$, est de

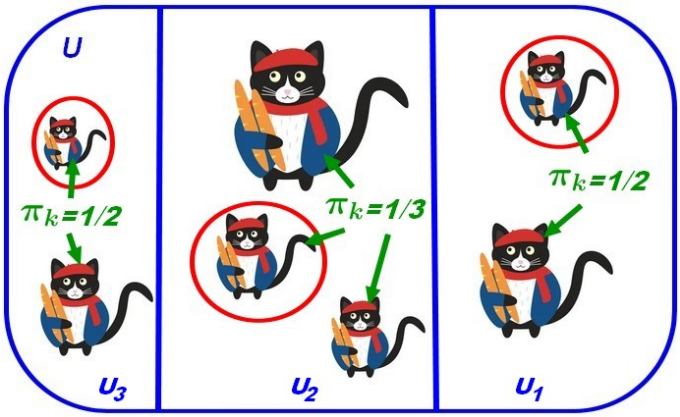

FIGURE 1.6 – Échantillon tiré via STASSR.

taille n_h, la probabilité d'inclusion de premier ordre pour un individu $k \in U$ est :

$$\pi_k = \frac{n_h}{N_h} \text{ si } k \in U_h.$$

La probabilité d'inclusion de second ordre pour deux individus k, $k' \neq k \in U$ est :

$$\pi_{kk'} = \begin{cases} \frac{n_h(n_h-1)}{N_h(N_h-1)} & \text{si } k, k' \in U_h, \\ \frac{n_h}{N_h}\frac{n_{h'}}{N_{h'}} & \text{si } k \in U_h \text{ et } k' \in U_{h'},\ h \neq h'. \end{cases}$$

Dans la figure 1.6, la population de 7 chats français est décomposée en trois strates U_1, U_2 et U_3 de taille respective 2, 3 et 2. Dans chaque strate on tire un échantillon de taille 1 par SASSR. Les probabilités d'inclusion de premier ordre sont donc 1/2 pour les individus appartenant aux strates 1 et 3, et 1/3 pour ceux appartenant à la strate 2.

Plan à deux degrés

Le plan SASSR, de Poisson ou les plans stratifiés, nécessitent d'avoir une base de sondage des individus. Quand cette base de sondage est difficile à obtenir, on peut faire appel aux plans à plusieurs degrés. Nous allons dans la suite décrire le plan à deux degrés.

La population U est découpée en N_I sous groupes U_i, $i \in \{1, \ldots, N_I\}$ de taille respective N_i, appelés Unités Primaires (UP). Les individus $k \in U_i$, $i \in \{1, \ldots, N_I\}$ sont appelés Unités Secondaires (US).

Le plan à deux degrés consiste à tirer un échantillon s_I d'UP avec les informations disponibles dans la base de sondage des UP. Pour chaque UP U_i contenue dans s_I, on récolte suffisamment d'information pour former une base de sondage adéquate pour U_i. Cette base de sondage permet de tirer un échantillon d'US s_i au sein de l'UP U_i. L'échantillon final s

est composé de tous les individus contenus dans les échantillons s_i, $i \in s_I$, $s = \cup_{i \in s_I} s_i$.

Pour tout $i, i' \in \{1, \ldots, N_I\}$, on note π_{Ii} la probabilité d'inclusion de l'UP U_i et $\pi_{Iii'}$ la probabilité d'inclusion des UP U_i et $U_{i'}$ dans un échantillon d'UP. Soient k et k' deux US contenues respectivement dans les UP U_i et $U_{i'}$. On note $\pi_{IIk|i}$ et $\pi_{IIkk'|i,i'}$ les probabilités conditionnelles d'inclusion dans s de, respectivement, l'US $k \in U_i$ et les deux US $k \in U_i$, $k' \in U_{i'}$ sachant que les UP U_i et $U_{i'}$ appartiennent à l'échantillon s_I. Si $i = i'$, on note $\pi_{IIkk'|i,i} = \pi_{IIkk'|i}$.

On note S_I l'ensemble des échantillons d'UP possibles et $p(s_I)$ la probabilité de tirer l'échantillon s_I. La probabilité d'inclusion dans l'échantillon final s de premier ordre pour un individu $k \in U$ est égale à :

$$\pi_k = \sum_{\substack{s_I \in S_I \\ U_i \in s_I}} p(s_I) \pi_{IIk|i} \text{ si } k \in U_i.$$

La probabilité d'inclusion dans l'échantillon final s de second ordre pour deux individus $k, k' \in U$ est égale à :

$$\pi_{kk'} = \sum_{\substack{s_I \in S_I \\ U_i, U_{i'} \in s_I}} p(s_I) \pi_{IIkk'|i,i'} \text{ si } k \in U_i, k' \in U_{i'}.$$

Pour un plan à deux degrés défini de façon générale comme ci-dessus, le calcul des probabilités d'inclusion des individus $k \in U$ peut s'avérer assez complexe. C'est pour cette raison que l'on suppose en général les conditions supplémentaires suivantes :

— le tirage des unités secondaires est effectué indépendamment au sein de chaque UP (indépendance),

— le plan de sondage utilisé pour tirer les unités secondaires ne dépend pas de l'échantillon s_I tiré au premier degré (invariance).

Sous ces conditions, les probabilités d'inclusion des individus $k \in U$ dans l'échantillon final s se simplifient. La probabilité d'inclusion de premier ordre pour un individu $k \in U$ devient :

$$\pi_k = \sum_{\substack{s_I \in S_I \\ U_i \in s_I}} p(s_I) \pi_{IIk|i} = \pi_{IIk|i} \sum_{\substack{s_I \in S_I \\ U_i \in s_I}} p(s_I) = \pi_{IIk|i} \pi_{Ii} \text{ si } k \in U_i,$$

car $\pi_{IIk|i}$ ne dépend pas de s_I par invariance. La probabilité d'inclusion de second ordre pour deux individus $k, k' \in U$ devient :

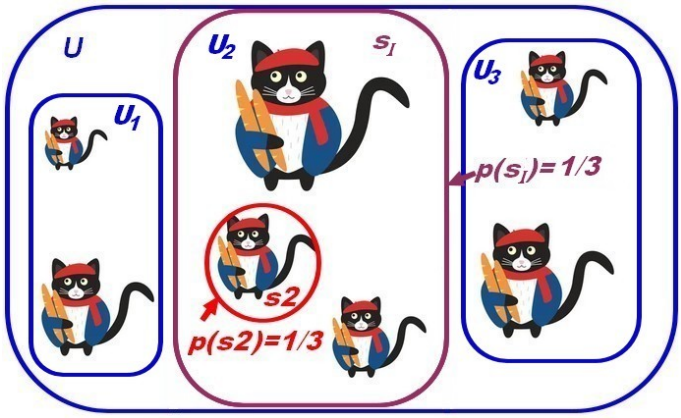

FIGURE 1.7 – Echantillons tirés via un plan à deux degrés avec SASSR à chaque degré.

— Si $i = i'$, alors par invariance, $\pi_{IIkk'|i}$ ne dépend pas de s_I :

$$\pi_{kk'} = \sum_{\substack{s_I \in S_I \\ U_i \in s_I}} p(s_I)\pi_{IIkk'|i} = \pi_{IIkk'|i} \sum_{\substack{s_I \in S_I \\ U_i \in s_I}} p(s_I) = \pi_{IIkk'|i}\pi_{Ii} \text{ si } k, k' \in U_i,$$

— si $i \neq i'$, alors par indépendance $\pi_{IIkk'|i,i'} = \pi_{IIk|i}\pi_{IIk'|i'}$ et par invariance, $\pi_{IIk|i}$ et $\pi_{IIk'|i'}$ ne dépendent pas de s_I, alors pour tout $k \in U_i, k' \in U_{i'}$, :

$$\pi_{kk'} = \sum_{\substack{s_I \in S_I \\ U_i, U_{i'} \in s_I}} p(s_I)\pi_{IIk|i}\pi_{IIk'|i'} = \pi_{IIk|i}\pi_{IIk'|i'} \sum_{\substack{s_I \in S_I \\ U_i, U_{i'} \in s_I}} p(s_I) = \pi_{IIk|i}\pi_{IIk'|i'}\pi_{Iii'}.$$

Soit un plan à deux degrés avec SASSR à chaque degré, avec n_I le nombre d'UP échantillonnées au premier degré et n_i le nombre d'US échantillonnées au deuxième degré au sein de l'UP U_i, $i \in \{1, \ldots, N_I\}$. Alors les probabilités d'inclusion de premier et second ordre pour les UP sont données par :

$$\pi_{Ii} = \frac{n_I}{N_I}, \ \pi_{Iii'} = \frac{n_I(n_I - 1)}{N_I(N_I - 1)}, \ i, i' \neq i \in \{1, \ldots, N_I\}.$$

Les probabilités conditionnelles d'inclusion de premier et second ordre pour les US sont données par :

$$\pi_{IIk|i} = \frac{n_i}{N_i}, \ \pi_{IIkk'|i} = \frac{n_i(n_i - 1)}{N_i(N_i - 1)}, \ k, k' \neq k \in U_i \text{ et } \pi_{IIkk'|i,i'} = \frac{n_i}{N_i}\frac{n_{i'}}{N_{i'}} \text{ si } k \in U_i, k' \in U_{i'}, i \neq i'.$$

Considérons maintenant l'exemple de la figure 1.7. La population est découpée en trois UP de taille 2 pour U_1 et U_3 et 3 pour U_2. Les chats français constituent les US. Un SASSR est utilisé à chaque degré, avec une UP échantillonnée au premier degré et une US échantillonnée

par UP de s_I. La probabilité d'observer une UP est donc de 1/3 pour toutes les UP. Dans notre exemple, l'UP U_2 est échantillonnée. Au sein de U_2, la probabilité d'observer un chat est de 1/3 pour chaque chat. La probabilité d'inclusion de premier ordre du chat échantillonné dans s_2 est donc 1/9.

1.1.3 Estimation

Soit $T_y = \sum_{k \in U} y_k$ le total de la variable d'intérêt y sur U. Seuls des estimateurs $\hat{T}_y$ du total T_y sont détaillés dans cette section. Ce choix est motivé par deux points : d'une part, les paramètres d'intérêt à La Poste sont des totaux, et d'autre part, les estimateurs de la moyenne $\bar{Y} = T_y/N$ se calculent facilement à partir de $\hat{T}_y$ si N est connue :

$$\hat{\bar{Y}} = \hat{T}_y/N.$$

L'objectif d'un sondage est de trouver une bonne stratégie, c'est à dire plan de sondage et estimateur, pour estimer un paramètre d'intérêt, le total dans notre cas. Soit $\hat{T}_y(s)$ la valeur prise par l'estimateur sur l'échantillon $s \in S$. On désigne par $\mathrm{E}[\hat{T}_y] = \sum_{s \in S} p(s)\hat{T}_y(s)$ l'espérance sous le plan d'un estimateur $\hat{T}_y$. La qualité d'un estimateur est généralement mesurée par son biais

$$\mathrm{Biais}(\hat{T}_y) = \mathrm{E}[\hat{T}_y] - T_y,$$

sa variance

$$\mathrm{Var}(\hat{T}_y) = \mathrm{E}[(\hat{T}_y - \mathrm{E}[\hat{T}_y])^2]$$

et l'Erreur Quadratique Moyenne (EQM) qui en découle

$$\mathrm{E}[(\hat{T}_y - T_y)^2] = \mathrm{Var}(\hat{T}_y) + \mathrm{Biais}^2(\hat{T}_y).$$

La précision d'un estimateur est donnée par sa variance. Si la variance de l'estimateur $\hat{T}_y$ est faible sous le plan de sondage $p(\cdot)$ utilisé, alors cette stratégie d'échantillonnage sera performante.

Dans cette section je présente les estimateurs les plus courants, qui sont nécessaires à la compréhension de mes travaux.

Estimateur de Horvitz-Thompson

L'estimateur par expansion, ou estimateur de Horvitz-Thompson (HT) du total T_y est l'estimateur le plus simple, introduit par Narain (1951) et Horvitz and Thompson (1952).

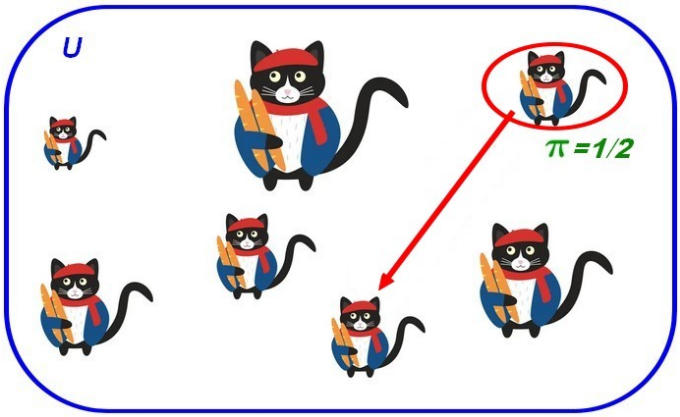

FIGURE 1.8 – Expansion d'un individu.

Cet estimateur ne nécessite que la connaissance des probabilités d'inclusion d'ordre un, π_k, $k \in U$.

Soit un individu $k \in U$, avec une probabilité $\pi_k \neq 0$ d'être sélectionné. Cet individu peut être vu comme le représentant de $1/\pi_k$ individus de la population. On appelle poids de sondage, noté d_k, la valeur $1/\pi_k$ pour $k \in U$. Dans la figure 1.8, le chat sélectionné a une probabilité d'inclusion de $1/2$, il représente donc 2 individus de la population qui ont des tailles similaires ; lui même et le chat indiqué par la flèche.

L'estimateur de HT étend les observations obtenues pour l'échantillon à l'ensemble de la population, comme montré dans la figure 1.9, en multipliant les valeurs y_k observées pour les individus $k \in s$ par leur poids de sondage d_k.

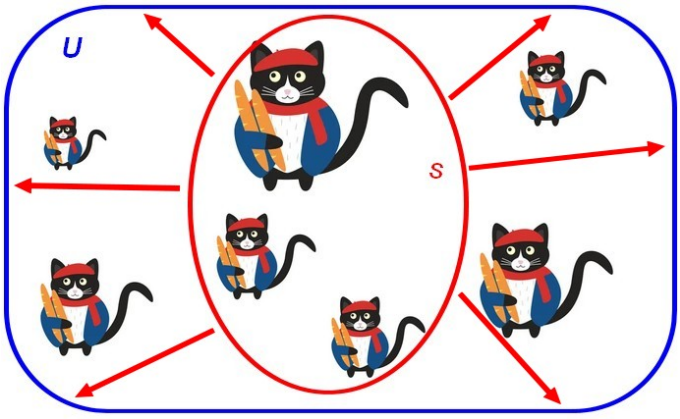

FIGURE 1.9 – Expansion de l'échantillon.

L'estimateur de HT d'un total $T_y = \sum_{k \in U} y_k$ est

$$\hat{T}_{HT,y} = \sum_{k \in s} d_k y_k. \tag{1.1}$$

Comme l'échantillon s est l'ensemble des indicatrices d'inclusion I_k, $k \in U$, telles que $I_k = 1$, l'estimateur de HT peut se réécrire

$$\hat{T}_{HT,y} = \sum_{k \in U} I_k d_k y_k.$$

Si $\pi_k \neq 0$ pour tout $k \in U$, l'estimateur de HT est sans biais. Comme seules les indicatrices d'inclusion sont aléatoires, on a :

$$\mathrm{E}[\hat{T}_{HT,y}] = \mathrm{E}[\sum_{k \in U} I_k d_k y_k] = \sum_{k \in U} \mathrm{E}[I_k] d_k y_k = \sum_{k \in U} \pi_k d_k y_k = \sum_{k \in U} y_k = T_y.$$

Sa variance est donnée par :

$$\begin{aligned} \mathrm{Var}(\hat{T}_{HT,y}) &= \mathrm{Var}(\sum_{k \in U} I_k d_k y_k) = \sum_{k \in U} \sum_{k' \in U} \mathrm{Cov}(I_k, I_{k'}) d_k d_{k'} y_k y_{k'} \\ &= \sum_{k \in U} \sum_{k' \in U} \frac{\pi_{kk'} - \pi_k \pi_{k'}}{\pi_k \pi_{k'}} y_k y_{k'}. \end{aligned} \tag{1.2}$$

La variance étant une somme double sur la population, elle ne peut pas être connue à partir de l'échantillon. On ne connaît donc pas la vraie précision d'un estimateur et il faut l'estimer.

Il existe plusieurs estimateurs pour la variance de l'estimateur de HT, mais le plus simple consiste à utiliser un estimateur par expansion :

$$\widehat{\mathrm{Var}}(\hat{T}_{HT,y}) = \sum_{k \in s} \sum_{k' \in s} \frac{1}{\pi_{kk'}} \frac{\pi_{kk'} - \pi_k \pi_{k'}}{\pi_k \pi_{k'}} y_k y_{k'}. \tag{1.3}$$

Cet estimateur de la variance est sans biais si $\pi_{kk'} \neq 0$ pour tout $k, k' \in U$.

Les résultats présentés ci dessus sont valables quelque soit le plan de sondage. Cependant elles se simplifient si un SASSR est utilisé, soit directement, soit dans un plan stratifié ou à deux degrés.

Plan SASSR

Si un SASSR de taille n est utilisé pour tirer l'échantillon, on a :

$$\hat{T}^{SAS}_{HT,y} = \sum_{k \in s} \frac{N}{n} y_k = N \bar{y},$$

où $\bar{y}$ est la moyenne de y dans l'échantillon. Sa variance est donnée par :

$$\mathrm{Var}(\hat{T}^{SAS}_{HT,y}) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n},$$

où $S_y^2 = (N-1)^{-1} \sum_{k \in U} (y_k - \bar{Y})^2$. Cette variance est estimée sans biais par :

$$\widehat{\text{Var}}(\hat{T}_{HT,y}^{SAS}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n},$$

où $s_y^2 = (n-1)^{-1} \sum_{k \in s} (y_k - \bar{y})^2$.

Plan STASSR

Soit une décomposition de U en H strates de taille N_h, $h = 1, \ldots, H$. On note $T_{yh} = \sum_{k \in U_h} y_k$ le total de y sur la strate U_h. Si un STASSR de taille n avec H échantillons s_h de taille n_h, $h = 1, \ldots, H$, telles que $n = \sum_{h=1}^{H} n_h$, est utilisé, alors l'estimateur de HT du total T_y est la somme des estimateurs des totaux T_{yh} :

$$\hat{T}_{HT,y}^{ST} = \sum_{h=1}^{H} \hat{T}_{HT,yh}^{SAS} = \sum_{h=1}^{H} \sum_{k \in s_h} \frac{N_h}{n_h} y_k = \sum_{h=1}^{H} N_h \bar{y}_h,$$

où $\hat{T}_{HT,yh}^{SAS}$ est l'estimateur SASSR du total T_{yh} et $\bar{y}_h$ est la moyenne de y dans l'échantillon s_h. Grâce à l'indépendance du tirage entre les strates, sa variance est donnée par :

$$\text{Var}(\hat{T}_{HT,y}^{ST}) = \sum_{h=1}^{H} \text{Var}(\hat{T}_{HT,yh}^{SAS}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{yh}^2}{n_h},$$

où $S_{yh}^2 = (N_h - 1)^{-1} \sum_{k \in U_h} (y_k - \bar{Y}_h)^2$ et $\bar{Y}_h$ est la moyenne de y dans la strate U_h. Cette variance est estimée sans biais par :

$$\widehat{\text{Var}}(\hat{T}_{HT,y}^{ST}) = \sum_{h=1}^{H} \widehat{\text{Var}}(\hat{T}_{HT,yh}^{SAS}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{yh}^2}{n_h},$$

où $s_{yh}^2 = (n_h - 1)^{-1} \sum_{k \in s_h} (y_k - \bar{y}_h)^2$.

Il est possible de décomposer S_y^2 comme suit :

$$S_y^2 = \frac{1}{N-1} \sum_{h=1}^{H} (N_h - 1) S_{yh}^2 + \frac{1}{N-1} \sum_{h=1}^{H} N_h (\bar{Y}_h - \bar{Y})^2 = S_{y,intra}^2 + S_{y,inter}^2,$$

avec $S_{y,intra}^2 = (1/(N-1)) \sum_{h=1}^{H} (N_h - 1) S_{yh}^2$ mesurant la variabilité au sein des strates (intra-strates) de la variable y, $S_{y,inter}^2 = (1/(N-1)) \sum_{h=1}^{H} N_h (\bar{Y}_h - \bar{Y})^2$ mesurant la variabilité entre les strates (inter-strates) des moyennes $\bar{Y}_h$, $h = 1, \ldots, H$ et $(N-1) S_y^2$ mesurant la variabilité

de y sur la population U. La variance de $\hat{T}_{HT,y}^{SAS}$ dépend de $(N-1)S_y^2$ et celle de $\hat{T}_{HT,y}^{ST}$ dépend de $S_{y,intra}^2$.

Sauf cas particuliers, il n'existe pas de propriété générale assurant un gain de précision du STASSR par rapport au SASSR quand un estimateur de HT est utilisé. Toutefois, le plan de sondage stratifié peut permettre un gain de précision par rapport au plan SASSR si la variabilité intra-strates est faible par rapport à la variabilité inter-strates.

Plan à deux degrés avec SASSR à chaque degré

Soit une décomposition de U en UP U_i, $i \in \{1, \ldots, N_I\}$. On note T_{yi} le total et $\bar{Y}_i = T_{yi}/N_i$ la moyenne de y dans l'UP U_i, $i \in \{1, \ldots, N_I\}$. Si un plan à deux degrés avec SASSR de taille n_I au premier degré et SASSR de taille n_i, $i \in \{1, \ldots, N_I\}$ au second degré est utilisé, alors l'estimateur de HT du total T_y devient :

$$\hat{T}_{HT,y}^{deg} = \sum_{U_i \in s_I} \sum_{k \in s_i} \frac{N_I}{n_I} \frac{N_i}{n_i} y_k = \sum_{U_i \in s_I} \frac{N_I}{n_I} \hat{T}_{HT,yi}$$

avec $\hat{T}_{HT,yi} = \sum_{k \in s_i} \frac{N_i}{n_i} y_k$ l'estimateur de HT conditionnel au premier degré du total T_{yi}, $i \in \{1, \ldots, N_I\}$.

La variance de $\hat{T}_{HT,y}^{deg}$ est donnée par :

$$\mathrm{Var}(\hat{T}_{HT,y}^{deg}) = V_{PSU} + V_{SSU},$$

avec $V_{PSU} = N_I^2 \left(1 - \frac{n_I}{N_I}\right) \frac{S_{yI}^2}{n_I}$, $S_{yI}^2 = (N_I - 1)^{-1} \sum_{i=1}^{N_I} \left(T_{yi} - \frac{T_y}{N_I}\right)^2$, $V_{SSU} = \frac{N_I}{n_I} \sum_{i=1}^{N_I} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{yi}^2}{n_i}$ et $S_{yi}^2 = (N_i - 1)^{-1} \sum_{k \in U_i} (y_k - \bar{Y}_i)^2$.

V_{PSU} mesure la variance due au premier degré (Primary Sampling Unit) et V_{SSU} mesure la variance due au deuxième degré (Secondary Sampling Unit).

Soit $\bar{y}_i = \sum_{k \in s_i} y_k / n_i$ la moyenne de y dans l'échantillon s_i, $i \in \{1, \ldots, N_I\}$. Cette variance est estimée sans biais par :

$$\widehat{\mathrm{Var}}(\hat{T}_{HT,y}^{deg}) = \hat{V}_{PSU} + \hat{V}_{SSU},$$

avec $\hat{V}_{PSU} = N_I^2 \left(1 - \frac{n_I}{N_I}\right) \frac{s_{yI}^2}{n_I}$, $s_{yI}^2 = (n_I - 1)^{-1} \sum_{i \in s_I} \left(\hat{T}_{HT,yi} - \left(\frac{\sum_{i \in s_I} \hat{T}_{HT,yi}}{N_I}\right)\right)^2$, $\hat{V}_{SSU} = \frac{N_I}{n_I} \sum_{i \in s_I} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{s_{yi}^2}{n_i}$ et $s_{yi}^2 = (n_i - 1)^{-1} \sum_{k \in s_i} (y_k - \bar{y}_i)^2$.

Estimateur de Hájek

L'estimateur de HT reste aléatoire même si la variable y est constante. Si toutes les valeurs y_k, $k \in U$ sont égales à une constante C, il semble logique que la valeur estimée pour T_y soit égale à NC, quel que soit l'échantillon tiré s. Or, la valeur de l'estimateur de HT du total T_y dans ce cas est égale à :

$$\hat{T}_{HT,C} = C \sum_{k \in s} d_k,$$

qui n'est généralement pas égale à NC et est aléatoire.

Ce problème peut être corrigé, si la valeur N est connue, en utilisant l'estimateur de Hájek, introduit par Hájek (1971) et donné par :

$$\hat{T}_{HA,y} = \frac{N}{\sum_{k \in s} d_k} \sum_{k \in s} d_k y_k.$$

Dans le cas où $y_k = C$ pour tout $k \in U$, l'estimateur de Hájek du total est :

$$\hat{T}_{HA,y} = C \frac{N}{\sum_{k \in s} d_k} \sum_{k \in s} d_k = NC.$$

Soit $\hat{N}_{HT} = \sum_{k \in s} d_k$ l'estimateur de HT du total $N = \sum_{k \in U} 1$. L'estimateur de Hájek peut être réécrit comme :

$$\hat{T}_{HA,y} = \frac{N}{\hat{N}_{HT}} \hat{T}_{HT,y}. \tag{1.4}$$

L'estimateur de Hájek du total n'est pas sans biais à cause de l'estimateur $\hat{N}_{HT}$ au dénominateur, et sa variance est plus complexe à calculer pour les mêmes raisons. On peut cependant calculer une variance approchée en utilisant des techniques de linéarisation valables pour n et N grands.

L'estimateur de Hájek est un cas particulier d'estimateur par le ratio (voir Särndal et al. (1989), Tillé (2019)).

Propriétés asymptotiques

Les propriétés asymptotiques des estimateurs utilisés en sondage ne peuvent pas s'étudier comme en statistique classique car la population U est finie. Il faut donc définir un contexte spécifique. Considérons le contexte asymptotique décrit par Isaki and Fuller (1982).

On considère une suite infinie de populations finies U_t de taille N_t, $t > 1$ telles que pour tout t, U_t est inclue dans U_{t+1}, $U_t \subset U_{t+1}$, ce qui implique que pour tout t, $N_t \leq N_{t+1}$. Dans chaque population U_t, on observe une variable Y_t ayant pour valeurs y_{kt}, $k \in U_t$. Pour tout

t, on définit un plan de sondage p_t, avec π_{kt} et $\pi_{kk't}$ les probabilités d'inclusion de premier et second ordre associées à p_t. Un échantillon s_t de taille n_t est tiré dans U_t selon le plan de sondage p_t, tel que pour tout t, $n_t < n_{t+1}$. Soit θ_t la valeur du paramètre d'intérêt pour la population U_t et $\hat{\theta}_t$ un estimateur de θ_t obtenu à partir de s_t dans la population t. Quand $t \to \infty$, alors $N_t \to \infty$ et $n_t \to \infty$. Dans la suite, l'indice t sera omis pour alléger les notations.

Dans cette section, nous allons présenter les propriétés asymptotiques de l'estimateur de HT de T_y puis de l'estimateur de Hájek de T_y.

On considère les hypothèses suivantes (Breidt and Opsomer, 2000) :

H1 On suppose que $\lim_{N\to\infty}\frac{n}{N} = f \in (0,1)$.

H2 On suppose que les probabilités d'inclusion vérifient les conditions suivantes :

(a) il existe une constante λ telle que $\min_{\mathrm{k}\in\mathrm{U}}\pi_{\mathrm{k}} \geq \lambda > 0$

(b) il existe une constante c_1 telle que $\limsup_{N\to\infty} n \max_{k\neq k'\in U}|\pi_{kk'} - \pi_k\pi_{k'}| \leq C_1 < \infty$ avec $C_1 > 0$;

H3 On suppose que $\lim_{N\to\infty} T_y/N < \infty$ et qu'il existe une constante positive C_2 telle que, pour tout N, $\sum_{k\in U} y_k^2/N \leq C_2 < \infty$.

Proposition 1.1.1. *Si les hypothèses précédentes sont vérifiées, alors il existe une constante positive C telle que*

$$nE\left(\frac{1}{N}(\hat{T}_{HT,y} - T_y)\right)^2 \leq C,$$

pour tout n.

Preuve de la Proposition 1.1.1. On a

$$\frac{n}{N^2}\mathrm{E}(|\hat{T}_{HT,y} - T_y|^2) = \frac{n}{N^2}\mathrm{Var}(\hat{T}_{HT,y}) = \frac{n}{N^2}\sum_{k\in U} y_k^2\left(\frac{1}{\pi_k} - 1\right) + \frac{n}{N^2}\sum_{k\in U}\sum_{k'\neq k\in U} y_k y_{k'}\frac{\pi_{kk'} - \pi_k\pi_{k'}}{\pi_k\pi_{k'}}.$$

Selon H2, on peut majorer $(1/\pi_k) - 1$ par λ^{-1} pour tout $k \in U$ et selon H3, on peut majorer $(1/N)\sum_{k\in U} y_k^2$ par C_2 :

$$
\begin{aligned}
&\frac{n}{N^2}\sum_{k\in U} y_k^2\left(\frac{1}{\pi_k}-1\right)+\frac{n}{N^2}\sum_{k\in U}\sum_{k'\neq k\in U} y_k y_{k'}\frac{\pi_{kk'}-\pi_k\pi_{k'}}{\pi_k\pi_{k'}}\\
&\leq \frac{nC_2}{N\lambda}+n\max_{k\neq k'\in U}|\pi_{kk'}-\pi_k\pi_{k'}|\frac{1}{\lambda^2}\left(\frac{1}{N}\sum_{k\in U}|y_k|\right)^2\\
&\leq \frac{nC_2}{N\lambda}+n\max_{k\neq k'\in U}|\pi_{kk'}-\pi_k\pi_{k'}|\frac{1}{\lambda^2}\left(\frac{1}{N}\sum_{k\in U}y_k^2\right) \text{ selon l'inégalité de Cauchy-Schwarz}\\
&\leq \frac{nC_2}{N\lambda}+n\max_{k\neq k'\in U}|\pi_{kk'}-\pi_k\pi_{k'}|\frac{1}{\lambda^2}C_2\\
&\leq C \text{ par les hypothèses H1 et H2 (b)}
\end{aligned}
$$

avec $C>0$. □

Corollaire 1.1.1. *Si les hypothèses précédentes sont vérifiées, alors $\hat{T}_{HT,y}$ est asymptotiquement sans biais pour T_y :*

$$\lim_{N\to\infty}\frac{1}{N}(E(\hat{T}_{HT,y})-T_y)=0,$$

et consistant pour T_y :

$$\lim_{N\to\infty}\mathbb{P}(|\hat{T}_{HT,y}-T_y|>N\varepsilon)=0 \text{ pour tout } \varepsilon>0.$$

En particulier,

$$\frac{1}{N}(\hat{T}_{HT,y}-T_y)=O_p\left(1/\sqrt{n}\right).$$

Preuve du corollaire 1.1.1. Soit $\varepsilon>0$. En utilisant l'inégalité de Markov, on obtient :

$$\mathbb{P}\left(\frac{1}{N}|\hat{T}_{HT,y}-T_y|\geq\varepsilon\right)\leq\frac{\frac{1}{N}\mathrm{E}(|\hat{T}_{HT,y}-T_y|)}{\varepsilon},$$

et l'inégalité de Cauchy-Schwarz

$$\mathrm{E}\left(\frac{1}{N}|\hat{T}_{HT,y}-T_y|\right)\leq\sqrt{\frac{1}{N^2}\mathrm{E}(|\hat{T}_{HT,y}-T_y|^2)}\leq\frac{C}{\sqrt{n}},$$

par la Proposition 1.1.1. On obtient donc :

$$\lim_{N\to\infty}\mathbb{P}(|\hat{T}_{HT,y}-T_y|>N\varepsilon)=0 \text{ pour tout } \varepsilon>0$$

et

$$\lim_{N\to\infty}\frac{1}{N}(\mathrm{E}(\hat{T}_{HT,y})-T_y)=0.$$

car $\mathrm{E}(\hat{T}_{HT,y} - T_y)/N \leq \mathrm{E}|\hat{T}_{HT,y} - T_y|/N$. □

La proposition suivante permet de linéariser l'estimateur de Hájek $\hat{T}_{HA,y} = \frac{N}{\sum_{k\in s} d_k} \sum_{k\in s} d_k y_k$ du total T_y pour obtenir une variance approximée et un estimateur de variance.

Proposition 1.1.2. *Si les hypothèses H1-H4 sont vérifiées, alors l'estimateur de Hájek de T_y peut être linéarisé comme suit :*

$$\frac{1}{N}\left(\hat{T}_{HA,y} - T_y\right) - \frac{1}{N}\left(\sum_{k\in s} d_k e_k - \sum_{k\in U} e_k\right) = O_p\left(1/n\right),$$

où $e_k = y_k - \bar{Y}$, $k \in U$ avec $\sum_{k\in U} e_k = 0$. On obtient en particulier grâce au corollaire 1.1.1 que

$$\frac{1}{N}\left(\hat{T}_{HA,y} - T_y\right) = O_p\left(1/\sqrt{n}\right).$$

Preuve de la Proposition 1.1.2. Soit :

$$D = \frac{1}{N}\left(\hat{T}_{HA,y} - T_y\right) - \frac{1}{N}\left(\sum_{k\in s} d_k e_k - \sum_{k\in U} e_k\right) = \frac{1}{N}\hat{T}_{HA,y} - \frac{1}{N}T_y - \frac{\hat{T}_{HT,y}}{N} + \bar{Y}\frac{\hat{N}_{HT}}{N}.$$

On a :

$$\begin{aligned} D &= \frac{\hat{T}_{HT,y}}{\hat{N}_{HT}} - \frac{\hat{T}_{HT,y}}{N} - \bar{Y} + \bar{Y}\frac{\hat{N}_{HT}}{N} \\ &= \frac{\hat{T}_{HT,y}}{\hat{N}_{HT}} - \bar{Y} + \bar{Y}\frac{\hat{N}_{HT}}{N} - \frac{\hat{T}_{HT,y}}{\hat{N}_{HT}}\frac{\hat{N}_{HT}}{N} \\ &= -\left(\frac{\hat{N}_{HT}}{N} - 1\right)\left(\frac{\hat{T}_{HT,y}}{\hat{N}_{HT}} - \bar{Y}\right) \end{aligned}$$

Selon le corollaire 1.1.1,

$$\frac{\hat{N}_{HT}}{N} - 1 = \frac{\hat{N}_{HT} - N}{N} = O_p\left(1/\sqrt{n}\right).$$

On peut réécrire

$$\begin{aligned} \frac{\hat{T}_{HT,y}}{\hat{N}_{HT}} - \bar{Y} &= \left(\frac{1}{\hat{N}_{HT}} - \frac{1}{N}\right)\hat{T}_{HT,y} + \frac{1}{N}\left(\hat{T}_{HT,y} - T_y\right) \\ &= \frac{N - \hat{N}_{HT}}{N}\frac{N}{\hat{N}_{HT}}\frac{\hat{T}_{HT,y}}{N} + \frac{1}{N}\left(\hat{T}_{HT,y} - T_y\right) \end{aligned}$$

avec $\frac{N - \hat{N}_{HT}}{N} = O_p\left(1/\sqrt{n}\right)$ et $\frac{1}{N}\left(\hat{T}_{HT,y} - T_y\right) = O_p\left(1/\sqrt{n}\right)$ selon le corollaire 1.1.1,

$\frac{N}{\hat{N}_{HT}} = O_p(1)$ selon les hypothèses H1 et H2a et $\frac{\hat{T}_{HT,y}}{N} = O_p(1)$ selon le corollaire 1.1.1 et l'hypothèse H3. On en déduit que

$$\frac{\hat{T}_{HT,y}}{\hat{N}_{HT}} - \bar{Y} = O_p\left(1/\sqrt{n}\right)$$

et donc que $D = O_p\left(1/n\right)$.

Comme $\frac{1}{N}\left(\sum_{k\in s} d_k e_k - \sum_{k\in U} e_k\right) = O_p\left(1/\sqrt{n}\right)$ d'après le corollaire 1.1.1

$$\frac{1}{N}\left(\hat{T}_{HA,y} - T_y\right) = D + \frac{1}{N}\left(\sum_{k\in s} d_k e_k - \sum_{k\in U} e_k\right) = O_p\left(1/\sqrt{n}\right).$$

□

La Proposition 1.1.2 permet d'approximer la variance de l'estimateur de Hájek $\hat{T}_{HA,y}$ de T_y par la variance de $\hat{T}_{HT,E} = \sum_{k\in s} d_k e_k$. On en déduit le corollaire suivant.

Corollaire 1.1.2. *Si les hypothèses H1, H2 ,H3 et H4 sont vérifiées, alors la variance de l'estimateur de Hájek $\hat{T}_{HA,y}$ du total T_y peut être approximée par*

$$AVar(\hat{T}_{HA,y}) = Var(\hat{T}_{HT,E}) = \sum_{k\in U}\sum_{k'\in U} \frac{\pi_{kk'} - \pi_k\pi_{k'}}{\pi_k\pi_{k'}} e_k e_{k'} \tag{1.5}$$

où la variable E prend les valeurs $e_k = y_k - \bar{Y}$ sur U.

Cette variance approximée peut être estimée si $\pi_{kk'} \neq 0$ pour tout $k, k' \in U$ par

$$\widehat{\mathrm{Var}}(\hat{T}_{HT,E}) = \sum_{k\in s}\sum_{k'\in s} \frac{\pi_{kk'} - \pi_k\pi_{k'}}{\pi_{kk'}\pi_k\pi_{k'}} \hat{e}_k \hat{e}_{k'}, \tag{1.6}$$

avec $\hat{e}_k = y_k - \hat{T}_{HT,y}/N$.

1.1.4 Sondage indirect

La base de sondage de la population d'intérêt n'est pas toujours disponible. Dans ce cas, un tirage direct dans la population d'intérêt est impossible et une solution est d'utiliser un sondage indirect.

Un sondage indirect consiste à utiliser une population U_F, dite population "Frame" de taille N_F pour laquelle on dispose d'une base de sondage et qui est reliée à la population d'intérêt U, de telle façon que tous les individus de U soient reliés au minimum à un individu de U_F. Cette condition permet de s'assurer que tous les individus de U peuvent être

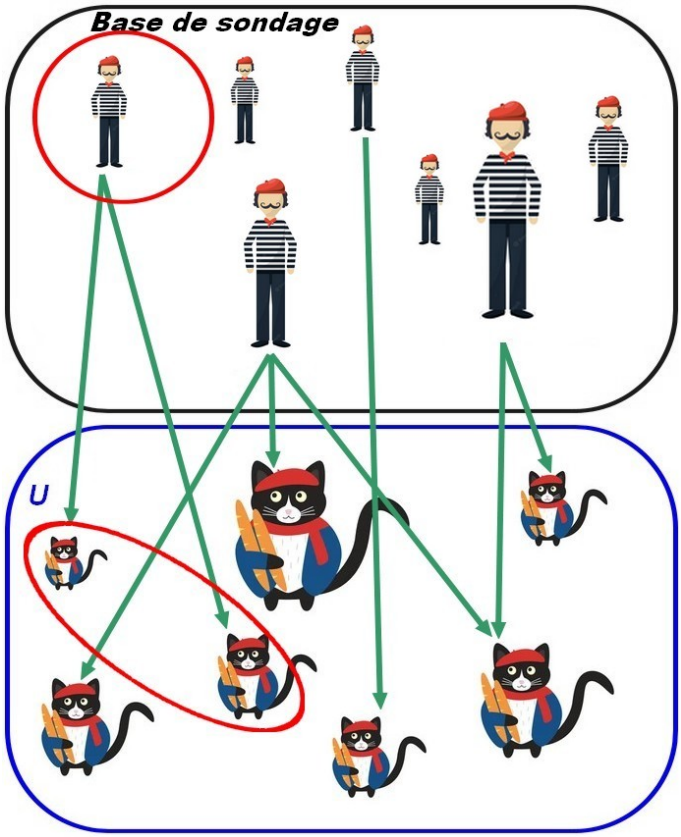

FIGURE 1.10 – Echantillons tirés via un plan indirect.

échantillonnés avec une probabilité non nulle. Un échantillon s_F est tiré dans U_F et l'échantillon s dans la population d'intérêt est composé de tous les individus reliés à s_F. Il n'est pas nécessaire que les individus de U_F soient tous reliés à au moins un individu de U, mais cela permet d'obtenir un échantillonnage plus efficace, car un individu de U_F non relié à U n'augmente pas la taille de l'échantillon s s'il est sélectionné.

Par exemple, dans la figure 1.10, il n'existe pas de base de sondage pour la population U des animaux de compagnie en France. On peut tirer indirectement un échantillon d'animaux de compagnie en utilisant la population des français : un échantillon de français est tiré et leurs animaux de compagnie constituent l'échantillon final. Les liens entre les populations sont indiqués en vert. On vérifie bien que tous les chats français sont reliés à au moins un français, mais un français n'est pas nécessairement relié à un chat. Le français tiré dans U_F est lié à deux chats, qui feront partie de l'échantillon de U.

Il est aussi possible d'utiliser un plan à deux degrés (voir section 1.1.2) quand la base de sondage de la population d'intérêt n'est pas disponible. Les individus échantillonnés lors d'un sondage à deux degrés sont des Unités Primaires (UP) de la population d'intérêt et ceux échantillonnés lors d'un sondage indirect sont des individus reliés à la population d'intérêt.

Le choix de la méthode utilisée dépend donc de l'information disponible.

Avec un plan de sondage indirect, la probabilité de tirer un individu $k \in U$ est complexe à calculer, car elle dépend des liens entre les deux populations. En effet, la probabilité de tirer $k \in U$ est égale à la probabilité de tirer au moins un individu i de U_F lié à k. Les poids de sondage calculés par rapport au plan de sondage dans U sont donc inconnus ou très difficilement calculables, contrairement aux poids de sondage calculés par rapport au plan de sondage dans U_F, et les estimateurs présentés en section 1.1.3 ne peuvent pas être utilisés tels quels.

Cependant, Deville and Lavallée (2006) ont proposé une méthode, la Méthode Généralisée de Partage des Poids (MGPP), qui contourne cette difficulté. Soit θ_{ik} un poids de lien entre $i \in U_F$ et $k \in U$, avec θ_{ik} prenant une valeur positive si i et k sont liés, 0 sinon. La valeur positive de θ_{ik} est choisie par le statisticien en fonction des informations disponibles. Soit $\tilde{\theta}_{ik} = \theta_{ik}/(\sum_{i \in U_F} \theta_{ik})$ le poids standardisé associé à θ_{ik}, c'est-à-dire tel que $\sum\limits_{i \in U_F} \tilde{\theta}_{ik} = 1$.

La standardisation permet de réécrire le total d'intérêt $T_y = \sum_{k \in U} y_k$ sur U comme un total $T_{\tilde{y}}$ sur U_F de la variable $\tilde{y}_i = \sum_{k \in U} \tilde{\theta}_{ik} y_k$:

$$T_y = \sum_{k \in U} y_k = \sum_{k \in U} \left(\sum_{i \in U_F} \tilde{\theta}_{ik} \right) y_k = \sum_{i \in U_F} \sum_{k \in U} \tilde{\theta}_{ik} y_k = \sum_{i \in U_F} \tilde{y}_i = T_{\tilde{y}}. \tag{1.7}$$

L'idée est de réécrire le total T_y comme un total d'une nouvelle variable sur la population U_F dans laquelle on effectue le tirage. On peut donc utiliser l'estimateur de HT du total $T_{\tilde{y}}$. L'estimateur de HT du total $T_{\tilde{y}}$ sur U_F est appelé estimateur MGPP du total T_y sur U :

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

Les seuls individus k de U pour lesquels la somme $\sum_{i \in s_F} \tilde{\theta}_{ik}/\pi_i$ n'est pas nulle sont ceux appartenant à s. L'estimateur MGPP du total T_y sur U peut donc s'écrire :

$$\hat{t}_{y1} = \sum_{k \in s} \left(\sum_{i \in s_F} \frac{\tilde{\theta}_{ik}}{\pi_i} \right) y_k. \tag{1.8}$$

Supposons que $i \in s_F$ et $k \in U$ soient liés. La Méthode Généralisée de Partage des Poids consiste à utiliser les poids standardisés pour partager le poids de sondage de i entre tous les individus, échantillonnés ou pas, reliés à k.

La figure 1.11 permet d'imager le principe du partage des poids et présente un cas très simplifié. Dans cet exemple, on s'intéresse aux chats français sur une zone géographique

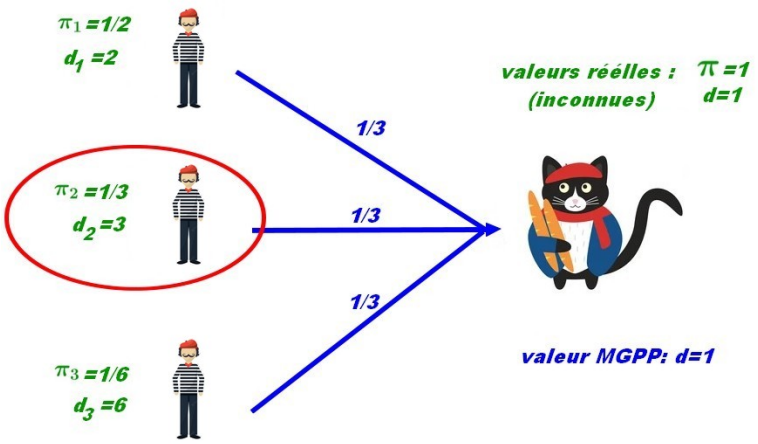

FIGURE 1.11 – Exemple de poids partagé.

réduite. La population d'intérêt est composée d'un chat et la population Frame est composée de trois personnes, qui sont toutes liées au chat. Il y a donc trois liens, indiqués en bleu. Le statisticien a décidé d'attribuer un poids de liens standardisé de 1/3 à chaque lien.

On tire un individu dans la population Frame avec les probabilités indiquées. Dans ce cas très simple, le chat est indirectement tiré quel que soit la personne échantillonnée et la probabilité d'observer le chat est de 1. Supposons qu'on tire la personne dans l'ellipse. L'individu tiré a un poids de 3. Comme il y a trois personnes permettant d'échantillonner le chat, ce poids doit être partagé entre les trois liens, c'est-à-dire divisé par trois. On obtient donc un poids de 1 pour le chat français. Si la première (resp. troisième) personne en partant du haut était échantillonnée, on obtiendrait un poids de 2/3 (resp. 2) pour le chat. Pour calculer les poids de sondage, il faut donc connaître le nombre de personnes reliées au chat échantillonné, c'est-à-dire 3. Dans le cas où il y a beaucoup de liens entre la population Frame et les individus de U, ce nombre peut être difficile à obtenir, comme dans le cas de La Poste.

L'estimateur MGPP a les propriétés d'un estimateur de HT, soit une absence de biais pour estimer T_y et une expression connue de la variance :

$$\mathrm{Var}(\hat{t}_{y1}) = \sum_{i\in U_F}\sum_{i'\in U_F}\Delta_{ii'}\sum_{k\in U}\tilde{\theta}_{ik}y_k\sum_{k'\in U}\tilde{\theta}_{i'k'}y_{k'}, \tag{1.9}$$

avec $\Delta_{ii'} = (\pi_{ii'} - \pi_i\pi_{i'})/\pi_i\pi_{i'}$ et $\Delta_i = \Delta_{ii}$. Cette variance dépend du choix de la valeur positive des liens pondérés, mais l'estimateur MGPP reste sans biais tant que les poids sont standardisés.

La précision d'un estimateur MGPP dépend non seulement des poids choisis, mais aussi et surtout des liens entre populations, qui sont fixés. Deville et Lavallée (2006) posent la

question du choix des poids de liens et cherchent des poids "optimaux" qui minimisent la variance de l'estimateur MGPP. Cependant, cette variance dépend aussi des valeurs prises par la variable d'intérêt et on a donc potentiellement autant de poids optimaux que de variables d'intérêt. Par la suite, on appelera poids optimaux, notés $\tilde{\theta}_{ik}^{opt}$, $i \in U_F$, $k \in U$, les poids standardisés qui minimisent la variance de l'estimateur MGPP pour toute variable d'intérêt, c'est-à-dire que pour toute variable d'intérêt y_k, $k \in U$, et tout poids standardisé $\tilde{\theta}_{ik}$, $i \in U_F$, $k \in U$:

$$\sum_{i\in U_F}\sum_{i'\in U_F}\Delta_{ii'}\sum_{k\in U}\tilde{\theta}_{ik}^{opt}y_k\sum_{k'\in U}\tilde{\theta}_{i'k'}^{opt}y_{k'} \leq \sum_{i\in U_F}\sum_{i'\in U_F}\Delta_{ii'}\sum_{k\in U}\tilde{\theta}_{ik}y_k\sum_{k'\in U}\tilde{\theta}_{i'k'}y_{k'}. \tag{1.10}$$

Liens "Tous-pour-Un"

On appelle liens de type Tous-pour-Un (TpU) les liens où un individu de U_F n'est lié qu'à un individu de U, mais un individu de U peut être lié à plusieurs individu de U_F. Par exemple, un français n'est relié qu'à un chat, mais un chat peut être lié à plusieurs français, comme dans la figure 1.11. Si le plan de sondage utilisé pour tirer dans U_F est un SASSR ou un plan de Poisson, Deville et Lavallée (2006) montrent qu'il existe des poids optimaux qui minimisent la variance de l'estimateur MGPP quelle que soit la variable d'intérêt si les liens sont de type "TpU" et qui sont donnés :

$$\text{pour le SASSR, par } \theta_{ik} = 1 \text{ si } i \in U_F \text{ et } k \in U \text{ sont liés, 0 sinon, et}$$

$$\text{pour le plan de Poisson, par } \theta_{ik} = \frac{1}{(1-\pi_i)\pi_i} \text{ si } i \in U_F \text{ et } k \in U \text{ sont liés, 0 sinon.}$$

Ces valeurs correspondent aux poids non standardisés. On appelle MGPP optimale et estimateur MGPP optimal la MGPP obtenue en utilisant les poids optimaux et l'estimateur résultant.

1.2 La Poste

La Poste est une société anonyme française assurant le transport d'objets postaux (lettres, colis) en France et à l'étranger. Les ancêtres de La Poste sont La Poste à chevaux, créée par Louis XI en 1477, qui effectuait le transport des messages royaux, et l'office des messagers royaux, qui transportait du courrier de particuliers.

Les premières boites aux lettres apparaissent avec la création de la "petite poste" à Paris en 1758-1760. Elle permet la distribution des lettres intra-muros. Avant, les lettres devaient être récupérées au bureau de poste. Le système est par la suite étendu aux grandes villes du

royaume (Bordeaux, Lyon, Nantes, Rouen, Nancy, Strasbourg, Marseille et Lille).

La révolution française transforme le système postal, et entre autres la protection du secret postal, avec le décret sur le secret et l'inviolabilité des lettres en 1791, puis l'ouverture aux services de transport de personnes ou biens aux particuliers en 1794. Le monopole postal est déclaré en 1801 et la direction générale des Postes est créée en 1804.

Au 18e siècle, La Poste se modernise et s'étend aux campagnes. Le retour à l'envoyeur, l'utilisation de bateaux à hélice ou de bureaux de poste ambulants, permettant le tri à bord de wagons des chemins de fer, les timbres-poste, l'uniforme et la carte postale sont créés sur cette période. Les femmes sont engagées à La Poste depuis 1877.

Au 20e siècle, l'essor continue avec la création et l'âge d'or de l'Aéropostale, la création de La Poste automobile rurale (qui remplace les bicyclettes), les distributeurs automatiques de timbres, les premières machines à affranchir... et les premières grèves.

Le 21e siècle voit l'ouverture progressive du marché postal à la concurrence. La Poste change de statut pour devenir SA en 2010.

L'histoire de La Poste est présentée en détail au Musée de La Poste (34 Bd de Vaugirard, 75015 Paris) ou sur leur site *https ://www.museedelaposte.fr/fr/histoire-de-la-poste*.

1.2.1 Plans de sondage SYCI 1 et 2

La Poste distribue plusieurs types de produits (lettres recommandées, lettres vertes, colis...), appelés objets postaux, ayant chacun des caractéristiques différentes, comme les dimensions de l'enveloppe, le poids, le type et le prix du timbre ou encore le délai maximum de livraison. Les objets postaux comprennent les communications (lettres...) et les objets de marchandise (colis...) et sont livrés lors des tournées de facteur. Une tournée est un ensemble d'adresses postales pour lesquelles un facteur va livrer du courrier.

La régulation de La Poste dans le cadre du Service Universel (SU) est assurée par l'ARCEP (Autorité de Régulation des Communications Électroniques, des Postes et de la distribution de la presse) qui requiert la mise à disposition de chiffres précis dans le domaine postal afin d'effectuer une régulation efficace. La Poste cherche donc à avoir une vision globale de la mise en distribution des objets. Entre autres, elle souhaite connaître le volume et la composition du trafic postal mensuel, trimestriel et annuel. La Poste s'intéresse aussi au traitement des objets et à l'organisation des tournées, en regardant notamment les délais de livraison, si un objet a été trié par machine de tri et le mode de transport utilisé pour la livraison.

La Poste ne pouvant pas observer toutes les tournées distribuées en France Métropolitaine, elle a mis en place depuis 1994 une étude par sondage, SYCI 1, (SYstème de Collecte

de l'Information), afin d'estimer la composition du trafic et assurer un suivi du traitement des objets postaux. SYCI est un observatoire des flux d'objets distribués et de l'organisation de la distribution.

Dans cette section, nous allons présenter le plan de sondage SYCI 1 utilisé entre 1994 et 2015. Des réorganisations des centres de tri ont poussé La Poste a revoir sa méthodologie en 2015. Nous présenterons le plan initialement prévu en 2015 et enfin le plan actuel SYCI 2.

SYCI 1

La population d'intérêt est composée des objets distribués un trimestre donné. Par la suite, on simplifiera cette population en se restreignant à un mois donné. La Poste cherchant aussi à observer l'organisation de la distribution, elle a décidé d'observer les tournées, aussi appelées SOrties-Jours (SOJ) ou tournées-jours.

Les tournées peuvent être vues comme un groupe d'objets, de taille trop importante pour que tous les objets d'une tournée puissent être observés. L'étude historique SYCI 1 consistait donc à tirer avec un plan à deux degrés (voir section 1.1.2) un échantillon de tournées puis un échantillon d'objets au sein de chaque tournée échantillonnée. Les objets observés sont appelés objets-échantillon.

Les plans de sondage utilisés pour tirer les tournées puis les objets au sein des tournées sont des STASSR (voir section 1.1.2). Le poids de sondage final d'un objet est donc le poids de sondage de sa tournée d'appartenance multiplié par son poids de sondage au sein de la tournée.

Ce plan de sondage repose sur l'hypothèse que la population des tournées est suffisamment stable pour que la base de sondage corresponde à la réalité du terrain. Cependant, les centres de distribution se réorganisent régulièrement pour permettre d'adapter la distribution selon le volume de courrier. Ces réorganisations ont rendu inadéquate la base de sondage des tournées. La Poste a donc mis en place en 2015 un nouveau plan de sondage, SYCI 2, qui modifie le tirage des tournées, mais garde le plan à deux degrés.

Plan SYCI 2 initialement prévu

Pour tirer les tournées, plusieurs solutions existent. On peut par exemple effectuer un tirage à deux degrés en tirant des bureaux puis des tournées au sein des bureaux. Cependant, cette méthode nécessite un nombre important d'enquêteurs et de moyens.

La Poste a décidé d'échantillonner les tournées via un sondage indirect (voir section 1.1.4), en utilisant la population des adresses-jours, qui est accessible, comme population Frame et la population des tournées comme population cible. Ce choix est motivé par la stabilité de

la population des adresses, qui n'est pas sensible aux réorganisations des bureaux. On tire un échantillon d'adresses et un échantillon de jours, puis les adresses-jours sont formées en associant aléatoirement un jour à une adresse. Le jour donné, l'enquêteur va voir le bureau qui s'occupe de l'adresse et repère la tournée qui dessert l'adresse tirée. Par simplicité, la population des adresses-jours sera désignée comme population des adresses dans ce qui suit.

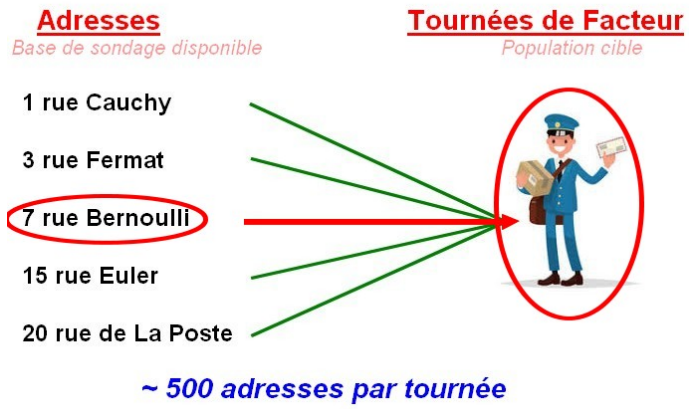

FIGURE 1.12 – MGPP initialement prévue par La Poste.

L'utilisation d'un sondage indirect va de pair avec l'utilisation d'une MGPP. Or, pour pouvoir utiliser la MGPP, il faut pouvoir calculer les poids standardisés (voir section 1.1.4). Pour cela il est nécessaire d'observer tous les liens entre la tournée échantillonnée et la population Frame, ce qui, pour La Poste, revient à pouvoir lister toutes les adresses de la tournée. Une tournée étant composée de 500 adresses en moyenne, comme indiqué dans la figure 1.12, récupérer toutes les adresses prend du temps. De plus, les adresses ne sont pas toujours correctement indiquées, ce qui rend leur identification complexe et nécessite l'aide du facteur. Le facteur ne pouvant pas prendre de retard pour répondre aux questions de l'enquêteur, il n'est généralement pas possible de récupérer les adresses d'une tournée. L'utilisation du sondage indirect, et de la MGPP associée, pose donc problème.

Plan SYCI 2

La solution adoptée par La Poste consiste à utiliser une deuxième population intermédiaire, liée à la fois à la population des adresses et des tournées. La population en question est la population des cases de tri, qui sont les cases permettant de trier les objets d'une tournée. Cette population est naturellement entre la population des adresses et la population des tournées.

Pour comprendre les choix de La Poste, regardons le fonctionnement d'un bureau de Poste avant le départ des facteurs en tournée. Tous les matins (en jour ouvré), les objets

FIGURE 1.13 – Un casier (à gauche) et une case de tri (à droite).

postaux à distribuer sont livrés dans les centres de tri. Ces objets sont ensuite triés dans des cases, appelées cases de tri, qui sont attribuées à une ou plusieurs adresses. Les objets à destination de ces adresses sont classés dans la case correspondante. Ces cases peuvent être de différents types, comme indiqué dans la figure 1.14, mais le principe reste le même quel que soit le type. Les cases sont regroupées en casiers similaires à celui de la figure 1.13. Les

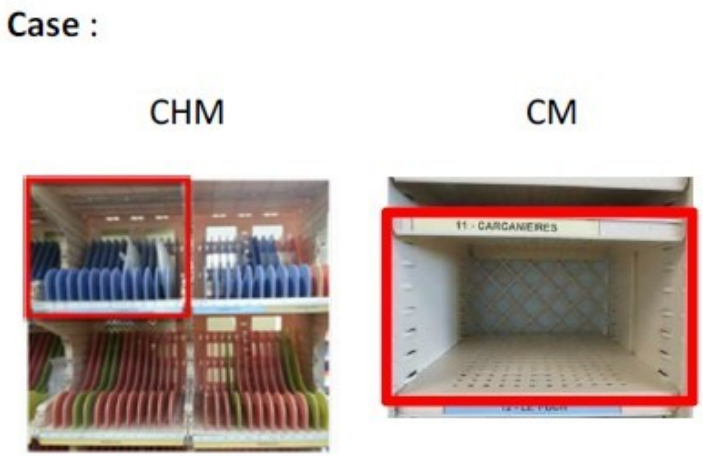

FIGURE 1.14 – Différents types de cases.

tournées sont formées à partir des casiers. Pour simplifier, on supposera par la suite qu'une tournée est composée de tous les objets contenus dans un casier et qu'un casier contient seulement les objets de la tournée correspondante.

Pour tirer un échantillon de tournées via la population des cases, deux tirages indirects (voir section 1.1.4) sont réalisés. Le premier permet d'obtenir un échantillon de cases à partir de la population des adresses et le second permet d'obtenir un échantillon de tournées à partir des cases (indirectement) tirées. On parle de sondage indirect double. L'échantillon final est composé des tournées liées aux cases indirectement tirées, qui sont elles-mêmes liées aux

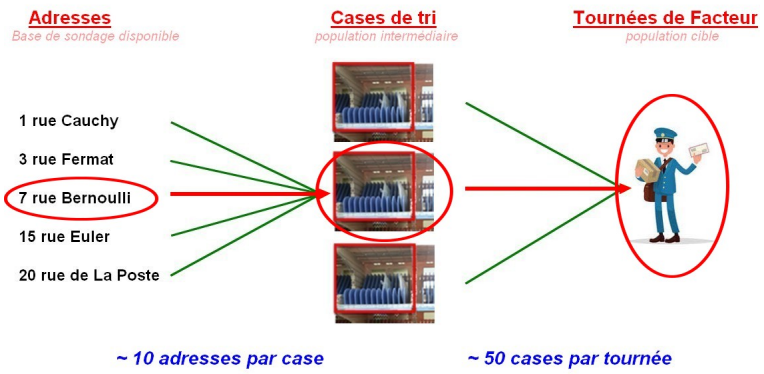

FIGURE 1.15 – MGPP double utilisée à La Poste.

adresses sélectionnées, comme indiqué dans la figure 1.15.

La population intermédiaire n'influe pas sur les liens entre la population Frame et la population cible. Une adresse fait toujours partie de la même tournée, que sa case de tri soit observée ou non. Tirer un échantillon par sondage indirect double revient donc à tirer un échantillon par sondage indirect, où les liens sont observés en deux temps plutôt que directement. Par la suite, le sondage indirect avec des liens observés en un seul temps sera appelé sondage indirect simple. Pour déterminer la tournée d'appartenance d'une adresse, il faut identifier sa case de tri puis la tournée, plutôt qu'identifier directement la tournée. Cette complexité d'observation des liens permet d'obtenir des estimateurs dont l'expression est plus complexe que ceux obtenus avec un sondage indirect simple.

L'estimateur est obtenu en appliquant une MGPP pour chaque ensemble de liens (voir section 1.1.4). On appelle donc cette méthode une MGPP double. Un premier ensemble de poids de liens est choisi pour les liens entre les adresses et les cases, et un second entre les cases et les tournées. Pour standardiser les poids de liens, il faut donc identifier toutes les adresses des cases échantillonnées et toutes les cases des tournées échantillonnées, comme indiqué dans la figure 1.15. Les liens rouges sont échantillonnés et pour pouvoir standardiser leur poids, il faut observer les liens verts, ce qui revient à observer 10 adresses pour une case et 50 cases pour une tournée en moyenne. Une tournée livrant 500 adresses en moyenne, cette méthode permet de réduire à 60 le nombre de liens observés par rapport à la MGPP initialement prévue, et est donc plus facile à mettre en place.

La MGPP double est un cas particulier de MGPP, comme on le verra dans la section 1.3. Cependant, la MGPP double utilisée à La Poste cause une perte de précision importante comparée aux estimateurs directs obtenus avec SYCI 1, comme le montre la figure 1.16. Sur cette figure, les estimations (normalisées pour des raisons de confidentialité) du trafic total sont indiquées en noir et les intervalles de confiance des estimateurs sont en rouge. On voit

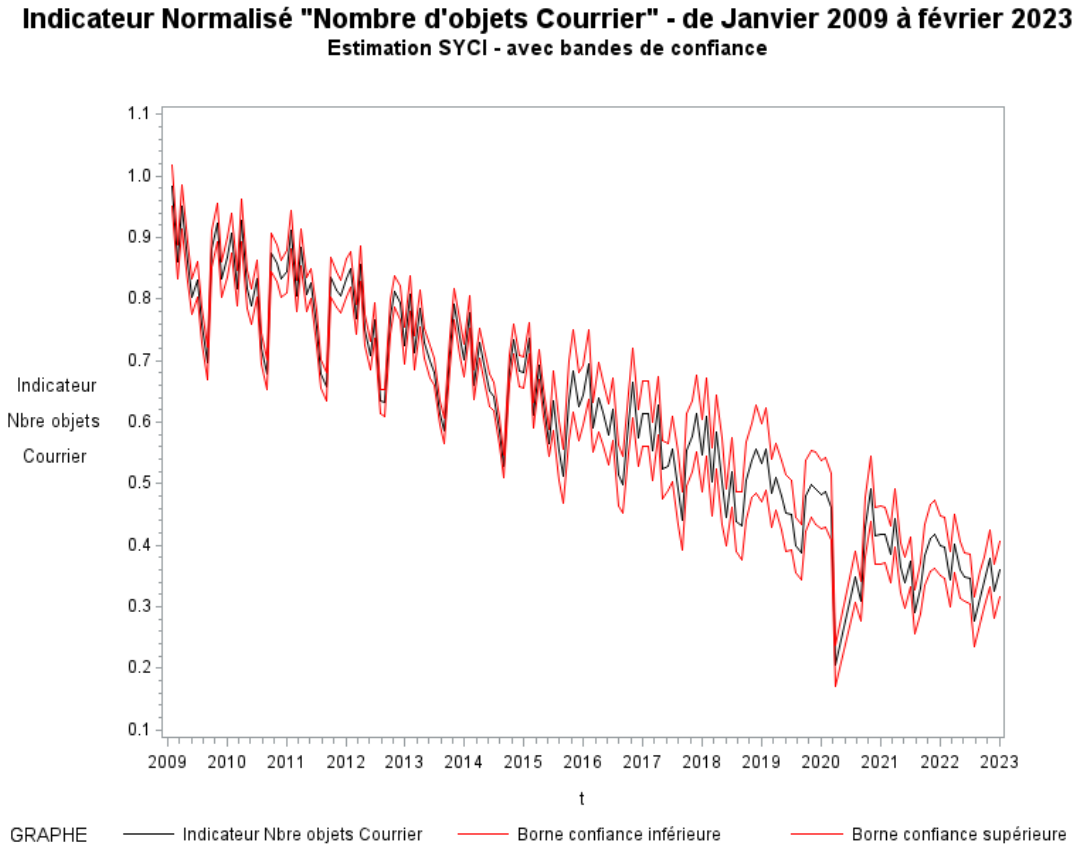

FIGURE 1.16 – Perte de précision des estimateurs de La Poste.

qu'en 2015, année de transition de SYCI 1 à SYCI 2, les intervalles de confiance s'élargissent. La MGPP n'ayant pas pu être mise en place, la comparaison entre les estimateurs MGPP et MGPP double ne peut pas être réalisée dans le cas de La Poste.

Les travaux menés lors de ma thèse s'inscrivent dans ce contexte de perte de précision due à la transition de SYCI 1 à SYCI 2. De plus, La Poste prévoit de réduire les tailles d'échantillons dans un futur proche, ce qui va dégrader la précision des estimateurs. Une partie de mon travail porte sur l'amélioration de la MGPP double pour améliorer la précision des estimateurs et limiter l'impact de la réduction de la taille des échantillons.

Dans mes travaux sur la MGPP (chapitre 2 et chapitre 3), le tirage à deux degrés des objets est omis. On considère par simplicité que la population d'intérêt est la population des tournées et que la variable d'intérêt est le nombre d'objets postaux distribués par une tournée. Le tirage s'effectue par sondage indirect double en utilisant la population des adresses comme population Frame.

La Poste effectue un travail important de traitement des données récoltées pour s'assurer de leur fiabilité. De plus, La Poste utilise des méthodes de repondération des estimateurs, comme le calage (Deville and Särndal, 1992) ou la winsorisation (Kokic and Bell, 1994).

Certains de ces traitements, notamment la winsorisation, biaisent les estimateurs. Bien que le biais causé par ces traitements soit une source de préoccupation pour La Poste, ce problème n'entre pas dans le cadre de la thèse et on étudiera dans la suite les estimateurs avant leur traitement par La Poste.

1.2.2 Données auxiliaires

La Poste possède des bases de données de diverses provenances, contenant des informations sur les populations utilisées pour le tirage : les adresses, les cases, les tournées et les objets. Dans cette section, on présente trois sources de données différentes : les données récoltées via SYCI 2, qui sont actuellement utilisées pour les estimations, les données permettant d'obtenir la base de sondage, dont une grande partie est inutilisée à l'heure actuelle, et les données récoltées par les machines de tri, dont l'utilisation fait partie de mes sujets de recherche et est actuellement testée à La Poste.

Données récupérées dans SYCI 2

Les données présentées ici sont les données récoltées par les enquêteurs dans le cadre de l'étude SYCI 2. Les données récoltées forment trois tables séparées : une table d'information sur les objets observés, une table sur les tournées observées et une table sur les cases observées.

La figure 1.17 montre une fiche de saisie objet. Les informations récoltées par les enquêteurs comprennent, entre autres, le type de produit (chronopost, lettre verte...), la date de dépôt de l'objet, le montant d'affranchissement, le poids, les dimensions et l'épaisseur du pli, la nature du contenant (film plastique pour les pubs ou revues, enveloppe à fenêtre...) et le contenu (s'il y a un objet, du papier...). Ces données composent la table des objets observés. En pratique, elles sont utilisées pour calculer les variables d'intérêt, qui sont souvent des regroupements de types de produit par exemple. Pour simplifier, on supposera par la suite que les variables d'intérêt sont directement récoltées via le formulaire 1.17.

La figure 1.18 montre une fiche de saisie SOJ. Il s'agit des informations récoltées sur la tournée, qui permettent de s'assurer du bon déroulement de l'enquête. La strate d'appartenance de la tournée et la date d'observation sont indiquées, ainsi que des indications sur l'état de la tournée qui peuvent affecter la saisie des informations. Par exemple, si la tournée n'a pas pu être observée entièrement.

Des informations permettant d'effectuer un suivi de l'organisation de la distribution des objets postaux sont aussi récoltées. Par exemple, le préparateur de la tournée n'est pas toujours celui qui la distribue et le site de préparation de la tournée n'est pas toujours le site de distribution. Ces informations sont donc récoltées, en plus de la période et du mode de

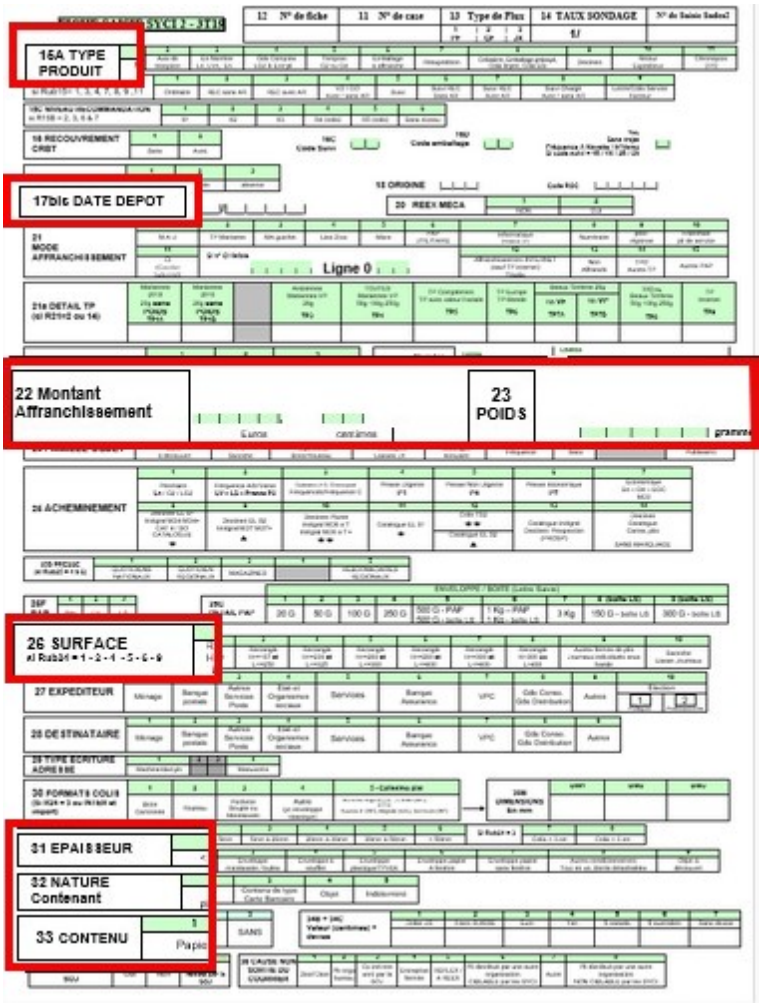

FIGURE 1.17 – Fiche de description des objets.

livraison.

Afin de pouvoir mettre en place une MGPP double, La Poste doit observer les liens entre les populations. Cette information est récupérée via la fiche 1.19. L'enquêteur indique les cases composant le casier de la tournée. Cette fiche permet aussi de récupérer des informations au niveau des cases, puisque, pour chaque case du casier, le type de case et ses dimensions sont récoltées.

La table des données sur les tournées contient les informations récoltées via la fiche 1.18. On y ajoute le nombre de cases dans la tournée, obtenu grâce à la fiche 1.19. Les autres informations collectées via 1.19 composent la table des données sur les cases.

Ces données sont vérifiées et validées après récolte. De plus, les enquêteurs suivent une formation afin de s'assurer de la fiabilité de leur captation. Les données récoltées via SYCI 2 sont donc considérées comme fiables.

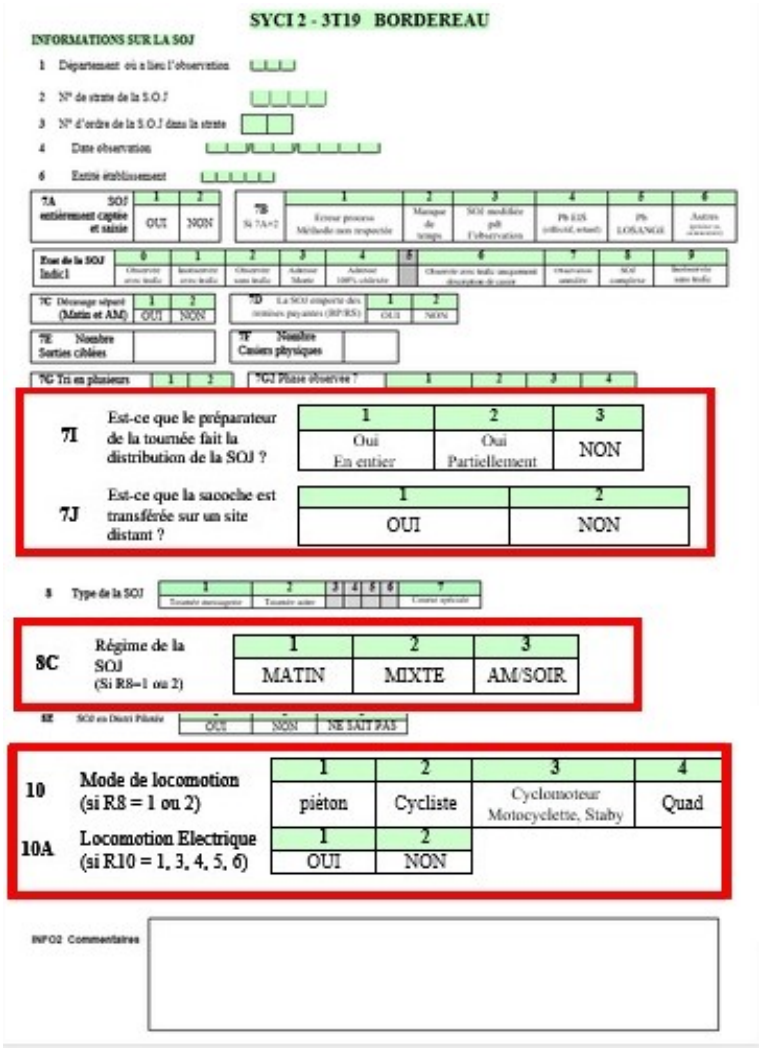

FIGURE 1.18 – Fiche de description des tournées.

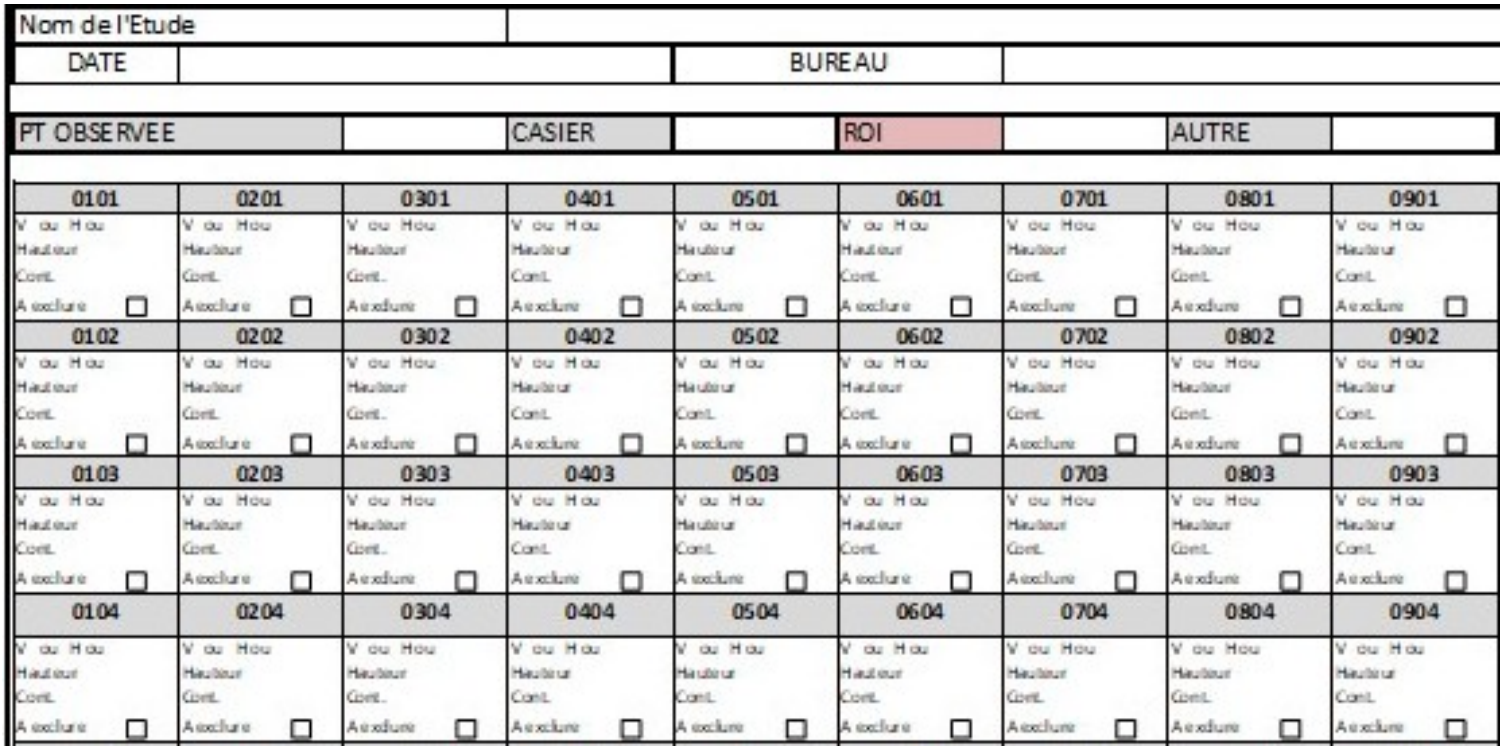

FIGURE 1.19 – Fiche de description des liens entre cases et tournées.

Référentiel des Adresses Organisées (RAO)

La table RAO contient des informations sur les adresses occupées en France Métropolitaine. Cette table, croisée avec les données de l'INSEE, permet d'obtenir une base de sondage pour la population des adresses. Cependant, elle contient des informations sur les tournées et cases d'appartenance des adresses, comme on peut le voir dans la figure 1.20.

Pour une adresse, RAO connaît son type (rue, numéro de rue, commune...) mais donne aussi l'identifiant de la tournée qui dessert cette adresse ainsi que l'identifiant de la case contenant les objets à destination de cette adresse. On remarque aussi que RAO permet de rechercher des informations directement sur un casier ou une tournée.

Pour chaque tournée et case, RAO contient la liste des adresses la composant. Malheureusement, ces informations ne prennent pas en compte les réorganisations journalières et ponctuelles des centres de tri et sont parfois manquantes. Les tournées et cases indiquées par RAO pour une adresse, si elles existent, ne sont généralement pas les tournées et cases observées sur le terrain.

Cependant, RAO reste une source importante de données sur les cases. Dans ma thèse, je propose une façon d'utiliser ces données pour améliorer la MGPP double actuellement utilisée à La Poste.

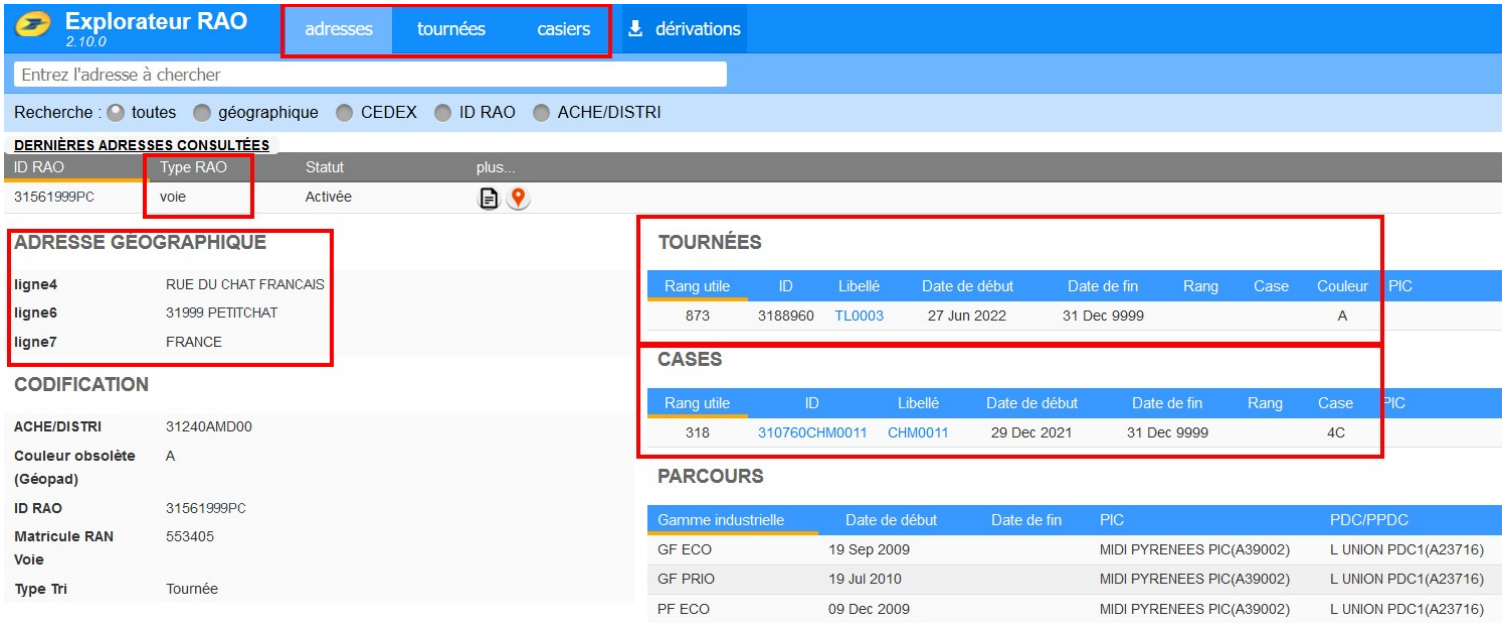

FIGURE 1.20 – Aperçu des données de RAO.

Traitement Automatisé de l'Enveloppe (TAE)

TAE est la table composée des données récoltées par machine de tri. Les machines prennent une photo de l'objet trié, puis utilisent des algorithmes de reconnaissance d'images pour lire les inscriptions de l'enveloppe.

A partir des photos, les machines récoltent plusieurs informations : l'adresse de destination, le type d'objet, la date et le montant d'affranchissement, la machine à affranchir utilisée (si utilisée) et la zone de départ. Ces informations sont récoltées dans le respect du secret des correspondances et du Règlement Général sur la Protection des Données. Les machines enregistrent aussi des informations sur le traitement, comme les dates de passage en machine de tri, la photo ou encore l'identifiant imprimé sur le contenant de l'objet (souvent une enveloppe) par la machine de tri. Cet identifiant est imprimé sur le contenant par les machines sous forme de code barre, similaire à celui montré dans la photo 1.22.

La description des objets réalisée par les machines est basée sur une photo Noir et Blanc de l'objet, comme on peut le voir dans la figure 1.21. De plus, la machine de tri ne peut prendre qu'une photo : recto ou verso. Les informations décrites dans le paragraphe précédent sont inscrites sur le recto et ne peuvent donc pas être obtenues par les machines pour tous les objets. Par exemple, une photo du verso d'un objet ne contient pas d'information sur le type, l'adresse de destination, l'affranchissement... Mais elle peut contenir des informations sur l'expéditeur. Les timbres "marianne", qui se reconnaissent à leur couleur (vert, rouge ou gris) ne sont pas facilement différentiables si la photo est en noir et blanc, comme montré

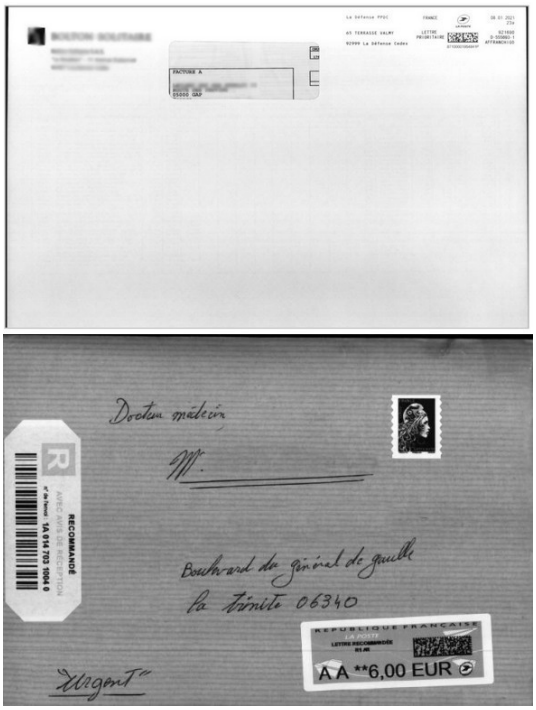

FIGURE 1.21 – Photos prises par les machines de tri.

dans la figure 1.23. Le prix du timbre, voire le type de pli, n'est alors pas reconnu par la machine. Ce dernier problème est voué à disparaître avec la suppression des timbres marianne rouge et noir.

D'autres informations, récupérées dans SYCI 2, ne peuvent pas être lues par les machines. Entre autres, le poids, l'épaisseur, le contenu... Ce qui s'explique par la différence d'objectif de TAE et SYCI. Les informations collectées dans TAE permettent de distribuer l'objet à la bonne adresse et dans les délais impartis, alors que les données collectées dans SYCI permettent d'estimer le trafic postal.

TAE est moins fiable que SYCI 2, de part sa sensibilité à la prise de photo. Une photo floue, trop sombre, trop claire, avec des objets agglomérés rendent la reconnaissance des informations par les machines erronée. A cela s'ajoutent les problèmes liés aux traitements et au transfert des données, qui peuvent être mal transmises d'une machine à une autre, voire perdues lors des coupures réseaux. Dans ce dernier cas, on observe dans SYCI 2 un objet, traité via machine de tri, mais absent de TAE. TAE souffre aussi de problème de doublons dû à la gestion des données, car un identifiant objet peut être réutilisé plusieurs fois.

TAE et SYCI récupèrent tous les deux des variables similaires, mais il est impossible de remplacer SYCI par TAE à l'heure actuelle pour observer le trafic, à cause des problèmes mentionnés dans cette section (données incomplètes et erronées). L'objectif principal de ma thèse est de trouver un moyen d'utiliser TAE, malgré ses inconvénients, pour améliorer la

FIGURE 1.22 – Code barre imprimé par la machine de tri.

FIGURE 1.23 – Timbres Marianne en couleur (haut) et en noir et blanc (bas).

précision des estimations réalisées via SYCI 2.

1.3 Contributions

L'objet de ma thèse est d'améliorer la précision des estimateurs du trafic postal. Par souci de simplicité nous ne prendrons pas en compte les procédures de calage ni de winsorisation utilisées par La Poste (voir section 1.2.1) dans l'étude des estimateurs.

L'étude de la précision des estimateurs considérés s'effectue selon deux axes. D'un côté, nous cherchons à réduire la variance des estimateurs actuels SYCI 2, considérés sans biais, notamment via des travaux sur la MGPP (voir section 1.1.4) et la MGPP double (voir section 1.2.1). D'un autre côté, nous développons des méthodes et des estimateurs qui utilisent

comme information auxiliaire les bases de données massives disponibles à La Poste. Dans cette section, je présente rapidement les travaux effectués pendant ma thèse. Dans l'ensemble de mes travaux de thèse, on ne considère que l'estimateur du trafic postal mensuel, qui est le total des objets postaux distribués en France pendant un mois donné.

1.3.1 Sondage indirect et MGPP

Cette section porte sur la réduction de la variance de l'estimateur MGPP du trafic total mensuel actuellement utilisé à La Poste. Je présente très brièvement les résultats principaux publiés dans l'article "*Many-to-One indirect sampling with application to the French postal traffic estimation*" (Annals of Applied Statistics (AoAS), 2023), dont le contenu figure dans le chapitre 2, ainsi que les résultats présentés dans le chapitre 3, "*Optimal Weights for double Many-To-One Generalized Weight Share Method*".

Chaptitre 2 : *Many-to-One indirect sampling with application to the French postal traffic estimation (AoAs)*

On a vu en section 1.2.1 que l'utilisation d'une MGPP double a diminué la précision des estimateurs SYCI 2, utilisés actuellement par La Poste, comparé à l'estimateur de HT SYCI 1 calculé à partir d'un échantillon direct de tournées et utilisé jusqu'en 2015 par La Poste. Dans cet article, je cherche à comprendre l'origine de cette perte de précision.

Pour faciliter la lecture, la MGPP définie en section 1.1.4 sera appelée MGPP simple et on dira qu'un estimateur du total T_y est direct s'il est obtenu à partir d'un échantillon sélectionné directement dans la population cible U. Dans l'article, la population d'intérêt est notée U_T (target population) mais la notation U sera conservée dans ce chapitre par souci de simplicité.

Dans un premier temps, je me suis penchée sur les conditions d'existence des poids de liens optimaux définis par l'équation (1.10), c'est à dire les poids de liens qui permettent de minimiser la variance, donnée par l'équation (1.9), de l'estimateur MGPP simple du total T_y pour une variable d'intérêt y quelconque. Comme mentionné en section 1.1.4, Deville and Lavallée (2006) cherchent des poids optimaux pour les plans SASSR et de Poisson, pour tout type de liens. Ils ont montré que ceux-ci existent si et seulement si les liens entre les populations sont de type "Tous-pour-Un" (TpU). Dans l'article Medous et al. (2023a), je montre que l'existence de poids optimaux pour la MGPP simple est une propriété des liens de type TpU qui peut être étendue à tous les plans de sondage satisfaisant la Δ-propriété définie ci-dessous.

Soit N_F la taille de la population Frame 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

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

$\boldsymbol{\Delta} = (\Delta_{ii'})_{i,i' \in U_F}$ la matrice de taille $N_F \times N_F$ et $\boldsymbol{\Delta}_{kk'} = (\Delta_{ii'})_{i \in U_{Fk}, i' \in U_{Fk'}}$ la sous-matrice de $\boldsymbol{\Delta}$ 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 $\boldsymbol{\Delta}_k = \boldsymbol{\Delta}_{kk}$. Dans la figure 1.24, l'individu k de la population cible U 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 $\Delta_{ii'}$, $i, i' = 1 \dots 3$, situées en haut à gauche de la matrice (en rouge). On peut similairement obtenir les trois autres matrices de la figure 1.24.

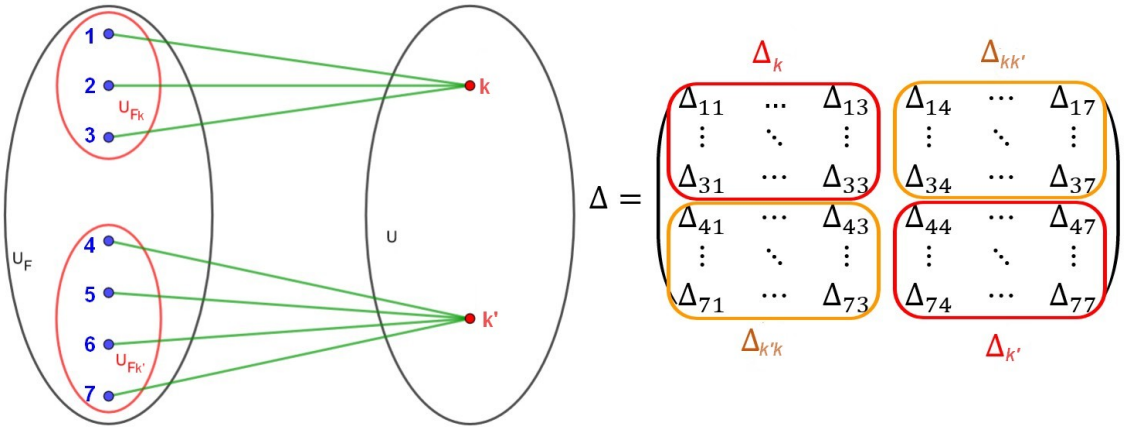

FIGURE 1.24 – Décomposition de la matrice $\boldsymbol{\Delta}$ dans un cas TpU.

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 1.24). On a alors (voir le graphique de droite de la figure 1.24)

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

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 $\boldsymbol{\Delta}$-propriété si

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

— pour $k \neq k' \in U$,

$$\boldsymbol{\Delta}_{kk'} = c_{kk'} \mathbb{1}_k \mathbb{1}_{k'}^t \quad \text{avec} \quad c_{kk'} \quad \text{ne dépendant ni de } i \text{ ni de } i'. \tag{1.11}$$

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., 2023a).

Je montre dans Medous et al. (2023a) 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 :

$$(\tilde{\theta}_{ik}^{opt})_{i \in U_{Fk}} = \mathbf{\Delta}_k^{-1} \mathbb{1}_k \left(\mathbb{1}_k^t \mathbf{\Delta}_k^{-1} \mathbb{1}_k \right)^{-1}, \text{ pour tout } k \in U_T, \tag{1.12}$$

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 :

$$\text{Var}(\hat{t}_{y1}) - \text{Var}(\hat{t}_{y1}^{opt}) = \sum_{k \in U_T} y_k^2 \text{Var}(\hat{t}_{\tilde{\theta}_k} - \hat{t}_{\tilde{\theta}_k^{opt}}) = \sum_{k \in U_T} y_k^2 \text{Var}(\hat{t}_{\tilde{\theta}_k - \tilde{\theta}_k^{opt}}), \tag{1.13}$$

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 (1.13) 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, comme illustré dans les études par simulation (voir section 2.4).

Comme discuté dans la section 1.2.1, 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, je me penche sur un cas particulier où un plan de Poisson est utilisé pour tirer l'échantillon direct s dans U 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, je 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 (voir Proposition 2.2.3).

Pour La Poste, U_F est la population des adresses et U 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 (voir section 1.2.1) 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. Comme mentionné en section 1.2.1, 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 décrite dans la section 1.2.1. Mon second objectif dans l'article Medous et al. (2023a) est d'introduire le principe général de la MGPP double et d'étudier ses propriétés.

Soit une population intermédiaire U_M, reliée à U_F et à U 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 (1.8) :

$$\hat{t}_{y2} = \sum_{k \in s} \left(\sum_{i \in s_F} \frac{1}{\pi_i} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} \right) y_k. \tag{1.14}$$

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. C'est le cas de La Poste, comme illustré par la figure 1.25. 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 $k \in U$.

Dans l'exemple de la figure 1.25, 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 1.25 est donc $1/5 * 1/3 = 1/15$. Dans le cas de La Poste,

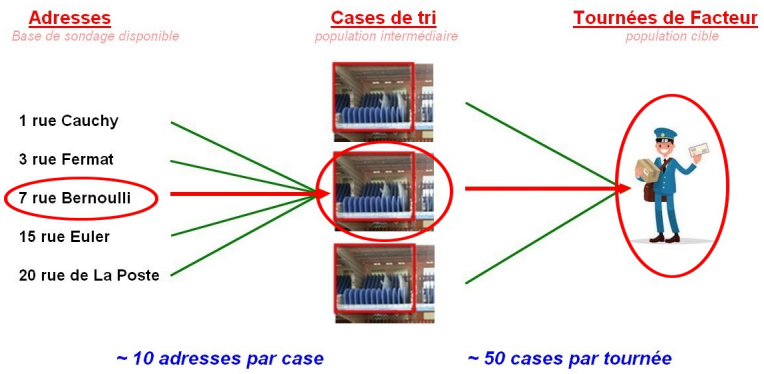

FIGURE 1.25 – TpU-TpU observé par La Poste.

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. (2023a), plusieurs méthodes de standardisation sont comparées : la standardisation de la MGPP simple (voir section 1.1.4), 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 1.26). 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 même nombre d'adresses, ce qui n'est pas

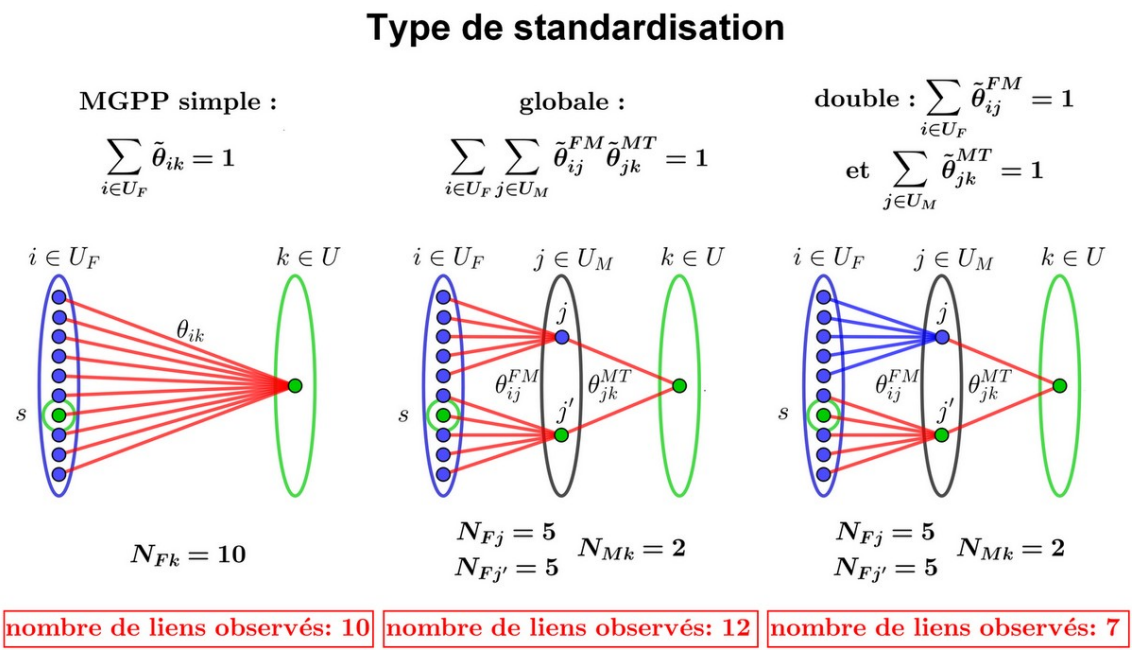

FIGURE 1.26 – Comparaison des méthodes de standardisation.

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 (voir section 1.2.1) par La Poste est donc responsable de la détérioration de la variance des estimateurs du trafic postal observée depuis 2015 (voir section 1.2.1). 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. Ce dernier point fait l'objet de l'article suivant.

Chapitre 3 : *Optimal Weights for double Many-To-One Generalized Weight Share Method*

Dans l'article Medous et al. (2023a), précédemment résumé, j'introduis le concept de MGPP double et je discute des gains en terme de liens observés quand les liens sont de type TpU. Je note aussi que la MGPP double utilisée par La Poste induit une perte de précision par rapport à la MGPP optimale.

Dans le chapitre 3, je me pose la question de l'existence de poids optimaux pour la MGPP double avec double standardisation, c'est à dire des poids standardisés appelés $\tilde{\theta}_{ij}^{FM,opt}$ et $\tilde{\theta}_{jk}^{MT,opt}$, $i \in U_F$, $j \in U_M$, $k \in U$, tels que la somme $\sum_{j\in U_M} \tilde{\theta}_{ij}^{FM,opt}\tilde{\theta}_{jk}^{MT,opt}$ satisfait l'équation (1.10) :

$$\begin{aligned}&\sum_{i\in U_F}\sum_{i'\in U_F}\Delta_{ii'}\sum_{k\in U}\sum_{j\in U_M}\tilde{\theta}_{ij}^{FM,opt}\tilde{\theta}_{jk}^{MT,opt}y_k\sum_{k'\in U}\sum_{j'\in U_M}\tilde{\theta}_{i'j'}^{FM,opt}\tilde{\theta}_{j'k'}^{MT,opt}y_{k'}\\ &\leq \sum_{i\in U_F}\sum_{i'\in U_F}\Delta_{ii'}\sum_{k\in U}\sum_{j\in U_M}\tilde{\theta}_{ij}^{FM}\tilde{\theta}_{jk}^{MT}y_k\sum_{k'\in U}\sum_{j'\in U_M}\tilde{\theta}_{i'j'}^{FM}\tilde{\theta}_{j'k'}^{MT}y_{k'},\end{aligned}$$

pour toute variable d'intérêt y_k, $k \in U$, et tout poids standardisé $\tilde{\theta}_{ij}^{FM}$ et $\tilde{\theta}_{jk}^{MT}$, $i \in U_F$, $j \in U_M$, $k \in U$. Ce problème est plus complexe que celui de l'existence de poids optimaux pour la MGPP simple, car il faut étudier deux ensembles de liens. Par souci de simplicité, on désignera la MGPP double donnée par l'équation (1.14), avec double standardisation, par MGPP double dans la suite de ce chapitre et dans le chapitre 3.

Je montre que des poids optimaux pour la MGPP double existent pour des liens de type TpU-TpU ("Tous-pour-Un-Tous-pour-Un"), pour des plans de sondage satisfaisant les conditions détaillées ci-dessous. Ces conditions sont plus fortes que la Δ-propriété pour U telle que définie dans Medous et al. (2023a). La première condition est la Δ-propriété entre U_F et U_M et la seconde, appelée c_k-condition, porte sur les liens entre U_M et U.

La Δ-propriété définie entre U_F et U dans Medous et al. (2023a) se définit similairement entre les populations U_F et U_M. Soit N_F la taille de la population Frame U_F et, pour $j \in U_M$, U_{Fj} la sous-population de U_F de taille N_{Fj} composée des individus reliés à j. On suppose

dans ce qui suit la population U_F ordonnée selon les sous populations U_{Fj}, $j \in U_M$, et on note $\mathbf{\Delta}^M_{jj'} = (\Delta_{ii'})_{i\in U_{Fj}, i'\in U_{Fj'}}$ la sous-matrice de $\mathbf{\Delta}$ de taille $N_{Fj} \times N_{Fj'}$ composée des éléments en position i, i' si i (resp. i') est lié à j (resp j'). Soit $\mathbb{1}^M_j$ le vecteur de taille N_{Fj} dont les composantes sont égales à 1. Si les liens entre U_F et U_M sont de type TpU, on définit la Δ-propriété pour U_M par :

— pour tout $j \in U_M$, $\mathbf{\Delta}^M_j$ est inversible,

— pour $j \neq j' \in U_M$,

$$\mathbf{\Delta}^M_{j,j'\neq j} = c_{jj'}\mathbb{1}^M_j \mathbb{1}^{Mt}_{j'} \quad \text{avec} \quad c_{jj'} \quad \text{ne dépendant ni de } i \text{ ni de } i'. \tag{1.15}$$

La Δ-propriété pour U_M permet d'obtenir la décomposition de Δ présentée dans la figure 1.27, avec $\Delta^T_{kk'}$, $k \in U$ les sous-matrices de Δ utilisées pour la Δ-propriété entre U_F et U (voir chapitre 2). Cette propriété n'implique pas la Δ-propriété pour U et on rajoute une condition sur les liens entre U_M et U pour obtenir des poids optimaux.

Pour $k \in U$, on définit par U_{Mk} la sous-population de U_M de taille N_{Mk} contenant tous les individus $j \in U_M$ liés à k. On suppose les individus de U_M ordonnés selon ces populations (voir la figure 1.27). Soient les sous-populations U_{Fk}, $k \in U$, de U_F, définies dans Medous et al. (2023a), avec U_{Fk} contenant les individus de U_F reliés à $k \in U$. Si les liens sont de type TpU-TpU, U_F est ordonnée selon les sous-population U_{Fj}, $j \in U_M$, et U_M est ordonnée selon les sous-populations U_{Mk}, $k \in U$, alors U_F est ordonnée selon les sous-populations U_{Fk}, $k \in U$.

Si les liens sont de type TpU-TpU, on dit qu'un plan de sondage satisfait la c_k-condition s'il satisfait la Δ-propriété pour U_M et que, pour tout j et $j' \in U_M$ liés respectivement à k et $k' \in U$ tels que $j \neq j'$, la constante $c_{jj'}$ donnée par la Δ-propriété pour U_M vérifie :

$$c_{jj'} = c_{kk'}$$

avec $c_{kk'}$ ne dépendant ni de j ni de j' et $c_{kk} = c_k$. La c_k-condition est vérifiée pour les plans de Poisson et SASSR, ainsi que, sous certaines conditions (détaillées dans le chapitre 3), pour le STASSR.

Si la c_k-condition est vérifiée, alors je montre qu'il existe des poids minimisant la variance de l'estimateur MGPP double pour toute variable d'intérêt et que ces poids optimaux prennent la valeur suivante :

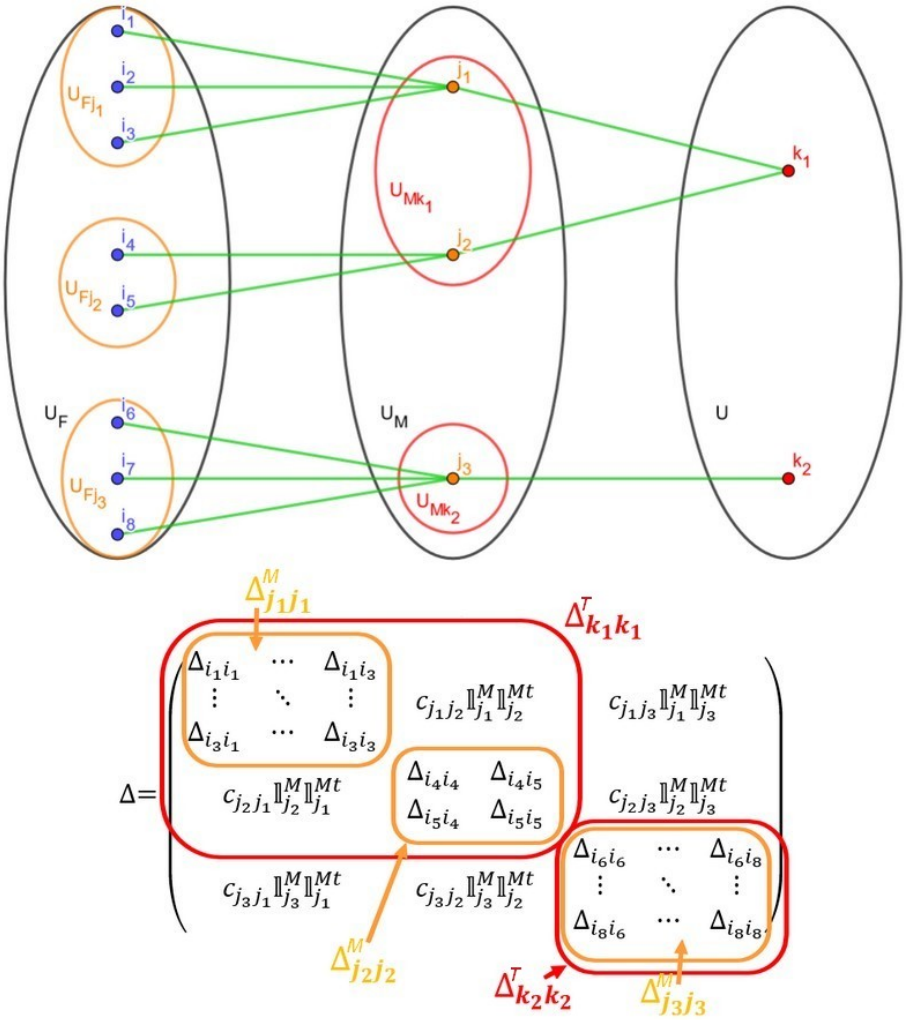

FIGURE 1.27 – Décomposition de la matrice Δ dans le cas TpU-TpU.

$$(\tilde{\theta}_{ij}^{FM,opt})_{i\in U_{Fj}} = (\boldsymbol{\Delta}_j^M)^{-1}\mathbb{1}_j^M \left(\mathbb{1}_j^{Mt}(\boldsymbol{\Delta}_j^M)^{-1}\mathbb{1}_j^M\right)^{-1}$$

$$\tilde{\theta}_{jk}^{MT,opt} = \begin{cases} \dfrac{\mathbb{1}_j^{Mt}(\boldsymbol{\Delta}_j^M)^{-1}\mathbb{1}_j^M/(1-c_k\mathbb{1}_j^{Mt}(\boldsymbol{\Delta}_j^M)^{-1}\mathbb{1}_j^M)}{\sum_{j'\in U_{Mk}}\mathbb{1}_{j'}^{Mt}(\boldsymbol{\Delta}_{j'}^M)^{-1}\mathbb{1}_{j'}^M/(1-c_k\mathbb{1}_{j'}^{Mt}(\boldsymbol{\Delta}_{j'}^M)^{-1}\mathbb{1}_{j'}^M)}, & \text{si } j \text{ est lié à } k, \\ 0 & \text{sinon} \end{cases} \tag{1.16}$$

où la valeur de c_k est donnée par la c_k-condition et $\tilde{\theta}_{ij}^{FM,opt}$, $i \in U_F$, $j \in U_M$, sont les poids optimaux pour la MGPP simple entre U_F et U_M, donnés par l'équation (1.12).

La différence entre la variance de l'estimateur MGPP double utilisant des poids $\tilde{\theta}_{ij}^{FM}$ et

$\tilde{\theta}_{jk}^{MT}$, $i \in U_F$, $j \in U_M$, $k \in U$, quelconques et la variance de l'estimateur MGPP double utilisant les poids optimaux définis par l'équation (1.16) est obtenue en remplaçant les poids $\tilde{\theta}_{ik}^{opt}$ par $\sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT}$ dans l'équation (1.13).

Cependant, si on compare un estimateur MGPP $\hat{t}_{y2}^{FM,opt}$ obtenu en utilisant les poids optimaux $\tilde{\theta}_{ij}^{FM,opt}$ donnés dans (1.16) et des poids quelconques $\tilde{\theta}_{jk}^{MT}$ avec l'estimateur MGPP double optimal, alors la différence de variances se simplifie comme suit :

$$\mathrm{Var}(\hat{t}_{y2}^{FM,opt}) - \mathrm{Var}(\hat{t}_{y2}^{opt}) = \sum_{k \in U_T} y_k^2 \sum_{j \in U_M} \left(\tilde{\theta}_{jk}^{MT} - \tilde{\theta}_{jk}^{MT,opt}\right)^2 \left(\frac{1}{\mathbb{1}_j^{Mt}(\mathbf{\Delta}_j^M)^{-1}\mathbb{1}_j^M} - c_k\right). \quad (1.17)$$

Les poids optimaux pour la MGPP double donnés dans (1.16) permettent d'obtenir la même précision que l'estimateur MGPP simple optimal tout en réduisant le nombre de liens observés. Cependant, ces poids ne sont calculables que si l'information $\mathbb{1}_j^{Mt}(\mathbf{\Delta}_j^M)^{-1}\mathbb{1}_j^M/(1 - c_k\mathbb{1}_j^{Mt}(\mathbf{\Delta}_j^M)^{-1}\mathbb{1}_j^M)$ est disponible pour $j \in U_M$. Pour le plan SASSR, les poids optimaux prennent une forme très simple :

$$(\tilde{\theta}_{ij}^{FM,opt})_{i \in U_{Fj}} = 1/N_{Fj} \text{ si } i \text{ et } j \text{ sont liés, 0 sinon, et}$$

$$\tilde{\theta}_{jk}^{MT,opt} = N_{FJ}/N_{Fk} \text{ si } j \text{ et } k \text{ sont liés, 0 sinon,}$$

avec N_{Fj} le nombre d'individus $i \in U_F$ liés à $j \in U_M$ et N_{Fk} le nombre d'individus $j \in U_M$ liés à $k \in U$. Si les liens sont de type TpU-TpU, $N_{Fk} = \sum_{j \in U_{Mk}} N_{Fj}$. Il faut donc connaître N_{Fj} pour tout $j \in U_M$ lié à k pour calculer $\tilde{\theta}_{jk}^{MT,opt}$.

Pour La Poste, N_{Fj} est le nombre d'adresses dans la case j et N_{Fk} le nombre d'adresses dans la tournée k. Pour calculer les poids $\tilde{\theta}_{jk}^{MT,opt}$, il faut connaître le nombre N_{Fj} d'adresses contenues dans la case j pour toutes les cases $j \in U_M$ liées à la tournée k. Cette information n'est pas disponible et il faut, pour l'obtenir, compter les adresses contenues dans les cases liées à la tournée k. Cela revient à compter toutes les adresses desservies par la tournée k et La Poste fait donc face au même problème qu'avec la MGPP simple.

Pour pallier ce problème, je propose des poids alternatifs pour la MGPP double, qui permettent de limiter la perte de précision des estimateurs comparés aux estimateurs MGPP double optimaux. En pratique, les poids optimaux $\tilde{\theta}_{ij}^{FM,opt}$, $i \in U_F$, $j \in U_M$ sont déjà calculés par La Poste dans SYCI 2, je me concentre donc sur des poids alternatifs aux poids optimaux $\tilde{\theta}_{jk}^{MT,opt}$.

Soit une variable auxiliaire x définie sur U_M, dont la valeur x_j est disponible pour tout

$j \in U_M$. On définit les poids alternatifs suivants pour tout $j \in U_M$ et $k \in U$:

$$\tilde{\theta}_{jk}^{MT,alt} = \begin{cases} \dfrac{x_j}{\sum_{j' \in U_{Mk}} x_{j'}}, & \text{si } j \text{ lié à } k, \\ \\ 0 & \text{sinon.} \end{cases}$$

En me basant sur la différence de variances donnée par (1.17), je propose deux façons de sélectionner la variable auxiliaire x utilisée pour calculer les poids alternatifs $\tilde{\theta}_{jk}^{MT,alt}$.

La première méthode consiste à choisir x corrélée au poids optimal non standardisé $(\theta_{jk}^{MT,opt})_{j \in U_M, k \in U}$. Dans le cas de La Poste, cela revient à sélectionner la variable x la plus corrélée au nombre d'adresses par case. La Poste observe le nombre d'adresses par case et d'autres informations sur les cases (voir section 1.2.2) pour les cases indirectement échantillonnées dans SYCI 2. La corrélation est donc facilement calculable à partir de l'échantillon ou, dans le cas de La Poste, à partir d'un historique de données. Cependant, cette méthode ne prend pas en compte l'impact de y sur la différence de variances et peut donc donner de mauvais résultats si y prend de grandes valeurs (voir section 3.5).

Les liens étant de type TpU, il n'y a qu'une unité $k \in U$ liée à $j \in U_M$ et il est possible de calculer les valeurs $\tilde{y}_j = y_k \tilde{\theta}_{jk}^{MT,opt}$ pour les unités de U_M. La seconde méthode proposée consiste à choisir x corrélée à la variable $\tilde{y}$ prenant les valeurs $\tilde{y}_j$ pour $j \in U_M$. Cette méthode permet de prendre en compte l'impact de y, mais la valeur $\tilde{\theta}_{jk}^{MT,opt}$ étant inconnue, l'hypothèse que la variable x est corrélée à $\tilde{y}$ est invérifiable en pratique. De plus, la variable x choisie dépend de la variable d'intérêt et il faut recalculer les poids alternatifs si la variable sélectionnée change selon les variables d'intérêt.

J'appuie cette discussion sur des simulations basées sur les données de La Poste (voir section 3.5.3). Cette étude par Monte-Carlo montre que la première méthode donne des estimateurs plus précis que la deuxième méthode quand les corrélations sont élevées, alors que la deuxième méthode donne des estimateurs plus précis quand les corrélations sont faibles. Elles montrent aussi que, dans le cas de La Poste, le seuil de corrélation pour que la deuxième méthode donne des estimateurs plus précis que la MGPP double actuelle est faible, alors qu'il est plus élevé avec la première méthode. Cette étude permet une meilleure compréhension du problème de réduction de la variance pour la MGPP double, mais il reste difficile d'améliorer la précision des estimateurs à La Poste à l'heure actuelle. Certaines bases de données (dont RAO, voir section 1.2.2) contiennent des variables auxiliaires qui semblent fortement corrélées au nombre d'adresses par case, comme un nombre d'adresses par case théorique, mais qui ne sont pas encore exploitables.

1.3.2 Intégration statistique de données

Une part importante de mes travaux de thèse a porté sur l'utilisation des bases de données massives dont dispose La Poste, avec en priorité la base TAE (voir section 1.2.2). Dans la section précédente, on a montré que des données auxiliaires peuvent être utilisées dans le calculs des poids de la MGPP double. Les données auxiliaires sont alors utilisées pour améliorer la méthodologie actuelle sans la modifier. Cependant la réduction de variance apportée par cette méthode est limitée, car on peut, au mieux, obtenir l'estimateur MGPP double optimal.

Dans cette section, je propose une nouvelle méthodologie basée sur l'utilisation des données disponibles dans TAE. Les résultats présentés seront publiés dans l'article *"QR Prediction for Statistical Data Integration"*(Survey Methodology, à paraître).

Chapitre 4 : *QR Prediction for Statistical Data Integration*

La littérature sur l'utilisation de bases de données massives en sondage considère le cas où les valeurs des variables d'intérêt sont connues pour tous les individus des bases de données non probabilistes (Kim and Tam, 2021; Beaumont, 2020). Or, les données de La Poste, TAE (voir Section 1.2.2), sont collectées par les machine de tri. Certaines variables d'intérêt considérées par La Poste, comme le type ou le montant d'affranchissement, ne sont pas reconnues par les machine et leurs valeurs pour les individus de TAE ne sont donc pas connues. Dans cet article, je commence par étudier certains estimateurs proposés par Kim and Tam (2021), puis je propose un estimateur du total T_y adapté aux données postales inspirés de l'estimateur QR introduit par Wright (1983) et Särndal and Wright (1984).

Soit s_P un échantillon probabiliste tiré avec un plan de sondage $p(.)$ et s_{NP} un échantillon non-probabiliste de taille N_{NP}. On note π_k, $k \in U$ (resp. $\pi_{kk'}$, $k, k' \in U$) les probabilités d'inclusion dans s_P de premier (resp. second) ordre et δ_k l'indicatrice d'inclusion dans s_{NP}, où $\delta_k = 1$ si k appartient à s_{NP}, 0 sinon. On suppose que la valeur de δ_k est connue pour tout individu k appartenant à s_P.

On suppose dans un premier temps que la variable d'intérêt y est connue pour tout individu de s_P et de s_{NP}. On note $T_{NP} = \sum_{k \in s_{NP}} y_k = \sum_{k \in U} \delta_k y_k$ le total de y sur s_{NP} et $T_C = T_y - T_{NP}$ le total de y sur $U - s_{NP}$. Si la taille de la population N est connue, Kim and Tam (2021) proposent l'estimateur suivant :

$$\hat{T}_{PDI} = T_{NP} + \hat{T}_C^{(\mathrm{Ha})}, \tag{1.18}$$

avec $\hat{T}_C^{(\mathrm{Ha})}$ l'estimateur de Hájek du total T_C (voir section 1.1.3) :

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

Je montre que la différence entre la variance de l'estimateur de HT du total $\hat{T}_{HT,y}$ obtenu à partir de l'échantillon s_P (voir section 1.1.3) et la variance asymptotique de $\hat{T}_{PDI}$ est donnée par :

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

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

L'estimateur $\hat{T}_{PDI}$ est asymptotiquement plus précis, en terme de variance, que l'estimateur de Horvitz-Thompson $\hat{T}_{HT,y}$. La Poste souhaite utiliser l'estimateur $\hat{T}_{PDI}$, en utilisant les échantillons probabilistes tirés dans le cadre de SYCI 2 et TAE comme échantillon non-probabiliste (voir 1.2.2), afin de réduire la variance de ses estimateurs actuels et limiter l'impact de la réduction future des tailles d'échantillons (voir section 1.2.1). Cependant, La Poste ne dispose pas la valeur de y pour les individus de Traitement Automatisé de l'Enveloppe (TAE). La Poste ne peut donc pas utiliser $\hat{T}_{PDI}$ tel quel.

Dans la deuxième partie de Medous et al. (2023b), on suppose que les valeurs de y sont connues pour les individus de s_P mais inconnues pour les individus de s_{NP}. On suppose aussi que l'intersection entre s_P et s_{NP} est non vide, ce qui est le cas à La Poste car la base de données TAE contient 90% des lettres distribuées (voir section 1.2.2). Dans cette partie, je cherche à adapter l'estimateur $\hat{T}_{PDI}$ aux données de La Poste.

On suppose qu'un vecteur $\boldsymbol{x}_k = (X_{k1}, \ldots, X_{kp})^\top$ de variables auxiliaires est disponible pour tout individu $k \in s_{NP}$ et que les valeurs de δ_k et $\delta_k \boldsymbol{x}_k$ sont connues pour tout individu $k \in s_P$.

L'idée présentée dans Medous et al. (2023b) est de remplacer les valeurs y_k, $k \in s_{NP}$ dans l'équation (1.18) par des prédictions $\hat{y}_k$, obtenues en modélisant y en fonction des variables auxiliaires sur l'intersection $s_p \cap s_{NP}$. On obtient alors un prédicteur $\hat{T}_{NP}$ du total T_{NP}.

On suppose le modèle suivant :

$$y_k = \boldsymbol{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \quad k \in s_{NP}, \tag{1.19}$$

où les erreurs ε_k sont indépendantes, d'espérance $\mathrm{E}_m(\varepsilon_k) = 0$ et de variance $\mathrm{Var}_m(\varepsilon_k)$ propor-

tionnelle à $\nu(\boldsymbol{x}_k) = v_k$ pour une constante positive v_k. L'indice m indique que l'on considère l'espérance et la variance sous le modèle 1.19.

Soit q_k, $k \in U$ des constantes positives connues. On définit le prédicteur de y_k, $k \in U$ suivant :

$$\hat{y}_k = \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}}$$

avec

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k \in s_P} q_k \delta_k \boldsymbol{x}_k \boldsymbol{x}_k^\top \right)^{-1} \left(\sum_{k \in s_P} q_k \delta_k \boldsymbol{x}_k y_k \right), \tag{1.20}$$

où la matrice $\sum_{k \in s_P} q_k \delta_k \boldsymbol{x}_k \boldsymbol{x}_k^\top$ est supposée inversible pour tout échantillon s_P. Je propose dans Medous et al. (2023b) d'utiliser un prédicteur de type QR, introduit par Wright (1983) :

$$\begin{aligned} \hat{T}_{NP}^{(\mathrm{QR})} &= \sum_{k \in U} \delta_k \hat{y}_k + \sum_{k \in s_P} r_k \delta_k (y_k - \hat{y}_k) \\ &= \sum_{k \in U} \delta_k \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}} + \sum_{k \in s_P} r_k \delta_k (y_k - \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}}), \end{aligned} \tag{1.21}$$

où r_k, $k \in U$ sont des constantes positives connues.

Ce prédicteur est sans biais sous le modèle 1.19 (voir chapitre 4). En pratique, on préfère détailler le biais du prédicteur QR sous le plan.

Afin de pouvoir étudier le comportement du prédicteur $\hat{T}_{NP}^{(\mathrm{QR})}$ sous l'inférence du plan de sondage, on suppose qu'il existe un vecteur $\boldsymbol{\lambda} \in \mathbf{R}^p$ tel que

$$(\mathrm{QR}) : \quad 1 - \pi_k r_k = \pi_k q_k \boldsymbol{x}_k^\top \boldsymbol{\lambda} \quad \text{pour tout} \quad k \in s_{NP}. \tag{1.22}$$

Cette condition, appelée (QR) condition dans Medous et al. (2023b), permet de simplifier l'expression du prédicteur QR comme suit :

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

avec

$$\hat{T}_{NP}^{(\mathrm{Q}\pi)} = \sum_{k \in U} \delta_k \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}} + \sum_{k \in s_P} d_k \delta_k (y_k - \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}}) \tag{1.23}$$

où $d_k = 1/\pi_k$.

Wright (1983) et Särndal and Wright (1984) ont montré que, sous certaines conditions, le prédicteur QR défini par (1.23) est asymptotiquement sans biais. Les conditions nécessaires,

ainsi que les preuves, sont détaillées dans le chapitre 4.

Le total T_y est estimé par :

$$\hat{T}_y^{(\mathrm{QR})} = \hat{T}_{NP}^{(\mathrm{QR})} + \hat{T}_C^{(\mathrm{Ha})}. \tag{1.24}$$

Dans le chapitre 4, je montre que $\hat{T}_y^{(\mathrm{QR})}$ a des propriétés de biais similaires à celles de $\hat{T}_{NP}^{(\mathrm{QR})}$.

Si la condition (QR) est vérifiée, et sous des hypothèses sur le plan de sondage et les variables y et $\boldsymbol{x}$, l'erreur d'estimation peut être approchée par :

$$\frac{1}{N}(\hat{T}^{(\mathrm{Q}\pi)} - T) - \frac{1}{N}\left(\sum_{k\in s_p} d_k(E_k + e_k) - \sum_{k\in U}(E_k + e_k)\right) = o_p\left(1/\sqrt{n}\right),$$

avec

$$E_k = \delta_k(y_k - \boldsymbol{x}_k^\top \tilde{\boldsymbol{\beta}}), \quad e_k = (1-\delta_k)\left(y_k - \frac{\sum_{k'\in U}(1-\delta_{k'})y_{k'}}{N - N_{NP}}\right),$$

et

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{k\in U} \pi_k q_k \delta_k \boldsymbol{x}_k \boldsymbol{x}_k^\top\right)^{-1} \sum_{k\in U} \pi_k q_k \delta_k \boldsymbol{x}_k y_k.$$

Cette approximation permet d'obtenir la variance approchée de l'estimateur QR :

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

Si $\pi_{kk'} > 0$ pour tout $k, k' \in U$, un estimateur de cette variance est donné par :

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

avec $\hat{E}_k = \delta_k(y_k - \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}})$ et $\hat{e}_k = (1-\delta_k)(y_k - \sum_{k'\in s_p} d_{k'}(1-\delta_{k'})y_{k'}/(N - N_{NP}))$.

Dans le chapitre 4, on discute du choix des constantes q_k et r_k et on propose trois options qui vérifient la condition (QR) donnée par (1.22) :

$$q_k = d_k v_k^{-1} \text{ et } r_k = d_k,$$

$$q_k = v_k^{-1} \text{ et } r_k = 1,$$

$$q_k = (d_k - 1)v_k^{-1} \text{ et } r_k = 1.$$

Les deux premières options permettent d'obtenir respectivement l'estimateur Model-Assisted et Model-Based (Särndal et al., 1989). La troisième option permet d'obtenir l'estimateur

cosmétique du total T_{NP}, introduit par Särndal and Wright (1984) :

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

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

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

Une étude sur populations simulées permet de mettre en avant les avantages et inconvénients de chaque option et de comparer leur efficacité (en terme de biais et variance) par rapport à l'estimateur de HT et l'estimateur $\hat{T}_{PDI}$ du total T_y.

Une deuxième étude utilisant les données historiques de La Poste montre des résultats similaires. Dans le cas de La Poste, l'utilisation d'un estimateur cosmétique du total T_{NP} semble la plus pertinente. Cependant, on note que l'utilisation de l'estimateur de Hájek du total T_C n'est pas la plus adaptée aux données de La Poste, d'une part à cause de sa sensibilité à la sélection de s_P (voir chapitre 4) et d'autre part parce que la taille N de la population à La Poste est inconnue. L'utilisation d'un estimateur de HT du total T_C est plus judicieuse pour La Poste et est actuellement à l'étude.

Sondage Indirect et Méthode Généralisée de Partage des Poids

Chapitre 2

Many-to-One indirect sampling with application to the French postal traffic estimation [1], modifié pour la thèse

Abstract

In social and economic surveys, it can be difficult to directly reach units of the target population, and indirect sampling is often advocated to solve this issue. In indirect sampling, the sample is drawn from a frame population that is linked to the target population, and estimation of target population parameters is typically achieved through the Generalized Weight Share Method (GWSM). This method provides a weight, for every unit of the target population, that depends on the one hand, on the sampling weights in the frame population and, on the other hand, on the link weights between the frame population and the target population. In the present study, we focus on the situation in which the units from the frame population are linked to one and only one unit from the target population (Many-to-One case). This situation is encountered at the French postal service where addresses are sampled instead of postman rounds. We aim at understanding of the impact of the link weights on the efficiency of the GWSM estimators. We derive variance expressions and optimality results for a large class of sampling designs. Moreover, we note that the Many-to-One case can lead to too many links to observe. We alleviate the problem by introducing an intermediate population and double indirect sampling. The question of the loss of precision in this situation is discussed in detail through theoretical results and simulations. These findings help to explain the loss of precision of double GWSM estimators observed recently at the French postal service.

1. Article de E.Medous, C.Goga, A. Ruiz-Gazen, J.F Beaumont, A. Dessertaine et P. Puech, 2023, publié dans *The Annals of Applied Statistics*, 17(1) :838-859

Keywords : Generalized Weight Share Method, Optimal link weights, Stratified sampling, Variance estimation.

2.1 Introduction

In France, at the postal service (La Poste), only part of the postal traffic goes though an automatized processing. The monthly postal traffic is unknown and is estimated through probability-based surveys. For many years, La Poste has drawn samples directly in the population of postman rounds, which is considered to be the population of interest or the target population. Since 2008, the organization of postman rounds has changed and is no longer stable over time. Sampling directly in the target population has become impossible, and the sampling design has evolved to an indirect sampling design where the frame population is the population of postal addresses.

Indirect sampling has been extensively studied in the survey sampling literature ; see Deville and Lavallée (2006), Lavallée (2007), Kiesl (2016) and Haziza and Beaumont (2017) for a general theory and reviews with many references inside. As described in Kiesl (2016), indirect sampling has many applications, including household panels : Kalton and Brick (1995); Rendtel and Harms (2009) and hard-to-reach populations : Deville and Maumy-Bertrand (2006) for tourism and De Vitiis et al. (2014) for the homeless population. The use of indirect sampling at La Poste, has been introduced in Dessertaine and Fluteaux (2004) and Lardin-Puech (2014). A useful estimation method in the context of indirect sampling is the Generalized Weight Share Method (GWSM), as detailed in Deville and Lavallée (2006). It consists of using the links that relate the frame and the target populations, and considering a total over the target population to be a total over the frame population. The use of standard methods, such as the Horvitz-Thompson estimator, is then possible and leads to the GWSM estimator. Kalton and Brick (1995), Deville and Lavallée (2006) and Kiesl (2016), among others, studied in detail the properties of the GWSM estimator, and in particular the question of the impact of the link structure and link weights on its variance. Optimality, in the sense of variance minimization with respect to the link weights for unbiased GWSM estimators, is discussed at length in Deville and Lavallée (2006). The conclusion is that optimal GWSM estimators, that do not depend on the variable of interest, cannot be derived for general link structure. In the present study, we propose to focus on a particular link structure described below that is of interest at La Poste, and go further into the understanding of indirect sampling.

At the French postal service, every postal address is linked to only one postman round

for a given day. This link structure is of a particular type, called Many-to-One (MtO), where each unit in the frame population is linked to one and only one unit in the target population. This situation is also encountered in households surveys where individuals are sampled instead of households. This link structure is studied in detail in the present paper. We derive the optimal GWSM estimator that minimizes the variance among the unbiased GWSM estimators, for a large class of indirect MtO sampling designs. This class includes Poisson sampling, simple random sampling without replacement, and stratified designs, including the design implemented at La Poste. Moreover, we derive a simple formula to evaluate the increase in variance when using a non-optimal GWSM estimator compared to the optimal one.

The weight share method is simple but requires that the links between the indirectly sampled units in the target population, and the frame population, are known. The problem faced by the French postal service with MtO links, is that every unit in the target population is linked to a very large number of units in the sampling frame. At La Poste, all addresses delivered during a sampled postman round must be known. On average, there are approximately 500 addresses per postman round, and it is not possible to enumerate all of the addresses in the morning, before the departure of the postman.

To get around this problem, La Poste has set up a double indirect sampling design, using the outgoing mail sorting boxes as an intermediate population. This method is much faster than simple indirect sampling. Only the addresses of the sampled boxes and the boxes of the sampled rounds are to be observed, which is approximately 60 items on average, compared to the 500 for simple indirect sampling. Given the situation at La Poste, double indirect sampling is an alternative to a time consuming simple indirect sampling design. This alternative is necessary to be able to collect the data. However, using this method, La Poste observed a deterioration in the precision of the estimators. The goal of the present paper is to understand the loss of precision observed at La Poste, but also to give guidelines for the implementation of an efficient double indirect sampling design.

In Section 2.2 of the present paper, we consider a large class of indirect MtO sampling designs. We derive the optimal GWSM estimator and give a simple expression of the difference in the variances between the optimal GWSM estimator and any non-optimal unbiased GWSM estimator. For Poisson sampling, we also prove that the optimal GWSM estimator is less precise than the direct estimator. The result on optimal GWSM estimator is used in Section 2.3, where we introduce and compare double indirect MtO sampling with simple indirect MtO sampling with optimal link weights. In the same section, we detail situations where there is a gain, in terms of the smaller number of links to observe, from using double indirect sampling compared to simple indirect sampling. In Section 2.4, we define several

setups and illustrate, through a Monte Carlo study, the impact of double indirect sampling on the precision of the estimators. Depending on the link structure, we observe that there could be no loss at all, or, on the contrary, an enormous loss of precision. In Section 2.5, we also give numerical results in a context similar to La Poste. These results allow us to explain the loss of precision observed at La Poste, when using a double indirect sampling design compared to a simple direct sampling design. Section 2.6 concludes the paper while the proofs are given in the Appendix.

2.2 Indirect sampling

2.2.1 GWSM

In some surveys, it is not possible to sample directly from the target population U_T. However, a sampling frame can exist for a population U_F, that is related to U_T in such a way that any unit in U_T is linked to at least one unit in U_F. Indirect sampling refers to selecting a sample s_F from U_F by using standard selection methods and derive estimators for parameters defined on U_T. In the case of the La Poste survey, the population of interest is made of postman rounds on a given day in France, but no sampling frame for the rounds exists. A sampling frame for postal addresses is however available, and each postman round contains at least one address.

Let us denote by N_T (resp. N_F) the size of U_T (resp. U_F) and by l_{ik} the link between $i \in U_F$ and $k \in U_T$, with $l_{ik} = 1$ if the units i and k are linked, and $l_{ik} = 0$ otherwise. Units from U_F can be linked in several ways to units from U_T (Deville and Lavallée, 2006). We can have "Many-to-One" (MtO) links as on the left panel of Figure 2.1, namely each unit from the frame population U_F is linked to only one unit from the target population U_T. We can have "One-to-Many" links as on the middle panel of Figure 2.1, namely each unit from U_T is linked to only one unit from U_F. Finally, we can have "Many-to-Many" (MtM) links as on the right panel of Figure 2.1, with units from U_F linked to several units in U_T and reciprocally. In the La Poste survey, an address almost always belongs to only one round and the links are MtO. Following Deville and Lavallée (2006), we start by making the assumption that the links between U_F and U_T can be observed for every unit in U_F and every unit in U_T.

Let y be the variable of interest measured on U_T, and let y_k be its value for the k-th unit in U_T. We are interested in estimating $t_y = \sum_{k \in U_T} y_k$, the total of y over U_T. A sample s_F is drawn from U_F according to a sampling design $p_F(\cdot)$. We can associate to s_F the vector $(I_1, \ldots, I_{N_F})'$ where I_i is the sample membership indicator of the individual i from

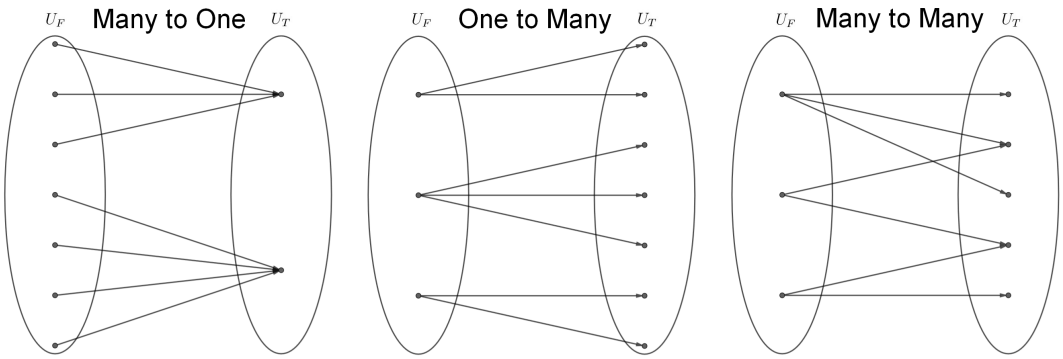

FIGURE 2.1 – The different types of links between U_F and U_T.

U_F defined as $I_i = 1$ if i is selected and $I_i = 0$ otherwise. We denote by $\pi_i = Pr(i \in s_F)$ the first-order inclusion probability of unit i and by $\pi_{ii'} = Pr(i, i' \in s_F)$ the second-order inclusion probability of units i and i'. We suppose that all of the units i have a positive inclusion probability $\pi_i > 0$ and we denote by $d_i = 1/\pi_i$ their sampling weights. Two standard sampling designs are considered in the present paper : Simple Random Sampling Without Replacement (SRSWOR) of size n_F, and Poisson design with inclusion probabilities π_i, $i \in U_F$. For SRSWOR, p_F assigns an equal probability to all without replacement samples of size n_F and zero otherwise. The sampling weights are equal to $d_i = N_F/n_F$ for all i in U_F. For Poisson sampling, the variables I_i's are independent and distributed as Bernoulli random variables with parameter π_i.

The sample s_F in U_F leads to a sample s_T in U_T, which is made of the units in U_T linked to at least one unit in s_F. However, the sampling design $p_T(\cdot)$ which governs the selection of s_T, as well as the associated first-order inclusion probabilities, may be difficult to derive (Deville and Lavallée, 2006). Fortunately, as we will see in the next paragraph, for the GWSM estimators only $p_F(\cdot)$ and the associated inclusion probabilities are needed for the estimation of t_y and $p_T(\cdot)$ will not be used.

Consider, for all $i \in U_F$ and $k \in U_T$, some non negative link weight θ_{ik} associated to the link l_{ik} between U_F and U_T, such that θ_{ik} is positive when $l_{ik} = 1$ and $\theta_{ik} = 0$ otherwise. We define the standardized link weights $\tilde{\theta}_{ik} = \theta_{ik}/\sum_{i' \in U_F} \theta_{i'k}$ which satisfy the constraint $\sum_{i \in U_F} \tilde{\theta}_{ik} = 1$. To compute the standardized link weights $\tilde{\theta}_{ik}$ for a given k in U_T, one needs to know $\sum_{i \in U_F} \theta_{ik}$, which implies that the units i in U_F linked with k must be known. We can take as an example $\theta_{ik} = l_{ik}$, and in this case, standardization implies that the number of units i in U_F linked with k is known. More general weights can also be considered ; see Deville and Lavallée (2006) and Haziza and Beaumont (2017). The total t_y can then be written as

the total on U_F of the variable $\sum_{k\in U_T} \tilde{\theta}_{ik} y_k, i \in U_F$, as follows :

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

The estimation of t_y can be obtained by considering standard estimators based on the sample s_F selected from U_F. The Horvitz-Thompson (HT) estimator of t_y is given by

$$\hat{t}_{y1} = \sum_{i\in s_F} d_i \left(\sum_{k\in U_T} \tilde{\theta}_{ik} y_k \right) \tag{2.1}$$

and it estimates unbiasedly the total t_y, provided that the link weights $\tilde{\theta}_{ik}, i \in U_F$, are standardized. This estimator is the Generalized Weight Share Method (GWSM) estimator, and it was studied by Deville and Lavallée (2006). It can also be written as follows :

$$\hat{t}_{y1} = \sum_{k\in U_T} \hat{t}_{\tilde{\theta}_k} y_k$$

where $\hat{t}_{\tilde{\theta}_k} = \sum_{i\in s_F} d_i \tilde{\theta}_{ik}$ is the HT estimator of the total $t_{\tilde{\theta}_k} = \sum_{i\in U_F} \tilde{\theta}_{ik} = 1$, for all k in U_T. To calculate $\hat{t}_{y1}$, we only need to standardize the link weights that correspond to the sampled units k in U_T.

Let us denote

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

The variance of the GWSM estimator $\hat{t}_{y1}$ is given by :

$$\mathrm{Var}(\hat{t}_{y1}) = \sum_{i\in U_F} \sum_{i'\in U_F} \Delta_{ii'} \sum_{k\in U_T} \tilde{\theta}_{ik} y_k \sum_{k'\in U_T} \tilde{\theta}_{i'k'} y_{k'}.$$

Interestingly, this variance can be rewritten as :

$$\mathrm{Var}(\hat{t}_{y1}) = \sum_{k\in U_T} \sum_{k'\in U_T} y_k y_{k'} \, \mathrm{Cov}(\hat{t}_{\tilde{\theta}_k}, \hat{t}_{\tilde{\theta}_{k'}}), \tag{2.2}$$

where

$$\mathrm{Cov}(\hat{t}_{\tilde{\theta}_k}, \hat{t}_{\tilde{\theta}_{k'}}) = \sum_{i\in U_F} \sum_{i'\in U_F} \Delta_{ii'} \, \tilde{\theta}_{ik} \tilde{\theta}_{i'k'}.$$

The variance expression (2.2) is similar to the variance in the case of a direct sampling design on U_T. However, while for direct sampling, the covariance involved in the double sum is between the sample membership indicators weighted by the sampling weights, for indirect

sampling, the covariance is between the HT estimators of the link weights totals.

Deville and Lavallée (2006) were interested in finding the optimal weights $\tilde{\theta}_{ik}^{opt}$ that minimize the variance $\mathrm{Var}(\hat{t}_{y1})$ for any survey variable y, namely

$$\mathrm{Var}(\hat{t}_{y1}^{opt}) \leq \mathrm{Var}(\hat{t}_{y1}), \text{for all } y, \tag{2.3}$$

where

$$\hat{t}_{y1}^{opt} = \sum_{i \in s_F} d_i \left(\sum_{k \in U_T} \tilde{\theta}_{ik}^{opt} y_k \right)$$

is the GWSM estimator obtained by using the optimal link weights $\tilde{\theta}_{ik}^{opt}, i \in U_F, k \in U_T$. Such a criterion is called *strong optimality* criterion and the optimal weights derived in this way, if they exist, should not depend on the survey variable y. However, the optimal link weights satisfying (2.3) might not exist, even for particular sampling designs such as Poisson or SRSWOR sampling designs (see Deville and Lavallée, 2006). Therefore, Deville and Lavallée (2006) suggested the *weak optimality* criterion which consists in finding the weak-optimal weights $\tilde{\theta}_{ik}^{wopt}$ which minimize the variance $\mathrm{Var}(\hat{t}_{y1})$ for the particular variables y such that $y_k = 1$ for a unit $k \in U_T$ and $y_{k'} = 0$ for $k' \neq k \in U_T$. We obtain for these particular variables :

$$\hat{t}_{y1} = \sum_{i \in s_F} d_i \tilde{\theta}_{ik} = \hat{t}_{\tilde{\theta}_k}.$$

Thus, the weak-optimal weights $\tilde{\theta}_{ik}^{wopt}$ are the weights that minimize the variance of $\hat{t}_{\tilde{\theta}_k}$:

$$(\tilde{\theta}_{ik}^{wopt})_{i \in U_F} = \arg \min_{\tilde{\theta}_{ik}, i \in U_F} \mathrm{Var}(\hat{t}_{\tilde{\theta}_k}), \quad \text{for all} \quad k \in U_T. \tag{2.4}$$

Deville and Lavallée (2006) derived weak-optimal link weights for Poisson and SRSWOR, and noticed that the weak optimality is a necessary condition for strong optimality.

2.2.2 MtO links and optimal weight links

The strong minimization problem previously mentioned becomes easier to handle when the links between U_F and U_T are Many-to-One. With MtO links, every unit from U_F is linked to only one unit from U_T, and we can order the units in U_F with respect to their common linked unit in U_T. For a given unit $k \in U_T$, let us denote by U_{Fk}, with size $N_{Fk} = \sum_{i \in U_F} l_{ik}$, the set of units i in U_F that are linked to k. In what follows, we consider that units in U_F are ordered according to these subpopulations. Let $\mathbf{\Delta} = (\Delta_{ii'})_{i,i' \in U_F}$ be the matrix of size $N_F \times N_F$. Thanks to this ordering, we can consider the submatrix $\mathbf{\Delta}_{kk'} = (\Delta_{ii'})_{i \in U_{Fk}, i' \in U_{Fk'}}$ of $\mathbf{\Delta}$ corresponding to elements in positions i and i' such that i (resp. i') is linked to k (resp.

k'), for all k (resp. k') in U_T. For simplicity, we denote $\mathbf{\Delta}_k$ the $\mathbf{\Delta}_{kk}$ square submatrix with size N_{Fk}. With MtO links, the submatrices $\mathbf{\Delta}_{kk'}$, $k, k' \in U_T$, form a partition of $\mathbf{\Delta}$, namely

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

Let $\mathbb{1}_k$ be the N_{Fk}-dimensional vector of ones. For MtO links, a sampling design is said to satisfy the $\mathbf{\Delta}$-property if, for all $k \in U_T$, $\mathbf{\Delta}_k$ is invertible and, for $k \neq k' \in U_T$, we have

$$\mathbf{\Delta}_{k,k' \neq k} = c_{kk'} \mathbb{1}_k \mathbb{1}_{k'}^t \quad \text{with} \quad c_{kk'} \quad \text{not depending on } i \text{ and } i'. \tag{2.5}$$

The $\mathbf{\Delta}$-property holds for Poisson sampling, SRSWOR and stratified SRSWOR under conditions detailed below.

For Poisson sampling from U_F with inclusion probabilities $\pi_i, i \in U_F$, $\mathbf{\Delta}_k$ is diagonal with positive terms, thus invertible, and $c_{kk'} = 0$ for all $k \neq k' \in U_T$. For SRSWOR of size n_F from U_F, $\mathbf{\Delta}_k$ is invertible as soon as $N_T > 1$, as can be seen in Deville and Lavallée (2006), page 174. If we denote $f = n_F/N_F$, we have

$$c_{kk'} = -\frac{1-f}{f}\frac{1}{N_F - 1},$$

which does not depend on i and i', for all $k \neq k' \in U_T$. For stratified SRSWOR with H strata of size N_h in U_F, $h = 1, \ldots, H$, let us denote $f_h = n_h/N_h$. The submatrix $\mathbf{\Delta}_{k,k' \neq k}$ cannot generally be written as $c_{kk'} \mathbb{1}_k \mathbb{1}_{k'}^t$, especially if k is linked to one unit i from stratum h and one unit i' from stratum $h' \neq h$. The $\mathbf{\Delta}$-property holds if we assume that, for all k in U_T, all units i linked to k belong to the same stratum h and if, for each stratum h, there are at least two units of U_T linked to h. Then $\mathbf{\Delta}_k$ is invertible, and we have

$$c_{kk'} = \begin{cases} -\dfrac{1-f_h}{f_h}\dfrac{1}{N_h - 1}, & \text{if all units in } U_F \text{ linked to } k \text{ and } k' \text{ are in the same stratum } h \\ \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $c_{kk'}$ does not depend on i and i', for all $k \neq k' \in U_T$.

The first part of Proposition 2.2.1 gives an expression of the variance of the GWSM estimator while the second part gives the optimal link weights, in the MtO case, for sampling designs that satisfy the $\mathbf{\Delta}$-property.

Proposition 2.2.1. *If the links are MtO and the sampling design satisfies the $\mathbf{\Delta}$-property,*

then :

$$Var(\hat{t}_{y1}) = \sum_{k\in U_T} y_k^2\, Var(\hat{t}_{\tilde{\theta}_k}) + \sum_{k\in U_T}\sum_{k'\neq k\in U_T} y_k y_{k'} c_{kk'}. \tag{2.6}$$

where $\hat{t}_{\tilde{\theta}_k} = \sum_{i\in s_F} d_i \tilde{\theta}_{ik}$ *is the HT estimator of the total* $t_{\tilde{\theta}_k} = 1$ *for all* $k \in U_T$.

Moreover, there is a unique set of optimal link weights verifying the strong-optimality criterion given in (2.3), given by

$$(\tilde{\theta}_{ik}^{opt})_{i\in U_{Fk}} = \mathbf{\Delta}_k^{-1}\mathbb{1}_k \left(\mathbb{1}_k^t \mathbf{\Delta}_k^{-1}\mathbb{1}_k\right)^{-1}, \text{ for all } k \in U_T.$$

The second term in the right-hand term of (2.6), does not depend on the link weights. As a consequence, minimizing the variance of $\hat{t}_{y1}$, regardless of the variable y is, is equivalent to minimizing $\text{Var}(\hat{t}_{\tilde{\theta}_k})$ for all k in U_T. Using the terminology introduced by Deville and Lavallée (2006) (see also relations (2.3) and (2.4)), this result means that, for MtO links and sampling designs that satisfy the $\mathbf{\Delta}$-property, *weak*-optimality of the link weights is equivalent to *strong*-optimality and the link weights will be simply called optimal in the following.

For Poisson sampling, we have $\Delta_{ii} = (1-\pi_i)/\pi_i$ and the optimal link weights are equal to :

$$\tilde{\theta}_{ik}^{\text{opt}} = \frac{l_{ik}/\Delta_{ii}}{\sum_{i'\in U_F} l_{i'k}/\Delta_{i'i'}}, \tag{2.7}$$

for all $i \in U_F$ and $k \in U_T$. For SRSWOR sampling, the optimal link weights are equal to $\tilde{\theta}_{ik}^{\text{opt}} = l_{ik}/\sum_{i'\in U_F} l_{i'k}$, for all $i \in U_F$ and $k \in U_T$. Details on the derivation of these optimal link weights can be found in Deville and Lavallée (2006). For stratified SRSWOR with the assumptions previously mentioned, it is easy to prove that the optimal link weights are the same as those for SRSWOR by following the proof in Deville and Lavallée (2006).

Consider now the GWSM estimator $\hat{t}_{y1}$ given in (2.1) and computed with some standardized link weights $\tilde{\theta}_{ik}, i \in U_F, k \in U_T$. Consider also the optimal GWSM estimator denoted by $\hat{t}_{y1}^{opt}$, and computed with the optimal link weights $\tilde{\theta}_{ik}^{opt}$. For MtO links and designs that satisfy the $\mathbf{\Delta}$-property, it is possible to derive a new formula for the loss of efficiency between $\hat{t}_{y1}$ and $\hat{t}_{y1}^{opt}$. This loss can be expressed as a simple function of the variances of the HT estimator of $t_{\tilde{\theta}_k - \tilde{\theta}_k^{opt}} = \sum_{i\in U_F}(\tilde{\theta}_{ik} - \tilde{\theta}_{ik}^{opt}) = 0$, k in U_T.

Proposition 2.2.2. *If the links are MtO and the sampling design satisfies the* $\mathbf{\Delta}$*-property, then the loss of efficiency compared with optimal link weights* $\tilde{\theta}_{ik}^{\text{opt}}$, $i \in U_F, k \in U_T$, *is given by :*

As mentioned before, the matrices $\mathbf{\Delta}$ derived with Poisson and SRSWOR designs satisfy

the $\mathbf{\Delta}$-property. Thus, we can compute easily the loss of efficiency between $\hat{t}_{y1}$ and $\hat{t}_{y1}^{opt}$ as formulated in the following corollary.

Corollary 2.2.1. *If the links are MtO and if the sampling designs are Poisson or SRSWOR, then the loss in efficiency between $\hat{t}_{y1}$ and $\hat{t}_{y1}^{opt}$ has the expression given in Proposition 2.2.2. For Poisson sampling, we have :*

$$Var(\hat{t}_{y1}) - Var(\hat{t}_{y1}^{opt}) = \sum_{k\in U_T} y_k^2 \sum_{i\in U_F} \frac{1-\pi_i}{\pi_i}(\tilde{\theta}_{ik} - \tilde{\theta}_{ik}^{opt})^2.$$

For SRSWOR, we have :

$$Var(\hat{t}_{y1}) - Var(\hat{t}_{y1}^{opt}) = c \sum_{k\in U_T} y_k^2 \sum_{i\in U_F} (\tilde{\theta}_{ik} - \tilde{\theta}_{ik}^{opt})^2,$$

where $c = N_F^2 \left(\frac{1}{n_F} - \frac{1}{N_F}\right) \frac{1}{N_F - 1}$.

The previous expression for SRSWOR can be easily adapted to stratified SRSWOR under the assumptions mentioned previously.

2.2.3 Comparison of direct and optimal indirect MtO sampling designs

It is not possible to compare theoretically the variances of simple indirect and direct estimators for general sampling designs even if we restrict ourselves to MtO links.

However, for Poisson sampling, we prove in Proposition 2.2.3 that the variance of the direct HT estimator is always smaller than the variance of the simple GWSM estimator when using optimal link weights. Let us consider a Poisson sampling design in U_F with first-order inclusion probabilities π_i, $i \in U_F$. Because of the independence between the inclusion indicators, this sampling design induces a Poisson sampling design on U_T. Accounting for the MtO links between U_F and U_T, we can calculate the probability of inclusion of every unit $k \in U_T$. We have that

$$\pi_k = P(\text{ at least one } i \in U_F \text{ linked to } k \text{ is sampled in } s_F) = 1 - \Pi_{i\in U_F}(1-\pi_i)^{l_{ik}}.$$

Proposition 2.2.3. *Let us consider a sample s_F drawn in a population U_F, using a Poisson sampling design with inclusion probabilities $0 < \pi_i < 1$. Let U_T be another population associated to U_F through MtO links l_{ik}. The sample s_T deduced from s_F using the MtO links between U_F and U_T can be considered as drawn in U_T, using Poisson sampling with inclusion probabilities $\pi_k = 1 - \Pi_{i\in U_F}(1-\pi_i)^{l_{ik}}$. The variance of the direct HT estimator, $\hat{t}_y = \sum_{k\in s_T} y_k/\pi_k$,*

is smaller than the variance of the GWSM estimator $\hat{t}_{y1}^{opt} = \sum_{i \in s_F} \sum_{k \in U_T} \tilde{\theta}_{ik}^{opt} y_k / \pi_i$ *calculated with optimal link weights given in Equation* (2.7).

As already stated, for Poisson sampling and MtO links, the optimal link weights lead to the smallest possible variance of the simple GWSM estimator for any variable of interest. So, under the assumptions of Proposition 2.2.3, the simple indirect estimator is always less precise than the direct estimator.

The MtO case is interesting because it is possible, at least for Poisson sampling, to compare the direct and the indirect sampling designs. Moreover, for several standard sampling designs, it is possible to define optimal link weights and to calculate the exact loss of precision when using non-optimal link weights. However, when the number of units in the frame population linked with a unit in the target population is large, all of the links might be not observable, and an MtO indirect sampling could be very costly or even unfeasible. This problem arises for the La Poste survey in which the number of addresses per postman round is 500, on average, and where it is not possible to enumerate all addresses before the departure of the postman in the morning. One solution is to use a double indirect sampling design as detailed in the next section.

2.3 Double indirect sampling

2.3.1 Double GWSM

Double indirect sampling or indirect sampling in two steps (see Deville and Lavallée, 2006) consists of introducing an intermediate population U_M in between the frame and the target populations, and using the same principle as for simple indirect sampling. There could be various reasons for introducing such a population. One reason could be that the target population U_T units are only reachable through U_M. Another reason could be to simplify derivations. For example, Deville and Lavallée (2006) introduces an artificial intermediate population to simplify the search of an optimal standardized link matrix. In the La Poste survey, the objective is rather to decrease the number of links to observe. At La Poste, the intermediate population is a population of mail sorting boxes. Every morning, postmen sort the letters into boxes and deliver the letters from their allocated boxes (see Figure 3.2). The population of boxes is used as an intermediate population U_M to link the addresses from the frame population U_F to the postman rounds from the target population U_T.

Let N_M be the size of the intermediate population U_M. Let l_{ij}^{FM} be the link between $i \in U_F$ and $j \in U_M$ and let l_{jk}^{MT} be the link between $j \in U_M$ and $k \in U_T$. A unit i from the

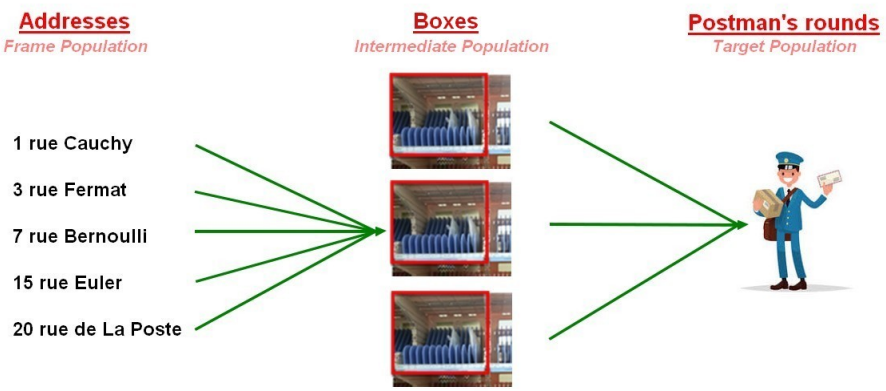

FIGURE 2.2 – The frame, intermediate and target populations at La Poste.

frame population U_F could be linked to a unit k from the target population U_T by means of the unit j from the intermediate population. The three populations U_F, U_M and U_T could be linked in various ways, as in simple indirect sampling with MtO, OtM and MtM links (see Figure 2.1).

As in Section 2.2, we consider non negative link weights θ_{ik} associated with the links l_{ik} between U_F and U_T such that θ_{ik} is positive when $l_{ik} = 1$ and $\theta_{ik} = 0$ otherwise. We consider also the non negative weights θ_{ij}^{FM} associated with the links l_{ij}^{FM} between U_F and U_M such that θ_{ij}^{FM} is positive when $l_{ij}^{FM} = 1$ and $\theta_{ij}^{FM} = 0$ otherwise. Finally, θ_{jk}^{MT} are non negative weights associated with the links between U_M and U_T and are defined in a similar way. With this double indirect sampling design, it can be seen that the links between units i from U_F and k from U_T are weighted by $\sum_{j \in U_M} \theta_{ij}^{FM} \theta_{jk}^{MT}$, and this link weight could be different from the link weight θ_{ik} used in simple indirect sampling.

Let $\tilde{\theta}_{ik}$ be the standardized link weight used in simple indirect sampling from U_F to U_T, namely $\sum_{i \in U_F} \tilde{\theta}_{ik} = 1$ for all $k \in U_T$. We denote by $\tilde{\theta}_{ij}^{FM}$, respectively $\tilde{\theta}_{jk}^{MT}$, the link weights between U_F and U_M, respectively U_M and U_T, such that the link weights used in double indirect sampling between the frame population U_F and the target population U_T are normalized, namely $\sum_{i \in U_F} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} = 1$ for all $k \in U_T$. Note that this standardization does not require us to standardize each set of links $\tilde{\theta}_{ij}^{FM}$ and $\tilde{\theta}_{jk}^{MT}$. However, to obtain these standardized link weights, for a given k in U_T, one needs to know the sum of the link weights of units from U_F linked to k passing by the intermediate population U_M, namely $\sum_{i \in U_F} \sum_{j \in U_M} \theta_{ij}^{FM} \theta_{jk}^{MT}$ should be known. The finite population total t_y of y can then be written as a total on the frame population U_F as follows :

$$t_y = \sum_{k \in U_T} y_k = \sum_{k \in U_T} \left(\sum_{i \in U_F} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} \right) y_k = \sum_{i \in U_F} \left(\sum_{k \in U_T} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} y_k \right).$$

An estimator of t_y can be derived easily by using the unbiased HT estimator as follows :

$$\hat{t}_{y2} = \sum_{i \in s_F} d_i \left(\sum_{k \in U_T} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} y_k \right). \tag{2.8}$$

We call $\hat{t}_{y2}$ the double GWSM estimator, and its variance is given by :

$$\mathrm{Var}(\hat{t}_{y2}) = \sum_{i,i' \in U_F} \Delta_{ii'} \sum_{k \in U_T} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} y_k \sum_{k' \in U_T} \sum_{j' \in U_M} \tilde{\theta}_{i'j'}^{FM} \tilde{\theta}_{j'k'}^{MT} y_{k'}.$$

2.3.2 MtO links

In this subsection, we focus on MtO links between the frame and the target populations. Comparing expressions (2.1) and (2.8), we can deduce that the double GWSM estimator $\hat{t}_{y2}$ can be viewed as a simple GWSM estimator with link weights $\tilde{\theta}_{ik} = \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT}$.

Assuming that the sampling design verifies the $\mathbf{\Delta}$-property, we can apply Proposition 2.2.2 to determine the differences of the variances between double and simple GWSM. Thus, Proposition 2.2.2 shows that, for any variable of interest y, the optimal simple GWSM estimator is always better than the double GWSM estimator. The loss of efficiency of the double GWSM estimator with respect to the optimal simple GWSM estimator depends on the variances of the HT estimators $\sum_{i \in s_F} d_i (\sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} - \tilde{\theta}_{ik}^{opt})$ for $k \in U_T$. This loss depends on the configuration of the link weights used in the double indirect sampling. If the double indirect sampling weights, $\sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT}$, are close to the optimal simple indirect sampling weights, $\tilde{\theta}_{ik}^{\mathrm{opt}}$, for all $i \in U_F$, then the use of a double GWSM estimator will cause a small loss in precision. Otherwise, the loss could be substantial.

In the following subsection, we describe a particular double indirect sampling design that requires fewer links than its simple indirect sampling counterpart while maintening the same precision.

2.3.3 MtO-MtO links and double standardization

Consider the MtO-MtO case where the links between the frame and the intermediate population are MtO, and the links between the intermediate and the target population are also MtO. This case implies that the links between U_F and U_T are MtO. In this situation, the double GWSM estimator has a simple expression, since a unit i from the frame population is linked to a single unit j from the intermediate population, which is itself linked to a single unit k from the target population. Thus, for a given sampled unit $i \in U_F$, the sums over the intermediate and the target populations in the double GWSM estimator $\hat{t}_{y2}$ given by (2.8),

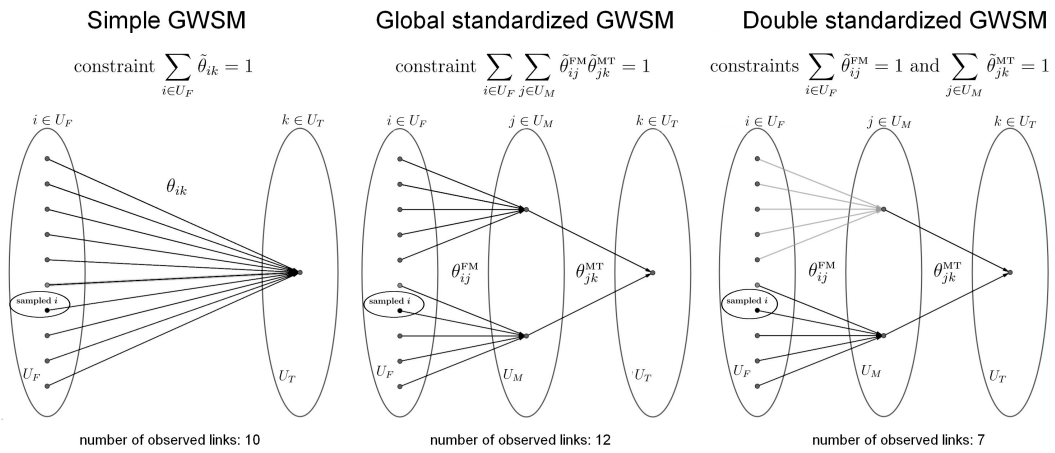

FIGURE 2.3 – Number of observed links for each standardization

contain only one non-zero element equal to $\tilde{\theta}^{FM}_{ij}\tilde{\theta}^{MT}_{jk}y_k$. To compute $\hat{t}_{y2}$, we need to compute only the standardized link weight $\tilde{\theta}^{FM}_{ij}\tilde{\theta}^{MT}_{jk}$ that corresponds to the unique unit $k \in U_T$ linked to the sampled unit $i \in U_F$ through the unique $j \in U_M$.

In this MtO-MtO setup, the choice of the standardization method arises and has an impact on the number of links to observe. One can consider link weights $\tilde{\theta}^{FM}_{ij}$ and $\tilde{\theta}^{MT}_{jk}$ that are either both standardized $\left(\sum_{i'\in U_F} \tilde{\theta}^{FM}_{i'j} = 1 \text{ and } \sum_{j'\in U_M} \tilde{\theta}^{MT}_{j'k} = 1\right)$, or globally standardized $\left(\sum_{i'\in U_F}\sum_{j'\in U_M} \tilde{\theta}^{FM}_{i'j'}\tilde{\theta}^{MT}_{j'k} = 1\right)$. Note that if both link weights are standardized (double standardization), then they are also globally standardized, but the converse is not true. As detailed below, using MtO-MtO with double standardization could allow for a reduction in the number of links to observe compared to the MtO-MtO with global standardization, or even compared to the simple MtO GWSM.

With double standardization, we can derive $\tilde{\theta}^{FM}_{ij}$ and $\tilde{\theta}^{MT}_{jk}$ separately. To compute $\tilde{\theta}^{FM}_{ij}$, the links between U_F and the units $j \in U_M$ indirectly sampled through $i \in s_F$ need to be observed, which gives the total number of links to observe equal to $N_{Fj} = \sum_{i'\in U_F} l^{FM}_{i'j}$. Similarly, to compute $\tilde{\theta}^{MT}_{jk}$, the links between U_M and the indirectly sampled $k \in U_T$ must be observed which leads to $N_{Mk} = \sum_{j'\in U_M} l^{MT}_{j'k}$ links to observe. Thus, for MtO-MtO GWSM with double standardization and for a given unit $i \in s_F$, the total number of links to observe is equal to $N_{Fj} + N_{Mk}$. The right plot of Fig. 2.3 shows the links to observe (black links) on a small example for MtO-MtO with double standardization.

For MtO-MtO with global standardization, we need to observe the links between U_F and the units $j \in U_M$ linked to the indirectly sampled units $k \in s_T$, namely we need to know $\sum_{i'\in U_F}\sum_{j'\in U_M} l^{FM}_{i'j'}l^{MT}_{j'k} = \sum_{j'\in U_M} N_{Fj'}l^{MT}_{j'k}$. We also need the number of links between U_M and the indirectly sampled $k \in U_T$, namely we need to know $N_{Mk} = \sum_{j'\in U_M} l^{MT}_{j'k}$. Thus, for MtO-MtO links with global standardization, we need to observe $\sum_{j'\in U_M} N_{Fj'}l^{MT}_{j'k} + N_{Mk}$

links. The middle plot of Fig. 2.3 shows the links to observe (black links), for a given unit $i \in s_F$, for MtO-MtO with global standardization.

Consider now simple MtO indirect sampling. A unit i from the frame population can be linked to a single unit k from the target population, and the sum over the target population in the simple GWSM estimator $\hat{t}_{y1}$ given by (2.1) contains only a non-zero element equal to $\tilde{\theta}_{ik} y_k$. To compute $\hat{t}_{y1}$, for each unit $i \in s_F$, we need to compute the standardized link weight $\tilde{\theta}_{ik}$ that corresponds to the unique unit $k \in U_T$ linked to $i \in s_F$. This circumstance implies that we need to observe the links between the indirectly sampled unit k of the target population and the frame population ; namely we need to observe a number of links equal to $N_{Fk} = \sum_{i' \in U_F} l_{i'k}$. The left plot of Fig. 2.3 shows the links to observe (black links) for a given unit $i \in s_F$ in MtO indirect sampling.

It is possible to compare the number of links between MtO-MtO with double standardization with simple MtO if we assume that $N_{Fk} = N_{Fj} N_{Mk}$, $N_{Fj} > 2$ and $N_{Mk} > 2$. We then have

$$N_{Fj} + N_{Mk} < N_{Fj} N_{Mk} = N_{Fk}, \tag{2.9}$$

and the double GWSM with double standardization always requires fewer links to observe than the simple GWSM. Moreover, the smallest number of links to observe with the double GWSM with double standardization is achieved when $N_{Fj} = N_{Mk} = N_{Fk}^{1/2}$, which is the most favorable situation for the double GWSM.

The assumption $N_{Fk} = N_{Fj} N_{Mk}$ is true if the numbers of links N_{Fj} are equal for all j linked to the same k in U_T. Indeed, if we let C_k denote a positive constant, that does not depend on j, such that $N_{Fj} = C_k$ for all units $j \in U_M$ linked to k in U_T, then,

$$N_{Fk} = \sum_{j \in U_M} l_{jk}^{MT} N_{Fj} = C_k \sum_{j \in U_M} l_{jk}^{MT} = C_k N_{Mk} = N_{Fj} N_{Mk}. \tag{2.10}$$

This remark is interesting as it stands in that when the number of links between U_F and U_M is the same for each unit in U_T, there will be fewer links to observe when using MtO-MtO with double standardization, regardless of what are the links between U_M and U_T (see Setups 1 and 3 in Section 2.4).

Furthermore, if the link weights are the link indicators, namely $\theta_{ij}^{FM} = l_{ij}^{FM}$, $\theta_{jk}^{MT} = l_{jk}^{MT}$, $\theta_{ik} = l_{ik}$, and $N_{Fk} = N_{Fj} N_{Mk}$, then the double GWSM and the simple GWSM estimators are equal. Indeed,

$$\sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} = \sum_{j \in U_M} \frac{l_{ij}^{FM}}{N_{Fj}} \frac{l_{jk}^{MT}}{N_{Mk}} = \frac{l_{ik}}{N_{Fk}} = \tilde{\theta}_{ik}. \tag{2.11}$$

In such a situation, the double GWSM with double standardization and the simple GWSM have the same precision, but the double GWSM ensures a gain in terms of the smaller number

of links to observe. However, if the condition $N_{Fk} = N_{Fj}N_{Mk}$ is not fulfilled, (see Setups 2 and 4 in Section 2.4 for further details), the double GWSM could be less efficient than the simple GWSM estimator.

For the double GWSM with global standardization, the number of links to observe is $\sum_{j' \in U_M} l^{MT}_{j'k} N_{Fj'} + N_{Mk} = N_{Fk} + N_{Mk}$, which is greater than the number of links to observe for both, the simple GWSM and the double GWSM with double standardization.

In Fig. 2.3, $N_{Fk} = 10$, $N_{Fj} = 5$ and $N_{Mk} = 2$. The number of links to observe is 10 for the simple GWSM (left plot of the figure), 12 for the double GWSM with global standardization (middle plot) and 7 for the double GWSM with double standardization (right plot). In the La Poste situation, where the double GWSM with double standardization is used, the gain is much larger because on average $N_{Fk} = 500$, $N_{Fj} = 10$ and $N_{Mk} = 50$. Thus, we have $N_{Fj} + N_{Mk} = 60$ links to observe for the double GWSM with double standardization while it is $N_{Fj}N_{Mk} = 500$ for the simple GWSM and $N_{Fk} + N_{Mk} = 550$ for the double GWSM with global standardization.

Double GWSM with global standardization is of poor interest in the context of La Poste. The loss of precision is not compensated by a gain in the number of links to observe. In what follows, we focus only on double GWSM with double standardization.

2.4 Simulation study

We have shown in Subsection 2.3.1 that, in the case of MtO links and sampling designs with the $\boldsymbol{\Delta}$-property, the loss of precision of the double GWSM compared to the optimal simple GWSM depends on the variance of the HT estimators $\sum_{i \in s_F} d_i (\sum_{j \in U_M} \tilde{\theta}^{FM}_{ij} \tilde{\theta}^{MT}_{jk} - \tilde{\theta}^{opt}_{ik})$ for all $k \in U_T$. In other words, the increase in the variance depends on the configuration of the double GWSM link weights $\sum_{j \in U_M} \tilde{\theta}^{FM}_{ij} \tilde{\theta}^{MT}_{jk}$, compared to the optimal simple link weights $\tilde{\theta}^{opt}_{ik}$. In this section, we conduct a Monte Carlo study to analyze the influence of the link weights on the efficiency of the double GWSM estimator.

2.4.1 Population and link setups

We have generated three populations, U_F, U_M and U_T as well as the links between them, to meet the framework assumed in Subsection 2.3.1 as well as at the La Poste situation. At La Poste, the target population U_T is the population of rounds, the frame population U_F is the population of addresses, and the intermediate population U_M is the population of boxes. All links are MtO, which means that an address can only be in one box and a box in only one round. There are on average 50 boxes per round and 500 addresses per round. The three

populations U_F, U_M and U_T were generated in the following way. The target population of rounds U_T of size $N_T = 6\,958$ was obtained from La Poste's historical data. Then, based on U_T, we generated the intermediate population of boxes U_M of size $N_M = 347\,900 = 50 \times 6\,958$, and the frame population of addresses U_F, of size $U_F = 3\,479\,000 = 500 \times 6\,958$, together with the links.

We are interested in estimating the total $t_y = \sum_{k \in U_T} y_k$ of a study variable y which is a particular measure of the postal traffic obtained from La Poste's historical data. For confidentiality reasons, the study variable y was transformed.

We considered different setups of MtO links between U_F and U_M and between U_M and U_T. For ease of comparison, we kept unchanged the links between U_F and U_T in all scenarios. In this double indirect sampling design, additional care must be taken to ensure that, if unit $i \in U_F$ is linked to $k \in U_T$ in simple indirect sampling case, then it must be linked to $k \in U_T$ for double indirect sampling also. There is a link between U_F (resp. U_F, U_M) and U_T (resp. U_M, U_T) when an address (resp. an address, a box) is part of a round (resp. a box, a round). The first set of links was generated between U_F and U_T, and is such that the number of addresses N_{Fk} in a round k of U_T is equal to 500, for all rounds $k \in U_T$. Thus, $N_{Fk} = \sum_{j \in U_M} l_{jk}^{MT} N_{Fj} = 500$, where N_{Fj} is the number of addresses in the box j. This configuration means that all units from the target population have the same number of links with the frame population. As detailed in Subsection 2.4.2, the advantage of this simplification is that, under some supplementary assumptions on the populations and on the inclusion probabilities that are true in our setting, the simple GWSM estimator has the same expression as a direct HT estimator. To study the effect of the links structure on the efficiency of the double GWSM estimator, we created four different setups of links with the intermediate population. The links between U_F and U_M (resp. U_M and U_T) were generated either uniformly or not. The links between U_F and the boxes j of U_M are called "uniform" when there is the same number of addresses, N_{Fj}, in all boxes j of U_M that are part of the same round k of U_T. They are called "non-uniform" when we generate one address in all boxes of U_M that are part of the same round of U_T, except for one box j_0 which contains the remaining addresses. This last situation is a type of extreme unbalanced case for the number of links between U_F and U_M. The links between U_M and U_T are "uniform" when there is the same number of boxes N_{Mk} per round. The "non-uniform" case is generated by considering two boxes for 6 286 rounds and 499 boxes for the remaining 672 rounds. Note that we cannot choose non-uniform links between U_M and U_T with one box or 500 boxes in a round together with non-uniform links between U_F and U_M similar to the ones proposed above. The reason is that we have set a constraint of 500 addresses per round. Under this constraint, having one box (resp. 500 boxes) in a round implies having 500 addresses (resp. one address) in each

box of the round, which corresponds to uniform links between the addresses and the boxes.

The four setups are detailed below (see also Figure 2.4 for graphical examples) :

— *Setup 1 : Uniform/Uniform* : the number of links between U_F (resp. U_M) and U_M (resp. U_T) is uniform. In this setup, the double GWSM and the simple GWSM estimators are equal, as proved in Subsection 2.3.3 (see (2.11)).

— *Setup 2 : non-uniform/Uniform* : the links between U_F and U_M are non-uniform while the links between U_M and U_T are uniform.

— *Setup 3 : Uniform/non-uniform* : the links between U_F and U_M are uniform while the links between U_M and U_T are non-uniform. In this setup, the N_{Fj} must be rounded as illustrated on Figure 2.4 (bottom left panel) where, for the second unit in U_T, there are 6 addresses to divide between 4 boxes, and thus, 2 boxes contain 1 address each, while the other 2 boxes contain 2 addresses each. Ignoring the rounding of the N_{Fj}, the relation $N_{Fk} = N_{Fj}N_{Mk}$ holds (see also equation (2.10)), which allows for the equality of the simple and double GWSM estimators.

— *Setup 4 : non-uniform/non-uniform* : the links between U_F and U_M, and between U_M and U_T are non-uniform.

2.4.2 Sampling designs and GWSM estimators

We consider two sampling designs that satisfy the $\mathbf{\Delta}$-property : the SRSWOR of sizes $n = 500$ and $n = 1\,000$, as well as the Bernoulli design (which is a Poisson design with equal inclusion probabilities) with expected sample sizes equal to 500 and 1 000.

Let y_k, $k \in U_T$, be a measure of the postal traffic in round k, as mentioned at the beginning of section 2.4.1. For each link setup, we compare the double GWSM estimator to the simple GWSM estimator, both computed on samples s_F drawn from the frame population U_F of addresses :

$$\hat{t}_{y1} = \sum_{i \in s_F} \frac{1}{\pi_i} \sum_{k \in U_T} \tilde{\theta}_{ik} y_k, \quad \hat{t}_{y2} = \sum_{i \in s_F} \frac{1}{\pi_i} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \sum_{k \in U_T} \tilde{\theta}_{jk}^{MT} y_k.$$

For the simple GWSM estimator, we consider the optimal link weights, which are equal to $\tilde{\theta}_{ik} = l_{ik}/N_{Fk}$ for SRSWOR and Bernoulli samplings. For the double GWSM, we consider $\tilde{\theta}_{ij}^{FM} = l_{ij}^{FM}/N_{Fk}$ and $\tilde{\theta}_{jk}^{MT} = l_{jk}^{MT}/N_{Mk}$.

The above simulation setting facilitates the comparison between the double GWSM estimator and the simple GWSM estimator by ensuring that only the double GWSM varies, while the simple GWSM remains fixed. In fact, the setting makes the simple GWSM very close to the direct HT estimator in the 4 setups. We compare also the double GWSM estimator to the direct HT estimator as calculated from samples s_T^* drawn directly from the

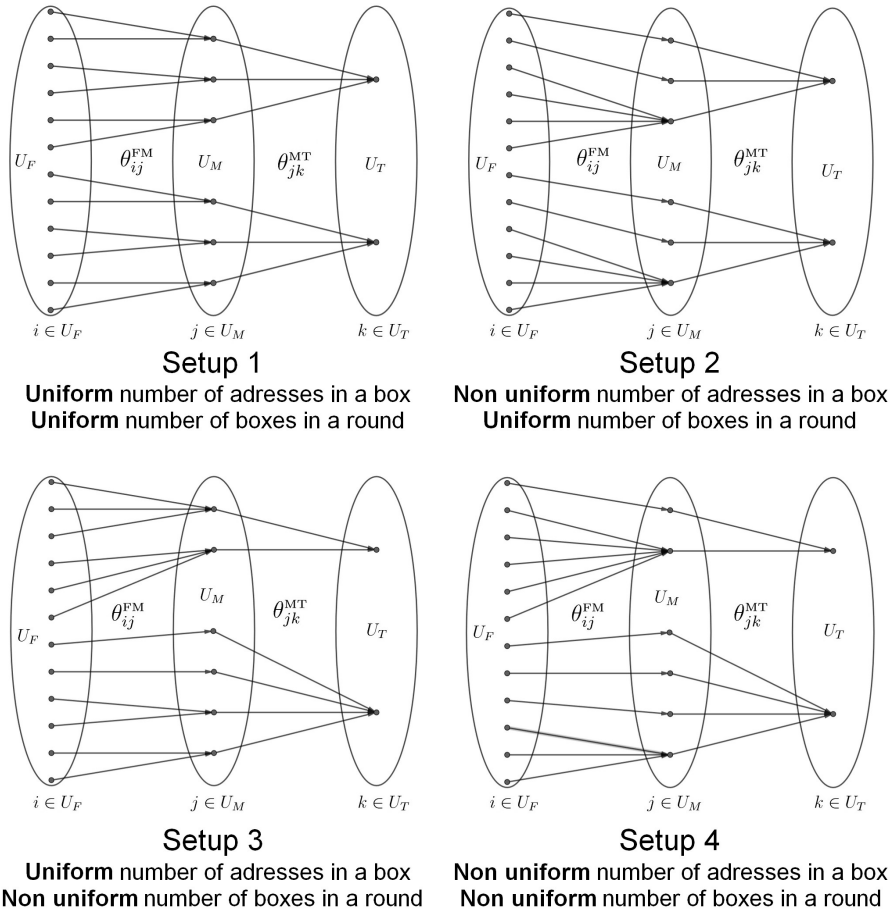

FIGURE 2.4 – The four setups

target population U_T of rounds using SRSWOR of size n_T :

$$\hat{t}_{HT} = \sum_{k \in s_T^*} \frac{y_k}{\pi_k}.$$

If the π_i are small, the probability of drawing two or more units from U_F that are linked to the same k in U_T is small and we can approximate π_k by $N_{Fk}\pi_i$. In this situation, every unit from the target population is generally linked with at most one unit from the sample s_F, and thus, for all $k \in s_T$, we have $\sum_{i \in s_F} l_{ik} \simeq 1$. This hypothesis is true at La Poste because given the large number of postal addresses in the frame population, we never sample several addresses from the same round. Thus, we have :

$$\begin{aligned}\hat{t}_{y1} &= \sum_{i \in s_F} \frac{1}{\pi_i} \sum_{k \in U_T} \tilde{\theta}_{ik} y_k = \sum_{k \in U_T} y_k \sum_{i \in s_F} \frac{1}{\pi_i} \tilde{\theta}_{ik} \\ &\simeq \sum_{k \in s_T} y_k \frac{N_{Fk}}{\pi_k} \frac{\sum_{i \in s_F} l_{ik}}{N_{Fk}} \simeq \sum_{k \in s_T} \frac{y_k}{\pi_k}\end{aligned}$$

which means that, in our setting, the simple GWSM estimator is approximately equivalent to the direct HT estimator for the sample s_T. Since this formula does not depend on the simulation setup, the simple GWSM does not vary between the setups.

2.4.3 Measures of efficiency and results

We have considered $R = 100\,000$ samples according to the SRSWOR and Bernoulli sampling designs with sizes or expected sizes 500 and 1 000. We have computed the Monte Carlo relative bias of the simple GWSM and the double GWSM estimators :

$$RB_{MC}(\hat{t}_{y1}) = 100 \times \frac{1}{R} \sum_{r=1}^{R} \frac{\hat{t}_{y1}^{(r)} - t_y}{t_y} \quad \text{and} \quad RB_{MC}(\hat{t}_{y2}) = 100 \times \frac{1}{R} \sum_{r=1}^{R} \frac{\hat{t}_{y2}^{(r)} - t_y}{t_y},$$

where $\hat{t}_{y1}^{(r)}$ and $\hat{t}_{y2}^{(r)}$ are the simple and the double GWSM estimates respectively, computed for the r-th sample, $r = 1, \ldots, R$. As a measure of efficiency, we have computed the Relative Root Mean Square Error (RRMSE) of $\hat{t}_{y2}$ with respect to $\hat{t}_{y1}$, and with respect to $\hat{t}_{HT}$:

$$RRMSE_{MC}(\hat{t}_{y1}) = \sqrt{\frac{MSE_{MC}(\hat{t}_{y2})}{MSE_{MC}(\hat{t}_{y1})}} \quad \text{and} \quad RRMSE_{MC}(\hat{t}_{HT}) = \sqrt{\frac{MSE_{MC}(\hat{t}_{y2})}{MSE_{MC}(\hat{t}_{HT})}}$$

where

$$MSE_{MC}(\hat{t}_{y2}) = R^{-1} \sum_{r=1}^{R} \left(\hat{t}_{y2}^{(r)} - R^{-1} \sum_{r=1}^{R} \hat{t}_{y2}^{(r)} \right)^2,$$

Design	Setup	n	$RB_{MC}(\hat{t}_{y2})$	$RRMSE_{MC}(\hat{t}_{y1})$	$RRMSE_{MC}(\hat{t}_{HT})$
SRSWOR 100 000 simulations	Setup 1	500	0.15	100.00	103.96
		1000	0.09	100.00	109.78
	Setup 2	500	-0.53	337.89	347.14
		1000	0.44	334.51	365.85
	Setup 3	500	0.07	100.01	105.35
		1000	0.12	100.02	106.12
	Setup 4	500	2.67	1173.36	1216.95
		1000	0.33	1093.90	1198.97

TABLE 2.1 – Relative bias and efficiency of the DGWSM estimate under different links setups.

and $MSE_{MC}(\hat{t}_{y1})$, and $MSE_{MC}(\hat{t}_{HT})$ are defined similarly.

Table 2.1 contains the simulation results for the SRSWOR design. Similar results were obtained for Bernoulli sampling and are not reported here. The results are also very comparable for both sampling sizes. As expected, both GWSM estimators have a low Monte Carlo relative bias in all setups.

Moreover, we observe small differences between the RRMSE of the double compared to the simple GWSM, and the double GWSM compared to the direct HT estimators in all setups. This result was expected since the simple GWSM expression corresponds to a direct HT expression (see details at the end of Subsection 2.4.2). Small differences arise because the samples s_T^*, drawn directly in the target population, differ from the samples s_T that are obtained through the samples s_F using the links between U_F and U_T.

For Setups 1 and 3, as shown in Subsection 2.3.3, $\hat{t}_{y1} = \hat{t}_{y2}$. Thus, $RRMSE_{MC}(\hat{t}_{y1}) = 100\%$ and $MSE_{MC}(\hat{t}_{y1}) = MSE_{MC}(\hat{t}_{y2})$. The small loss of precision between the GWSM estimators in Setup 3 occurs because the relation $N_{Fk} = N_{Fj}N_{Mk}$ is not exactly satisfied due to rounding errors.

For Setups 2 and 4, the equation $N_{Fk} = N_{Fj}N_{Mk}$ does not hold at all. In Setup 2, if the sampled address i is alone in the box j, then $N_{Fj}N_{Mk} = 1 * 50$ which is far from $N_{Fk} = 500$. If i is in the box containing 451 addresses, then $N_{Fj}N_{Mk} = 451 * 50$ which is also far from N_{Fk}. In Setup 4, the difference between N_{Fk} and $N_{Fj}N_{Mk}$ is even larger because we also let the N_{Mk} vary. We note an important loss of precision of the double GWSM estimator compared to the simple GWSM estimator in Setups 2 and 4.

The precision of the double GWSM estimator depends on how close the values $N_{Fj}N_{Mk}$ and N_{Fk} are, for every j linked to the same k. As proved in equation (2.10), the uniform link structure between U_F and U_M implies that $N_{Fk} = N_{Fj}N_{Mk}$, regardless of the link structure between U_M and U_T. This remark helps to explain the good results for Setups 1 and 3 and

the poor results for Setups 2 and 4.

It is also interesting to compare the number of links to observe for the two GWSM estimators in each setup. The simple GWSM estimator requires the observation of 500 links per sampled round in each setup. To compute the double GWSM estimator, there are 60 links to observe on average per sampled round in Setup 1, 457 in Setup 2, 276 in Setup 3 and 500 in Setup 4 (the averages are rounded values). For Setups 1 and 3, the equation $N_{Fk} = N_{Fj}N_{Mk}$ (almost) holds, and thus, there will always be a gain in the number of links to observe (see equation (2.9)). It can be noted that this gain is even larger if the N_{Mk} are uniform. For Setup 2, the gain in the number of links is limited, while there is no gain in Setup 4.

The simulations illustrate that the link structure between the three populations has a large impact on the double GWSM estimator in terms of the precision, but also in terms of the number of links to observe. In the ideal situation of Setup 1, there is a clear advantage of using double indirect sampling for an MtO-MtO situation, while it is not at all recommended in situations like the one illustrated in Setup 4.

2.5 Application to the French post data

Before 2008, La Poste sampled directly the postmen rounds to estimate the monthly postal traffic. After a reorganization of the post offices in 2008, the population of rounds became incomplete and La Poste had to use indirect sampling through the frame population of addresses. Because of the large number of links to observe, a simple GWSM estimator was not possible. La Poste had to consider a double MtO-MtO indirect sampling design and a double GWSM estimator with double standardization. The use of double indirect sampling, compared to the previous direct sampling method, led to a precision loss of the estimators. The estimated standard deviations of the estimators were increased by a factor between 2 and 3. To complete the theoretical results of Sections 2.2 and 2.3, we propose, through simulations and in a setup similar to La Poste, to evaluate the loss of precision due to using double indirect sampling, and to check if the calculated loss is of the same order as the loss observed in reality.

In this application, we focus on simple GWSM and double GWSM, which are both computed on a sample of addresses, and a direct HT estimator computed on a sample of rounds. The samples are drawn according to SRSWOR designs. The sampling design at La Poste is more complex and involves a stratification based on a typology of the post offices. This stratification of the postal addresses ensures that a round cannot belong to two different strata of addresses, and that every stratum contains at least two rounds. Thus,

Design	n	GWSM	RB $y=1$	RRMSE $y=1$ rel. to S	RB y =traffic	RRMSE y =traffic rel. to S	RRMSE y =traffic rel. to direct
SRSWOR	500	S	-0.01	100	0.02	100	137.96
100 000	500	D	-0.03	219.00	0.08	166.69	229.98
simulations	1000	S	0.00	100	-0.01	100	144.55
	1000	D	0.00	220.89	-0.07	163.53	236.39

TABLE 2.2 – Relative bias and comparison of RMSE, in percentages, for the double (D) GWSM, the simple (S) GWSM and the direct estimates in a setup comparable to La Poste.

the $\boldsymbol{\Delta}$-property holds for this sampling design (see details in Subsection 2.2.2 on stratified SRSWOR). The estimators are also more complex and involve calibration and winzorisation. Considering such complex designs and estimators is beyond the scope of the present study, and we focus on the SRSWOR sampling design with direct HT, and with simple and double indirect GWSM estimators.

We do not only look at the loss caused by the use of a double GWSM compared to a simple GWSM. We also examine the loss caused by the use of double indirect sampling compared to a direct sampling. The objective is to capture the total loss in the precision observed at La Poste when changing their sampling design from direct to doubly indirect.

The setup for the simulations below is close to the La Poste setup in the sense that the number of addresses in a box, and the number of boxes in a round were generated using observed distributions from La Poste data. Compared to the four setups in Section 2.4, the number of addresses in a round was not fixed at 500, but computed using the number of addresses in a box and the number of boxes in a round. The number of addresses in a box varies from 1 to 29 with two modes at 1 and 13, with rare observations at approximately 40 and 120. The number of boxes in a round varies from 28 to 73 with two modes at approximately 35 and 70, and rare observations between 100 and 1000. The average number of addresses in a box in this setup is 14, the average number of boxes in a round is 60 and the average number of addresses in a round is 841. This setup is close to Setup 4 in Section 2.4, but it has less variability in the number of links, with the number of addresses between 30 and 70 and the number of boxes between 1 and 29, while the number of addresses and boxes varies between 2 and 499 in Setup 4.

We consider two study variables y, and we are interested in estimating their totals on the target population of rounds. The first study variable is equal to 1 for all units, which gives a total over the target population equal to N_T, while the second variable is a confidential measure of postal traffic obtained from La Poste data. For the selection of indirect sample, we consider simple random sampling without replacement with respective sizes of 500 and

1000 selected from the frame population of addresses. For the selection of direct sample, We also consider a simple random sampling without replacement with respective sizes 500 and 1000 selected in the target population of rounds.

As in Section 2.4, we compute, for $R = 100\,000$ simulations, the Monte Carlo relative bias (RB) as a percentage of the simple (S) and double (D) GWSM estimators together with their mean square errors. For both variables ($y = 1$ and y =traffic), we use the RRMSE in percentage, to compare the double to the simple GWSM (see the two RRMSE columns relative to S in Table 2.2). For y =traffic, we also compare the double GWSM to the direct estimator (see the RRMSE relative to direct in the last column of Table 2.2). We note that, for $y = 1$, the MSE of the direct estimator is zero since the estimator is calibrated on the size of the population. Thus, the RRMSE of the double GWSM compared to the direct is infinite and not reported. We notice that there is almost no difference between the results obtained for the two sample sizes. As expected, given that all estimators are unbiased, the relative biases of the simple and double GWSM estimators in Table 2.2 are small for both variables of interest. Moreover, we observe a loss in precision by using the double GWSM estimator instead of the simple GWSM estimator for both variables. This loss is less important than those observed in Setups 2 and 4, since the differences in the weights between the double GWSM and the simple GWSM are smaller here and have less variability than in Setups 2 and 4. In Table 2.1 (see the last two columns), for Setup 2 (resp. Setup 4), the standard deviations are multiplied by a factor between 3 and 4 (resp. 10 and 13). In Table 2.2, the standard deviations are multiplied by a factor between 1 and 2 (resp. 2 and 3) for the "traffic" (resp. 1) variable when comparing D and S. The additional loss of precision between the S and the direct HT estimators is not negligible, and gives a factor between 1 and 2 for the "traffic"variable (1.38 for the sample size n_F=500 and 1.45 for $n_F = 1000$). In total, the standard deviations increase by a factor between 2 and 3 (approximately 2.3) when changing from direct to double indirect sampling. Interestingly, the loss of precision that we observe is of the same order as the loss observed in practice at La Poste.

2.6 Conclusion

The MtO situation in indirect sampling allows us to obtain optimal link weights for some classical sampling designs such as simple random sampling without replacement, Poisson sampling and stratified SRSWOR. In this context, it is also possible to derive an exact expression for the loss of precision when the link weights are not optimal. This expression shows that the increase of variance of the GWSM estimator depends on how far the link weights estimators are from the optimal link weights estimators. When the number of links

to observe is large, it is possible to introduce a double indirect sampling design that allows us to reduce the number of links to observe when using a double standardization. As illustrated by our simulations, the double GWSM with double standardization proves to be especially interesting in some specific cases. It allows for a reduction in the number of observed links while maintaining the level of precision of a simple GWSM. However, it can be less useful in other cases, with a considerable loss of precision and not an important reduction of the number of links to observe. In the La Poste situation, there is a clear reduction in the number of addresses to observe per round, but at the cost of a large loss of precision. One perspective to improve on the precision of the estimators at La Poste is to keep the double indirect sampling design but use simple indirect GWSM and to predict the unobserved link weights as proposed for example by Xu and Lavallée (2009) and Falorsi et al. (2019). Indeed, double indirect sampling helps on saving costs and should be maintained. However, the use of the double GWSM with double standardization may lead to a significant loss of precision. Thus, a perspective that is currently under study at La Poste is to predict the number of unobserved links that are needed for the simple GWSM by using a model along with auxiliary information available at the level of the intermediate population of boxes.

2.7 Appendix

Proof of Proposition 2.2.1

We compute first the variance of the simple GWSM assuming MtO links and a sampling design that satisfy the $\boldsymbol{\Delta}$-property. We follow Deville and Lavallée (2006) and use matrix notations. Let $\tilde{\boldsymbol{\Theta}}_k = (\tilde{\theta}_{ik})_{i \in U_F}$ be the N_F-dimensional vector of standardized link weights. With MtO links, we can write $\tilde{\boldsymbol{\Theta}}_k$ as follows :

$$\tilde{\boldsymbol{\Theta}}_k = \begin{pmatrix} \mathbf{0} \\ \tilde{\boldsymbol{\theta}}_k \\ \mathbf{0} \end{pmatrix}, \quad k \in U_T, \tag{2.12}$$

where $\tilde{\boldsymbol{\theta}}_k = (\tilde{\theta}_{ik})_{i \in U_{Fk}}$ is the N_{Fk}-dimensional vector of positive weighted links with $N_{Fk} = \sum_{i \in U_F} l_{ik}$, the number of units i in U_F linked to k in U_T. The variance of the GWSM estimator

$\hat{t}_{y1}$ given in (2.2) can be written as follows :

$$\begin{aligned}\mathrm{Var}(\hat{t}_{y1}) &= \sum_{k\in U_T}\sum_{k'\in U_T} y_k y_{k'}\mathrm{Cov}(\hat{t}_{\tilde{\theta}_k},\hat{t}_{\tilde{\theta}_{k'}}) \\ &= \sum_{k\in U_T}\sum_{k'\in U_T} y_k y_{k'}\tilde{\boldsymbol{\Theta}}_k^t\boldsymbol{\Delta}\tilde{\boldsymbol{\Theta}}_{k'} \\ &= \sum_{k\in U_T}\sum_{k'\in U_T} y_k y_{k'}\tilde{\boldsymbol{\theta}}_k^t\boldsymbol{\Delta}_{kk'}\tilde{\boldsymbol{\theta}}_{k'}, \end{aligned} \tag{2.13}$$

where $\boldsymbol{\Delta} = \left(\frac{\pi_{ii'}-\pi_i\pi_{i'}}{\pi_i\pi_{i'}}\right)_{i,i'\in U_F}$ and $\boldsymbol{\Delta}_{kk'} = (\Delta_{ii'})_{i \text{ linked to } k, i' \text{ linked to } k'}$ is the submatrix of $\boldsymbol{\Delta}$ of size $N_{Fk}\times N_{Fk'}$ corresponding to elements in positions i and i' such that i is linked to k and i' to k'.

The variance of $\hat{t}_{y1}$ is given in (2.13), and the $\boldsymbol{\Delta}$ property states that, for any units k and k' in U_T, we have $\boldsymbol{\Delta}_{kk'} = c_{kk'}\mathbb{1}_k^t\mathbb{1}_{k'}$. By using also the fact that, for all $k\in U_T$, $\tilde{\boldsymbol{\theta}}_k$ satisfies the standardization constraint :

$$\mathbb{1}_k^t\tilde{\boldsymbol{\theta}}_k = 1 \quad \text{for all} \quad k\in U_T, \tag{2.14}$$

where $\mathbb{1}_k$ is the N_{Fk}-dimensional vector of ones, we have :

$$\begin{aligned}\mathrm{Var}(\hat{t}_{y1}) &= \sum_{k\in U_T} y_k^2\tilde{\boldsymbol{\theta}}_k^t\boldsymbol{\Delta}_k\tilde{\boldsymbol{\theta}}_k + \sum_{k\in U_T}\sum_{k'\neq k\in U_T} y_k y_{k'}\tilde{\boldsymbol{\theta}}_k^t\boldsymbol{\Delta}_{kk'}\tilde{\boldsymbol{\theta}}_{k'} \\ &= \sum_{k\in U_T} y_k^2\tilde{\boldsymbol{\theta}}_k^t\boldsymbol{\Delta}_k\tilde{\boldsymbol{\theta}}_k + \sum_{k\in U_T}\sum_{k'\neq k\in U_T} y_k y_{k'}c_{kk'}\tilde{\boldsymbol{\theta}}_k^t\mathbb{1}_k\mathbb{1}_{k'}^t\tilde{\boldsymbol{\theta}}_{k'} \\ &= \sum_{k\in U_T} y_k^2\mathrm{Var}(\hat{t}_{\tilde{\theta}_k}) + \sum_{k\in U_T}\sum_{k'\neq k\in U_T} y_k y_{k'}c_{kk'}. \end{aligned} \tag{2.15}$$

This is exactly (2.6) and finishes the proof of the first part of Proposition 2.2.1.

For the derivation of the optimal link weights, the proof is similar to the proof given in Deville and Lavallée (2006) (section 6.2) for deriving the optimal weighted links. In our situation, the links between the frame population U_F and the target population U_T are of type MtO. Thus, we don't need to use a factorization step as in Deville and Lavallée (2006).

Our aim is to find the link weights $\tilde{\boldsymbol{\theta}}_k^{opt}, k\in U_T$ that minimize

$$\mathrm{Var}(\hat{t}_{y1}) = \sum_{k\in U_T} y_k^2\tilde{\boldsymbol{\theta}}_k^t\boldsymbol{\Delta}_k\tilde{\boldsymbol{\theta}}_k + \sum_{k\in U_T}\sum_{k'\neq k\in U_T} y_k y_{k'}c_{kk'}$$

under the standardization constraint (2.14). Since Δ_k, $k\in U_T$ is invertible, this problem is the minimisation of a positive quadratic form under linear constraint. Thus a solution exists

and is unique. The variance $\text{Var}(\hat{t}_{y1})$ is minimized for vectors $\tilde{\boldsymbol{\theta}}_k$ verifying the following equation (see Deville and Lavallée, 2006, equation 6.4) :

$$y_k^2 \boldsymbol{\Delta}_k \tilde{\boldsymbol{\theta}}_k = \lambda_k \mathbb{1}_k, \quad k \in U_T, \tag{2.16}$$

where $\lambda_k, k \in U_T$ the Lagrange multipliers. Let us show that the optimal weights are given by :

$$\tilde{\boldsymbol{\theta}}_k^{opt} = \left(\mathbb{1}_k^t \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k\right)^{-1} \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k, \quad k \in U_T,$$

where $\boldsymbol{\Delta}_k = \boldsymbol{\Delta}_{kk}$. Equation (2.16) implies

$$\tilde{\boldsymbol{\theta}}_k \;=\; y_k^{-2} \boldsymbol{\Delta}_k^{-1} \lambda_k \mathbb{1}_k, \quad k \in U_T. \tag{2.17}$$

Multiplying by $\mathbb{1}_k^t$ the equation (2.17) and using again the standardization constraints (2.14), we obtain the following expression for the Lagrange multipliers :

$$\lambda_k \;=\; y_k^2 \left(\mathbb{1}_k^t \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k\right)^{-1}, \quad k \in U_T. \tag{2.18}$$

Finally, by plugging the expression of λ_k given in (2.18) in the expression of $\tilde{\boldsymbol{\theta}}_k$ from (2.17), we obtain :

$$\begin{aligned}
\tilde{\boldsymbol{\theta}}_k^{opt} &= y_k^{-2} \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k \lambda_k \\
&= y_k^{-2} \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k y_k^2 \left(\mathbb{1}_k^t \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k\right)^{-1} \\
&= \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k \left(\mathbb{1}_k^t \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k\right)^{-1}.
\end{aligned}$$

This finishes the proof of 2.2.1. □

Proof of Proposition 2.2.2

Result (2.15) holds for any standardized set of weights. Thus, the result also holds for the set of optimal weights $\tilde{\theta}_{ik}^{opt}$, and we can write :

$$\text{Var}(\hat{t}_{y1}^{opt}) = \sum_{k \in U_T} y_k^2 \text{Var}(\hat{t}_{\tilde{\theta}_k^{opt}}) + \sum_{k \in U_T} \sum_{k' \neq k \in U_T} y_k y_{k'} c_{kk'}. \tag{2.19}$$

Using equations (2.15) and (2.19), we get :

$$\text{Var}(\hat{t}_{y1}) - \text{Var}(\hat{t}_{y1}^{opt}) \;= \sum_{k \in U_T} y_k^2 \left(\text{Var}(\hat{t}_{\tilde{\theta}_k}) - \text{Var}(\hat{t}_{\tilde{\theta}_k^{opt}})\right).$$

Let us show that for the optimal weights derived in Proposition 2.2.1, we have :

$$\mathrm{Var}(\hat{t}_{\tilde{\theta}_k}) - \mathrm{Var}(\hat{t}_{\tilde{\theta}_k^{opt}}) = \mathrm{Var}(\hat{t}_{\tilde{\theta}_k} - \hat{t}_{\tilde{\theta}_k^{opt}}).$$

The optimal set of weights is given by $\tilde{\boldsymbol{\theta}}_k^{opt} = \boldsymbol{\Delta}_k^{-1}\mathbb{1}_k \left(\mathbb{1}_k^t \boldsymbol{\Delta}_k^{-1}\mathbb{1}_k\right)^{-1}, k \in U_T$.

$$\begin{aligned}\mathrm{Var}(\hat{t}_{\tilde{\theta}_k}) - \mathrm{Var}(\hat{t}_{\tilde{\theta}_k^{opt}}) &= \tilde{\boldsymbol{\theta}}_k^t \boldsymbol{\Delta}_k \tilde{\boldsymbol{\theta}}_k - (\tilde{\boldsymbol{\theta}}_k^{opt})^t \boldsymbol{\Delta}_k \tilde{\boldsymbol{\theta}}_k^{opt} \\ &= \left(\tilde{\boldsymbol{\theta}}_k - \tilde{\boldsymbol{\theta}}_k^{opt}\right)^t \boldsymbol{\Delta}_k \left(\tilde{\boldsymbol{\theta}}_k - \tilde{\boldsymbol{\theta}}_k^{opt}\right) + \tilde{\boldsymbol{\theta}}_k^t \Delta_k \tilde{\boldsymbol{\theta}}_k^{opt} + (\tilde{\boldsymbol{\theta}}_k^{opt})^t \boldsymbol{\Delta}_k \tilde{\boldsymbol{\theta}}_k \\ &\quad -2(\tilde{\boldsymbol{\theta}}_k^{opt})^t \boldsymbol{\Delta}_k \tilde{\boldsymbol{\theta}}_k^{opt}.\end{aligned}$$

Now, $\tilde{\boldsymbol{\theta}}_k^t \boldsymbol{\Delta}_k \tilde{\boldsymbol{\theta}}_k^{opt} = (\tilde{\boldsymbol{\theta}}_k^{opt})^t \boldsymbol{\Delta}_k \tilde{\boldsymbol{\theta}}_k$ since they are real quantities. Straightforward calculations using the standardization constraint (2.14), give us that $\tilde{\boldsymbol{\theta}}_k^t \boldsymbol{\Delta}_k \tilde{\boldsymbol{\theta}}_k^{opt} = \left(\mathbb{1}_k^t \boldsymbol{\Delta}_k^{-1}\mathbb{1}_k\right)^{-1}$. Moreover, $(\tilde{\boldsymbol{\theta}}_k^{opt})^t \boldsymbol{\Delta}_k \tilde{\boldsymbol{\theta}}_k^{opt} = \left(\mathbb{1}_k^t \boldsymbol{\Delta}_k^{-1}\mathbb{1}_k\right)^{-1}$, and we finally get that

$$\begin{aligned}\mathrm{Var}(\hat{t}_{\tilde{\theta}_k}) - \mathrm{Var}(\hat{t}_{\tilde{\theta}_k^{opt}}) &= \left(\tilde{\boldsymbol{\theta}}_k - \tilde{\boldsymbol{\theta}}_k^{opt}\right)^t \boldsymbol{\Delta}_k \left(\tilde{\boldsymbol{\theta}}_k - \tilde{\boldsymbol{\theta}}_k^{opt}\right) \\ &= \mathrm{Var}(\hat{t}_{\tilde{\theta}_k} - \hat{t}_{\tilde{\theta}_k^{opt}}) = \mathrm{Var}(\hat{t}_{\tilde{\theta}_k - \tilde{\theta}_k^{opt}})\end{aligned}$$

which ends the proof of Proposition 2.2.2. □

Proof of Proposition 2.2.3

For all $i \in U_F$, we recall that we assume $0 < \pi_i < 1$, and $\pi_k = 1 - \Pi_{i \in U_F}(1-\pi_i)^{l_{ik}}$. For Poisson sampling,

$$\tilde{\theta}_{ik}^{opt} = \frac{l_{ik}\frac{\pi_i}{1-\pi_i}}{\sum_{i' \in U_F} l_{i'k}\frac{\pi_{i'}}{1-\pi_{i'}}}.$$

Using the variance expressions for Poisson sampling, we have that proving $\mathrm{Var}(\hat{t}_y) < \mathrm{Var}(\hat{t}_{y1}^{opt})$ is equivalent to prove

$$\sum_{k \in U_T} y_k^2 \frac{1-\pi_k}{\pi_k} < \sum_{k \in U_T} y_k^2 \sum_{i \in U_F} \frac{1-\pi_i}{\pi_i}(\tilde{\theta}_{ik}^{opt})^2.$$

This inequality is true if, for all $k \in U_T$, we have :

$$\frac{1-\pi_k}{\pi_k} < \sum_{i \in U_F} \frac{1-\pi_i}{\pi_i}(\tilde{\theta}_{ik}^{opt})^2$$

which is true as proved next.

$$\begin{aligned}
\frac{1-\pi_k}{\pi_k} < \sum_{i\in U_F} \frac{1-\pi_i}{\pi_i}(\tilde{\theta}_{ik}^{opt})^2 &\Leftrightarrow \frac{\Pi_{i\in U_F}(1-\pi_i)^{l_{ik}}}{1-\Pi_{i\in U_F}(1-\pi_i)^{l_{ik}}} < \sum_{i\in U_F} \frac{1-\pi_i}{\pi_i}(\tilde{\theta}_{ik}^{opt})^2 \\
&\Leftrightarrow \sum_{i\in U_F} l_{ik}\frac{\pi_i}{1-\pi_i} < \frac{1-\Pi_{i\in U_F}(1-\pi_i)^{l_{ik}}}{\Pi_{i\in U_F}(1-\pi_i)^{l_{ik}}} \\
&\Leftrightarrow 1+\sum_{i\in U_F} l_{ik}\frac{\pi_i}{1-\pi_i} < \frac{1}{\Pi_{i\in U_F}(1-\pi_i)^{l_{ik}}} \\
&\Leftrightarrow 1+\sum_{i\in U_F} l_{ik}\frac{\pi_i}{1-\pi_i} < \Pi_{i\in U_F}\left(1+\frac{\pi_i}{1-\pi_i}\right)^{l_{ik}}.
\end{aligned}$$

This last inequality is true since $\frac{\pi_i}{1-\pi_i} > 0$, and this finishes the proof of Proposition 2.2.3. □

Chapitre 3

Optimal Weights for double Many-To-One Generalized Weight Share Method

Abstract

In probabilistic surveys, the sampling frame of the target population is not always available. If there is a frame population linked to the target population, a sample can be drawn in the target population through indirect sampling. The sampling weights of the target population units can be determined using the Generalized Weight Share Method (GWSM). These sampling weights take into account the sampling weights of the frame population units, together with some link weights that are assigned by the statistician to the links between the frame and the target populations. This chapter focuses on "many-to-one" links where it is possible to derive a set of optimal link weights that minimize the variance of the GWSM estimator, under constraints on the sampling design which are not very restrictive (Medous et al., 2023a). To get unbiased estimators of finite population totals, the link weights have to be standardized, which implies that the links between the frame population and the sample units in the target population have to be known or observed. If the number of links is large, the link weights are thus difficult to retrieve. A solution is to consider an intermediate population, linked to both the frame and the target populations, and to use a double indirect sampling. The sampling weights can be determined using a double GWSM, first between the frame and the intermediate populations, and then between the intermediate and the target populations. When both link weights are standardized (double standardization), the double GWSM allows for a reduction in the number of observed links. This method is therefore easier to apply than the GWSM, but it may deteriorate the precision of the estimator. The objective of the present paper is to derive doubly standardized link

weights that are optimal in the sense that the variance of the double GWSM is minimized. When such weights cannot be computed, as is the case in the French postal traffic survey, alternative weights can be used to improve the precision of the double GWSM estimator. Different alternative weights are compared through Monte Carlo simulations and an application to the French postal traffic estimation is discussed.

Keywords : complex sampling design, finite population, generalized weight share method, indirect sampling, surveys.

3.1 Introduction

For economic and marketing purposes, the French postal service (La Poste) is interested in obtaining information not only on the monthly postal traffic but also on the types of letters being sent. As certain letters cannot be processed automatically due to either confidentiality concerns or challenges in identifying certain types of letters, La Poste conducts a probability-based survey of postman rounds to estimate the monthly postal traffic. However, due to frequent reorganizations of postal offices, the population of postman rounds is too unstable to be directly observed and La Poste uses the more stable population of addresses as a proxy to indirectly draw samples from the population of rounds, as detailed in Medous et al. (2023a).

Indirect sampling consists in drawing a sample from a frame population linked to the target population and observing all target units linked to this sample. This sampling method requires that all units in the target population are linked to at least one unit of the frame population. In La Poste example, the target population comprises postman rounds departing in a particular month, while the frame population consists of all pairs of French addresses and days of the month, referred to as "address-days." A round is considered linked to an address-day if the postman delivers letters to that address on that specific day. In this example, all units in the target population are linked to the frame population since each postman round serves at least one address on a particular day. Furthermore, a single address can only be served by one round at most on a given day, making the links of the Many-to-One (MtO) type.

Indirect sampling raises a challenge in computing sampling weights of the target population units, but this can be addressed by using the Generalized Weight Share Method (GWSM). GWSM was introduced by Deville and Lavallée (2006) and Lavallée (2007) and was inspired by the weight sharing method used by Ernst (1986) and Kalton and Brick (1995) for longitudinal household studies. It consists of allocating weights to the links bet-

ween frame and target populations, which are then used to compute the final weights. GWSM has been widely studied in the literature, particularly for its utility in sampling hard-to-reach populations, as demonstrated in Deville and Lavallée (2006) and De Vitiis et al. (2014).

Deville and Lavallée (2006) discuss the existence of link weights minimizing the variance of the GWSM estimator for any variable of interest and demonstrate that such optimal link weights exist for Simple Random Sampling Without Replacement and Poisson sampling, if the links are of type MtO. Medous et al. (2023a) extend the results of Deville and Lavallée (2006) to a more general setting and show that optimal link weights exist for the complex sampling design used by La Poste.

However, in MtO setups, a large number of links has to be observed in order to implement the GWSM. In la Poste case, GWSM requires to identify all addresses per round, which can be relatively high. To reduce this number, La Poste has adopted the strategy of drawing double indirect samples using the intermediate population of sorting boxes and double GWSM estimators of finite population totals. According to Medous et al. (2023a), employing double GWSM for MtO-MtO links (MtO links between frame and intermediate and between intermediate and target populations) can significantly reduce the required number of observed links when compared to a single GWSM. However, this reduction comes at the expense of imposing constraints on the link weights, which requires to reformulate the optimal weights described by Deville and Lavallée (2006) and Medous et al. (2023a).

This work intends to investigate the existence of optimal weights that minimize the variance of double GWSM estimators regardless of the variable of interest while respecting the weights constraints defined by Medous et al. (2023a) to ensure a reduction in the number of observed links compared to simple GWSM estimators. We show that, in the case of La Poste, optimal weights for the double GWSM exist but cannot be computed, as the required information is difficult to obtain. Consequently, the current double GWSM estimator used by La Poste suffers from a significant loss of efficiency compared to an optimal double GWSM estimator. To limit this loss of efficiency, we suggest the use of alternative weights computed using auxiliary information.

Section 3.2 provides a brief overview of simple and double indirect sampling and their corresponding simple and double GWSM estimators, along with their variances. In section 3.3, we derive optimal double GWSM weights for MtO-MtO links. We discuss the conditions for the existence of such weights and provide their values for common sampling designs, such as Simple Random Sampling without Replacement or Poisson sampling. In section 3.4, we address the situation where optimal double GWSM weights cannot be computed and presents alternative weight options, including those currently used at La Poste. We examine the loss of precision between an optimal double GWSM estimator and a double GWSM estimator

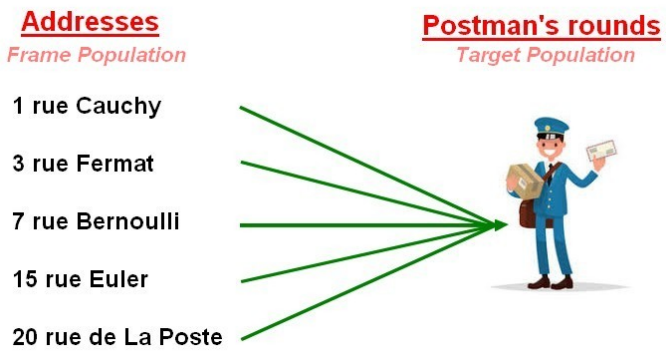

FIGURE 3.1 – Example of MtO links at La Poste.

using alternative weights, and provide advice on choosing the best alternative weights. In Section 3.5, we report results of a simulation study that mimics La Poste situation, where we test various link weights possibilities. Finally, we provide details on the alternative weights currently being tested at La Poste and those planned for the future. Proofs are gathered in the Appendix.

3.2 Double Generalized Weight Share Method

3.2.1 Indirect sampling and simple GWSM estimator

La Poste estimates its monthly traffic by observing postman rounds. However, the organization of post offices makes it difficult to know which rounds will depart in advance. Direct sampling methods cannot be used to draw a sample of rounds and La Poste has to use alternative sampling methods such as indirect sampling. Indirect sampling involves selecting a sample from a frame population that is linked to the target population.

Let U_T be the target population of size N_T and U_F be a sampling frame of size N_F linked to U_T. Following Deville and Lavallée (2006), we assume that every link between U_F and U_T is observable and that each unit of U_T is linked to at least one unit of U_F.

Various types of links between populations exist and have been studied in the literature ; see Deville (1999), Deville and Lavallée (2006), Medous et al. (2023a), among others. In this paper, we concentrate on a specific type of links, represented in figure 3.1, where each unit in U_F is connected to at most one unit in U_T. This type of links is known as the "Many-to-One" (MtO) link.

We denote by y_k the value of the variable of interest y for unit k from U_T and focus on the estimation of the total $t_y = \sum_{k \in U_T} y_k$ over U_T. A sample s_F is drawn from U_F using a sampling design $p_F(.)$. The indirect sample s_T from U_T is composed of units from U_T linked

to at least a unit in s_F. We denote by $\pi_i = \mathbb{P}(i \in s_F)$ the first-order inclusion probability of unit $i \in U_F$ and by $\pi_{ii'} = \mathbb{P}(i, i' \in s_F)$ the second-order inclusion probability of units i and $i' \in U_F$. We assume that $\pi_i > 0$ for all units i in U_F and note $d_i = 1/\pi_i > 0$ the sampling weight of unit i. We denote $\Delta_{ii'} = \dfrac{\pi_{ii'} - \pi_i \pi_{i'}}{\pi_i \pi_{i'}}, \quad i, i' \in U_F$.

Three sampling designs $p_F(.)$ are considered below : Poisson sampling, Simple Random Sampling Without Replacement (SRSWOR) of size n_F and Stratified Simple Random Sampling Without Replacement (SSRSWOR) with H strata of size N_h, $h = 1 \dots H$. Poisson sampling consists in performing an independent Bernoulli trial of parameter π_i for each $i \in U_F$. SRSWOR of size n_F consists in assigning equal and positive probabilities to each of the without replacement samples of size n_F and a null probability otherwise. The sampling weights are equal to $d_i = N_F/n_F$ for all i in U_F. Finally, SSRSWOR with H strata of size N_h, $h = 1 \dots H$, consists in performing an independent SRSWOR of size n_h in each stratum h, $h = 1 \dots H$. The sampling weights are equal to $d_i = N_h/n_h$ for all i in strata h, $h = 1 \dots H$.

As mentionned in Deville and Lavallée (2006), the main difficulty with indirect sampling is the computation of the sampling weights of units $k \in U_T$. In order to deal with this challenge, we assign a link weight θ_{ik} to each link between $i \in U_F$ and $k \in U_T$, with $\theta_{ik} > 0$ when i and k are linked and $\theta_{ik} = 0$ otherwise. We define the standardized link weights by $\tilde{\theta}_{ik} = \theta_{ik} / \sum_{i' \in U_F} \theta_{i'k}$ for $i \in U_F$, $k \in U_T$, such that $\sum_{i \in U_F} \tilde{\theta}_{ik} = 1$ for all $k \in U_T$. The simple Generalized Weight Share Method (GWSM) estimator of t_y, introduced by Deville and Lavallée (2006), is given by :

$$\hat{t}_{y1} = \sum_{i \in s_F} d_i \left(\sum_{k \in U_T} \tilde{\theta}_{ik} y_k \right) \tag{3.1}$$

and its variance by :

$$\mathrm{Var}(\hat{t}_{y1}) = \sum_{i \in U_F} \sum_{i' \in U_F} \Delta_{ii'} \sum_{k \in U_T} \tilde{\theta}_{ik} y_k \sum_{k' \in U_T} \tilde{\theta}_{i'k'} y_{k'}. \tag{3.2}$$

It can be shown that the simple GWSM estimator is unbiased as long as the link weights are standardized. However, as we can see in equation (3.2), the variance of the simple GWSM estimator depends on the standardized link weights. Deville and Lavallée (2006) and Lavallée (2007) seek for optimal standardized link weights, namely link weights minimizing the variance given by equation (3.2) for any variable of interest y. When the links are of type MtO and the samples are drawn using SRSWOR or Poisson sampling, Deville and Lavallée (2006) and Lavallée (2007) demonstrate the existence and uniqueness of a set of optimal standardized link weights. Medous et al. (2023a) extend this result to a wider range of sam-

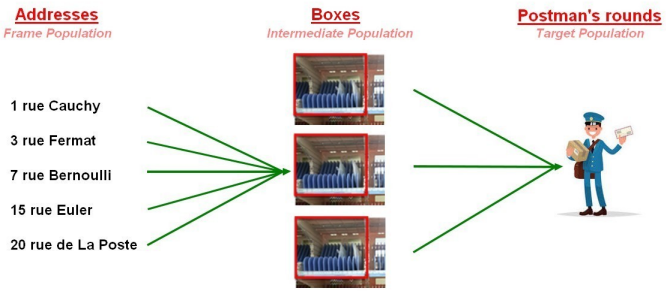

FIGURE 3.2 – Example of MtO-MtO at La Poste.

pling designs satisfying a specific condition. In the following, we refer to the optimal simple GWSM estimator as the simple GWSM estimator obtained with the optimal standardized link weights.

MtO links allow optimal weights to be derived, but the large number of observed links required to compute the standardized link weights, whether they are optimal or not, can raise implementation challenges, as it is the case of La Poste (see Medous et al. (2023a)). To deal with this issue, one can use double indirect sampling as suggested in Deville and Lavallée (2006) and Medous et al. (2023a).

3.2.2 Double GWSM estimator

Double indirect sampling consists of introducing a second intermediate population U_M, linked to both the frame population U_F and the target population U_T. Using the links between these populations, two indirect samples can be obtained, one from U_M and the other from U_T. The estimators of totals on U_T are then obtained by applying the GWSM twice, first between U_F and U_M, and then between U_M and U_T. This method is known as double GWSM (Medous et al., 2023a).

Let N_M be the size of the intermediate population U_M. Following Medous et al. (2023a), we make the assumption that the links between U_F and U_M, respectively U_M and U_T, are both of type MtO. This type of links is referred to as MtO-MtO, which is a specific case of MtO.

As in section 3.2.1, a sample s_F is drawn from U_F with a sampling design $p_F(.)$. The indirect sample s_M from U_M (resp. s_T from U_T) is composed of units from U_M linked to at least a unit $i \in s_F$ (resp. of units from U_T linked to at least a unit $j \in s_M$). This sampling technique is called double indirect sampling and is currently used in La Poste survey, with an intermediate population of sorting boxes. These boxes are used to sort the mail before the beginning of a postman round.

For each $i \in U_F$ and $j \in U_M$ (resp. $j \in U_M$, $k \in U_T$), we assign a link weight θ_{ij}^{FM} (resp.

θ_{jk}^{MT}) with $\theta_{ij}^{FM} > 0$ when i and j are linked and $\theta_{ij}^{FM} = 0$ otherwise (resp. $\theta_{jk}^{MT} > 0$ when j and k are linked and $\theta_{jk}^{MT} = 0$ otherwise). Let $\tilde{\theta}_{ij}^{FM}$ (resp. $\tilde{\theta}_{jk}^{MT}$) be the standardized link weight between $i \in U_F$ and $j \in U_M$ (resp. $j \in U_M$ and $k \in U_T$) such that $\sum_{i \in U_F} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} = 1$. In what follows, we call FM weights (resp. MT weights) the set of standardized weights between U_F and U_M (resp. between U_M and U_T).

The double GWSM estimator of t_y is given by

$$\hat{t}_{y2} = \sum_{i \in s_F} d_i \left(\sum_{k \in U_T} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} y_k \right). \tag{3.3}$$

The double GWSM estimator is a simple GWSM with standardized link weights $\tilde{\theta}_{ik}$, $i \in U_F$, $k \in U_T$, such that

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

The condition $\sum_{i \in U_F} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} = 1$, $k \in U_T$, produces unbiased double GWSM estimators and is verified for various standardization methods for FM and MT link weights. However, not all standardization methods allow for a reduction of the number of observed links. Medous et al. (2023a) show that a reduction of the number of observed links occurs when we consider FM and MT weights such that $\sum_{i \in U_F} \tilde{\theta}_{ij}^{FM} = 1$, $j \in U_M$, and $\sum_{j \in U_M} \tilde{\theta}_{jk}^{MT} = 1$, $j \in U_T$. This is known as double standardization and we say that the link weights are twice (or double) standardized.

The double GWSM with double standardization offers an advantage when it comes to the number of links to observe. Yet, the optimal link weights defined for simple GWSM have to be rewritten to ensure that they satisfy the double standardization constraint. Without this reformulation of the optimal simple weights, the variance of the double GWSM estimator with double standardization can be higher than the variance of the optimal simple GWSM estimator, as shown by Medous et al. (2023a). In order to get an estimator with both a small number of links to observe and a small variance, we look for twice standardized link weights that minimize the variance of the double GWSM estimator for all variable of interest y. Thus we look for a set of FM weights and MT weights such that $\sum_{i \in U_F} \tilde{\theta}_{ij}^{FM} = 1$, $j \in U_M$, and $\sum_{j \in U_M} \tilde{\theta}_{jk}^{MT} = 1$, $j \in U_T$, that minimizes the variance given by equation (3.2) for any variable of interest, with $\tilde{\theta}_{ik} = \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT}$. In what follows, we will refer to those weights as the optimal weights for double GWSM estimators.

In what follows, we only consider twice standardized FM and MT weights.

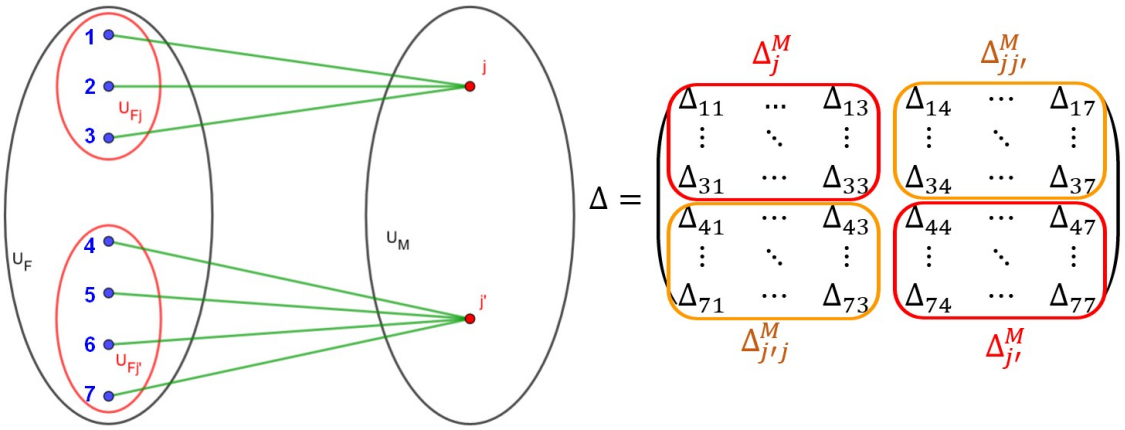

FIGURE 3.3 – U_F ordered according to U_{Fj}, $j \in U_M$.

3.3 Optimality of double GWSM

Medous et al. (2023a) show that optimal weights for the simple GWSM estimator exist for sampling designs satisfying a condition that depends on the links between U_F and U_T. Likewise, we can show that optimal weights for double GWSM estimator with double standardization exist for sampling designs that meet two conditions : one depending on the links between U_F and U_M, and another depending on the links between U_M and U_T.

Let us consider first the links between U_F and U_M. For a given unit $j \in U_M$, we denote by U_{Fj} the set of size N_{Fj} containing all units i in U_F that are linked to j. If the links between U_F and U_M are MtO, these subpopulations form a partition of U_F. In what follows, we assume that units in U_F are ordered according to the subpopulations U_{Fj}, $j \in U_M$, as in figure 3.3. Let $\mathbf{\Delta} = (\Delta_{ii'})_{i,i' \in U_F}$ be the covariance matrix of size $N_F \times N_F$ and $\mathbf{\Delta}^M_{jj'} = (\Delta_{ii'})_{i \in U_{Fj}, i' \in U_{Fj'}}$ be the submatrix of $\mathbf{\Delta}$ corresponding to the elements in position i and i' such that i (resp. i') is linked to j (resp. to j'). We denote by $\mathbf{\Delta}^M_j$ the square submatrix $\mathbf{\Delta}^M_{jj}$ of size $N_{Fj} \times N_{Fj}$. If the links between U_F and U_M are MtO, then the submatrices $\mathbf{\Delta}^M_{jj'}$ form a partition of $\mathbf{\Delta}$ as in figure 3.3. Let us denote $\mathbb{1}^M_j$ the N_{Fj}-dimensional vector of ones and by u^t the transpose of any vector u.

Definition 3.3.1 (Δ-property for U_M). For MtO links between U_F and U_M, a sampling design is said to satisfy the Δ-property for U_M if, for all $j \in U_M$, $\mathbf{\Delta}^M_j$ is invertible and, for $j \neq j' \in U_M$, we have

$$\mathbf{\Delta}_{j,j' \neq j} = c_{jj'} \mathbb{1}^M_j \mathbb{1}^{Mt}_{j'} \quad \text{with} \quad c_{jj'} \quad \text{not depending on } i \text{ and } i'. \tag{3.4}$$

Following Medous et al. (2023a), we can show that the Δ-property for U_M holds for

Poisson sampling and SRSWOR with the following constants $c_{jj'}$, $j, j' \in U_M$, $j \neq j'$. For Poisson sampling from U_F with inclusion probabilities π_i for $i \in U_F$,

$$c_{jj'} = 0 \text{ for all } j \neq j' \in U_T. \tag{3.5}$$

For SRSWOR of size n_F from U_F,

$$c_{jj'} = -\frac{1-f}{f}\frac{1}{N_F - 1} \tag{3.6}$$

For stratified SRSWOR with H strata of size N_h in U_F, $h = 1, ..., H$, the Δ-property for U_M holds if the following condition is verified :

(C1) for all $j \in U_M$, all units $i \in U_F$ linked to j belong to the same stratum and for each stratum h, there are at least two units in U_M linked to h.

Then

$$c_{jj'} = \frac{1-f_h}{f_h}\frac{1}{N_h - 1}, \tag{3.7}$$

if all units $i, i' \in U_F$ linked to respectively j and j' belong to the same stratum and 0 otherwise, for all $j \neq j' \in U_M$.

Let us denote $\lambda_j = \mathbb{1}_j^{Mt}(\mathbf{\Delta}_j^M)^{-1}\mathbb{1}_j^M$, $j \in U_M$. The Δ-property for U_M allows us to derive the following proposition on the FM weights.

Proposition 3.3.1. *If the links between U_F and U_M are MtO and the sampling design satisfies the Δ-property for U_M, then there is a unique set of optimal FM weights $\tilde{\theta}_{ij}^{FM,opt}$, $i \in U_F$, $j \in U_M$ minimizing the variance of $\hat{t}_{y2}$ for any variable of interest and any set of weights $\tilde{\theta}_{jk}^{MT}$, $j \in U_M$, $k \in U_T$. These weights are given by :*

$$(\tilde{\theta}_{ij}^{FM,opt})_{i \in U_{Fj}} = \frac{(\mathbf{\Delta}_j^M)^{-1}\mathbb{1}_j^M}{\mathbb{1}_j^{Mt}(\mathbf{\Delta}_j^M)^{-1}\mathbb{1}_j^M} = \frac{1}{\lambda_j}(\mathbf{\Delta}_j^M)^{-1}\mathbb{1}_j^M, \text{ if } i \text{ is linked to } j, \text{ and } 0 \text{ otherwise.}$$

Interestingly, the FM weights given by Proposition 3.3.1 do not depend on the MT weights used. Let us now consider the links between U_M and U_T. For a given unit $k \in U_T$, let us denote by U_{Mk} the set of size N_{Mk} containing all units j in U_M that are linked to k. In the MtO-MtO setup, a unit j from U_M can only be linked to one and only one unit k from U_T and the subpopulations U_{Mk}, $k \in U_T$, form a partition of U_M (see figure 3.4). In what follows we assume that the units in U_M are ordered according to the subpopulations U_{Mk}, $k \in U_T$, as in figure 3.4.

We can define the following condition.

Definition 3.3.2 (c_k-condition). If the links are MtO-MtO and the Δ-property holds for U_M,

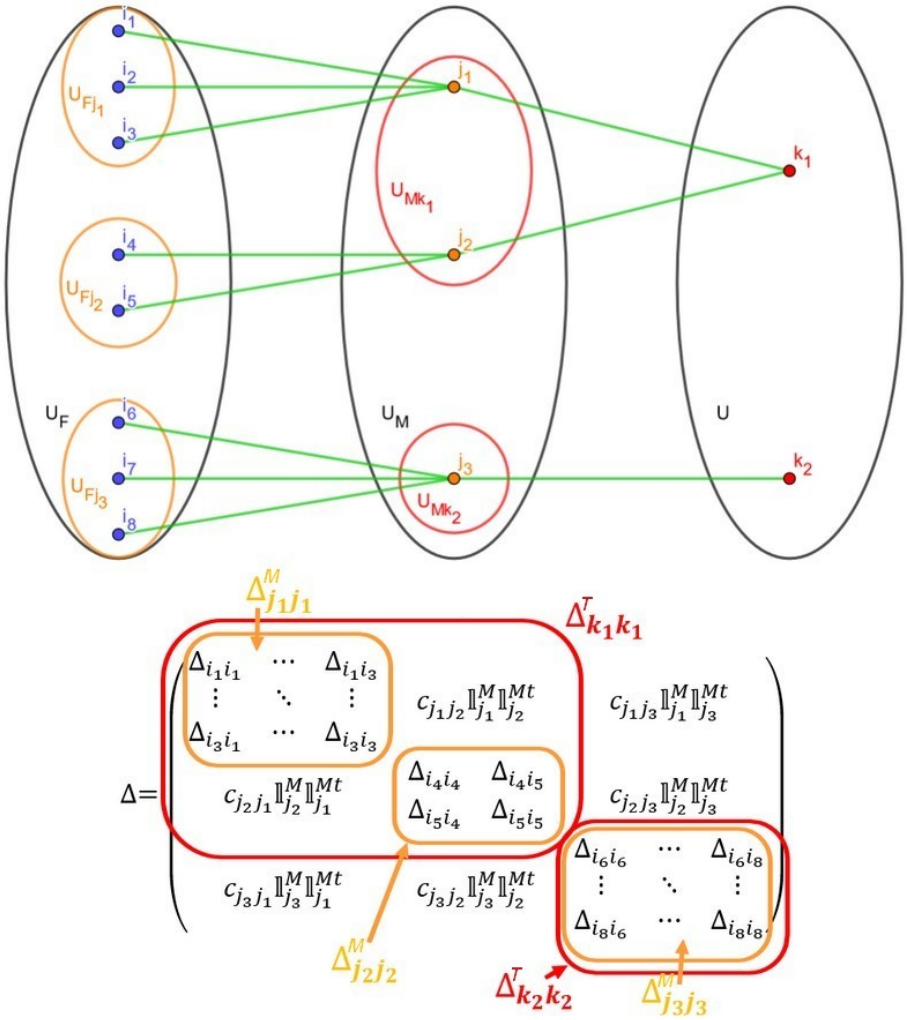

FIGURE 3.4 – U_F ordered according to U_{Fj}, $j \in U_M$ and U_{Mk}, $k \in U_T$, with $\mathbf{\Delta}^T_{kk}$ the submatrix of $\mathbf{\Delta}$ corresponding to the element in position i and i' such that i and i' are linked to k.

$\mathbf{\Delta}$ is said to verify the c_k-condition if, for all j (resp. j') $\in U_M$ linked to k (resp. k') $\in U_T$ such that $j \neq j'$, the constants $c_{jj'}$ given by the Δ-property for U_M verify

$$c_{jj'} = c_{kk'},$$

with $c_{kk'}$ not depending on j and j'.

Let us denote c_k, $k \in U_T$ the value of $c_{kk'}$ for $k' = k$. The c_k-condition holds for Poisson sampling and SRSWOR. For Poisson sampling and SRSWOR, the values of $c_{jj'}$, $j \neq j' \in U_M$ given by the Δ-property for U_M are constant and do not depend on j or j'. Thus the c_k-property holds and the values $c_{kk'}$, $k \neq k' \in U_T$, are equal to the values $c_{jj'}$, $j \neq j' \in U_M$ given respectively in equations (3.5) and (3.6). For stratified SRSWOR with H strata of size N_h in U_F, $h = 1, \ldots, H$, the c_k-condition holds if the following condition is verified :

(C2) for all $k \in U_T$, all units $i \in U_F$ linked to k belong to the same stratum.

Then the values $c_{kk'}$, $k \neq k' \in U_T$, are equal to the values $c_{jj'}$, $j \neq j' \in U_M$ given in equation (3.7).

Let $\tilde{\mathbf{\Delta}}$ be the matrix of size $N_M \times N_M$ composed of the values

$$\tilde{\Delta}_{jj'} = \sum_{i \in U_F} \sum_{i' \in U_F} \Delta_{ii'} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{i'j'}^{FM}, j, j' \in U_M. \tag{3.8}$$

Let $\tilde{\mathbf{\Delta}}_k$ be the submatrix of $\tilde{\mathbf{\Delta}}$ corresponding to the elements in position j and j' such that j and j' are linked to k. Let $\tilde{\mathbb{1}}_k$ be the N_{Mk} dimensional vector of ones. If the Δ-property for U_M and the c_k-condition are verified, we can derive the following proposition on the MT weights.

Proposition 3.3.2. *If the links between U_F and U_T are MtO-MtO, the sampling design satisfies the Δ-property for U_M, the c_k-condition is verified, and the FM weights $\tilde{\theta}_{ij}^{FM}$, $i \in U_F$, $j \in U_M$, verify :*

$$(\tilde{\theta}_{ij}^{FM})_{i \in U_{Fj}}^t \Delta_j^M (\tilde{\theta}_{i'j}^{FM})_{i' \in U_{Fj}} - c_k \neq 0 \text{ for all } j \in U_{Mk}$$
$$\text{and } c_k \tilde{\mathbb{1}}_k^t diag\left(\left((\tilde{\theta}_{ij}^{FM})_{i \in U_{Fj}}^t \Delta_j^M (\tilde{\theta}_{i'j}^{FM})_{i' \in U_{Fj}} - c_k \right)_{j \in U_{Mk}}^{-1} \right) \tilde{\mathbb{1}}_k \neq -1,$$

then there is a unique set of MT weights $\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM})$, $j \in U_M$, $k \in U_t$, functions of $(\tilde{\theta}_{ij}^{FM})_{i \in U_{Fj}}$ and called θ^{FM}-optimal MT weights, minimizing the variance of $\hat{t}_{y2}$ for any variable of interest. These weights are given by :

$$\left(\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM}) \right)_{j \in U_{Mk}} = \tilde{\mathbf{\Delta}}_k^{-1} \tilde{\mathbb{1}}_k \left(\tilde{\mathbb{1}}_k^t \tilde{\mathbf{\Delta}}_k^{-1} \tilde{\mathbb{1}}_k \right)^{-1}, \text{ if } j \text{ is linked to } k, \text{ and } 0 \text{ otherwise.}$$

If we compute the θ^{FM}-optimal MT weights for the optimal FM weights given by Proposition 3.3.1, we get a unique set of FM and MT weights that minimize the variance of the double GWSM estimator for any variable of interest as detailled in Proposition 3.3.3.

Proposition 3.3.3. *If the links are MtO-MtO, the Δ-property holds for U_M, the c_k-condition is verified and for all $j \in U_{Mk}$, $k \in U_T$, the following requirement is verified,*

$$\text{(inversibility conditions) } c_k \lambda_j \neq 1 \text{ and } c_k \sum_{j \in U_{Mk}} \lambda_j / (1 - c_k \lambda_j) \neq -1,$$

with the value of c_k given by the c_k-condition, then the θ^{FM}-optimal MT weights for the optimal FM weights $\tilde{\theta}_{ij}^{FM,opt}$, $i \in U_F$, $j \in U_M$, are given by :

$$\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}) = \frac{\lambda_j/(1-c_k\lambda_j)}{\sum_{j'\in U_{Mk}}\lambda_{j'}/(1-c_k\lambda_{j'})}, \text{ if } j \text{ is linked to } k, \text{ and } 0 \text{ otherwise.}$$

Moreover, we can show that the optimal weights for the simple GWSM estimator $\tilde{\theta}_{ik}^{opt}$ *can be rewritten for all* $i \in U_F$, $k \in U_T$ *as :*

$$\tilde{\theta}_{ik}^{opt} = \sum_{j\in U_M} \tilde{\theta}_{ij}^{FM,opt}\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}).$$

The proofs of Propositions 3.3.1, 3.3.2 and 3.3.3 are given in the Appendix.

We show in Proposition 3.3.3 the existence of optimal link weights for double GWSM estimators with double standardization. Using these weights allows for a reduction in the number of observed links compared to a simple GWSM while preserving the accuracy of the optimal simple GWSM estimator. In what follows, the double GWSM estimator with double standardization obtained by using optimal FM and $\theta^{FM,opt}$-optimal MT weights is called optimal double GWSM estimator and denoted $\hat{t}_{y2}^{opt}$, and the set of optimal FM weights and $\theta^{FM,opt}$-optimal MT weights is referred to as optimal double GWSM weights.

Let us derive the optimal FM weights and the $\theta^{FM,opt}$-optimal MT weights for Poisson sampling, SRSWOR and stratified SRSWOR. We denote by l_{ij}^{FM} (resp. l_{jk}^{MT}) the link between unit $i \in U_F$ and unit $j \in U_M$ (resp. between $j \in U_M$ and unit $k \in U_T$) with $l_{ij}^{FM} = 1$ if i and j are linked and 0 otherwise (resp. $l_{jk}^{MT} = 1$ if j and k are linked and 0 otherwise). For Poisson sampling, we have $c_k = 0$ for all $k \in U_T$, thus the inversibility conditions hold. Let $\Delta_{ii} = (1-\pi_i)/\pi_i$ for $i \in U_F$. The optimal double GWSM weights are given by

$$\tilde{\theta}_{ij}^{FM,opt} = \frac{l_{ij}^{FM}/\Delta_{ii}}{\sum_{i'\in U_F} l_{i'j}^{FM}/\Delta_{i'i'}} \text{ and } \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}) = \frac{l_{jk}^{MT}\sum_{i\in U_F} l_{ij}^{FM}/\Delta_{ii}}{\sum_{j'\in U_M} l_{j'k}^{MT}\sum_{i'\in U_F} l_{i'j'}^{FM}/\Delta_{i'i'}},$$

for all $j \in U_M$ and $k \in U_T$. For SRSWOR sampling, it can be shown, following Deville and Lavallée (2006), that $\lambda_j = -(1/c_k)\, N_{Fj}/(N_F - N_{Fj})$, $j \in U_{Mk}$, $k \in U_T$. Then $c_k\lambda_j \neq 1$ and

$$c_k \sum_{j\in U_{Mk}} \lambda_j/(1-c_k\lambda_j) = -\sum_{j\in U_{Mk}} N_{Fj}/N_F.$$

For any unit k from U_T, let us denote U_{Fk} of size N_{Fk} the set of units in U_F linked to k. When the links are of type MtO-MtO, $N_{Fk} = \sum_{j\in U_{Mk}} N_{Fj}$ for all $k \in U_T$ and $N_F = \sum_{k\in U_T} N_{Fk}$. If $N_T > 1$, then $-N_{Fk}/N_F \neq -1$ for all $k \in U_T$ and the inversibility conditions hold. The

optimal double GWSM weights are given by

$$\tilde{\theta}_{ij}^{FM,opt} = l_{ij}^{FM}/N_{Fj} \text{ and } \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}) = N_{Fj} l_{jk}^{MT} / \sum_{j' \in U_F} (N_{Fj'} l_{j'k}^{MT})$$

for all $j \in U_M$ and $k \in U_T$. Details on the derivation of the value of $(\mathbf{\Delta}_j^M)^{-1}\mathbb{1}_j^M$ can be found in Deville and Lavallée (2006). The remaining computation is straightforward.
Following the reasoning in Medous et al. (2023a), it can be shown that the optimal link weights $\tilde{\theta}_{ij}^{FM,opt}$ and $\theta^{FM,opt}$-optimal link weights $\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$ for stratified SRSWOR are the same as those for SRSWOR if the previously mentioned conditions **(C1)** and **(C2)** are verified.

Following Medous et al. (2023a), we can show that, if the links are MtO-MtO, the Δ-property holds for U_M and the c_k-condition is verified, the loss of efficiency between an estimator $\hat{t}_{y2}$ given by (3.3) and $\hat{t}_{y2}^{opt}$ is given by :

$$\text{Var}(\hat{t}_{y2}) - \text{Var}(\hat{t}_{y2}^{opt}) = \sum_{k \in U_T} y_k^2 \text{Var}(\hat{t}_{\tilde{\theta}_k} - \hat{t}_{\tilde{\theta}_k^{opt}}) = \sum_{k \in U_T} y_k^2 \text{Var}(\hat{t}_{\tilde{\theta}_k - \tilde{\theta}_k^{opt}}),$$

where

$$\hat{t}_{\tilde{\theta}_k^{opt}} = \sum_{i \in s_F} d_i \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM,opt} \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$$

is the HT estimator of the total

$$t_{\tilde{\theta}_k^{opt}} = \sum_{i \in U_F} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM,opt} \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}) = 1,$$

and

$$\hat{t}_{\tilde{\theta}_k - \tilde{\theta}_k^{opt}} = \sum_{i \in s_F} d_i \left(\sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} - \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM,opt} \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}) \right)$$

is the HT estimator of the total $t_{\tilde{\theta}_k - \tilde{\theta}_k^{opt}} = 0$, for all $k \in U_T$.

Consider now the double GWSM estimator with double standardization $\hat{t}_{y2}^{FM,opt}$ computed with the optimal FM weights $\tilde{\theta}_{ij}^{FM,opt}$, $i \in U_F$, $j \in U_M$ and any weights $\tilde{\theta}_{jk}^{MT}$, $k \in U_T$. The loss of efficiency between the optimal double GWSM estimator $\hat{t}_{y2}^{opt}$ and the double GWSM estimator $\hat{t}_{y2}^{FM,opt}$ can be expressed as a function of the differences between MT weights.

Proposition 3.3.4. *If the links are MtO-MtO, the Δ-property holds for U_M and the c_k-condition is verified, then the loss of efficiency between the optimal double GWSM estimator*

$\hat{t}_{y2}^{opt}$ and the double GWSM estimator $\hat{t}_{y2}^{FM,opt}$ is given by :

$$Var(\hat{t}_{y2}^{FM,opt}) - Var(\hat{t}_{y2}^{opt}) = \sum_{k\in U_T} y_k^2 \sum_{j\in U_M} \left(\tilde{\theta}_{jk}^{MT} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})\right)^2 \left(\frac{1}{\lambda_j} - c_k\right)$$

with the value of c_k given by the c_k-condition.

Since Poisson sampling, SRSWOR and, under conditions **(C1)** and **(C2)**, stratified SRWOR verify the Δ-property for U_M and the c_k-condition, we can compute the loss of efficiency between $\hat{t}_{y2}^{opt}$ and $\hat{t}_{y2}^{FM,opt}$ for these sampling designs as follows.

Corollary 3.3.3. *If the links are MtO-MtO and the sampling designs are Poisson, SRSWOR, or stratified SRSWOR under conditions* **(C1)** *and* **(C2)**, *then the loss of efficiency between $\hat{t}_{y2}^{opt}$ and $\hat{t}_{y2}^{FM,opt}$ can be expressed using Proposition 3.3.4 .*

For Poisson sampling, we have :

$$Var(\hat{t}_{y2}^{FM,opt}) - Var(\hat{t}_{y2}^{opt}) = \sum_{k\in U_T} y_k^2 \sum_{j\in U_M} \left(\tilde{\theta}_{jk}^{MT} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})\right)^2 \frac{1}{\sum_{i\in U_F} l_{ij}/\Delta_{ii}}.$$

For SRSWOR of size n_F from U_F, let us denote $f = n_F/N_F$. We have :

$$\begin{aligned} &Var(\hat{t}_{y2}^{FM,opt}) - Var(\hat{t}_{y2}^{opt}) \\ &= \sum_{k\in U_T} y_k^2 \sum_{j\in U_M} \left(\tilde{\theta}_{jk}^{MT} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})\right)^2 \left(\frac{1}{n_F}\frac{N_F - n_F}{N_F - 1}\frac{N_F - N_{Fj}}{N_{Fj}} + \frac{1-f}{f}\frac{1}{N_F - 1}\right). \end{aligned}$$

For stratified SRSWOR with H strata of size N_h in U_F, $h = 1, \ldots, H$, let us denote $f_h = n_h/N_h$, $h = 1, \ldots, H$. Proposition 3.3.4 holds if the previously mentioned conditions **(C1)** *and* **(C2)** *are verified. Then we have :*

$$\begin{aligned} &Var(\hat{t}_{y2}^{FM,opt}) - Var(\hat{t}_{y2}^{opt}) \\ &= \sum_{h=1}^{H} \sum_{k\in U_T} y_k^2 \sum_{j\in U_M} \left(\tilde{\theta}_{jk}^{MT} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})\right)^2 \left(\frac{1}{n_h}\frac{N_h - n_h}{N_h - 1}\frac{N_h - N_{Fj}}{N_{Fj}} + \frac{1-f_h}{f_h}\frac{1}{N_h - 1}\right). \end{aligned}$$

The expressions of the difference of variances for SRSWOR and stratified SRSWOR can be approximated by simpler expressions as follows.

Corollary 3.3.4. *If Proposition 3.3.4 holds, then, under conditions detailed in the Appendix, we have :*

For SRSWOR of size n_F from U_F resulting in an indirect sample of size n_T from U_T :

$$n_T \left(\frac{Var(\hat{t}_{y2}^{FM,opt})}{N_T^2} - \frac{Var(\hat{t}_{y2}^{opt})}{N_T^2} \right)$$
$$\approx \frac{n_T}{N_T} \sum_{k \in U_T} y_k^2 \sum_{j \in U_M} \left(\tilde{\theta}_{jk}^{MT} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}) \right)^2 \left(\frac{1}{n_F} \frac{N_F - n_F}{N_F - 1} \frac{N_F - N_{Fj}}{N_{Fj}} \right).$$

For stratified SRSWOR with H strata of size N_h in U_F, $h = 1, \ldots, H$:

$$n_T \left(\frac{Var(\hat{t}_{y2}^{FM,opt})}{N_T^2} - \frac{Var(\hat{t}_{y2}^{opt})}{N_T^2} \right)$$
$$\approx \frac{n_T}{N_T} \sum_{h=1}^{H} \sum_{k \in U_T} y_k^2 \sum_{j \in U_M} \left(\tilde{\theta}_{jk}^{MT} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}) \right)^2 \left(\frac{1}{n_h} \frac{N_h - n_h}{N_h - 1} \frac{N_h - N_{Fj}}{N_{Fj}} \right).$$

In La Poste survey, the sampling design is complex but can be approximated by a stratified SRSWOR verifying both the Δ-property and c_k-condition. As a result, optimal FM weights and $\theta^{FM,opt}$-optimal MT weights exist. Although the optimal FM weights are currently being used at La Poste, the $\theta^{FM,opt}$-optimal MT weights cannot be computed which results in a loss of precision compared to the optimal double GWSM estimator as discussed in Medous et al. (2023a).

3.4 Alternative to double GWSM optimal link weights

In this section, we address the situation where the $\theta^{FM,opt}$-optimal MT weights cannot be computed. We propose alternative weights and discuss their practical computation.

3.4.1 Alternative link weights using auxiliary information

In section 3.3, we demonstrate that there exists a set of optimal FM weights and $\theta^{FM,opt}$-optimal MT weights that minimize the variance of double GWSM estimators under certain conditions, irrespective of the variable of interest. To compute the optimal FM weights, the knowledge of all links between U_F and indirectly sampled units in U_M, as well as the sampling design on U_F, is required. Usually, this information is available in practice as the sampling design is selected by the statistician. Moreover, in cases where the links are MtO-MtO, the number of links between U_F and units in U_M may be reasonably low. For instance, in the case of La Poste, a sorting box typically contains 10 addresses.

However, computing the $\theta^{FM,opt}$-optimal MT weights between the indirectly sampled

units $j \in U_M$ and $k \in U_T$ can be challenging in practice as it requires knowing the value of $\lambda_{j'} = \mathbb{1}_{j'}^{Mt}(\mathbf{\Delta}_{j'}^{M})^{-1}\mathbb{1}_{j'}^{M}$ for all units $j' \in U_M$ linked to k. In La Poste case, for instance, λ_j is the number of addresses in box j'. Unfortunately, the number of addresses contained in each box is not always explicitly stated, and counting them individually is time-consuming, as mentioned in Medous et al. (2023a). Therefore, it is not possible to compute the $\theta^{FM,opt}$-optimal MT weights for La Poste survey. However, it is not uncommon for auxiliary information on U_M to be available. For example, La Poste has access to datasets containing a list of characteristics for all boxes, including their dimensions and an approximate count of the number of addresses they contain. If such information is available, an alternative MT weight can be used in place of the $\theta^{FM,opt}$-optimal MT weights.

If auxiliary information x_j is available for all $j \in U_M$, then, for all $k \in U_T$, we can compute the following alternative MT weight

$$\tilde{\theta}_{jk}^{MT,alt} = \frac{x_j}{\sum_{j' \in U_{Mk}} x_{j'}}, \text{ if } j \text{ is linked to } k, \text{ and 0 otherwise.} \tag{3.9}$$

In what follows, the GWSM estimator computed using $\tilde{\theta}_{ik}^{MT,alt}$ is called alternative double GWSM estimator and is denoted by $\hat{t}_{y2}^{alt}$. In practice, in order to compute an alternative double GWSM estimate, it is enough to know x_j, $j \in U_{Mk}$ for all $k \in U_T$ indirectly sampled.

It is possible to choose x_j, $j \in U_M$, constant for all $j \in U_M$. In that case the weight given by (3.9) reduces to the usual MT weight $\tilde{\theta}_{jk}^{MT,alt} = 1/N_{Mk}$, $k \in U_T$, $j \in U_{Mk}$. The usual MT weights are the easiest MT weights to compute, as they only require to know the size of the set U_{Mk} for a given $k \in U_T$. However, these weights generally have poor performance when compared to the $\theta^{FM,opt}$-optimal MT weights. Further information regarding the double GWSM estimator using usual MT weights can be found in Medous et al. (2023a).

3.4.2 How to choose the best auxiliary variable

In practice, it is not uncommon to have more than one auxiliary variable available for all $j \in U_M$. To minimize the loss of efficiency of the alternative double GWSM estimator, we aim to select the variable x_j, $j \in U_M$, that results in the smallest difference between the alternative MT weights and the $\theta^{FM,opt}$-optimal MT weights. We propose two approaches to identify such a variable.

One approach to select the best auxiliary variable x_j for each $j \in U_M$ is to choose a variable that is highly correlated with the non-standardized $\theta^{FM,opt}$-optimal MT weight $\theta_{jk}^{MT}(\theta^{FM,opt})$ for all $j \in U_{Mk}$, $k \in U_T$. This method allows to compute the correlations between the non-standardized $\theta^{FM,opt}$-optimal MT weight and the available auxiliary va-

riables. For instance, in La Poste case, the non-standardized $\theta^{FM,opt}$-optimal MT weight is the number of addresses in a box, which is known for all indirectly sampled boxes. The survey provides additional variables, such as the location and dimensions of the boxes, which can be used to determine the relationship between the number of addresses and the auxiliary variables. However, the loss of efficiency between the optimal and alternative double GWSM also depends on the value of the variable of interest, as demonstrated in Proposition 3.3.4. In other words, when the variable of interest takes high values, the difference in variance between the optimal and alternative double GWSM estimators can still be substantial, even when the difference in weights is small. This suggests that the correlation between the chosen auxiliary variable and the non-standardized $\theta^{FM,opt}$-optimal MT weight may not be strong enough to minimize the loss of efficiency.

The second approach involves using a variable x_j that is correlated with the variable $y_j = y_k \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$ for all $j \in U_{Mk}$. This approach takes into account the impact of the variable of interest and typically results in better weights than the first approach for lower correlation, as demonstrated in section 3.5. However, this approach requires to obtain the $\theta^{FM,opt}$-optimal MT weights for many indirectly sampled boxes, which can be challenging. In practice, the $\theta^{FM,opt}$-optimal MT weights are unknown and the choice of the auxiliary variable is based on expert knowledge. Moreover, since the alternative MT weights obtained depend on the values y_k, $k \in U_T$, they have to be computed for each variable of interest, unlike the first approach which gives a unique set of weights suitable for all variables of interest.

When the number of auxiliary variables available on U_M is large, it can be unrealistic to expect that there exists a single variable x_j that is strongly correlated with either $\theta_{jk}^{MT,opt}$ or y_j for all $j \in U_{Mk}$ and $k \in U_T$. Instead, it is more practical to assume that x_j is highly correlated with $\theta_{jk}^{MT,opt}$ or y_j for only a subset of units $k \in U_T$. To enable more flexibility in the choice of auxiliary variables, an alternative standardized weight can be computed by using different variables for different subsets of U_T.

Definition 3.4.1 (subset weights). Let the population U_T be divided into G subsets g, $g = 1, \ldots, G$, such that, for each subset g, a variable x_{gj} is available for all $j \in U_M$ linked to the units $k \in g$. Then, for all subset g, the following subset weights can be used as alternative MT weights :

$$\tilde{\theta}_{jk}^{MT} = \frac{x_{gj}}{\sum_{j \in U_{Mk}} x_{gj}},\ j \in U_{Mk},\ k \in g.$$

Using subset weights allows for greater precision in selecting the alternative MT weights as both solutions can be applied to the same estimator.

3.5 Simulations

The simulations presented in the following section aim to demonstrate the performance of the alternative double GWSM estimators based on the choice of auxiliary variables used to compute the alternative MT weights. The conclusion of this section is meant as a practical guide for computing alternative MT weights.

3.5.1 Populations

The populations used in the simulations are generated to replicate the setup of La Poste, where U_F represents the population of addresses, U_M the population of boxes, and U_T the population of rounds. To ensure the validity of the results presented in sections 3.3 and 3.4, the links between populations are MtO-MtO. An address i is linked to a box j (resp. j is linked to a round k) if j contains letters to i (resp. if k deliver the letters contained in j).

La Poste has access to extensive historical data on rounds, including information such as traffic volume and details on sampled boxes within each round. In our simulations, we consider a target population U_T of 6,750 rounds selected from this historical data. The frame population U_F of size 5,816,907 and the intermediate population U_M of size 400,011 are obtained by generating numbers of addresses in a box and numbers of boxes in a round from distributions observed in La Poste's historical data.

Our objective is to estimate the total of three variables of interest on the target population U_T. The first variable of interest is the traffic t, which is derived from La Poste data history after modification for confidentiality reasons. The second variable of interest, u, follows a uniform distribution. The third variable of interest, q, is proportional to the inverse of the traffic. The bounds of the uniform distributions of variable u and the coefficient of proportionality of variable q are selected in such a way that all variables have the same mean as the traffic variable.

Samples from the target population U_T are obtained through double indirect sampling, with samples from the frame population U_F drawn using SRSWOR. In that case, we saw in Sections 3.3 and 3.4 that, for a box $j \in U_M$, the $\theta^{FM,opt}$-optimal MT weights before standardization reduces to the number N_{Fj} of addresses in box j and the optimal FM weights reduces to $1/N_{Fj}$. In the simulations, we compare five double GWSM estimators, each with optimal FM weights and different MT weights. The five considered MT weights are computed by using the following variables x_j, $j \in U_M$:

1. the number of addresses in a box N_{Fj}, which gives the $\theta^{FM,opt}$-optimal MT weights $\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}) = \frac{N_{Fj}}{N_{Fk}}$,
2. a constant variable D, $D \in \mathbb{R}$, which gives the usual MT weight $D_k = \frac{1}{N_{Mk}}$,

3. a variable *alt* N_{Fj} correlated with N_{Fj},
4. a variable *alt* y_j correlated with $y_j = y_k \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$,
5. and a variable *alt* $\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$ correlated with the optimal weights.

We aim to observe the behavior of the estimators obtained using *alt* N_{Fj}, *alt* y_j, and *alt* $\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$ when the correlation to respectively N_{Fj}, y_j, and $\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$ varies. Five correlation values, namely 0.1, 0.3, 0.5, 0.7, and 0.9, are considered for each variable. To achieve this, for each $j \in U_M$, a different value of *alt* N_{Fj} (resp. *alt* y_j, *alt* $\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$) is generated for each correlation. The values of *alt* N_{Fj} (resp. *alt* y_j, *alt* $\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$), $j \in U_M$, are obtained by taking a random value between $0.5N_{Fj}$ and $1.5N_{Fj}$ (resp. $0.5y_j$ and $1.5y_j$, $0.5\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$ and $1.5\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$) using a uniform distribution. Then, for each $j \in U_M$, a uniformly distributed random number is added to alter the correlation value.

3.5.2 Parameters and results

We draw $R = 10,000$ samples of size 1,000 in the frame population using the SRSWOR sampling design. For each r-th sample, $r = 1, \ldots, R$, we compute the value of the five double GWSM estimators considered in subsection 3.5.1. Let us denote $\hat{t}_{y2}^{(r)}$ a double GWSM estimate computed for the r-th sample, $r = 1, \ldots, R$. For each double GWSM estimator $\hat{t}_{y2}$ introduced in subsection 3.5.1 we have computed the Monte Carlo relative bias :

$$RB_{MC}(\hat{t}_{y2}) = 100 \times \frac{1}{R} \sum_{r=1}^{R} \frac{\hat{t}_{y2}^{(r)} - t_y}{t_y},$$

and the Relative Root Mean Square Error (RRMSE) with respect to the optimal double GWSM estimator $\hat{t}_{y2}^{opt}$:

$$RRMSE_{MC}(\hat{t}_{y2}) = \sqrt{\frac{MSE_{MC}(\hat{t}_{y2})}{MSE_{MC}(\hat{t}_{y2}^{opt})}}$$

where

$$MSE_{MC}(\hat{t}_{y2}) = R^{-1} \sum_{r=1}^{R} \left(\hat{t}_{y2}^{(r)} - R^{-1} \sum_{r=1}^{R} \hat{t}_{y2}^{(r)} \right)^2 .$$

Figures 3.5, 3.6 and 3.7 display the values of the RRMSE of each double GWSM estimator for different correlations, with respectively variables t, u and q as the variable of interest. As the absolute value of the relative bias of all estimators considered remains below 0.3%, it is not reported.

The estimators obtained with *alt* $\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$ are consistently less precise than the estimators obtained with *alt* N_{Fj} in all figures, regardless of the correlation. Noting that the

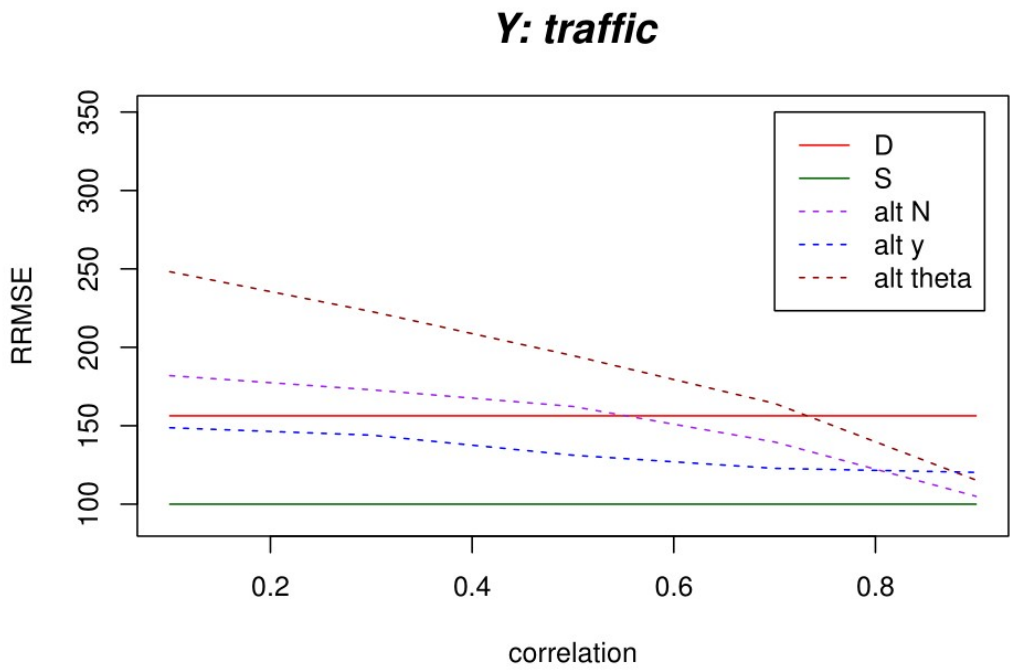

FIGURE 3.5 – Relative efficiency of the alternative double GWSM estimates compared to optimal double GWSM estimates of the total traffic.

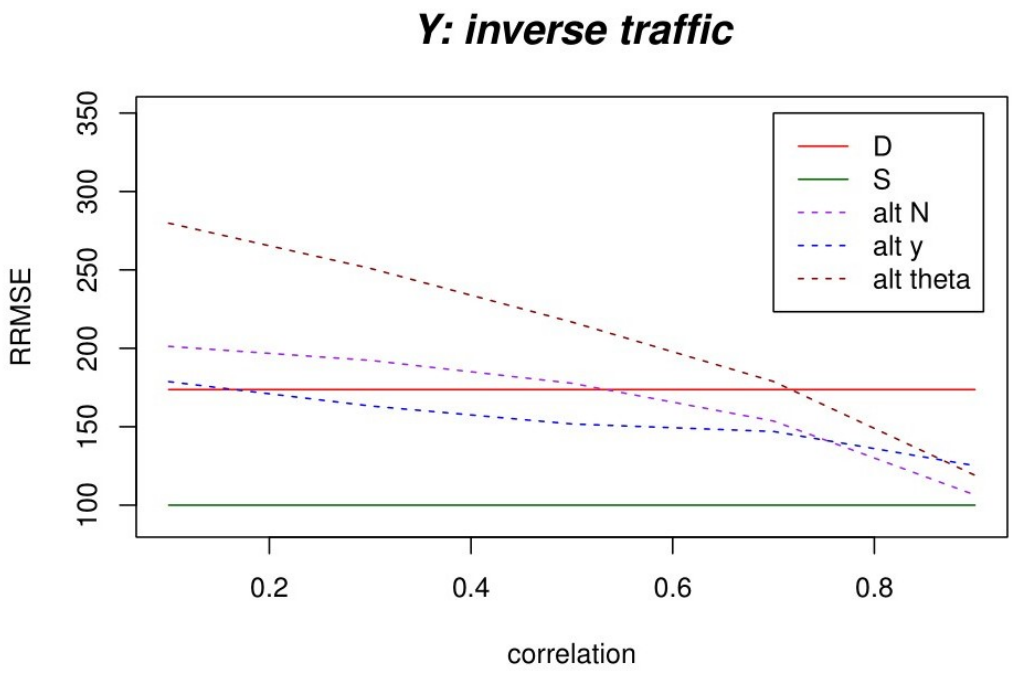

FIGURE 3.6 – Relative efficiency of the alternative double GWSM estimates compared to optimal double GWSM estimates of the total of the inverse traffic.

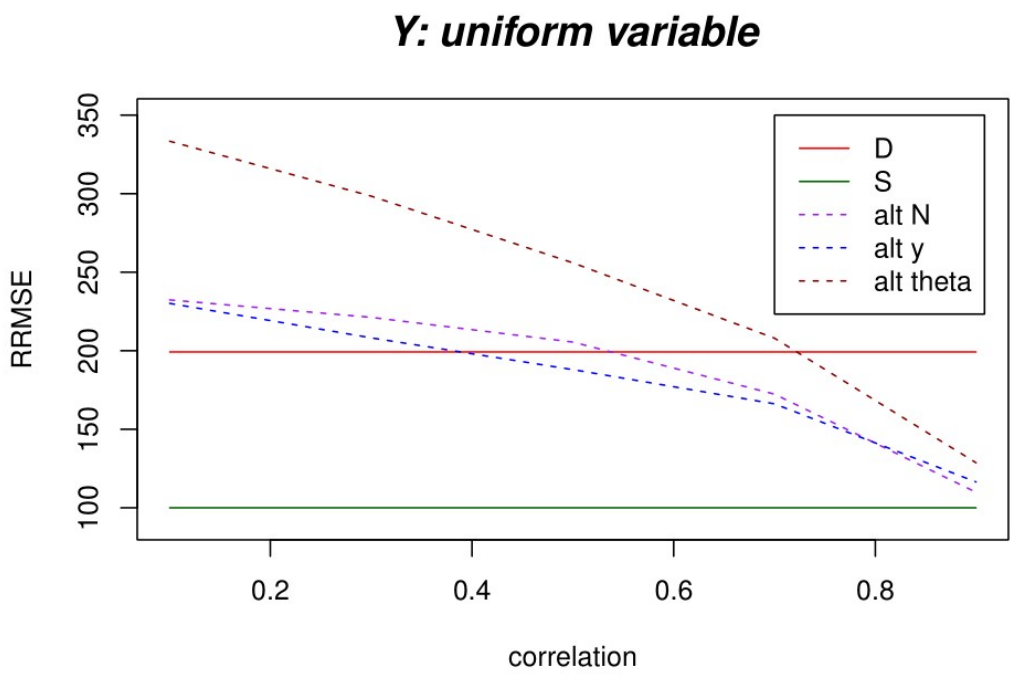

FIGURE 3.7 – Relative efficiency of the alternative double GWSM estimates compared to optimal double GWSM estimates of the total of the uniform variable.

computation of *alt* $\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$ requires knowledge of the $\theta^{FM,opt}$-optimal MT weights N_{Fj}/N_{Fk} for some cases j, making it more costly to use than *alt* N_{Fj}, we do not recommend using *alt* $\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$ to compute alternative MT weights.

The precision of the estimators obtained with *alt* N_{Fj} is consistently the best in all figures when the correlation is higher than 0.8. However, in figures 3.6 and 3.7, when the correlation is low, the usual estimator performs better. When the correlation is reasonably high (above 0.1 for figure 3.6 and 0.4 for figure 3.7) yet below 0.8, the estimator obtained with *alt* y_j is the best. In some cases, as in figure 3.5, the estimator obtained with *alt* y_j consistently performs better than the usual estimator.

The results suggest that while the usual double GWSM estimator may not be the most efficient alternative in terms of variance, its loss in efficiency compared to the optimal double GWSM estimator is still acceptable, with a multiplier ranging from 1.6 to 2. However, it is important to note that these simulations were designed specifically for situations similar to La Poste, where the loss in efficiency between the usual and optimal double GWSM estimators is around 1.6 (as discussed in Medous et al. (2023a)).

To summarize, we recommend the use of alternative MT weights obtained with *alt* N_{Fj} due to their low computation cost and close proximity to the $\theta^{FM,opt}$-optimal MT weights, provided that the correlation between *alt* N_{Fj} and N_{Fj} is high. Alternatively, when the

correlation between all potential *alt* N_{Fj} and N_{Fj} is too low, it may be beneficial to choose a variable *alt* y_j with a reasonable correlation to y_j instead. However, this method requires to know the value of y_j for many units $j \in s_M$ and for all variables of interest, which can lead to an increase in observation and computation costs. The usual double GWSM estimator is the simplest to compute and show an acceptable loss of precision in our simulations. Medous et al. (2023a) study the impact of the distribution of links between populations on the precision of the usual double GWSM and identify a few situations where the usual double GWSM estimator performs poorly compared to the optimal simple GWSM estimator.

3.5.3 Perspectives for La Poste application

Currently, La Poste calculates the total quarterly traffic carried by rounds through double indirect sampling and double GWSM, using a frame population of addresses and an intermediate population of boxes. The samples in the frame population are obtained through stratified SRSWOR, satisfying both the Δ-property and c_k-condition, such that there exists a set of optimal FM weights and $\theta^{FM,opt}$-optimal MT weights given by Proposition 3.3.3.

In La Poste case, N_{Fj} is the number of addresses in the box j and N_{Fk} is the number of addresses in the round k. The optimal FM weight for a box j is $1/n_{Fj}$ and the $\theta^{FM,opt}$-optimal MT weight between j and round k is N_{Fj}/N_{Fk} if k and j are linked, 0 otherwise. Unfortunately, in La Poste case, the number of addresses N_{Fk} in round k is unknown for all $k \in U_T$ and the $\theta^{FM,opt}$-optimal MT weights cannot be computed for La Poste estimator. Thus, La Poste currently uses optimal FM weights and usual MT weights to compute its double GWSM estimator. However, in the future, the sample size will be reduced, leading to a loss of precision in the estimators. Therefore, La Poste aims to enhance their MT weights to limit this loss of precision.

La Poste intends to use available auxiliary information on the population of boxes to compute alternative MT weights. This information includes dimensions of the boxes, which are observed for all boxes linked to an indirectly sampled round. However, using this variable to compute alternative MT weights did not result in a significant improvement in precision compared to the usual MT weights.

La Poste is currently exploring the use of another observed variable, namely the traffic within the box, which is available for all boxes of an indirectly sampled round. This variable is assumed to be correlated with the variable y_j defined in section 3.4 for a set of round that is yet to be identified. Work is ongoing to evaluate the potential of using this variable to compute alternative MT weights and improve the precision of the estimators.

As an additional source of information, La Poste has access to the RAO dataset, which

contains a non-updated number of addresses in boxes. While this dataset shows potential for computing alternative MT weights, it currently suffers from inconsistencies and cannot be used at the moment.

3.6 Conclusion

The use of indirect sampling coupled with double GWSM can reduce the number of observed links, but this comes at the cost of a potential decrease in accuracy of the double GWSM estimator, as discussed in Medous et al. (2023a). In the present paper, we demonstrate the existence of optimal weights for double GWSM that can be used to reduce costs without deteriorating the precision of the estimators relative to an optimal simple GWSM estimator. We have also proposed alternative weights for situations where the optimal weights are not computable, and have demonstrated their relative efficiency in a case similar to La Poste example. To gain a deeper understanding of how alternative weights work, it may be necessary to develop models for both the variable of interest (Särndal et al. (1992)) and the links (Xu and Lavallée (2009)). However, this topic is not in the scope of our paper, which is primarily focused on optimal double GWSM.

3.7 Appendix

Propositions from Medous et al. (2023a)

For any unit k from U_T, let us denote U_{Fk} of size N_{Fk} the set of units in U_F linked to k. For MtO links, the sets U_{Fk}, $k \in U_T$ form a partition of U_F and we can assume units in U_F are ordered with respect to the sub-populations U_{Fk}. We denote by $\mathbf{\Delta} = (\Delta_{ii'})_{i,i' \in U_F}$ the covariance matrix of size $N_F \times N_F$. For given k, k' from U_T, let $\mathbf{\Delta}_{kk'} = (\Delta_{ii'})_{i \in U_{Fk}, i' \in U_{Fk'}}$ be the submatrix of $\mathbf{\Delta}$ corresponding to elements in position i (resp. i') if i (resp. i') is linked to k (resp. k') and let $\mathbf{\Delta}_k$ be the submatrix $\mathbf{\Delta}_{kk}$. For MtO links, these submatrices form a partition of $\mathbf{\Delta}$:

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

Let us denote by $\mathbb{1}_k$ the vector of ones of size N_{Fk} and by A^t the transpose of a given matrix or vector A.

The following definition was introduced by Medous et al. (2023a) to characterize all sampling designs for which a set of optimal standardized link weight exists.

Definition 3.7.1 (Δ-property for U_T). For MtO links, a sampling design is said to satisfy the

Δ-property for U_T if, for any k in U_T, $\mathbf{\Delta}_k$ is invertible and for any $k \neq k'$ in U_T,

$$\mathbf{\Delta}_{k,k' \neq k} = c_{kk'} \mathbb{1}_k \mathbb{1}_{k'}^t \quad \text{with} \quad c_{kk'} \quad \text{not depending on } i \text{ and } i'.$$

The following proposition, introduced by Medous et al. (2023a), gives the expression of the set of optimal link weights for all designs satisfying the Δ-property.

Proposition 3.7.1 (Medous et al. (2023a))**.** *If the links are MtO and the sampling design satisfies the $\mathbf{\Delta}$-property for U_T, then there exists a unique set of standardized link weights minimizing the variance of $\hat{t}_{y1}$ for all variable of interest y given by :*

$$(\tilde{\theta}_{ik}^{opt})_{i \in U_{Fk}} = \mathbf{\Delta}_k^{-1} \mathbb{1}_k \left(\mathbb{1}_k^t \mathbf{\Delta}_k^{-1} \mathbb{1}_k \right)^{-1}, \text{ for all } k \in U_T.$$

This proposition holds as long as the matrix Δ verify the Δ-property for U_T, but Δ does not need to be a covariance matrix. Let us consider a simple GWSM estimator $\hat{t}_{y1}$, given by equation (3.1), obtained using some standardized link weights $\tilde{\theta}_{ik}, i \in U_F, k \in U_T$ and the optimal simple GWSM estimator $\hat{t}_{y1}^{opt}$ obtained using optimal link wieghts. The following proposition, introduced by Medous et al. (2023a), gives the loss of efficiency between $\hat{t}_{y1}^{opt}$ and $\hat{t}_{y1}$ when the sampling design satisfy the Δ-property for U_T.

Proposition 3.7.2 (Medous et al. (2023a))**.** *If the links are MtO and the sampling design satisfies the $\mathbf{\Delta}$-property for U_T, then the loss of efficiency compared with optimal link weights $\tilde{\theta}_{ik}^{\text{opt}}$, $i \in U_F, k \in U_T$, is given by :*

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

where $\hat{t}_{\tilde{\theta}_k^{opt}} = \sum_{i \in s_F} d_i \tilde{\theta}_{ik}^{opt}$ is the Horvitz-Thompson estimator of the total $t_{\tilde{\theta}_k^{opt}} = \sum_{i \in U_F} \tilde{\theta}_{ik}^{opt} = 1$, and $\hat{t}_{\tilde{\theta}_k - \tilde{\theta}_k^{opt}} = \sum_{i \in s_F} d_i (\tilde{\theta}_{ik} - \tilde{\theta}_{ik}^{opt})$ is the Horvitz-Thompson estimator of the total $t_{\tilde{\theta}_k - \tilde{\theta}_k^{opt}} = 0$, for all $k \in U_T$.

Sherman-Morrison formula

This formula, introduced by Sherman and Morrison (1950), is used in the proof of Propositions 3.3.2 and 3.3.3. Let A be a square matrix of size $n \times n$, $n \in \mathbb{N}$, and u, v two vectors of size n. If A is invertible and $v^t A^{-1} u \neq -1$, then $A + uv^t$ is invertible and

$$\left(A + uv^t \right)^{-1} = A^{-1} - \frac{A^{-1} u v^t A^{-1}}{1 + v^t A^{-1} u}.$$

Proof of Proposition 3.3.1

Let us consider the double GWSM estimator $\hat{t}_{y2}$ computed for any variable of interest y and any set of weights $\tilde{\theta}_{jk}^{MT}$, $j \in U_M$, $k \in U_T$. The double GWSM estimator can be rewritten as a GWSM estimator between U_F and U_M of a variable $z_j = \sum_{k \in U_T} \tilde{\theta}_{jk}^{MT} y_k$:

$$\hat{t}_{y2} = \sum_{i \in s_F} d_i \left(\sum_{k \in U_T} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} y_k \right) = \sum_{i \in s_F} d_i \left(\sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} z_j \right).$$

Since the Δ-property holds for U_M, we can apply Proposition 3.7.1 to the GWSM estimator $\sum_{i \in s_F} d_i \left(\sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} z_j \right)$ of the total $\sum_{j \in U_M} z_j$. We obtain the unique set of optimal weights between U_F and U_M :

$$(\tilde{\theta}_{ij}^{FM,opt})_{i \in U_{Fj}} = \left(\mathbf{\Delta}_j^M \right)^{-1} \mathbb{1}_j^M \left(\mathbb{1}_j^{Mt} \left(\mathbf{\Delta}_j^M \right)^{-1} \mathbb{1}_j^M \right)^{-1}, \text{ for all } j \in U_M,$$

which concludes the proof. □

Proof of Proposition 3.3.2

Let us consider any set of FM weights $\tilde{\theta}_{ij}^{FM}$, $i \in U_F$, $j \in U_M$. The double GWSM estimator computed can be rewritten as a simple GWSM estimator between U_M and U_T with sampling weights $\tilde{d}_j = \sum_{i \in U_{Fj}} d_i \tilde{\theta}_{ij}^{FM}$:

$$\hat{t}_{y2} = \sum_{i \in s_F} d_i \left(\sum_{k \in U_T} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} y_k \right) = \sum_{j \in s_M} \tilde{d}_j \left(\sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} z_j \right).$$

We can get the matrix $\tilde{\mathbf{\Delta}} = \left(\tilde{\Delta}_{jj'} \right)_{j,j' \in U_M}$ of size $N_M \times N_M$, as defined in equation (3.8), by rewriting the variance of $\hat{t}_{y2}$:

$$\begin{aligned}
\text{Var}(\hat{t}_{y1}) &= \sum_{i \in U_F} \sum_{i' \in U_F} \Delta_{ii'} \sum_{k \in U_T} \sum_{j \in U_M} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{jk}^{MT} y_k \sum_{k' \in U_T} \sum_{j' \in U_M} \tilde{\theta}_{i'j'}^{FM} \tilde{\theta}_{j'k'}^{MT} y_{k'} \\
&= \sum_{j \in U_M} \sum_{j' \in U_M} \left(\sum_{i \in U_F} \sum_{i' \in U_F} \Delta_{ii'} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{i'j'}^{FM} \right) \sum_{k \in U_T} \tilde{\theta}_{jk}^{MT} y_k \sum_{k' \in U_T} \tilde{\theta}_{j'k'}^{MT} y_{k'} \\
&= \sum_{j \in U_M} \sum_{j' \in U_M} \tilde{\Delta}_{jj'} \sum_{k \in U_T} \tilde{\theta}_{jk}^{MT} y_k \sum_{k' \in U_T} \tilde{\theta}_{j'k'}^{MT} y_{k'}
\end{aligned}$$

with $\tilde{\Delta}_{jj'} = \sum_{i \in U_F} \sum_{i' \in U_F} \Delta_{ii'} \tilde{\theta}_{ij}^{FM} \tilde{\theta}_{i'j'}^{FM}$.

Let us show that the Δ-property for U_T holds for $\tilde{\mathbf{\Delta}}$. Let $\tilde{\mathbf{\Delta}}_k$ be the submatrix of $\tilde{\mathbf{\Delta}}$ corresponding to the elements in position j and j' such that j and j' are linked to k. We

have :

$$\tilde{\mathbf{\Delta}}_k = \left(\sum_{i \in U_F} \sum_{i' \in U_F} \Delta_{ii'} \tilde{\theta}^{FM}_{ij} \tilde{\theta}^{FM}_{i'j'} \right)_{j,j' \in U_{Mk}} = \left((\tilde{\theta}^{FM}_{ij})^t_{i \in U_{Fj}} \mathbf{\Delta}^M_{jj'} (\tilde{\theta}^{FM}_{i'j'})_{i' \in U_{Fj'}} \right)_{j,j' \in U_{Mk}}$$

Since the Δ-property for U_M and the c_k-condition are verified, $\mathbf{\Delta}^M_{jj'} = \mathbf{\Delta}^M_j$ if $j = j'$ and $c_k \mathbb{1}^M_j \mathbb{1}^{Mt}_{j'}$ otherwise, and, for all $j, j' \in U_{Mk}$,

$$(\tilde{\theta}^{FM}_{ij})^t_{i \in U_{Fj}} \mathbf{\Delta}^M_{jj'} (\tilde{\theta}^{FM}_{i'j'})_{i' \in U_{Fj'}} = \begin{cases} (\tilde{\theta}^{FM}_{ij})^t_{i \in U_{Fj}} \Delta^M_j (\tilde{\theta}^{FM}_{i'j})_{i' \in U_{Fj}}, & \text{if } j = j', \\ c_k (\tilde{\theta}^{FM}_{ij})^t_{i \in U_{Fj}} \mathbb{1}^M_j \mathbb{1}^{Mt}_{j'} (\tilde{\theta}^{FM}_{i'j'})_{i' \in U_{Fj'}} = c_k & \text{otherwise,} \end{cases}$$

because, for all $j \in U_M$, $(\tilde{\theta}^{FM}_{ij})^t_{i \in U_{Fj}} \mathbb{1}^M_j = 1$ by standardisation.

$\tilde{\mathbf{\Delta}}_k$ is the square matrix containing the values of $(\tilde{\theta}^{FM}_{ij})^t_{i \in U_{Fj}} \mathbf{\Delta}^M_{jj'} (\tilde{\theta}^{FM}_{i'j'})_{i' \in U_{Fj'}}$ for $j, j' \in U_{Mk}$. Since $(\tilde{\theta}^{FM}_{ij})^t_{i \in U_{Fj}} \Delta^M_j (\tilde{\theta}^{FM}_{i'j})_{i' \in U_{Fj}}$, $j \in U_{Mk}$, and c_k are real numbers, this means that $\tilde{\mathbf{\Delta}}_k$ is a square matrix with the values $(\tilde{\theta}^{FM}_{ij})^t_{i \in U_{Fj}} \Delta^M_j (\tilde{\theta}^{FM}_{i'j})_{i' \in U_{Fj}}$, $j \in U_{Mk}$, on its diagonal and c_k everywhere else. Let $\tilde{\mathbb{1}}_k$ be the N_{Mk} dimensional vector of ones and, for any vector u, let us denote $diag(u)$ the diagonal matrix containing the elements of u. We can rewrite $\tilde{\mathbf{\Delta}}_k$, $k \in U_T$, as follows :

$$\tilde{\mathbf{\Delta}}_k = diag \left(\left((\tilde{\theta}^{FM}_{ij})^t_{i \in U_{Fj}} \Delta^M_j (\tilde{\theta}^{FM}_{i'j})_{i' \in U_{Fj}} - c_k \right)_{j \in U_{Mk}} \right) + c_k \tilde{\mathbb{1}}_k \tilde{\mathbb{1}}^t_k.$$

Using the Sherman-Morrison formula, we can show that $\tilde{\mathbf{\Delta}}_k$ is invertible if and only if (iif) $(\tilde{\theta}^{FM}_{ij})^t_{i \in U_{Fj}} \Delta^M_j (\tilde{\theta}^{FM}_{i'j})_{i' \in U_{Fj}} - c_k \neq 0$ for all $j \in U_{Mk}$ and
$c_k \tilde{\mathbb{1}}^t_k diag \left(\left((\tilde{\theta}^{FM}_{ij})^t_{i \in U_{Fj}} \Delta^M_j (\tilde{\theta}^{FM}_{i'j})_{i' \in U_{Fj}} - c_k \right)_{j \in U_{Mk}} \right)^{-1} \tilde{\mathbb{1}}_k \neq -1.$

Let $\tilde{\mathbf{\Delta}}_{kk'}$, $k \neq k' \in U_T$ be the submatrix of $\tilde{\mathbf{\Delta}}$ of size $N_{Mk} \times N_{Mk'}$ corresponding to the elements in position j and j' such that j (resp. j') is linked to k (resp. k'). Since the links are of type MtO-MtO, $k \neq k'$ implies that, for all $j \in U_{Mk}$ and $j' \in U_{Mk'}$, $j \neq j'$. Thus, by using the c_k-condition, we have :

$$\begin{aligned} \tilde{\mathbf{\Delta}}_{kk'} &= \left((\tilde{\theta}^{FM}_{ij})^t_{i \in U_{Fj}} \mathbf{\Delta}^M_{jj'} (\tilde{\theta}^{FM}_{i'j'})_{i' \in U_{Fj'}} \right)_{\substack{j \in U_{Mk} \\ j' \in U_{Mk'}}} \\ &= \left(c_{kk'} (\tilde{\theta}^{FM}_{ij})^t_{i \in U_{Fj}} \mathbb{1}^M_j \mathbb{1}^{Mt}_{j'} (\tilde{\theta}^{FM}_{i'j'})_{i' \in U_{Fj'}} \right)_{\substack{j \in U_{Mk} \\ j' \in U_{Mk'}}} = c_{kk'} \tilde{\mathbb{1}}_k \tilde{\mathbb{1}}^t_{k'}, \end{aligned}$$

which proves that $\tilde{\mathbf{\Delta}}$ satisfies the Δ-property for U_T.

Then we can apply Proposition 3.7.1 to the GWSM estimator $\sum_{j \in s_M} \tilde{d}_j \left(\sum_{k \in U_T} \tilde{\theta}^{MT}_{jk} y_k \right)$

and we obtain the unique set of optimal weights between U_M and U_T :

$$\left(\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM})\right)_{j\in U_{Mk}} = \tilde{\mathbf{\Delta}}_k^{-1}\tilde{\mathbb{1}}_k\left(\tilde{\mathbb{1}}_k^t\tilde{\mathbf{\Delta}}_k^{-1}\tilde{\mathbb{1}}_k\right)^{-1}, \text{ for all } j\in U_M,$$

which concludes the proof. □

Proof of Proposition 3.3.3

We assume that the Δ-property for U_M and the c_k-condition hold. Let us first compute the value of $\tilde{\mathbf{\Delta}}_k$, $k\in U_T$, defined above for the optimal FM weights given by Proposition 3.3.1, $(\tilde{\theta}_{ij}^{FM,opt})_{i\in U_{Fj}} = \left(\mathbf{\Delta}_j^M\right)^{-1}\mathbb{1}_j^M\lambda_j^{-1}$, for all $j\in U_M$, with $\lambda_j = \mathbb{1}_j^{Mt}\left(\mathbf{\Delta}_j^M\right)^{-1}\mathbb{1}_j^M$:

$$\begin{aligned}\tilde{\mathbf{\Delta}}_k &= \left((\tilde{\theta}_{ij}^{FM,opt})_{i\in U_{Fj}}^t\mathbf{\Delta}_{jj'}^M(\tilde{\theta}_{i'j'}^{FM,opt})_{i'\in U_{Fj'}}\right)_{j,j'\in U_{Mk}}\\ &= \left(\frac{\mathbb{1}_j^{Mt}\left(\mathbf{\Delta}_j^M\right)^{-1}}{\lambda_j}\mathbf{\Delta}_{jj'}^M\frac{\left(\mathbf{\Delta}_{j'}^M\right)^{-1}\mathbb{1}_{j'}^M}{\lambda_{j'}}\right)_{j,j'\in U_{Mk}}\\ &= diag\left((1/\lambda_j - c_k)_{j\in U_{Mk}}\right) + c_k\tilde{\mathbb{1}}_k\tilde{\mathbb{1}}_k^t\end{aligned}$$

since for $j,j'\in U_{Mk}$, $\mathbb{1}_j^{Mt}\left(\mathbf{\Delta}_j^M\right)^{-1}\mathbf{\Delta}_{jj'}^M\left(\mathbf{\Delta}_{j'}^M\right)^{-1}\mathbb{1}_{j'}^M = 1/\lambda_j$ if $j=j'$, and c_k otherwise, with the value of c_k, $k\in U_T$, given by the c_k-condition.

Since $\mathbf{\Delta}_j^M$, $j\in U_M$, is invertible thanks to the Δ-property for U_M and $1/\lambda_j - c_k = (1-c_k\lambda_j)/\lambda_j$, $j\in U_M$, is a real number, we get by using the Sherman-Morrison formula that $\tilde{\mathbf{\Delta}}_k$ is invertible iif :

$$\begin{aligned}&c_k\tilde{\mathbb{1}}_k^t diag\left((1/\lambda_j - c_k)_{j\in U_{Mk}}\right)^{-1}\tilde{\mathbb{1}}_k \neq -1\\ \iff\ &c_k\sum_{j\in U_{Mk}}\lambda_j/\left(1-c_k\lambda_j\right) \neq -1 \text{ and } c_k\lambda_j\neq 1 \text{ for all } j\in U_{Mk}\end{aligned}$$

and we have :

$$\tilde{\mathbf{\Delta}}_k^{-1} = \tilde{D}^{-1} - \frac{c_k\tilde{D}^{-1}\tilde{\mathbb{1}}_k\tilde{\mathbb{1}}_k^t\tilde{D}^{-1}}{1+c_k\mathbb{1}_k^t\tilde{D}^{-1}\tilde{\mathbb{1}}_k},$$

with $\tilde{D}^{-1} = diag\left((1/\lambda_j - c_k)_{j\in U_{Mk}}\right)^{-1} = diag\left((\lambda_j/(1-c_k\lambda_j)_{j\in U_{Mk}}\right)$. Then we can compute

$$\tilde{\mathbf{\Delta}}_k^{-1}\tilde{\mathbb{1}}_k = \tilde{D}^{-1}\tilde{\mathbb{1}}_k - \frac{c_k\tilde{D}^{-1}\tilde{\mathbb{1}}_k\tilde{\mathbb{1}}_k^t\tilde{D}^{-1}\tilde{\mathbb{1}}_k}{1+c_k\mathbb{1}_k^t\tilde{D}^{-1}\tilde{\mathbb{1}}_k} = \left(1-\frac{c_k\mathbb{1}_k^t\tilde{D}^{-1}\tilde{\mathbb{1}}_k}{1+c_k\mathbb{1}_k^t\tilde{D}^{-1}\tilde{\mathbb{1}}_k}\right)\tilde{D}^{-1}\tilde{\mathbb{1}}_k$$

and

$$\tilde{\mathbb{1}}_k^t\tilde{\mathbf{\Delta}}_k^{-1}\tilde{\mathbb{1}}_k = \left(1-\frac{c_k\mathbb{1}_k^t\tilde{D}^{-1}\tilde{\mathbb{1}}_k}{1+c_k\mathbb{1}_k^t\tilde{D}^{-1}\tilde{\mathbb{1}}_k}\right)\tilde{\mathbb{1}}_k^t\tilde{D}^{-1}\tilde{\mathbb{1}}_k.$$

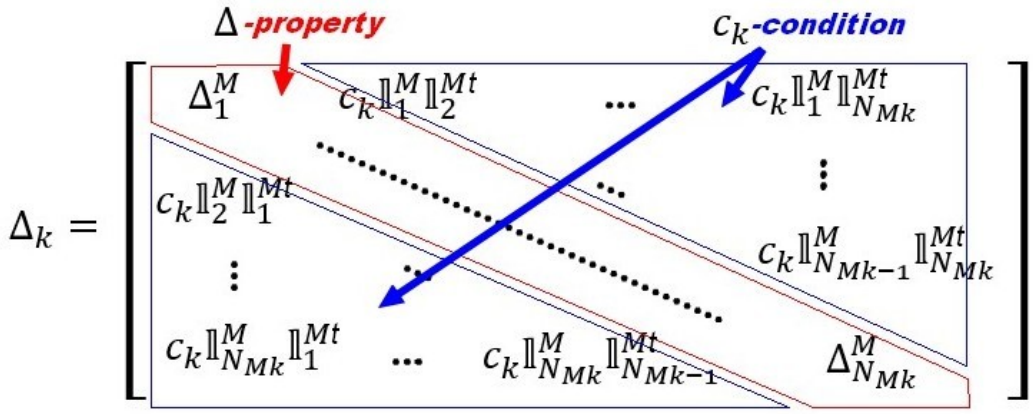

FIGURE 3.8 – Expression of $\mathbf{\Delta}_k$ when the Δ-property holds for U_M and the c_k-condition is verified.

Finally, we get :

$$\begin{aligned}\left(\tilde{\theta}_{jk}^{MT}(\theta^{FM,opt})\right)_{j\in U_{Mk}} &= \tilde{\mathbf{\Delta}}_k^{-1}\tilde{\mathbb{1}}_k\left(\tilde{\mathbb{1}}_k^t\tilde{\mathbf{\Delta}}_k^{-1}\tilde{\mathbb{1}}_k\right)^{-1}\\ &= \frac{\tilde{D}^{-1}\tilde{\mathbb{1}}_k}{\tilde{\mathbb{1}}_k^t\tilde{D}^{-1}\tilde{\mathbb{1}}_k} = \left(\frac{\lambda_j/\left(1-c_k\lambda_j\right)}{\sum_{j'\in U_{Mk}}\lambda_{j'}/\left(1-c_k\lambda_{j'}\right)}\right)_{j\in U_{Mk}},\end{aligned}$$

since $\tilde{D}^{-1} = diag\left((\lambda_j/(1-c_k\lambda_j)_{j\in U_{Mk}}\right)$.

We will now show that the optimal weights for simple GWSM estimators given by Proposition 3.7.1 can be rewritten as an expression of the optimal FM weights given by Proposition 3.3.1 and the $\theta^{FM,opt}$-optimal MT weights if the Δ-property for U_M and the c_k-condition hold, and if $c_k\sum_{j\in U_{Mk}}\lambda_j/\left(1-c_k\lambda_j\right)\neq 1$ and $c_k\lambda_j\neq 1$ for all $j\in U_{Mk}$.

Let us compute the value of

$$(\tilde{\theta}_{ik}^{opt})_{i\in U_{Fk}} = \mathbf{\Delta}_k^{-1}\mathbb{1}_k\left(\mathbb{1}_k^t\mathbf{\Delta}_k^{-1}\mathbb{1}_k\right)^{-1},\ \text{for all } k\in U_T$$

when the links are MtO-MtO, the Δ-property holds for U_M and the c_k-condition is verified.

For a set of square matrices $(A_j)_{j\in U_{Mk}}$, let $diag\left((A_j)_{j\in U_{Mk}}\right)$ be the block diagonal matrix whose main-diagonal blocks are the matrices $(A_j)_{j\in U_{Mk}}$. When the links are MtO-MtO, the Δ-property holds for U_M and the c_k-condition is verified, the matrix $\mathbf{\Delta}_k$ defined in section 3.2.1 is given in figure 3.8 and can be rewritten for all $k\in U_T$ as

$$\mathbf{\Delta}_k = diag\left(\left(\mathbf{\Delta}_j^M - c_k\mathbb{1}_j^M\mathbb{1}_j^{Mt}\right)_{j\in U_{Mk}}\right) + c_k\mathbb{1}_k\mathbb{1}_k^t.$$

We will first show that $\mathbf{\Delta}_k$ is invertible iif $c_k\sum_{j\in U_{Mk}}\lambda_j/\left(1-c_k\lambda_j\right)\neq 1$ and $c_k\lambda_j\neq$

1 for all $j \in U_{Mk}$. Let us denote $D = diag\left(\left(\mathbf{\Delta}_j^M - c_k \mathbb{1}_j^M \mathbb{1}_j^{Mt}\right)_{j\in U_{Mk}}\right)$ the square matrix of size N_{Fk}. Using the Sherman-Morrison formula, we can show that $\mathbf{\Delta}_k$ is invertible iif D is invertible and $c_k \mathbb{1}_k^t D^{-1} \mathbb{1}_k \neq -1$.

Since $D^{-1} = diag\left(\left(\mathbf{\Delta}_j^M - c_k \mathbb{1}_j^M \mathbb{1}_j^{Mt}\right)_{j\in U_{Mk}}\right)^{-1} = diag\left(\left(\mathbf{\Delta}_j^M - c_k \mathbb{1}_j^M \mathbb{1}_j^{Mt}\right)^{-1}_{j\in U_{Mk}}\right)$, D is invertible iif $\mathbf{\Delta}_j^M - c_k \mathbb{1}_j^M \mathbb{1}_j^{Mt}$ is invertible for all $j \in U_{Mk}$.
Since $\mathbf{\Delta}_j^M$, $j \in U_M$, is invertible thanks to the Δ-property for U_M, we can show using the Sherman-Morrison formula that $\mathbf{\Delta}_j^M - c_k \mathbb{1}_j^M \mathbb{1}_j^{Mt}$ is invertible iif $c_k \mathbb{1}_j^{Mt} \left(\mathbf{\Delta}_j^M\right)^{-1} \mathbb{1}_j^M = c_k \lambda_j \neq 1$, and we have :

$$\begin{aligned}\left(\mathbf{\Delta}_j^M - c_k \mathbb{1}_j^M \mathbb{1}_j^{Mt}\right)^{-1} &= \left(\mathbf{\Delta}_j^M\right)^{-1} + \frac{c_k \left(\mathbf{\Delta}_j^M\right)^{-1} \mathbb{1}_j^M \mathbb{1}_j^{Mt} \left(\mathbf{\Delta}_j^M\right)^{-1}}{1 - c_k \mathbb{1}_j^{Mt} \left(\mathbf{\Delta}_j^M\right)^{-1} \mathbb{1}_j^M} \\ &= \left(\mathbf{\Delta}_j^M\right)^{-1} + \frac{c_k \left(\mathbf{\Delta}_j^M\right)^{-1} \mathbb{1}_j^M \mathbb{1}_j^{Mt} \left(\mathbf{\Delta}_j^M\right)^{-1}}{1 - c_k \lambda_j}.\end{aligned}$$

Then we can compute D^{-1} :

$$D^{-1} = diag\left(\left(\left(\mathbf{\Delta}_j^M\right)^{-1} + \frac{c_k \left(\mathbf{\Delta}_j^M\right)^{-1} \mathbb{1}_j^M \mathbb{1}_j^{Mt} \left(\mathbf{\Delta}_j^M\right)^{-1}}{1 - c_k \lambda_j}\right)_{j\in U_{Mk}}\right).$$

Since $\mathbb{1}_k$ is the vector of size N_{Fk} composed of all vectors of size N_{Fj} $\mathbb{1}_j^M$, $j \in U_{Mk}$, we can rewrite $D^{-1}\mathbb{1}_k$ as :

$$D^{-1}\mathbb{1}_k = \left(\left(1 + \frac{c_k \lambda_j}{1 - c_k \lambda_j}\right)\left(\mathbf{\Delta}_j^M\right)^{-1} \mathbb{1}_j^M\right)_{j\in U_{Mk}} = \left(\frac{\left(\mathbf{\Delta}_j^M\right)^{-1} \mathbb{1}_j^M}{1 - c_k \lambda_j}\right)_{j\in U_{Mk}}$$

with $\left(\left(\mathbf{\Delta}_j^M\right)^{-1} \mathbb{1}_j^M\right)(1 - c_k\lambda_j)^{-1}$, $j \in U_{Mk}$, vectors of size N_{Fj}. We compute $\mathbb{1}_k^t D^{-1} \mathbb{1}_k$:

$$\mathbb{1}_k^t D^{-1} \mathbb{1}_k = \sum_{j\in U_{Mk}} \frac{\mathbb{1}_j^{Mt} \left(\mathbf{\Delta}_j^M\right)^{-1} \mathbb{1}_j^M}{1 - c_k \lambda_j} = \sum_{j\in U_{Mk}} \frac{\lambda_j}{1 - c_k \lambda_j}.$$

Then $\mathbf{\Delta}_k^{-1}$ is invertible iif $c_k \mathbb{1}_k^t D^{-1} \mathbb{1}_k = c_k \sum_{j\in U_{Mk}} \frac{\lambda_j}{1 - c_k \lambda_j} \neq -1$ and $c_k \lambda_j \neq 1$ for all $j \in U_{Mk}$ and its inverse is given by :

$$\mathbf{\Delta}_k^{-1} = D^{-1} - \frac{c_k D^{-1} \mathbb{1}_k \mathbb{1}_k^t D^{-1}}{1 + c_k \mathbb{1}_k^t D^{-1} \mathbb{1}_k}.$$

We can use Proposition 3.7.1 to compute the optimal weights $\tilde{\theta}_{ik}^{opt}$, $i \in U_F$, $k \in U_T$:

$$\begin{aligned}
\left(\tilde{\theta}_{ik}^{opt}\right)_{i\in U_{Fk}} &= \mathbf{\Delta}_k^{-1}\mathbb{1}_k \left(\mathbb{1}_k^t \mathbf{\Delta}_k^{-1}\mathbb{1}_k\right)^{-1} \\
&= \left(1 - \frac{c_k \mathbb{1}_k^t D^{-1}\mathbb{1}_k}{1 + c_k \mathbb{1}_k^t D^{-1}\mathbb{1}_k}\right) D^{-1}\mathbb{1}_k \times \left(\left(1 - \frac{c_k \mathbb{1}_k^t D^{-1}\mathbb{1}_k}{1 + c_k \mathbb{1}_k^t D^{-1}\mathbb{1}_k}\right) \mathbb{1}_k^t D^{-1}\mathbb{1}_k\right)^{-1} \\
&= \frac{D^{-1}\mathbb{1}_k}{\mathbb{1}_k^t D^{-1}\mathbb{1}_k}.
\end{aligned}$$

By plugging the values of $D^{-1}\mathbb{1}_k$ and $\mathbb{1}_k^t D^{-1}\mathbb{1}_k$ computed above, we get :

$$\begin{aligned}
\left(\tilde{\theta}_{ik}^{opt}\right)_{i\in U_{Fk}} &= \left(\frac{\left(\mathbf{\Delta}_j^M\right)^{-1}\mathbb{1}_j^M}{1 - c_k\lambda_j}\right)_{j\in U_{Mk}} \times \left(\sum_{j'\in U_{Mk}} \frac{\lambda_{j'}}{1 - c_k\lambda_{j'}}\right)^{-1} \\
&= \left(\frac{\left(\mathbf{\Delta}_j^M\right)^{-1}\mathbb{1}_j^M / \left(1 - c_k\lambda_j\right)}{\sum_{j'\in U_{Mk}} \lambda_{j'} / \left(1 - c_k\lambda_{j'}\right)}\right)_{j\in U_{Mk}} \\
&= \left(\frac{\left(\left(\mathbf{\Delta}_j^M\right)^{-1}\mathbb{1}_j^M / \lambda_j\right) \times \left(\lambda_j / \left(1 - c_k\lambda_j\right)\right)}{\sum_{j'\in U_{Mk}} \lambda_{j'} / \left(1 - c_k\lambda_{j'}\right)}\right)_{j\in U_{Mk}} \\
&= \left(\frac{\left(\mathbf{\Delta}_j^M\right)^{-1}\mathbb{1}_j^M}{\lambda_j}\right)_{j\in U_{Mk}} \times \frac{\lambda_j / \left(1 - c_k\lambda_j\right)}{\sum_{j'\in U_{Mk}} \lambda_{j'} / \left(1 - c_k\lambda_{j'}\right)} \\
&= \left(\tilde{\boldsymbol{\theta}}_j^{FM}\right)_{j\in U_{Mk}} \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})
\end{aligned}$$

with $\tilde{\boldsymbol{\theta}}_j^{FM,opt} = \left(\tilde{\theta}_{ij}^{FM,opt}\right)_{i\in U_{Fj}}$, $j \in U_{Mk}$, vectors of size N_{Fj}. Recalling that $\sum_{j\in U_M} \tilde{\theta}_{ij}^{FM}\tilde{\theta}_{jk}^{MT}$ has only one non null element when the links are MtO-MtO, this concludes the proof. □

Proof of Proposition 3.3.4

Using Proposition 3.7.2, the difference of variances between $\hat{t}_{y2}^{MT,opt}$ and $\hat{t}_{y2}^{MT,alt}$ can be rewritten as

$$\begin{aligned}
\mathrm{Var}(\hat{t}_{y2}^{alt}) - \mathrm{Var}(\hat{t}_{y2}^{opt}) &= \sum_{k\in U_T} y_k^2 \left(\sum_{i\in U_F} \sum_{i'\in U_F} \Delta_{ii'} (\tilde{\theta}_{ik}^{alt} - \tilde{\theta}_{ik}^{opt})(\tilde{\theta}_{i'k}^{alt} - \tilde{\theta}_{i'k}^{opt}) \right) \\
&= \sum_{k\in U_T} y_k^2 \sum_{i\in U_F} \sum_{i'\in U_F} \Delta_{ii'} \sum_{j\in U_M} \left((\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}))\tilde{\theta}_{ij}^{FM,opt} \right) \times \\
&\quad \sum_{j'\in U_M} \left((\tilde{\theta}_{j'k}^{MT,alt} - \tilde{\theta}_{j'k}^{MT,opt})\tilde{\theta}_{i'j'}^{FM,opt} \right) \\
&= \sum_{k\in U_T} y_k^2 \sum_{j\in U_M} (\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})) \sum_{j'\in U_M} (\tilde{\theta}_{j'k}^{MT,alt} - \tilde{\theta}_{j'k}^{MT,opt}) \times \\
&\quad \sum_{i\in U_F} \sum_{i'\in U_F} \Delta_{ii'} \tilde{\theta}_{ij}^{FM,opt} \tilde{\theta}_{i'j'}^{FM,opt}
\end{aligned}$$

For all $j \in U_M$, we denote by $\boldsymbol{\theta}_j^{MT,opt}$ the vector of optimal FM weights defined in Proposition 3.3.3. Then we can rewrite

$$\begin{aligned}
\sum_{i\in U_F} \sum_{i'\in U_F} \Delta_{ii'} \tilde{\theta}_{ij}^{FM,opt} \tilde{\theta}_{i'j'}^{FM,opt} &= \left(\boldsymbol{\theta}_j^{MT,opt}\right)^t \boldsymbol{\Delta}_{jj'}^M \boldsymbol{\theta}_j^{MT,opt} \\
&= \frac{\mathbb{1}_j^{Mt} \left(\boldsymbol{\Delta}_j^M\right)^{-1}}{\mathbb{1}_j^{Mt} \left(\boldsymbol{\Delta}_j^M\right)^{-1} \mathbb{1}_j^M} \boldsymbol{\Delta}_{jj'}^M \frac{\left(\boldsymbol{\Delta}_{j'}^M\right)^{-1} \mathbb{1}_{j'}^M}{\mathbb{1}_{j'}^{Mt} \left(\boldsymbol{\Delta}_{j'}^M\right)^{-1} \mathbb{1}_{j'}^M}.
\end{aligned}$$

Since the sampling design satisfies the Δ-property for U_M and the c_k-condition, $\boldsymbol{\Delta}_{jj'} = c_{jj'} \mathbb{1}_j^{Mt} \mathbb{1}_{j'} = c_k \mathbb{1}_j^{Mt} \mathbb{1}_{j'}$ if $j \neq j'$ both linked to k and

$$\sum_{i\in U_F} \sum_{i'\in U_F} \Delta_{ii'} \tilde{\theta}_{ij}^{FM,opt} \tilde{\theta}_{i'j'}^{FM,opt} = \begin{cases} \dfrac{1}{\mathbb{1}_j^{Mt} \left(\boldsymbol{\Delta}_j^M\right)^{-1} \mathbb{1}_j^M}, & \text{if } j = j', \\ c_k & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned}\mathrm{Var}(\hat{t}_{y2}^{alt}) - \mathrm{Var}(\hat{t}_{y2}^{opt}) = &\sum_{k\in U_T} y_k^2 \sum_{j\in U_M} (\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}))^2 \frac{1}{\mathbb{1}_j^{Mt}\left(\mathbf{\Delta}_j^M\right)^{-1}\mathbb{1}_j^M} + \\ &\sum_{k\in U_T} y_k^2 \sum_{j\in U_M} (\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})) \sum_{j'\in U_M, j'\neq j} (\tilde{\theta}_{j'k}^{MT,alt} - \tilde{\theta}_{j'k}^{MT,opt}) c_k \\ = &\sum_{k\in U_T} y_k^2 \sum_{j\in U_M} (\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}))^2 \frac{1}{\mathbb{1}_j^{Mt}\left(\mathbf{\Delta}_j^M\right)^{-1}\mathbb{1}_j^M} + \\ &\sum_{k\in U_T} y_k^2 \sum_{j\in U_M} (\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})) \sum_{j'\in U_M} (\tilde{\theta}_{j'k}^{MT,alt} - \tilde{\theta}_{j'k}^{MT,opt}) c_k - \\ &\sum_{k\in U_T} y_k^2 \sum_{j\in U_M} (\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}))^2 c_k.\end{aligned}$$

Since for all $k \in U_T$ the link weights $\tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})$ and $\tilde{\theta}_{jk}^{MT,alt}$ between $j \in U_{Mk}$ and $k \in U_T$ are standardized, we have $\sum_{j\in U_M}(\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})) = 0$. Then

$$\mathrm{Var}(\hat{t}_{y2}^{alt}) - \mathrm{Var}(\hat{t}_{y2}^{opt}) = \sum_{k\in U_T} y_k^2 \sum_{j\in U_M} (\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt}))^2 \left(\frac{1}{\mathbb{1}_j^{Mt}\left(\mathbf{\Delta}_j^M\right)^{-1}\mathbb{1}_j^M} - c_k \right).$$

□

Proof of Corollary 3.3.4

To prove the corollary, we have to consider estimator of the mean over U_T, since the total over U_T may not converge. Let us denote $f_N = \frac{n_F}{N_F}$. We consider the asymptotic framework considered in Isaki and Fuller (1982) for U_F, U_M and U_T simultaneously, with MtO-MtO links between populations. More specifically, we assume that the population sizes of N_F, N_M and N_T, and the sample sizes n_F, n_M and n_T go to infinity while the size of N_{Mk} for all $k \in U_T$ and N_{Fj} for all $j \in U_M$ is bounded. In the example of La Poste survey, this entails that the number of boxes in a round or the number of addresses in a box does not go to infinity. Every time a new round is added to U_T, all boxes and addresses linked to this round are added to U_M and U_F respectively without impacting the preexisting rounds.

We assume the following :

H1 $\lim_{N\to\infty} f_N = \pi \in (0,1)$,

H2 N_F, N_T, N_M, n_F and $n_T > 1$ and that $\frac{N_F}{N_T} = O(1)$ with the Landau notation O,

H3 there is a positive constant C such that, for all N_T and all N_M, $\frac{1}{N_T}\sum_{k\in U} y_k^2 N_{Mk} < C$.

Corollary 3.3.3 gives the expression of the difference of variances of double GWSM estimator of the mean over U_T for samples in U_F drawn using SRSWOR :

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{N_T}\hat{t}_{y2}^{alt}\right) - \mathrm{Var}\left(\frac{1}{N_T}\hat{t}_{y2}^{opt}\right) \\
&= \sum_{k\in U_T}\frac{1}{N_T^2}y_k^2 \sum_{j\in U_M}\left(\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})\right)^2\left(\frac{1}{n_F}\frac{N_F-n_F}{N_F-1}\frac{N_F-N_{Fj}}{N_{Fj}} + \frac{1-f_N}{f_N}\frac{1}{N_F-1}\right).
\end{aligned}$$

To show that we can approximate this expression with

$$\sum_{k\in U_T}\frac{1}{N_T^2}y_k^2 \sum_{j\in U_M}\left(\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})\right)^2\left(\frac{1}{n_F}\frac{N_F-n_F}{N_F-1}\frac{N_F-N_{Fj}}{N_{Fj}}\right)$$

we have to show that the difference

$$A_N = \sum_{k\in U_T}\frac{1}{N_T^2}y_k^2 \sum_{j\in U_M}\left(\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})\right)^2\left(\frac{1-f_N}{f_N}\frac{1}{N_F-1}\right)$$

is such that $n_T|A_N|$ converges to 0.

For all $j \in U_M$ and $k \in U_T$, the standardized weights are smaller than 1 thus $|\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})| < 1$ and $\sum_{j\in U_M}\left(\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})\right)^2 < N_{Mk}$. We get that :

$$\begin{aligned}
n_T|A_N| &= n_T \sum_{k\in U_T}\frac{1}{N_T^2}y_k^2 \sum_{j\in U_M}\left(\tilde{\theta}_{jk}^{MT,alt} - \tilde{\theta}_{jk}^{MT,opt}(\theta^{FM,opt})\right)^2\left(\frac{1-f_N}{f_N}\frac{1}{N_F-1}\right) \\
&\leq n_T \sum_{k\in U_T}\frac{1}{N_T^2}y_k^2 N_{Mk}\left(\frac{1-f_N}{f_N}\frac{1}{N_F-1}\right) \\
&\leq \frac{n_T}{N_T}\left(\frac{1}{N_T}\sum_{k\in U_T} y_k^2 N_{Mk}\right)\left(\frac{1-f_N}{f_N}\frac{1}{N_F-1}\right) \\
&\leq \left(\frac{n_T}{N_T}\frac{1-f_N}{f_N}\right)\left(\frac{1}{N_T}\sum_{k\in U_T} y_k^2 N_{Mk}\right)\frac{1}{N_F-1} \\
&\leq \left(\frac{1-f_N}{f_N}\right)\left(\frac{1}{N_T}\sum_{k\in U_T} y_k^2 N_{Mk}\right)\frac{1}{N_F-1}
\end{aligned}$$

Thus

$$\lim_{N\to\infty} n_T |A_N| \leq \lim_{N\to\infty} \left(\frac{1-f_N}{f_N}\right) \frac{N_F}{N_T} \left(\frac{1}{N_T} \sum_{k\in U_T} y_k^2 N_{Mk}\right) \frac{1}{N_F - 1} = 0,$$

using **H1**, **H2** and **H3**, whih concludes the proof.

For stratified SRSWOR with H strata of size N_h in U_F, $h = 1, \ldots, H$, we denote by h_T (resp. h_M) the strata of size N_{Th} in U_T (resp. of size N_{Mh}in U_M) composed of all units $k \in U_T$ (resp. $j \in U_M$) linked to strata h in U_F. Since the links are of type MtO-MtO, the strata in U_M and U_T do not overlap. If **H1**, **H2** and **H3** holds for each stratum in U_F, U_M and U_T, then the previous reasoning can be done for each stratum, which concludes the proof. □

Intégration statistique de données (ISD)

Chapitre 4

QR Prediction for Statistical Data Integration [1]

Abstract

In this paper, we investigate how a big non-probability database can be used to improve estimates of finite population totals from a small probability sample through data integration techniques. In the situation where the study variable is observed in both data sources, Kim and Tam (2021) proposed two design-consistent estimators that can be justified through dual frame survey theory. First, we provide conditions ensuring that these estimators are more efficient than the Horvitz-Thompson estimator when the probability sample is selected using either Poisson sampling or simple random sampling without replacement. Then, we study the class of QR predictors, introduced by Särndal and Wright (1984), to handle the less common case where the non-probability database contains no study variable but auxiliary variables. We also require that the non-probability database is large and can be linked to the probability sample. We provide conditions ensuring that the QR predictor is asymptotically design-unbiased. We derive its asymptotic design variance and provide a consistent design-based variance estimator. We compare the design properties of different predictors, in the class of QR predictors, through a simulation study. This class includes a model-based predictor, a model-assisted estimator and a cosmetic estimator. In our simulation setups, the cosmetic estimator performed slightly better than the model-assisted estimator. These findings are confirmed by an application to La Poste data, which also illustrates that the properties of the cosmetic estimator are preserved irrespective of the observed non-probability sample.

1. Article de E.Medous, C.Goga, A. Ruiz-Gazen, J.F Beaumont, A. Dessertaine et P. Puech, à paraître dans *Techniques d'Enquêtes/Survey Methodology*, 2023

Keywords : cosmetic estimator, dual frame, GREG estimator, non-probability sample, probability sample, variance estimator.

4.1 Introduction

In the field of economics and social sciences, surveys are usually based on probability sampling methods. At the French postal service (La Poste) for example, the postal traffic is estimated through quarterly probability surveys. Controlling the sampling design allows for design-based inference without resorting to modeling of the study variables ; this feature is attractive to many survey statisticians. Neyman (1934) is usually known as the founding paper of probability sampling theory. Since then, the literature on this topic has grown rapidly with an interplay between theory and practice (see Rao (2005) for the most important contributions).

Recently, survey statisticians have observed a decline in response rates together with an increase of the survey costs, which make probability sampling more challenging. In addition, large non-probability samples, such as administrative data or web-based surveys, become available often at low cost (see, e.g., Beaumont (2020) and Rao (2021) for more details). These observations are also true at La Poste where, for cost reasons, the size of probability samples is bound to decrease while a big database containing the automatically processed postal mail is available. Even if non-probability samples are associated with unknown selection mechanisms and may suffer from selection bias and measurement errors, they provide timely information on the population of interest. This context leads survey statisticians to study the integration or combination of data from probability and non-probability samples.

The literature on data integration in survey sampling has grown rapidly recently, and the reader may refer to several reviews on the subject (see Beaumont (2020), Yang and Kim (2020), Rao (2021), Kim (2022) and Wu (2022)). If we focus on the problem of combining probability and non-probability samples, the different data integration methods can be divided into three groups depending on whether the study variable is observed in the probability sample only, in the non-probability sample only, or in both samples (see e.g. Rao (2021)). Most methods tackle the problem of the study variable observed in the non-probability sample only, e.g. Kim (2022). In this context, the objective is to address the selection bias by combining data from the non-probability sample with auxiliary data available in a probability sample.

At La Poste, the problem is rather that the study variables (such as the different types of mails sent) are only available in the probability sample whereas auxiliary information is only available in the non-probability database. Such a context is rather rare in practice and

has therefore not been studied in detail so far. The aim of the present paper is to study this particular context thoroughly. The method we recommend and study in detail is applicable nicely if (i) the overlap between the probability and the non-probability samples is ideally large but at least non-empty, and (ii) it is possible to accurately match observations from both samples. At La Poste, the non-probability sample represents more than 80% of the population ; as a result, its intersection with the probability sample is large. The matching, however, is a difficult task that La Poste is still investigating.

In the situation where the study variables are measured in both samples, Kim and Tam (2021) propose a design-based dual frame approach to improve the efficiency of the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), which uses the probability sample only. The total of the study variable over the whole population is estimated by summing the true total over the non-probability sample and an estimator of the total over the complementary of the non-probability sample. Kim and Tam (2021) propose several estimators that can be deduced from a calibration perspective.

In Section 4.2, we revisit the approach of Kim and Tam (2021) and derive general results on the efficiency of their proposed dual frame estimators. In the situation where the study variable is not measured in the non-probability sample, we propose to replace the true unknown total over the non-probability sample by some prediction. In Section 4.3, we adapt the general class of QR predictors, introduced in Wright (1983), to data integration. This class of estimators includes the well-known model-assisted (GREG) and model-based estimators, but also the cosmetic estimator (Särndal and Wright, 1984). We first exhibit a condition under which the QR predictors can be written in a projection form. We then derive a QR condition such that these predictors are identical to model-assisted predictors. In Section 4.4, we look at the asymptotic properties of the QR estimators. We show that they are asymptotically unbiased under the model and the sampling design. We also prove that, under the QR condition, the predictors are asymptotically design-unbiased. We derive their asymptotic design variance and provide a design-consistent variance estimator. In Section 4.5, we use Monte Carlo simulations to compare several QR predictors and show that the cosmetic estimator is a good compromise for several setups. In Section 4.6, we consider an application to La Poste data and illustrate the impact of the non-probability sample on the estimators. Finally we conclude and give perspectives in Section 4.7.

4.2 Study variable observed in both samples

We are interested in estimating the population total $T = \sum_{k \in U} y_k$, where y_k is the value of the variable of interest Y for unit k of the population U. A probability sample s_P is drawn

from U using a sampling design $p(s_P|\boldsymbol{Z})$, where the population matrix $\boldsymbol{Z}$ contains design information such as strata identifiers. The sample inclusion indicator, I_k, $k \in U$, takes the value 1 if unit k is selected in s_P, and 0 otherwise. The probability that a given population unit k is selected in s_P is $\pi_k = \mathrm{E}_p(I_k|\boldsymbol{Z})$. We assume in the present section that the variable of interest Y is observed for each unit of the probability sample but also for each unit in the non-probability sample $s_{NP} \subset U$. The inclusion indicator in s_{NP} for the population unit $k \in U$ is denoted as δ_k (i.e., $\delta_k = 1$, if $k \in s_{NP}$, and $\delta_k = 0$, otherwise). We assume that δ_k is available for each unit of the probability sample s_P. Denote by N (resp. N_{NP}) the size of U (resp. s_{NP}) and by n the expected size of s_P. Let $\hat{T}_{HT} = \sum_{k \in s_P} d_k y_k$ be the well-known expansion or Horvitz-Thompson estimator of T with the sampling weights $d_k = 1/\pi_k$. If $\pi_k > 0$, for all $k \in U$, $\hat{T}_{HT}$ is a design-unbiased estimator of T.

The non-probability sample s_{NP} is usually a cheap and large source of data. Its selection mechanism is unknown, and its selection bias cannot be ignored when making inference. On the other hand, the probability sample s_P is assumed representative (without selection bias), yet often expensive and of relatively small size. By combining information from the two samples, we can expect to find an estimator more precise than the expansion estimator obtained using s_P.

Kim and Tam (2021) propose two estimators using combined data from s_P and s_{NP} and we propose to revisit the properties of these estimators. The total can be decomposed as :

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

where $T_{NP} = \sum_{k \in s_{NP}} y_k = \sum_{k \in U} \delta_k y_k$ and $T_C = \sum_{k \in U - s_{NP}} y_k = \sum_{k \in U} (1 - \delta_k) y_k$. Since y_k is measured for all units of s_{NP}, T_{NP} is known, and we only have to estimate T_C. Kim and Tam (2021) propose the following estimator :

$$\hat{T}_{DI} = T_{NP} + \sum_{k \in s_P} d_k (1 - \delta_k) y_k, \tag{4.1}$$

where T_C is estimated using the expansion estimator. As pointed out by Beaumont (2020), this can be viewed as a dual frame problem, with frames U and s_{NP}, where the sample s_P is randomly selected from U and a census is taken from s_{NP}. In this context of two sampling frames, $\hat{T}_{DI}$ is an estimator that results from a direct application of the method proposed by Bankier (1986). One may think that $\hat{T}_{DI}$ is more efficient than $\hat{T}_{HT}$, especially if the size of the non-probability sample is large, but this is not true in general. The following Proposition shows that, while the variance of $\hat{T}_{DI}$ is always smaller than the variance of $\hat{T}_{HT}$ for Poisson sampling, the property is only true under a condition on the study variable for

simple random sampling without replacement.

Proposition 4.2.1. *(i) For Poisson sampling, the variance of $\hat{T}_{DI}$ is less than or equal to the variance of $\hat{T}_{HT}$.*

(ii) For simple random sampling without replacement, the variance of $\hat{T}_{DI}$ is less than or equal to the variance of $\hat{T}_{HT}$ if and only if

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

where $\bar{Y}_U = \frac{1}{N}\sum_{k\in U} y_k$ is the mean of Y over U, $\bar{Y}_{NP} = \frac{1}{N_{NP}}\sum_{k\in U}\delta_k y_k$ is the mean of Y over s_{NP}, and $CV_{NP} = \sqrt{S_{Y,NP}^2}/\bar{Y}_{NP}$ is the coefficient of variation of Y in s_{NP}, with $S_{Y,NP}^2 = \frac{1}{N_{NP}-1}\sum_{k\in U}\delta_k(y_k-\bar{Y}_{NP})^2$.

The proof of Proposition 4.2.1 is given in the appendix. Intuitively, the result (ii) of Proposition 4.2.1 can be explained by the fact that the size of s_P is fixed for simple random sampling without replacement in the expression of $\hat{T}_{HT}$ while the size of $s_P \cap U - s_{NP}$ is random for $\hat{T}_{DI}$. In other words, the estimator $\hat{T}_{DI}$ is calibrated on N_{NP} and T_{NP}, but not on N while $\hat{T}_{HT}$ is calibrated on N.

If the size of the population U is known, Kim and Tam (2021) propose to improve $\hat{T}_{DI}$ by using the following estimator :

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

where

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

is a Hájek-type estimator of the total T_C. Kim and Tam (2021) proved that $\hat{T}_{PDI}$ is a Generalized Regression (GREG) estimator calibrated on N, N_{NP} and T_{NP}. Its expression can be further generalized by including additional calibration variables.

Following Kim and Tam (2021), it is possible to use the linearization approach and derive the approximate variance of $\hat{T}_{PDI}$, denoted as $\mathrm{AVar}(\hat{T}_{PDI})$. For Poisson sampling, the independence of the inclusion indicators reduces the comparison of $\hat{T}_{PDI}$ and $\hat{T}_{DI}$ to the comparison of Horvitz-Thompson and Hájek estimators of the total $T_C = \sum_{k\in U}(1-\delta_k)y_k$. The Hájek estimator can be significantly more efficient than the Horvitz-Thompson estimator but it is not true in general (see, e.g., Särndal et al. (1992))

For Poisson sampling, the estimator $\hat{T}_{PDI}$ can be substantially more efficient than the Horvitz-Thompson estimator $\hat{T}_{HT}$, as illustrated in our simulation study in Section 4.5. For

simple random sampling without replacement, the approximate variance of $\hat{T}_{PDI}$ can be compared to the variance of $\hat{T}_{HT}$ in more general conditions than in Kim and Tam (2021). Proposition 4.2.2 below shows that the approximate variance of $\hat{T}_{PDI}$ is smaller than the variance of $\hat{T}_{HT}$ for simple random sampling without replacement, and gives the expression of the difference between the variances.

Proposition 4.2.2. *For simple random sampling without replacement,*

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

where $\bar{Y}_U = \frac{1}{N}\sum_{k\in U} y_k$ *is the mean of* Y *over* U, *and* $\bar{Y}_C = \frac{1}{N - N_{NP}}\sum_{k\in U}(1-\delta_k)y_k$ *is the mean of* Y *over* $U - s_{NP}$.

In the present section, the study variable Y is assumed to be measured in both samples, s_P and s_{NP}. In the next section, we alleviate this assumption by considering that the study variable is not known in the non-probability sample. This situation is the one encountered at La Poste where not all variables of interest are measured in the automatically processed postal mail. The big non-probability database is based on an image recognition process and covers around 80% of the postal mails. This database contains some relevant auxiliary information such as the departure dates from the sending post office. However, such data are subject to selection bias (e.g., mails with atypical shape are not automatically processed), and measurement errors (e.g., errors in barcode scanning during the image recognition process). In such a situation, we propose to use the intersection between the big database and the probability sample, where the auxiliary variables together with the study variable are available, and predict the unknown y_k for $k \in s_{NP} - s_P$.

4.3 Prediction estimators for study variable unobserved in the non-probability sample

Recall that the finite population total of Y can be decomposed as $T = T_{NP} + T_C$. The total T_C is estimated as in Section 4.2 by the Hájek-type estimator $\hat{T}_C^{(\mathrm{Ha})}$. In the present section, y_k is unknown for $k \in s_{NP}$, and contrarily to Section 4.2, the total T_{NP} has to be estimated. In order to do so, we introduce a working model for Y and the general QR class of predictors of T_{NP} that does not require y_k to be known for units in s_{NP}. We assume that a vector of auxiliary variables $\boldsymbol{x}_k = (X_{k1}, \ldots, X_{kp})^\top$ is available for each unit k of a non-probability sample $s_{NP} \subset U$. We also assume that δ_k and $\delta_k \boldsymbol{x}_k$ are available for each

unit k of the probability sample s_P. Table 4.1 gives a summary of the characteristics of the data we consider in the remainder of this paper.

Sample	y_k measured	δ_k available	known selection mechanism	Auxiliary variables available
s_P	Yes	Yes	Yes	No
s_{NP}	No	Yes	No	Yes

TABLE 4.1 – Data characteristics in the data integration context of Section 4.3.

The variable Y is not available in s_{NP} and we cannot use anymore $\hat{T}_{PDI}$ since the total $T_{NP} = \sum_{k\in U} \delta_k y_k$ is unknown. The idea behind the class of estimators introduced in this section is to predict y_k for $k \in s_{NP}$ by using regression modelling between Y and the auxiliary variables, and then predict T_{NP}. We assume the following working model between the study variable Y and the vector of auxiliary variables $\boldsymbol{x}_k$:

$$y_k = \boldsymbol{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \quad k \in s_{NP}, \tag{4.2}$$

where the errors ε_k are independent with expectation $\mathrm{E}_m(\varepsilon_k) = 0$ and variance $\mathrm{Var}_m(\varepsilon_k)$ proportional to $\nu(\boldsymbol{x}_k) = v_k$ for some known positive constants v_k. The subscript m indicates that the expectation and variance are taken with respect to model (4.2) conditionally on observed auxiliary variables $\boldsymbol{x}_k$, $k \in s_{NP}$. Note that model (4.2) only needs to hold for units in the non-probability sample. A model for Y does not need to be explicitly specified for units $k \in U - s_{NP}$ as we always make inferences conditional on y_k, $k \in U - s_{NP}$.

We define a predictor $\hat{y}_k$ of y_k for $k \in s_{NP}$ by $\hat{y}_k = \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}}$ with

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k\in s_P} q_k \delta_k \boldsymbol{x}_k \boldsymbol{x}_k^\top \right)^{-1} \left(\sum_{k\in s_P} q_k \delta_k \boldsymbol{x}_k y_k \right), \tag{4.3}$$

where q_k are known positive constants for $k \in s_{NP}$. We assume that the $p \times p$ dimensional matrix $\sum_{k\in s_P} q_k \delta_k \boldsymbol{x}_k \boldsymbol{x}_k^\top$ is nonsingular for all possible samples s_P.

We propose to estimate $T_{NP} = \sum_{k\in U} \delta_k y_k$ by a *QR predictor* as suggested in Wright (1983) :

$$\begin{aligned} \hat{T}_{NP}^{(\mathrm{QR})} &= \sum_{k\in U} \delta_k \hat{y}_k + \sum_{k\in s_P} r_k \delta_k (y_k - \hat{y}_k) \\ &= \sum_{k\in U} \delta_k \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}} + \sum_{k\in s_P} r_k \delta_k (y_k - \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}}), \end{aligned} \tag{4.4}$$

where $r_k \geq 0$ are predefined constants. The initials Q and R refer to the constants q_k and

r_k. The final estimator of T is then given by

$$\hat{T}^{(\mathrm{QR})} = \hat{T}_{NP}^{(\mathrm{QR})} + \hat{T}_{C}^{(\mathrm{Ha})}. \tag{4.5}$$

Various choices of q_k and r_k yield predictors $\hat{T}_{NP}^{(\mathrm{QR})}$ with familiar forms as detailed below.

1. For $q_k = d_k v_k^{-1}$ and $r_k = d_k$, we obtain the model-assisted or GREG-type estimator :

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

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

2. For $q_k = v_k^{-1}$ and $r_k = 1$, we obtain the model-based type estimator :

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

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

3. For $q_k = (d_k - 1) v_k^{-1}$ and $r_k = 1$, we obtain the cosmetic-type estimator (Särndal and Wright, 1984; Brewer, 1999) :

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

   where $\hat{y}_k^{(\mathrm{Cos})} = \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}}^{(\mathrm{Cos})}$ with

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

Let us derive some properties for this class of QR predictors. Proposition 4.3.1 gives a general condition on the constants q_k and r_k such that the QR predictor can be defined as a sum of predictions over the population. Proposition 4.3.2 gives another general condition on the constants q_k and r_k such that the QR predictor is a model-assisted type estimator. The proofs are given in the Appendix.

Proposition 4.3.1. *(projection form) Consider the QR predictor $\hat{T}_{NP}^{(\mathrm{QR})}$ given by (4.4). Under the condition that there exists a vector $\boldsymbol{\mu} \in \mathbf{R}^p$ such that*

$$(Proj) : \quad \boldsymbol{\mu}^\top \boldsymbol{x}_k q_k = r_k \quad \textit{for all} \quad k \in s_{NP}, \tag{4.6}$$

we have $\sum_{k\in s_P} r_k\delta_k(y_k - \hat{y}_k) = 0$. *In this case,* $\hat{T}_{NP}^{(\mathrm{QR})}$ *can be written in the projection form :*

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

The model-assisted estimator $\hat{T}_{NP}^{(\mathrm{MA})}$ and model-based estimator $\hat{T}_{NP}^{(\mathrm{MB})}$ satisfy Condition (Proj) if there exists a vector $\boldsymbol{\mu} \in \mathbf{R}^p$ such that $\boldsymbol{\mu}^\top \boldsymbol{x}_k = v_k$ for all $k \in s_{NP}$. This condition is satisfied when v_k is one of the auxiliary variables in the model. If $v_k = 1$, it is satisfied provided that the intercept is included in the model. Condition (Proj) holds for $\hat{T}_{NP}^{(\mathrm{Cos})}$ if $\boldsymbol{\mu}^\top \boldsymbol{x}_k = v_k(d_k - 1)^{-1}$ for all $k \in s_{NP}$. A consequence of Proposition 4.3.1 is that, for equal probability sampling design such as simple random sampling without replacement, the model-assisted, the model-based and the cosmetic estimators are all equal.

Using Theorem 2 from Wright (1983), we derive the following proposition. For r_k satisfying Condition (QR) below and any given q_k, the QR predictor of T_{NP} is identical to the model-assisted predictor of T_{NP} with the same q_k.

Proposition 4.3.2. *Suppose that the constants* r_k *and* q_k *are such that there exists some vector* $\boldsymbol{\lambda} \in \mathbf{R}^p$ *such that*

$$(QR): \quad 1 - \pi_k r_k = \pi_k q_k \boldsymbol{x}_k^\top \boldsymbol{\lambda} \quad \textit{for all} \quad k \in s_{NP}. \tag{4.7}$$

Then :

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

where

$$\hat{T}_{NP}^{(\mathrm{Q}\pi)} = \sum_{k\in U} \delta_k \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}} + \sum_{k\in s_P} d_k\delta_k(y_k - \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}}) \tag{4.8}$$

is the model-assisted type predictor of T_{NP} *with* $\hat{\boldsymbol{\beta}}$ *given by (4.3).*

Following Wright (1983), we note that the (QR) condition always holds for $\hat{T}_{NP}^{(\mathrm{MA})}$. This condition also holds for the model-based estimator $\hat{T}_{NP}^{(\mathrm{MB})}$ if and only if there exists a vector $\boldsymbol{\lambda} \in \mathbf{R}^p$ such that $v_k(d_k - 1) = \boldsymbol{x}_k^\top \boldsymbol{\lambda}$, for all $k \in s_{NP}$. This condition is true if we take $v_k(d_k - 1)$ among the auxiliary variables $\boldsymbol{x}_k$. Condition (QR) holds for the cosmetic estimator $\hat{T}_{NP}^{(\mathrm{Cos})}$ if and only if there exists a vector $\boldsymbol{\lambda} \in \mathbf{R}^p$ such that $v_k = \boldsymbol{x}_k^\top \boldsymbol{\lambda}$, for all $k \in s_{NP}$. This condition is true if v_k is included in the vector of auxiliary variables.

4.4 Asymptotic properties and variance estimation of QR predictors

Let us consider the class of QR predictors $\hat{T}^{(\mathrm{QR})}$ given in (4.5) and start by studying the prediction error $\hat{T}^{(\mathrm{QR})} - T$ under the model (4.2) and the sampling design $p(\cdot)$. We have

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

The first right-hand term depends on the model and the sampling design, while the second right-hand term only depends on the sampling design. Assuming that the sampling design is not informative with respect to the model (4.2), we can prove that the expectation of $\hat{T}_{NP}^{(\mathrm{QR})} - T_{NP}$ computed with respect to the model is equal to 0. Indeed, the model bias of $\hat{T}_{NP}^{(\mathrm{QR})}$ is given by :

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

with $\hat{\boldsymbol{\beta}} = \left(\sum_{k\in s_P} q_k \delta_k \boldsymbol{x}_k \boldsymbol{x}_k^\top\right)^{-1} \left(\sum_{k\in s_P} q_k \delta_k \boldsymbol{x}_k y_k\right)$. Under the model (4.2), $\mathrm{E}_m(y_k) = \boldsymbol{x}_k^\top \boldsymbol{\beta}$ for all $k \in s_{NP}$, $\mathrm{E}_m(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\mathrm{E}_m(\boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}} - y_k) = 0$, implying that

$$\mathrm{E}_m(\hat{T}_{NP}^{(\mathrm{QR})} - T_{NP}) = 0. \tag{4.9}$$

The estimator $\hat{T}_{C}^{(\mathrm{Ha})}$ is a Hájek-type estimator of T_C and is not design-unbiased for T_C. Following Särndal (1980), we rather look at its asymptotic design-unbiasedness. An estimator $\hat{T}$ is said to be asymptotically design-unbiased for the finite population total T if $\lim_{N\to\infty} N^{-1}[\mathrm{E}_p(\hat{T}) - T] = 0$, where $\mathrm{E}_p(\cdot)$ is the expectation computed with respect to the sampling design. The asymptotic framework from Isaki and Fuller (1982) can be considered to allow for the population and sample sizes to grow to infinity. Assuming that the probability of observing an empty intersection set $s_P \cap s_{NP}$ is negligible, then $\hat{T}_{C}^{(\mathrm{Ha})}$ is asymptotically design-unbiased for T_C. Combining this property with relation (4.9), and considering a non-informative sampling design, we get that $\hat{T}^{(\mathrm{QR})}$ is asymptotically mp-unbiased for T under model (4.2).

4.4.1 Bias properties of $\hat{T}^{(\mathrm{Q}\pi)}$

Let us consider now the QR class of predictors that satisfy the (QR) condition given by (4.7). For this class of predictors, the final estimator of T is

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

and the standardized total error is given by :

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

The estimator $\hat{T}^{(\mathrm{Q}\pi)}$ is not exactly design-unbiased because of the nonlinearity of $\hat{\boldsymbol{\beta}}$ and of the Hájek estimator $\hat{T}_{C}^{(\mathrm{Ha})}$. Wright (1983) proved that the (QR) condition given in Proposition 4.3.2 is a sufficient condition for $\hat{T}_{NP}^{(\mathrm{Q}\pi)}$ to be asymptotically design-unbiased for T_{NP}, provided that

$$\lim_{N\to\infty} \frac{1}{N}\mathrm{E}_p\left[\left(\sum_{k\in U}\delta_k \boldsymbol{x}_k - \sum_{k\in s_P} d_k\delta_k \boldsymbol{x}_k\right)^{\top}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})\right] = 0, \tag{4.10}$$

where $\tilde{\boldsymbol{\beta}} = (\sum_{k\in U}\pi_k q_k \delta_k \boldsymbol{x}_k \boldsymbol{x}_k^{\top})^{-1}\sum_{k\in U}\pi_k q_k \delta_k \boldsymbol{x}_k y_k$, and assuming that $\sum_{k\in U}\pi_k q_k \delta_k \boldsymbol{x}_k \boldsymbol{x}_k^{\top}$ is nonsingular. Following Breidt and Opsomer (2000), if the sampling fraction n/N converges to a constant different from 0, assuming mild conditions on the first and second-order inclusion probabilities of the sampling design, and on the auxiliary information vectors $\boldsymbol{x}_k$ for all $k \in s_{NP}$, it can be shown that :

$$\lim_{N\to\infty} \mathrm{E}_p||N^{-1}(\sum_{k\in U}\delta_k \boldsymbol{x}_k - \sum_{k\in s_P} d_k\delta_k \boldsymbol{x}_k)||^2 = 0,$$

where $||\cdot||$ is the usual Euclidian norm. Equation (4.10) follows by assuming that the regression coefficient estimator satisfies $\lim_{N\to\infty}\mathrm{E}_p||\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}||^2 = 0$ (see Cardot et al. (2013) for more details). The estimator $\hat{T}_{C}^{(\mathrm{Ha})}$ is a Hájek-type estimator which can be shown to be asymptotically design-unbiased for T_C if the probability to observe the empty set for $s_P \cap s_{NP}$ is negligible. We conclude that the QR predictor $\hat{T}^{(\mathrm{Q}\pi)}$ is asymptotically design-unbiased for T.

4.4.2 Asymptotic variance and variance estimation of $\hat{T}^{(\mathrm{Q}\pi)}$

Because the QR estimator $\hat{T}^{(\mathrm{Q}\pi)}$ is asymptotically design-unbiased, we estimate its asymptotic design variance rather than its design mean square error. We can write the standardized

total error as

$$\frac{1}{N}(\hat{T}^{(\mathrm{Q}\pi)} - T) = \frac{1}{N}\left(\sum_{k\in s_p} d_k(E_k + e_k) - \sum_{k\in U}(E_k + e_k)\right) + R_1 + R_2,$$

where

$$E_k = \delta_k(y_k - \boldsymbol{x}_k^\top \tilde{\boldsymbol{\beta}}), \quad e_k = (1-\delta_k)\left(y_k - \frac{\sum_{k'\in U}(1-\delta_{k'})y_{k'}}{N - N_{NP}}\right),$$

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

and

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

As in Section 4.4.1, we assume the usual conditions on the sampling fraction n/N, on the first and second-order inclusion probabilities, on the variable of interest y_k and on the auxiliary information vector $\boldsymbol{x}_k$. We use the Landau notations big O_p and little o_p. If $||\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}||^2 = o_p(1)$ and $\left(\sum_{k\in s_p} d_k(1-\delta_k)/N\right)^{-1} = O_p(1)$, then $R_1 = o_p(n^{-1/2})$, $R_2 = O_p(n^{-1})$ and the standardized total error can be approximated by

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

where the right-hand side of the above expression is of order $O_p(n^{-1/2})$. Since $\sum_{k\in s_p} d_k(E_k+e_k)$ is the Horvitz-Thompson estimator of the total $\sum_{k\in U}(E_k+e_k)$, the asymptotic variance of the QR estimator $\hat{T}^{(\mathrm{Q}\pi)}$ is given by :

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

Assuming that $\pi_{kl} > 0$ for all $k, l \in U$, an estimator of the asymptotic variance is given by

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

where $\hat{E}_k = \delta_k(y_k - \boldsymbol{x}_k^\top\hat{\boldsymbol{\beta}})$ and $\hat{e}_k = (1-\delta_k)(y_k - \sum_{k'\in s_p} d_{k'}(1-\delta_{k'})y_{k'}/(N - N_{NP}))$. We can show, under assumptions detailed in Breidt and Opsomer (2000) and Goga et al. (2009),

that this estimator is design-consistent for $\text{AVar}(\hat{T}^{(\text{Q}\pi)})$ in the sense that $N^{-2}n(\hat{V}(\hat{T}^{(\text{Q}\pi)}) - \text{AVar}(\hat{T}^{(\text{Q}\pi)})) = o_p(1)$.

4.5 Simulations

In this section, we show the results of a Monte-Carlo study that compares the efficiency of three special cases of the QR predictor, $\hat{T}^{(\text{QR})} = \hat{T}_{NP}^{(\text{QR})} + \hat{T}_{C}^{(\text{Ha})}$, given in Section 4.3, namely the model-assisted, the model-based and the cosmetic estimators, assuming that $v_k = 1$ in model (4.2). We also compare these estimators with the expansion estimator and the PDI estimator defined in Section 4.2. To illustrate that the relative superiority of estimators depends on the data structure, we define three different setups based on different artificial populations. As mentioned in Section 4.3, if the probability samples are drawn using simple random sampling without replacement, the three QR estimators are all equal. Therefore, we focus on Poisson sampling with inclusion probabilities proportional to an auxiliary variable.

4.5.1 Populations and setups

The variables are generated using Gamma distributions to ensure their positiveness. Similar simulation results were obtained with Gaussian distributions but are not reported below. All populations have a size $N = 1,000$. We generate two auxiliary variables X_1 and X_2, where X_1 (resp X_2) follows a Gamma distribution with mean $\nu_1 = 20$ (resp $\nu_2 = 30$) and standard deviation (Std) $\sigma_1 = 15$ (resp $\sigma_2 = 20$). We use different models to generate the variable Y for all population units. For each model, $Y|X_1, X_2$ follows a Gamma distribution with mean $\mu_{Y|X_1,X_2}$ and constant variance $\sigma^2_{Y|X_1,X_2}$, which depend on the model.

1. For Model 1, $\mu_{Y|X_1,X_2}$ is a linear function of X_1 and X_2 :

$$\mu_{Y|X_1,X_2} = a_0 + a_1X_1 + a_2X_2.$$

2. For Model 2, $\mu_{Y|X_1,X_2}$ is a quadratic function of X_1 and a linear function of X_2 :

$$\mu_{Y|X_1,X_2} = b_0 + b_1(X_1 - \bar{X}_1)^2 + b_2X_2 \text{ with } \bar{X}_1 \text{ the mean of } X_1 \text{ over } U.$$

3. For Model 3, $\mu_{Y|X_1,X_2}$ is a linear function of X_2 :

$$\mu_{Y|X_1,X_2} = c_0 + c_2X_2.$$

To make the results comparable between the three models, we determine the constants a_0,

a_1, a_2, b_0, b_1, b_2, c_0,c_2, and $\sigma^2_{Y|X_1,X_2}$ in such a way that the following characteristics are the same :

— the unconditional mean μ and variance σ^2 of the variable Y,

— the coefficient of determination of the model, denoted as R^2,

— the ratio of variances for the explanatory variables :

$$\gamma = \text{Var}(a_1X_1)/\text{Var}(a_2X_2) = \text{Var}(b_1(X_1 - \bar{X}_1)^2)/\text{Var}(b_2X_2).$$

This ratio is only relevant for models 1 and 2 since X_1 is not included in Model 3.

In the following, we set $\mu = 100$, $\sigma^2 = 100$, and $\gamma = 0.5$. In Section 4.5.2, the R^2 value is either fixed to 0.8 or varies between 0.1 and 0.96. The main characteristics of the three population models are summarized in Table 4.2.

Model	Mean of (X_1, X_2)	Std of (X_1, X_2)	Mean of $Y\|X_1, X_2$	R^2
1	(20,30)	(15,20)	$\mu_Y = a_0 + a_1X_1 + a_2X_2$	equal between populations
2			$\mu_Y = b_0+b_1(X_1-\bar{X}_1)^2+b_2X_2$	
3			$\mu_Y = c_0 + c_2X_2$	

TABLE 4.2 – Population models with $\mu = 100$, σ^2=100, and $\gamma = 0.5$.

A non-probability sample of size 900 is drawn using simple random sampling without replacement and is the same for all populations. The probability samples are drawn using Poisson sampling with expected size 200 or 50 and probabilities proportional to X_1. We consider three setups. In each setup, we generate $Y|X_1, X_2$ using one of the three different population models, and we compute $\hat{y}_k, k \in s_{NP}$ for different QR predictors. The variables used as explanatory variables in the prediction models differ between setups as follows :

1. Setup 1 : Informative case. Population Model 1 is used to generate population Y values and only X_2 is used as explanatory variable in the prediction model along with the intercept.
2. Setup 2 : Quadratic case. Population Model 2 is used to generate population Y values and both auxiliary variables X_1 and X_2 are used as explanatory variables in the prediction model along with the intercept.
3. Setup 3 : Non-informative case. Population Model 3 is used to generate population Y values and only X_2 is used as explanatory variable in the prediction model along with the intercept.

For the informative and quadratic setups, the prediction model differs from the population model for Y, while the correct model is used in the non-informative setup. Table 4.3 gives a summary of the three setups.

Setup	Population	Variables used in prediction	Model correctly specified
Informative	$\mu_Y = a_0 + a_1 X_1 + a_2 X_2$	$\boldsymbol{x}_k^\top = (1, x_{2k})$	No
Quadratic	$\mu_Y = b_0 + b_1 (X_1 - \bar{X}_1)^2 + b_2 X_2$	$\boldsymbol{x}_k^\top = (1, x_{1k}, x_{2k})$	No
Non-informative	$\mu_Y = c_0 + c_2 X_2$	$\boldsymbol{x}_k^\top = (1, x_{2k})$	Yes

TABLE 4.3 – Studied setups.

4.5.2 Results

Let us consider the three setups defined above and compare the following estimators :

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

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

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

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

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

For each setup, $L = 10,000$ probability samples s_P are drawn according to Poisson sampling as detailed above and several Monte Carlo measures are computed. We compute the Monte Carlo relative bias of a given estimator $\hat{T}$ (either $\hat{T}_{HT}$, $\hat{T}^{(\mathrm{MB})}$, $\hat{T}^{(\mathrm{MA})}$, $\hat{T}^{(\mathrm{Cos})}$ or $\hat{T}_{PDI}$) as

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

where $\hat{T}^{(l)}$ is an estimate of T computed for the l-th sample, $l = 1, \ldots, L$.

As a measure of efficiency, we compute the Monte Carlo relative mean square error (RMSE) of an estimator $\hat{T}$ (relative to $\hat{T}^{(\mathrm{Cos})}$) :

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

where

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

We also compute the Monte Carlo relative variance (RVar) of an estimator $\hat{T}$ (relative to

Population parameters	Setup	Monte Carlo measures	$\hat{T}_{HT}$	$\hat{T}^{(\mathrm{MB})}$	$\hat{T}^{(\mathrm{MA})}$	$\hat{T}^{(\mathrm{Cos})}$	$\hat{T}_{PDI}$
$\mu = 100$ $\sigma^2 = 100$ $R^2 = 0.8$ $\gamma = 0.5$	Setup 1	RB_{MC}	-0.13	3.34	0.11	0.11	0.03
		RVar_{MC}	23,566.93	55.62	114.06	100.00	20.97
		RMSE_{MC}	22,897.58	2,715.21	113.91	100.00	20.65
	Setup 2	RB_{MC}	-0.07	-1.65	-0.06	-0.05	0.02
		RVar_{MC}	36,947.99	84.94	118.21	100.00	23.17
		RMSE_{MC}	36,638.27	1,056.44	118.42	100.00	23.15
	Setup 3	RB_{MC}	0.03	-0.01	0.01	0.01	0.01
		RVar_{MC}	41,088.93	58.38	100.49	100.00	33.47
		RMSE_{MC}	41,080.51	58.39	100.48	100.00	33.51

TABLE 4.4 – Relative bias (in % of the true value), percent relative (to $\hat{T}^{(\mathrm{Cos})}$) variance and MSE of the different estimators for the three different setups ; the expected size of the probability sample is 200 and the size of the non-probability sample is 900.

$\hat{T}^{(\mathrm{Cos})}$) :

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

where

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

Table 4.4 contains the simulation results for the three setups when $R^2 = 0.8$. In all setups, we confirm that both $\hat{T}_{PDI}$ and $\hat{T}_{HT}$ have a small Monte Carlo bias, as expected. In terms of MSE, $\hat{T}_{PDI}$ is the most precise estimator, while $\hat{T}_{HT}$ is the least precise estimator among all estimators. This result is expected since the expansion estimator does not make use of any auxiliary information, while $\hat{T}_{PDI}$ takes into account the true values of the study variable y_k for $k \in s_{NP}$; i.e., it takes into account the true values of Y for 900 units out of the 1,000 population units. In our context, where the study variable is not observed in s_{NP}, the estimator $\hat{T}_{PDI}$ is however not computable and serves more as a gold standard.

The Monte Carlo bias of $\hat{T}^{(\mathrm{MA})}$ and $\hat{T}^{(\mathrm{Cos})}$ is negligible in the three setups whereas $\hat{T}^{(\mathrm{MB})}$ is biased in the informative and quadratic setups. In these two setups, the prediction model differs from the population model used to generate Y values. In the non-informative setup, where the prediction model is correctly specified, the bias of $\hat{T}^{(\mathrm{MB})}$ is also negligible. The estimator $\hat{T}^{(\mathrm{MA})}$ has the largest variance of the QR predictors in the informative and quadratic setups, while $\hat{T}^{(\mathrm{MB})}$ has the smallest variance in all setups. In the quadratic setup, the variance of $\hat{T}^{(\mathrm{MB})}$ is similar to the variance of $\hat{T}^{(\mathrm{Cos})}$ but $\hat{T}^{(\mathrm{MB})}$ has the highest MSE amongst the QR predictors in both informative and quadratic setups. This means that the bias of $\hat{T}^{(\mathrm{MB})}$ degrades its MSE a lot despite its small variance. In the non-informative setup,

$\hat{T}^{(\mathrm{MB})}$ has the lowest MSE amongst the QR predictors. We can see in Table 4.4 that this comes from the absence of bias for $\hat{T}^{(\mathrm{MB})}$ in this setup together with its small variance.

In the informative and quadratic setups, $\hat{T}^{(\mathrm{Cos})}$ is more precise in term of variance than $\hat{T}^{(\mathrm{MA})}$. The estimators $\hat{T}^{(\mathrm{MA})}$ and $\hat{T}^{(\mathrm{Cos})}$ are similar in the non-informative setup. Both estimators use weighted regression with slightly different weights (d_k for $\hat{T}^{(\mathrm{MA})}$ and $d_k - 1$ for $\hat{T}^{(\mathrm{Cos})}$).

To summarize, when the prediction model is incorrectly specified, as in the informative and quadratic setups, both $\hat{T}^{(\mathrm{MA})}$ and $\hat{T}^{(\mathrm{Cos})}$ are significantly more efficient than $\hat{T}^{(\mathrm{MB})}$ because of the bias of $\hat{T}^{(\mathrm{MB})}$, even though the bias is not large. On the opposite, if the model is correctly specified but the design weights and Y are uncorrelated, as in the non-informative setup, $\hat{T}^{(\mathrm{MB})}$ is better than $\hat{T}^{(\mathrm{MA})}$ and $\hat{T}^{(\mathrm{Cos})}$ in terms of MSE. In all setups, $\hat{T}^{(\mathrm{Cos})}$ is more efficient or similar to $\hat{T}^{(\mathrm{MA})}$.

To better understand the impact of the R^2 on the results, we also plot, on the y-axis of Figures 4.1, 4.2 and 4.3, the RMSE_{MC} for 10 different values of R^2 on the x-axis : 0.1, 0.2, ..., 0.9, 0.96. In order to do that, we generate for each setup ten populations, one for each R^2 value. Figure 4.1 (resp. Figure 4.2 and 4.3) gives the results for Setup 1 (resp. 2 and 3) with the sample size equal to 200 (resp. 50) on the left (resp. right) column plots. On all plots, the curves correspond to the different estimators with a red curve at 100 for $\hat{T}^{(\mathrm{Cos})}$ (since the RMSE are relative to $\hat{T}^{(\mathrm{Cos})}$) and different colors for $\hat{T}_{HT}$, $\hat{T}^{(\mathrm{MA})}$, $\hat{T}^{(\mathrm{MB})}$ and $\hat{T}_{PDI}$. The plots on the top row of the figures include all the estimators while for the second row (and third row for Figures 4.1 and 4.2), $\hat{T}_{HT}$ (and $\hat{T}^{(\mathrm{MB})}$ for Figures 4.1 and 4.2) is removed in order to zoom in and ease the comparison between $\hat{T}^{(\mathrm{Cos})}$, $\hat{T}^{(\mathrm{MA})}$, $\hat{T}^{(\mathrm{MB})}$ and $\hat{T}_{PDI}$. The scale on the y-axis is kept fixed for the two columns (sample sizes).

As expected, $\hat{T}_{PDI}$ is by far the best estimator with the smallest MSE in all setups. In all figures, $\hat{T}_{HT}$ has a very bad relative MSE, especially when R^2 is high. Note that in fact the absolute MSE of $\hat{T}_{HT}$ remains stable when R^2 increases (results not reported), while the MSE of the other estimators improves. This result is expected because $\hat{T}_{HT}$ does not depend on the distribution of $Y|X_1, X_2$, but depends on μ and σ^2 which are constant across the populations.

Figure 4.1 (resp. 4.2) shows the evolution of RMSE_{MC} with respect to R^2 in the informative setup (resp. quadratic setup) for sample s_P of expected size 200 (left column) and 50 (right column). In these two setups, not only is $\hat{T}^{(\mathrm{Cos})}$ better than $\hat{T}^{(\mathrm{MB})}$ or $\hat{T}^{(\mathrm{MA})}$, as seen in Table 4.4, but its gain compared to its competitors increases the most with R^2. The precision of $\hat{T}^{(\mathrm{MA})}$ also increases, but at a slightly slower pace. The MSE of $\hat{T}^{(\mathrm{MB})}$ worsens with R^2 because the prediction model differs too much from the population model in these setups. This fact implies a larger bias of $\hat{T}^{(\mathrm{MB})}$ when R^2 increases. For informative and quadratic

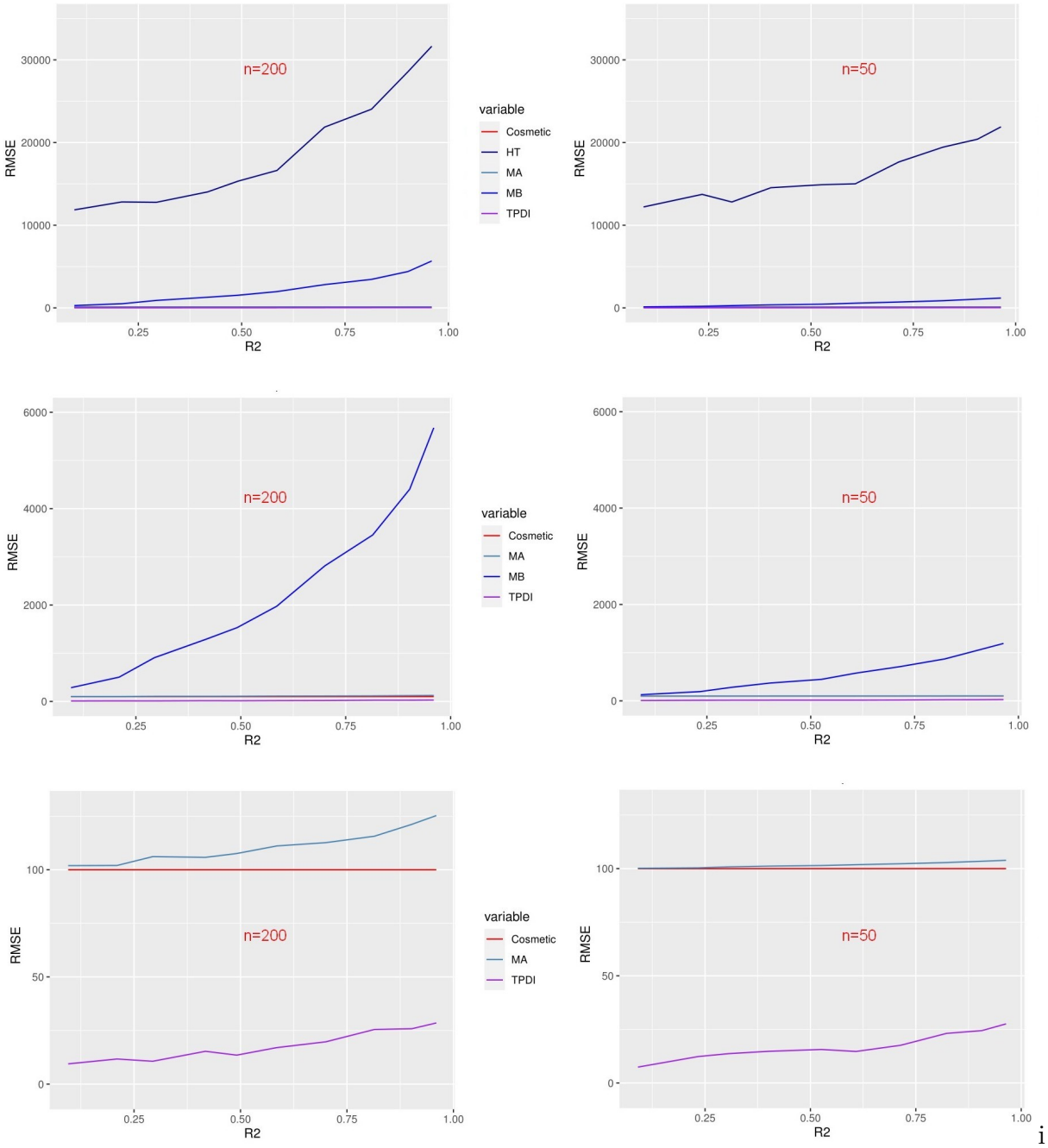

FIGURE 4.1 – Relative MSE (in %), with the MSE of the cosmetic estimator as the baseline, versus R^2 in the informative setup

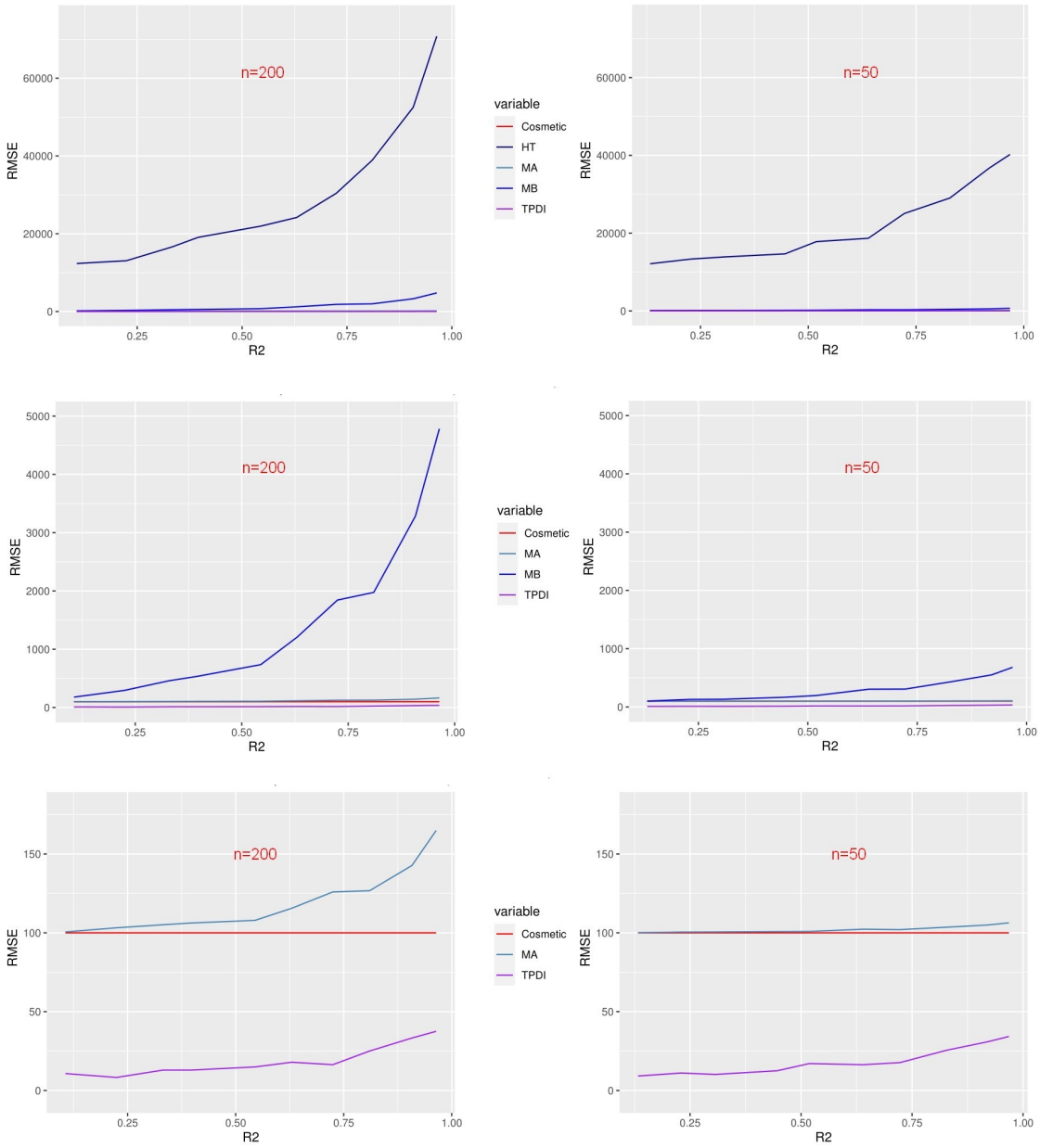

FIGURE 4.2 – Relative MSE (in %), with the MSE of the cosmetic estimator as the baseline, versus R^2 in the quadratic setup

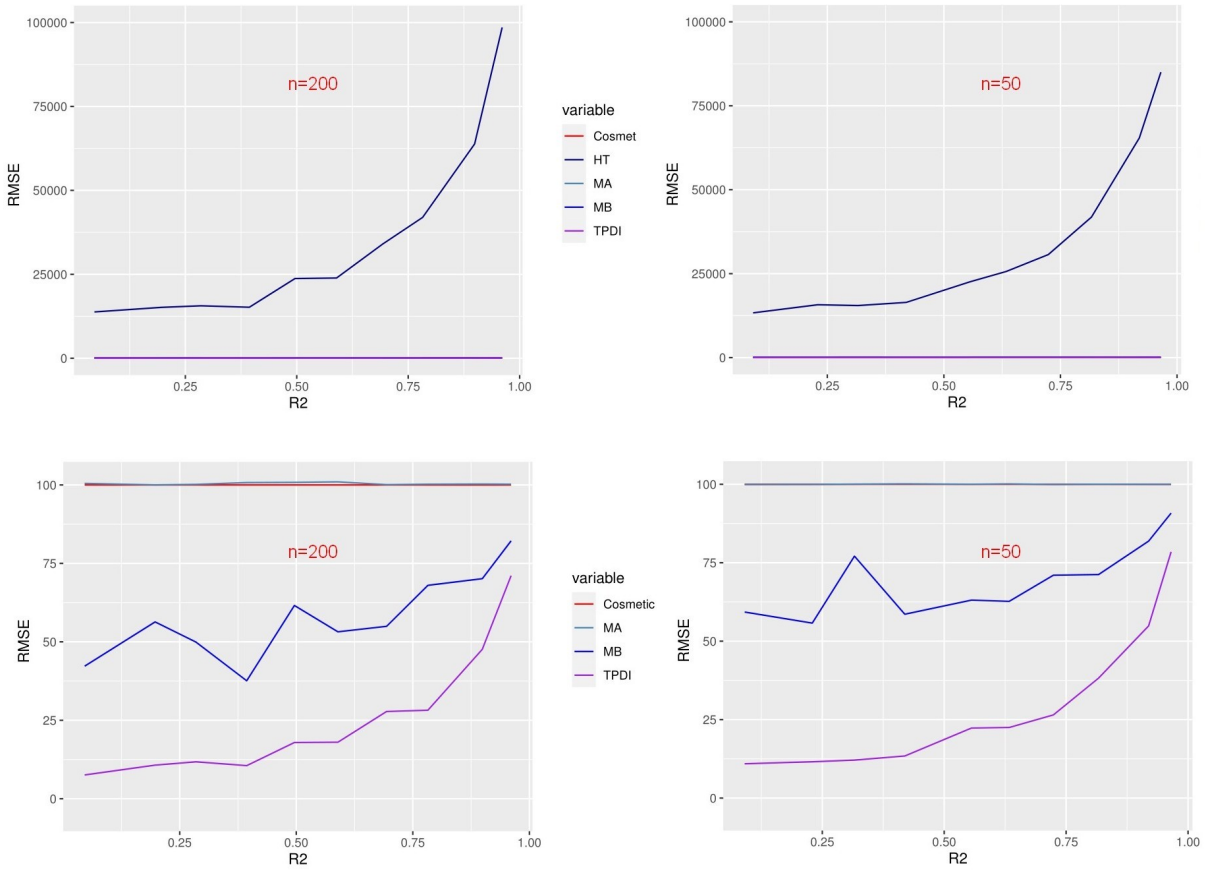

FIGURE 4.3 – Relative MSE (in %), with the MSE of the cosmetic estimator as the baseline, versus R^2 in the non-informative setup

setups, a smaller size reduces the difference between $\text{RMSE}_{MC}(\hat{T})$ of QR predictors.

Figure 4.3 shows the evolution of the RMSE_{MC} with respect to the R^2 in the non-informative setup. This time, $\hat{T}^{(\text{MB})}$ does not lose precision when R^2 increases because the prediction model is the same as the population model. All QR predictors show an increase in precision with R^2, with $\hat{T}^{(\text{Cos})}$ and $\hat{T}^{(\text{MA})}$ having similar precision for all values of R^2. In this setup, the plots are comparable for the two sample sizes, because the model is correctly specified for all prediction models.

To sum up, if the prediction model is misspecified, the cosmetic estimator is the best choice in our setups. It has the smallest MSE amongst all QR predictors, and its precision increases faster with R^2 than the other estimators. The advantage of $\hat{T}^{(\text{Cos})}$ over $\hat{T}^{(\text{MA})}$ might disappear in a scenario where the probability sample size would be a smaller fraction of the population size. The estimator $\hat{T}^{(\text{MB})}$ is biased and has the largest MSE, even for smaller values of R^2. If the model is correctly specified, and Y is not correlated to X_1 while the first-order inclusion probabilities are proportional to X_1, $\hat{T}^{(\text{MB})}$ is the best choice in terms

of MSE. However, the efficiency gain achieved by choosing $\hat{T}^{(\mathrm{MB})}$ over $\hat{T}^{(\mathrm{Cos})}$ in this third setup is significantly smaller than the efficiency loss observed when choosing $\hat{T}^{(\mathrm{MB})}$ over $\hat{T}^{(\mathrm{Cos})}$ in the first two setups. We thus recommend the choice of the Cosmetic estimator as a good compromise in all setups, followed closely by the model-assisted estimator. Similar observations can be made from real data, as shown in the next section.

4.6 Application to La Poste data

4.6.1 Data presentation

In France, more than 90% of letters sent are sorted out using automatic sorting machines. The information collected by the machines is based on pictures of the letters and form the non-probability database. La Poste has also access to a probability sample and wants to use both databases to estimate monthly totals of different types of letters using the data integration methods proposed above. Some letters such as "*lettres vertes*" (letters with a green stamp and ecologically transported) are not recognized by the sorting machines, which only take black and white pictures, meaning that the variable of interest is not available in the non-probability database. The auxiliary information associated with the letters automatically sorted is not easily linked to the probability sample. This issue is currently being investigated at La Poste. Thus, the illustration below is based on data from previous surveys made over the years at La Poste.

Data from these surveys are collected from postmen rounds and contain in particular the number of *lettres vertes*, the number of letters named "*produits 1b*" (which include different types of letters, including *lettres vertes*) and the total number of letters in the round. The idea is to mimic the situation at La Poste where the number of *lettres vertes* is available in the probability sample but not on the non-probability database whereas the number of *produits 1b* and the total number of letters are available in the non-probability database but not in the probability sample.

The goal of this section is on one hand to see if the conclusion drawn with simulated populations holds with real data and on the other hand to see if the selection method for the non-probability sample has an impact on the estimators. The population of interest consists in 11,906 rounds from historical data. In the following, we assume that the variable of interest is the number of *lettres vertes* and the explanatory variable X_1 (resp. X_2) is the total number of letters (resp. the number of *produits 1b*).

We consider three configurations. For each configuration, a non-probability sample of size 9,524 is drawn from the population ; the non-probability sampling fraction is thus 80%. In

S_{NP}	Monte Carlo measures	$\hat{T}_{HT}$	$\hat{T}^{(\mathrm{MB})}$	$\hat{T}^{(\mathrm{MA})}$	$\hat{T}^{(\mathrm{Cos})}$	$\hat{T}_{PDI}$
SRSWOR	RB_{MC}	0.08	8.30	0.08	0.08	0.06
	RVar_{MC}	253.67	122.66	102.21	100.00	87.41
	RMSE_{MC}	252.95	5,489.24	102.18	100.00	87.31
Highest Y values	RB_{MC}	0.03	5.51	0.03	0.03	0.01
	RVar_{MC}	1,528.57	197.54	101.82	100.00	3.78
	RMSE_{MC}	1,528.57	13,676.65	101.87	100.00	3.82
Lowest Y values	RB_{MC}	-0.02	1.23	0.13	0.13	0.41
	RVar_{MC}	197.09	100.67	100.26	100.00	93.17
	RMSE_{MC}	195.36	184.83	100.24	100.00	92.95

TABLE 4.5 – Relative bias (in % of the true value), percent relative (to $\hat{T}^{(\mathrm{Cos})}$) variance and MSE of the different estimators for the three different non-probability samples ; the expected size of the probability sample is 2,000 and the size of the non-probability sample is 9,524.

the first configuration, the non-probability sample is drawn using SRSWOR. For the second (resp. third) configuration, the non-probability sample contains the 9,524 rounds with highest (resp. lowest) values of Y. Only X_2 is used as explanatory variable in the prediction model along with the intercept.

We compare the same estimators as in section 4.5.2. For each of the 3 non-probability samples, we draw $L = 1,000$ probability samples s_P of expected size 2,000 using Poisson sampling with probabilities proportional to X_1. The same Monte-Carlo measures as those in Section 4.5.2 are computed.

4.6.2 Results

Table 4.5 contains the simulation results for La Poste data for the three non-probability samples. For all configurations, the prediction model appears to be significantly misspecified, which causes $\hat{T}^{(\mathrm{MB})}$ to be less efficient, both in terms of variance and MSE, than in the simulations of Section 4.5.2. Different diagnostic statistics could be computed beforehand to highlight the misspecification problem and alternative prediction models could be proposed. The assessment of such models is not in the scope of this paper.

For all configurations, the results are similar to those obtained in the informative and quadratic setups of Section 4.5.2 with $\hat{T}^{(\mathrm{MA})}$ and $\hat{T}^{(\mathrm{Cos})}$ having similar efficiency and being both more efficient than $\hat{T}^{(\mathrm{MB})}$, which is the only biased estimator. It can be noted that, although the choice of s_{NP} does not impact the relative efficiency of $\hat{T}^{(\mathrm{MA})}$, it impacts the relative precision of the HT and PDI estimators.

In all configurations, s_P is drawn using Poisson sampling and the variance of the data

S_{NP}	$\hat{T}_{HT}$	$\hat{T}_{NP}^{(\mathrm{MA})}$	$\hat{T}_{NP}^{(\mathrm{Cos})}$	$\hat{T}_{NP}^{(\mathrm{Ha})}$	$\hat{T}_{C}^{(\mathrm{Ha})}$
Highest Y values	101.97	111.39	110.38	101.61	0.92
Lowest Y values	104.61	73.07	69.03	15.43	181.29

TABLE 4.6 – Percent relative variance of the estimators when s_{NP} contains the highest or lowest Y values, relative to their variance when s_{NP} is drawn with SRSWOR.

integration estimators can be simplified as follows :

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

with $\hat{T}_{NP}$ the predictor of the total T_{NP} and $\hat{T}_{C}^{(\mathrm{Ha})}$ the Hájek estimator of the total T_C. To further understand the impact of the selection of s_{NP} on the estimators, we study the variance of $\hat{T}_{NP}^{(\mathrm{MA})}$, $\hat{T}_{NP}^{(\mathrm{Cos})}$ and $\hat{T}_{C}^{(\mathrm{Ha})}$, and the variance of the Hájek estimator $\hat{T}_{NP}^{(\mathrm{Ha})} = N_{NP} \sum_{k \in s_P} d_k \delta_k y_k / \sum_{k \in s_P} d_k \delta_k$ of the total T_{NP}.

Table 4.6 contains the relative variance of $\hat{T}_{HT}$, $\hat{T}_{NP}^{(\mathrm{MA})}$, $\hat{T}_{NP}^{(\mathrm{Cos})}$, $\hat{T}_{C}^{(\mathrm{Ha})}$ and $\hat{T}_{NP}^{(\mathrm{Ha})}$ for the second and third configurations, when s_{NP} contains the highest or lowest values of Y, relative to their variance in the first configuration (SRSWOR). As expected, the precision of the HT estimator does not depend on s_{NP} and is fixed for the three configurations. In the second configuration, $\hat{T}_{C}^{(\mathrm{Ha})}$ has a much smaller variance than in the other configurations, since only the smallest values of Y are left in $U - s_{NP}$. Similarly, its variance is greater in the third configuration, when only the largest values of Y are left in $U - s_{NP}$. A similar reasoning explains the ratio of variances between configurations for $\hat{T}_{NP}^{(\mathrm{MA})}$, $\hat{T}_{NP}^{(\mathrm{Cos})}$ and $\hat{T}_{NP}^{(\mathrm{Ha})}$. The important point is that both $\hat{T}_{NP}^{(\mathrm{MA})}$ and $\hat{T}_{NP}^{(\mathrm{Cos})}$ are not very sensitive to s_{NP} in terms of variance.

We could argue that the non-probability samples selected in the second and third configurations are not realistic. Yet, the results observed for more realistic non-probability samples, selected using Poisson sampling proportional to X_1 or Y, were not significantly different from the results observed when s_{NP} is drawn using SRSWOR and are not reported. Similarly, using both auxiliary variables in the prediction model only impacts the bias of $\hat{T}^{(\mathrm{MB})}$, and the results are again not reported.

To further improve the precision, other estimators of T_C may be used. As mentioned before, the data integration methods proposed in this paper can be used when the overlap between the probability and the non-probability samples is non-empty and, ideally, large. In this context, we recommend, based on our empirical results, the choice of the cosmetic or the model-assisted estimators.

4.7 Conclusion

Most of the literature on data integration in finite population tackles the problem of unobserved study variable in the probability sample. In this paper, we have proposed to fill the gap and considered the problem of unobserved study variable in the non-probability sample, assuming that it is observed in the probability sample and that auxiliary information is available in both samples. We have defined a general class of prediction estimators, based on the already known QR class, which includes the model-assisted, model-based and cosmetic estimators, and studied theoretically their bias and variance properties. We have also derived a variance estimator and compared the three types of estimators with the usual Horvitz-Thompson estimator in different simulation setups, both in terms of bias and MSE, and concluded that the cosmetic estimator is a good compromise in general.

The main conclusion of our experiments is that significant efficiency gains can be achieved by leveraging a big non-probability database that contains auxiliary information associated with the main study variables. For large domains, the efficiency gains obtained from using model-assisted estimators, including the cosmetic estimator, may be sufficient to obtain high-quality estimates of the population parameters of interest. For smaller domains, these estimators may not achieve precision targets. However, they could be used as direct estimates in a small area estimation model, such as the well-known Fay-Herriot area level model. This model requires area level auxiliary information. The big non-probability database would be a natural candidate for providing the auxiliary information required for producing small area estimates. Small area estimation methods often yield significant precision gains over direct estimators at the expense of introducing model assumptions.

Appendix

Proof of Proposition 4.2.1

We recall that $\hat{T}_{DI} = \sum_{k\in U} \delta_k y_k + \sum_{k\in s_P}(1-\delta_k)d_k y_k$ and $\hat{T}_{HT} = \sum_{k\in s_P} d_k y_k = \sum_{k\in s_P} \delta_k d_k y_k + \sum_{k\in s_P}(1-\delta_k)d_k y_k$. Thus, we have :

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

(i) For Poisson sampling, we have :

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

and

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

which proves the first part of the proposition.
(ii) For simple random sampling without replacement, let $\bar{Y}_U = \sum_{k\in U} y_k/N$, $\bar{Y}_{NP} = \sum_{k\in U}\delta_k y_k/N_{NP}$, $S^2_{Y,NP} = \sum_{k\in U}\delta_k(y_k-\bar{Y}_{NP})^2/(N_{NP}-1)$ and $CV^2_{NP} = S^2_{Y,NP}/\bar{Y}^2_{NP}$. Using some simple calculus, we have :

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

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

and thus

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

We conclude that $\mathrm{Var}(\hat{T}_{HT})$ is larger than or equal to $\mathrm{Var}(\hat{T}_{DI})$ if and only if

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

which is equivalent to :

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

and proves the second part of the proposition. □

Proof of Proposition 4.2.2

We have :

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

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

where

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

Using some basic but tedious calculus, we obtain :

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

□

Proof of Proposition 4.3.1

Let $\mathbf{R}_{s_P} = \mathrm{diag}(r_k\delta_k)_{k\in s_P}$, $\mathbf{X}_{s_P} = (\boldsymbol{x}_k^\top)_{k\in s_P}$, $\mathbf{y}_{s_P} = (y_k)_{k\in s_P}$ and $\mathbf{Q}^\top_{xs_P} = \mathbf{X}^\top_{s_P}\mathrm{diag}(q_k\delta_k)_{k\in s_P}$. Then $\hat{\boldsymbol{\beta}} = (\mathbf{Q}^\top_{xs_P}\mathbf{X}_{s_P})^{-1}\mathbf{Q}^\top_{xs_P}\mathbf{y}_{s_P}$. We can write the sum $\sum_{k\in s_P} r_k\delta_k(y_k-\hat{y}_k)$ in a matrix form as follows :

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

where $\mathbf{1}_{s_P}$ is a vector of ones with dimension the size of s_P. If the condition $\boldsymbol{\mu}^\top\boldsymbol{x}_k q_k - r_k = 0$ is fulfilled for all $k\in s_{NP}$, then $\delta_k(\boldsymbol{\mu}^\top\boldsymbol{x}_k q_k - r_k) = 0$ for all $k\in s_P$ and as a consequence, $\boldsymbol{\mu}^\top\mathbf{Q}^\top_{xs_P} = \mathbf{1}^\top_{s_P}\mathbf{R}_{s_P}$. We get then $\mathbf{1}^\top_{s_P}\mathbf{R}_{s_P}(\mathbf{y}_{s_P}-\mathbf{X}_{s_P}\hat{\boldsymbol{\beta}}) = 0$.

Proof of Proposition 4.3.2

We have

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

4.8 Supplementary material

This section details the results presented in section 4.4.2.

4.8.1 Assumptions

To the assumptions already mentioned in the paper :

$$||\hat{\beta} - \tilde{\beta}||^2 = o_p(1)$$

and

$$\left(\sum_{k \in s_p} d_k(1-\delta_k)/N\right)^{-1} = O_p(1),$$

we add the following assumptions to ensure the convergence of the HT estimator as discussed in section 1.1.3 :

1. We assume that $\lim_{N\to\infty} \frac{n}{N} \in (0,1)$.
2. We assume the following assumptions on the inclusion probabilities :
   $\min_{k\in U}\pi_k \geq \lambda > 0$ and $\limsup_{N\to\infty} n \max_{k\neq l\in U}|\pi_{kl} - \pi_k\pi_l| < C_1 < \infty$;
3. We assume that there is a positive constant C_2 such that, for all N, $\frac{1}{N}\sum_{k\in U} y_k^2 < C_2$.
4. We assume that there is a positive constant C_3 such that, for all $k \in U$, $||\boldsymbol{x}_k||^2 < C_3$.

4.8.2 Development of the results presented in section 4.4.2

In this section, we show that we can write the standardized total error as

$$\frac{1}{N}(\hat{T}^{(Q\pi)} - T) = \frac{1}{N}\left(\sum_{k\in s_p} d_k(E_k + e_k) - \sum_{k\in U}(E_k + e_k)\right) + R_1 + R_2.$$

By definition, the standardized total error is given by :

$$\frac{1}{N}(\hat{T}^{(Q\pi)} - T) = \frac{1}{N}(\hat{T}_{NP}^{(Q\pi)} - T_{NP}) + \frac{1}{N}(\hat{T}_C^{(Ha)} - T_C).$$

For any variable z_k, $k \in U$, let $T_z = \sum_{k\in U} z_k$ be the total of z over U and $\hat{t}_z = \sum_{k\in U} d_k z_k$ be the HT estimator of T_z.

Let us first look at $\frac{1}{N}(\hat{T}_{NP}^{(\mathrm{Q}\pi)} - T_{NP})$. The QR estimator $\hat{T}_{NP}^{(\mathrm{Q}\pi)}$ can be rewritten

$$\hat{T}_{NP}^{(\mathrm{Q}\pi)} = \sum_{k\in s_P} d_k\delta_k y_k - \left(\sum_{k\in s_p} d_k\delta_k \boldsymbol{x}_k^t - \sum_{k\in U}\delta_k \boldsymbol{x}_k^t\right)\hat{\beta} = \hat{t}_{\delta y} - \left(\hat{t}_{\delta\boldsymbol{x}} - T_{\delta\boldsymbol{x}}\right)^t\hat{\beta}.$$

Then

$$\begin{aligned}
\frac{1}{N}(\hat{T}_{NP}^{(\mathrm{Q}\pi)} - T_{NP}) &= \frac{1}{N}(\hat{t}_{\delta y} - T_{NP}) - \frac{1}{N}\left(\hat{t}_{\delta\boldsymbol{x}} - T_{\delta\boldsymbol{x}}\right)^t(\hat{\beta}-\tilde{\beta}) - \frac{1}{N}\left(\hat{t}_{\delta\boldsymbol{x}} - T_{\delta\boldsymbol{x}}\right)^t\tilde{\beta} \\
&= \frac{1}{N}(\hat{t}_{\delta y} - \hat{t}_{\delta\boldsymbol{x}}^t\tilde{\beta}) - \frac{1}{N}\left(T_{NP} - T_{\delta\boldsymbol{x}}^t\tilde{\beta}\right) + R_1 \\
&= \frac{1}{N}\left(\sum_{k\in s_p} d_k\delta_k(y_k - \boldsymbol{x}_k^t\tilde{\beta}) - \sum_{k\in U}\delta_k(y_k - \boldsymbol{x}_k^t\tilde{\beta})\right) + R_1 \\
&= \frac{1}{N}\left(\hat{t}_E - T_E\right) + R_1
\end{aligned}$$

with $R_1 = -\frac{1}{N}\left(\hat{t}_{\delta\boldsymbol{x}} - T_{\delta\boldsymbol{x}}\right)^t(\hat{\beta}-\tilde{\beta})$ and for all $k\in U$ $E_k = \delta_k(y_k - \boldsymbol{x}_k^t\tilde{\beta})$. Under the assumed conditions 1, 2 and 4, the HT estimator $\hat{t}_{\delta\boldsymbol{x}} = \sum_{k\in U} d_k\delta_k\boldsymbol{x}_k$ satisfy $\frac{1}{N}(\hat{t}_{\delta\boldsymbol{x}} - T_{\delta\boldsymbol{x}}) = O_p\left(1/\sqrt{n}\right)$. Since we assumed that $||\hat{\beta}-\tilde{\beta}||^2 = o_p(1)$, then $R_1 = o_p\left(1/\sqrt{n}\right)$.

Let us now look at $\frac{1}{N}(\hat{T}_C^{(\mathrm{Ha})} - T_C)$:

$$\frac{1}{N}(\hat{T}_C^{(\mathrm{Ha})} - T_C) = \frac{1}{N}T_{1-\delta}\left(\frac{\hat{t}_{(1-\delta)y}}{\hat{t}_{1-\delta}} - \frac{T_{(1-\delta)y}}{T_{1-\delta}}\right).$$

The difference of ratio can be rewritten as :

$$\begin{aligned}
\frac{\hat{t}_{(1-\delta)y}}{\hat{t}_{1-\delta}} - \frac{T_{(1-\delta)y}}{T_{1-\delta}} &= \frac{1}{\hat{t}_{1-\delta}}\left(\hat{t}_{(1-\delta)y} - \frac{T_{(1-\delta)y}}{T_{1-\delta}}\hat{t}_{1-\delta}\right) = \frac{1}{\hat{t}_{1-\delta}}\hat{t}_e = \frac{1}{\hat{t}_{1-\delta}}\left(\hat{t}_e - T_e\right) \\
&= \left(\frac{1}{\hat{t}_{1-\delta}} - \frac{1}{T_{1-\delta}}\right)\left(\hat{t}_e - T_e\right) + \frac{1}{T_{1-\delta}}\left(\hat{t}_e - T_e\right) \\
&= \frac{N}{T_{1-\delta}}R_2 + \frac{N}{T_{1-\delta}}\frac{1}{N}\left(\hat{t}_e - T_e\right)
\end{aligned}$$

with for all $k\in U$, $e_k = (1-\delta_k)\left(y_k - \frac{T_{(1-\delta)y}}{T_{1-\delta}}\right) = (1-\delta_k)\left(y_k - \frac{T_{(1-\delta)y}}{N-N_{NP}}\right)$, $T_e = \sum_{k\in U} e_k = 0$ and
$R_2 = \left(\frac{(\hat{t}_{1-\delta} - T_{1-\delta})/N}{\hat{t}_{1-\delta}/N}\right)\frac{1}{N}\left(\hat{t}_e - T_e\right)$. Since $(\hat{t}_e - T_e)/N$ and $(\hat{t}_{1-\delta} - T_{1-\delta})/N$ are both $O_p\left(1/\sqrt{n}\right)$ under the assumed conditions, $R_2 = O_p\left(1/n\right)$ if $\frac{1}{\hat{t}_{1-\delta}/N} = O_p(1)$.

Finally we get

$$\begin{aligned}\frac{1}{N}(\hat{T}^{(\mathrm{Q}\pi)} - T) &= \frac{1}{N}(\hat{T}_{NP}^{(\mathrm{Q}\pi)} - T_{NP}) + \frac{1}{N}(\hat{T}_{C}^{(\mathrm{Ha})} - T_C) \\ &= \frac{1}{N}\left(\hat{t}_E - T_E\right) + \frac{1}{N}\left(\hat{t}_e - T_e\right) + R_1 + R_2 \\ &= \frac{1}{N}\left(\hat{t}_E - T_E\right) + \frac{1}{N}\left(\hat{t}_e - T_e\right) + o_p\left(1/\sqrt{n}\right).\end{aligned}$$

□

Conclusions et perspectives

Dans cette thèse, j'ai cherché à améliorer la précision en terme de variance des estimateurs du trafic postal, d'une part en améliorant la méthode de partage des poids utilisée à La Poste et d'autre part en utilisant les larges sources de données auxiliaires dont dispose La Poste.

Je me suis d'abord penchée sur l'étude des estimateurs du total utilisés par La Poste afin de comprendre la perte de précision observée depuis quelques années et j'ai montré que l'augmentation de la variance était majoritairement due à l'utilisation de poids MGPP non optimaux. Ce travail a mené à des propositions d'amélioration de la pondération des estimateurs actuels via la modélisation du nombre d'adresses par case en fonction des données auxiliaires. Cette modélisation est à l'étude à La Poste et son effet sur la précision des estimateurs reste à confirmer.

Les travaux présentés dans les chapitres 2 et 3 ont mis en valeur deux avantages liés à l'utilisation d'une MGPP quand les liens sont de type TpU : d'une part, l'existence de poids optimaux et d'autre part, dans le cas TpU-TpU, la possibilité de réduire le nombre de liens à observer via une MGPP double. Il est intéressant de se demander si ces avantages peuvent être étendus à d'autres situations que les cas TpU et TpU-TpU.

Dans mes travaux, je ne m'intéresse pas à l'existence de poids optimaux indépendants de la variable d'intérêt pour les cas non TpU car l'existence de poids optimaux dans ce cas est complexe. Deville et Lavallée montrent que, pour les plans de Poisson et SASSR, ces poids n'existent pas. Cependant, il est possible d'obtenir des poids qui minimisent la variance des estimateurs MGPP dans des cas non TpU en modifiant la population d'intérêt de manière à obtenir des liens de type TpU.

Une façon de faire est d'utiliser une population de clusters, comme indiqué dans la figure 4.4. Dans cette figure, on regroupe les individus de U reliés à au moins un individu commun de U_F. Par exemple, U_1 et U_2 sont tous deux reliés à F_3, ils sont donc regroupés. On crée ensuite une population U' composée des clusters ainsi formés. L'individu U'_1 est donc le cluster composé de U_1 et U_2, et dont la valeur de la variable d'intérêt est $y'_{U'_1} = y_{U_1} + y_{U_2}$.

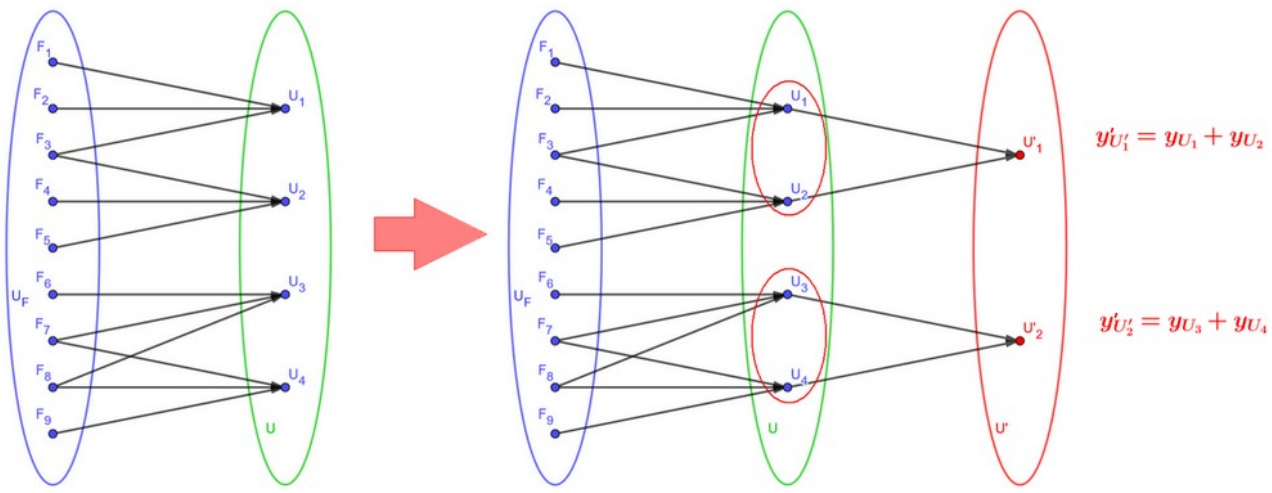

FIGURE 4.4 – Utilisation d'une population U' composée de clusters.

Le total T_y peut être réécrit comme le total $T_{y'}$ de y' sur U'. Les liens entre U_F et U' étant de type TpU, comme indiqué dans la figure 4.5, la propriété d'optimalité décrite dans le chapitre 2 de ce manuscrit s'applique si le plan de sondage satisfait la Δ-propriété pour U'. Cependant, cette méthode augmente de manière significative le nombre de liens à observer.

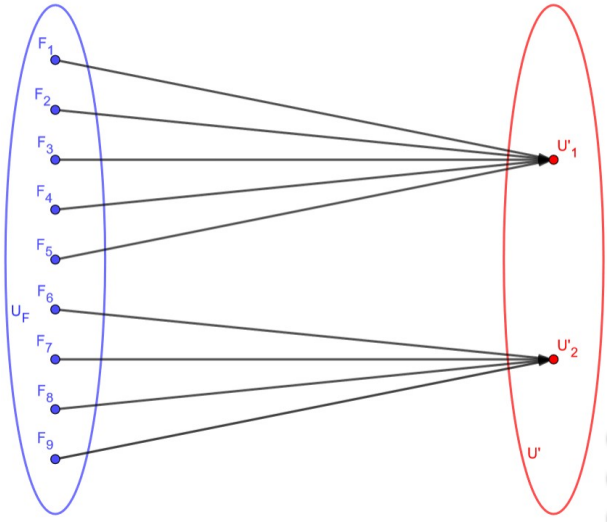

FIGURE 4.5 – Liens entre U_F et U'.

On a vu dans le chapitre 2 que la MGPP double avec double standardisation permettait de réduire le nombre de liens à identifier. Ce nombre peut être encore diminué en considérant plus de deux populations intermédiaires, telles que les liens entre chaque population soit de type TpU. L'estimateur MGPP du total T_y s'obtient alors similairement à l'estimateur MGPP double, en standardisant les liens entre chaque populations. L'existence de poids

optimaux pour ces liens, ainsi que le gain (ou perte) en nombre de liens observés selon le nombre de populations intermédiaires, est une des perspective de la thèse que j'ai commencé à étudier.

Il serait aussi possible de considérer les liens qui n'ont pas pu être identifiés comme non-répondants. Les techniques de traitement de la non-réponse s'appliquent alors, comme détaillé dans Xu and Lavallée (2009). Une idée pourrait être de s'inspirer des graphes aléatoires en attribuant à chaque couple $(i, k) \in U_F \times U$ une probabilité d'être liés.

Dans un deuxième temps, je me suis penchée sur l'intégration de bases de données massives pour améliorer les estimations postales, dans le cas où les variables d'intérêt ne sont pas disponibles dans les bases non probabilistes. J'ai proposé des améliorations théoriques qui dépendent de la capacité à identifier de manière unique un individu dans les différentes bases de données et de la qualité des variables auxiliaires. Cependant, La Poste fait face à des challenges industriels liés à l'accessibilité et à la qualité des bases de données massives, rendant l'application des mesures proposées délicate.

Un des problèmes vient de la construction du modèle. La méthode proposée dans ma thèse consiste à utiliser un modèle linéaire pour prédire la valeur des variables d'intérêt pour les individus contenus dans la base de données massive. Or, l'étude des corrélations des données disponibles à La Poste montre que celles-ci sont trop faibles pour justifier l'utilisation d'un tel modèle.

Ce problème pourrait être contourné en utilisant un modèle non paramétrique qui serait moins sensible à la qualité (et quantité) des données auxiliaires. L'idée est d'obtenir un modèle et de calculer des prédictions, qui puissent être utilisées pour toutes les variables d'intérêt. Dans ma thèse, j'ai testé, sur données simulées, le comportement des estimateurs QR avec des modèles de type forêts aléatoires. Ce travail est prometteur mais à approfondir.

De son côté, La Poste explore différentes méthodes de calage afin d'améliorer la précision de ses estimateurs du total, mais ces méthodes n'ont pas été étudiées dans le cadre de ma thèse.

Bibliographie

Ardilly, P. (2006). *Les techniques de sondage*. Editions TECHNIP.

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). Are probability surveys bound to disappear for the production of official statistics ? *Survey Methodology*, 46(1) :1–29.

Breidt, F.-J. and Opsomer, J.-D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4) :1023–1053.

Brewer, K. (1999). Cosmetic calibration with unequal probability sampling. *Survey Methodology*, 25(2) :205–212.

Cardot, H., Goga, C., and 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.

De Vitiis, C., Falorsi, S., Inglese, F., Masi, A., Pannuzi, N., and Russo, M. (2014). A methodological approach based on indirect sampling to survey the homeless population. *Rivista di statisticaufficiale*, 1(2) :9–30.

Dessertaine, A. and Fluteaux, L. (2004). Utilisation de la méthode généralisée du partage des poids dans le cadre des estimations de flux de courrier à la poste. *In : Ardilly, Pascal, (ed.) Echantillonnage et méthodes d'enquêtes, Science Sup.*

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. (1999). Simultaneous calibration of several surveys. In *Proceedings of Statistics Canada Symposium 99 of Statistics Canada*, pages 207–212.

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

Deville, J.-C. and Lavallée, P. (2006). Sondage indirect : les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, 32(2) :185.

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

Ernst, L. R. (1986). *Weighting issues for longitudinal household and family estimates*. US Bureau of the Census.

Falorsi, P. D., Righi, P., and Lavallée, P. (2019). Cost optimal sampling for the integrated observation of different populations. *Survey methodology*, 45(3) :485–511.

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

Hájek, J. (1971). Comment on a paper by D. Basu. In *Foundations of Statistical Inference (eds. V.P. Godambe and D.A. Sprott)*, page 236. Toronto : Holt, Rinehart and Winston.

Hájek, J. (1981). *Sampling from a finite population*. Statistics : Textbooks and Monographs. Marcel Dekker, New York.

Haziza, D. and Beaumont, J.-F. (2017). Construction of weights in surveys : A review. *Statistical Science*, Vol. 32(2) :206–226.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47 :663–685.

Isaki, C.-T. and Fuller, W.-A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.*, 77 :49–61.

Kalton, G. and Brick, J. M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21(2) :33–34.

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

Kim, J. K. (2022). A gentle introduction to data integration in survey sampling. *The survey statistician*, 85 :19–29.

Kim, J. K., Park, S., Chen, Y., and 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.

Kim, J.-K. and Tam, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89(2) :382–401.

Kokic, P. and Bell, P. (1994). Optimal winsorizing cutoffs for a stratified finite population estimator. *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-*, 10 :419–419.

Lardin-Puech, P. (2014). Estimation du trafic de courrier distribué en france métropolitaine par sondage indirect. In *the proceedings of the 8ème colloque francophone sur les sondages, Dijon*. http ://paperssondages14.sfds.asso.fr/submission_100.pdf.

Lavallée, P. (2007). *Indirect sampling*. Springer Science & Business Media.

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

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.-F., Dessertaine, A., and Puech, P. (To be published, 2023b). Qr prediction for statistical data integration. *Survey Methodology*.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3(2) :169–175.

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 :558–606.

Rao, J. (2005). Interplay between sample survey theory and practice : An appraisal. *Survey Methodology*, 31(2) :117.

Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 83(1) :242–272.

Rendtel, U. and Harms, T. (2009). Weighting and calibration for household panels. *Methodology of longitudinal surveys*, pages 265–286.

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

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. Springer-Verlag, New York.

Särndal, C.-E., Swensson, B., and Wretman, J. H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3) :527–537.

Särndal, C.-E. and Wright, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian J. of Statistics*, 11 :146–156.

Sherman, J. and Morrison, W. J. (1950). Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix. *The Annals of Mathematical Statistics*, 21(1) :124 – 127.

Tillé, Y. (2019). *Théorie des sondages : échantillonnage et estimation en populations finies*. Dunod, Paris.

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

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48(2) :283–311.

Xu, X. and Lavallée, P. (2009). Treatments for link nonresponse in indirect sampling. *Survey Methodology*, 35(2) :153–164.

Yang, S. and 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.

Résumé La Poste française a mis en place une étude par sondage pour effectuer un suivi de la distribution des objets postaux en France Métropolitaine. Le but des travaux réalisés lors de cette thèse est d'améliorer la précision des estimateurs du trafic total postal. La Poste souhaite observer des tournées de facteurs, mais ne peut pas les échantillonner directement. Elle a donc mis en place un tirage indirect et utilise une Méthode Généralisée de Partage des Poids (MGPP) double, inspirée de la MGPP simple proposée par Deville and Lavallée (2006), pour calculer les estimations du trafic postal. Dans un premier temps, nous étudions dans le cadre de La Poste l'existence d'une MGPP simple optimale, au sens que la variance des estimateurs est minimale, et la comparons aux estimateurs postaux actuels. La MGPP double, telle qu'utilisée à La Poste, est plus facile à mettre en place que la MGPP simple, mais est moins précise que la MGPP simple optimale. Nous nous penchons donc sur la question d'une MGPP double optimale dans le cas de La Poste et proposons des alternatives basées sur l'utilisation de données auxiliaires. Dans un deuxième temps, nous nous intéressons à l'Intégration Statistique de Données (ISD), qui consiste à utiliser un échantillon non-probabiliste, généralement une base de données massives, et un échantillon probabiliste pour construire des estimateurs de totaux. La littérature sur l'ISD traite majoritairement de cas où la variable d'intérêt est disponible dans la base de données massives. Les données massives de La Poste, intitulées Traitement Automatisé de l'Enveloppe (TAE), sont récoltées sur les objets triés par machine et ne contiennent pas d'information sur certaines variables d'intérêt. Des méthodologies adaptées au cas particulier de La Poste doivent donc être développées. Nous proposons une amélioration des estimateurs présentés dans Kim and Tam (2021) en utilisant une famille de prédicteurs, dits prédicteurs QR, proposée par Wright (1983), pour prédire le total des variables d'intérêt sur TAE. Nous utilisons un estimateur de Hájek pour estimer le total des variables d'intérêt sur le complémentaire de TAE. Nous étudions les propriétés théoriques des estimateurs QR et préconisons l'utilisation de l'estimateur cosmétique qui est un cas particulier d'estimateur QR. Les résultats présentés dans cette thèse sont illustrés par des études Monte-Carlo basées sur des données simulées et sur des données historiques postales, afin d'évaluer le gain de précision potentiel apporté par l'utilisation des différents estimateurs que nous proposons.

Mots Clés : données non probabilistes, estimateur QR, Intégration Statistique de Données, Méthode Généralisée de Partage des Poids, minimisation de la variance, poids de sondage optimaux, population finie, propriétés asymptotiques, sondage.

Abstract The French Postal Service (La Poste) has set up a sample survey to monitor mail distribution in Metropolitan France. The aim of the work carried out in this thesis is to improve the accuracy of estimators of total postal traffic. La Poste wants to observe postman rounds, but cannot sample them directly. It has therefore set up an indirect sampling and uses a double Generalized Weight Share Method (GWSM), inspired by the simple GWSM proposed by Deville and Lavallée (2006), to compute the estimates of the postal trafic. First, we study in the situation of La Poste the existence of a simple optimal GWSM, in the sense that the variance of the estimators is minimal, and compare it with current postal estimators. The double GWSM, as used at La Poste, is easier to implement than the simple GWSM, but is less accurate than the optimal simple GWSM. We therefore address the question of an optimal double GWSM in La Poste case and propose alternatives based on the use of auxiliary data. Secondly, we focus on Statistical Data Integration (SDI), which consists in using a non-probability sample, mainly a massive database, and a probability sample to compute estimators of totals. The literature on SDI deals mainly with cases where the variable of interest is available in the massive database used. La Poste massive database, entitled Traitement Automatisé de l'Enveloppe (TAE), is collected on mails sorted by machine and does not contain information for several variables of interest. Methodologies adapted to the specific case of La Poste must therefore be developed. We propose an amelioration of the estimators presented in Kim and Tam (2021) by using a family of predictors, known as QR predictors, proposed by Wright (1983), to predict the total of the variables of interest in TAE. We propose a Hájek estimator to estimate the total of the variables of interest in the complementary set of TAE. We study the theoretical properties of QR estimators and recommend the use of the cosmetic estimator, which is a special case of QR estimator. The results presented in this thesis are illustrated by Monte-Carlo studies based on simulated data and historical postal data, in order to assess the potential gain in accuracy brought by the use of the different proposed estimators.

Key Words : asymptotic properties, finite population, Generalized Weight Sharing Method, non-probabilistic data, optimal sampling weights, QR estimator, sampling method, Statistical Data Integration, variance minimisation.