WORKING PAPERS

N° 1269

June 2022

"Many-to-One indirect sampling with application to the French postal traffic estimation"

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

Toulouse School of Economics

Many-to-One indirect sampling with application to the French postal traffic estimation

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

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

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.

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

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 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 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 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 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 6 concludes the paper while the proofs are given in the Appendix.

2 Indirect sampling

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 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 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 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.

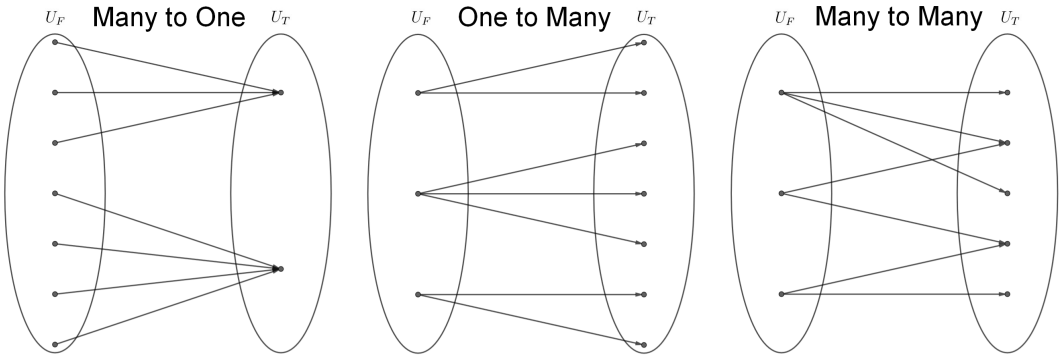

Figure 1: The different types of links between U_F and U_T.

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 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

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

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:

$$\mathrm{Var}(\hat{t}_{y1}) = \sum_{k \in U_T} \sum_{k' \in U_T} y_k y_{k'} \, \mathrm{Cov}(\hat{t}_{\tilde{\theta}_k}, \hat{t}_{\tilde{\theta}_{k'}}), \tag{2}$$

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) 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

$$\mathrm{Var}(\hat{t}_{y1}^{opt}) \leq \mathrm{Var}(\hat{t}_{y1}), \text{for all } y, \tag{3}$$

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 (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 $\hat{t}_{\tilde{\theta}_k}$:

$$(\tilde{\theta}_{ik}^{wopt})_{i\in U_F} = \arg \min_{\tilde{\theta}_{ik}, i\in U_F} \mathrm{Var}(\hat{t}_{\tilde{\theta}_k}), \quad \text{for all} \quad k \in U_T. \tag{4}$$

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 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

$$\mathbf{\Delta}_{k,k'\neq k} = c_{kk'}\mathbb{1}_k\mathbb{1}_{k'}^t \quad \text{with} \quad c_{kk'} \quad \text{not depending on } i \text{ and } i'. \tag{5}$$

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 2.1 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 2.1. *If the links are MtO and the sampling design satisfies the $\mathbf{\Delta}$-property, then:*

$$Var(\hat{t}_{y1}) = \sum_{k\in U_T} y_k^2\, Var(\hat{t}_{\tilde{\theta}_k}) + \sum_{k\in U_T}\sum_{k'\neq k\in U_T} y_k y_{k'} c_{kk'}. \tag{6}$$

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, the optimal link weights verifying the strong-optimality criterion given in (3) are 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 (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 $\mathrm{Var}(\hat{t}_{\tilde{\theta}_k})$ for all k in U_T. Using the terminology introduced by Deville and Lavallée (2006) (see also relations (3) and (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:

$$\tilde{\theta}_{ik}^{\mathrm{opt}} = \frac{l_{ik}/\Delta_{ii}}{\sum_{i'\in U_F} l_{i'k}/\Delta_{i'i'}}, \tag{7}$$

for all $i \in U_F$ and $k \in U_T$. For SRSWOR sampling, the optimal link weights are equal to $\tilde{\theta}_{ik}^{\mathrm{opt}} = l_{ik}/\sum_{i'\in U_F} l_{i'k}$, 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 (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. *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:*

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

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

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 2.1. *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. 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.

2.3 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 2.3 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 2.3. *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 (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 2.3, 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.

3 Double indirect sampling

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 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.

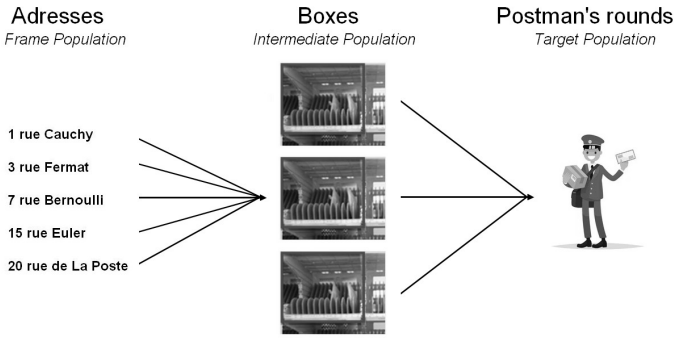

Figure 2: The frame, intermediate and target populations at La Poste.

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 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 1).

As in Section 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:

$$\hat{t}_{y2} = \sum_{i\in s_F} d_i\left(\sum_{k\in U_T}\sum_{j\in U_M}\tilde{\theta}_{ij}^{FM}\tilde{\theta}_{jk}^{MT}y_k\right). \tag{8}$$

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

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

3.2 MtO links

In this subsection, we focus on MtO links between the frame and the target populations. Comparing expressions (1) and (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 to determine the differences of the variances between double and simple GWSM. Thus, Proposition 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.

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 (8), contain only one non-zero element equal to $\tilde{\theta}_{ij}^{FM}\tilde{\theta}_{jk}^{MT}y_k$. To compute $\hat{t}_{y2}$, we need to compute only the standardized link weight $\tilde{\theta}_{ij}^{FM}\tilde{\theta}_{jk}^{MT}$ 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 $\tilde{\theta}_{ij}^{FM}$ and $\tilde{\theta}_{jk}^{MT}$ that are either both standardized $\left(\sum_{i' \in U_F} \tilde{\theta}_{i'j}^{FM} = 1 \text{ and } \sum_{j' \in U_M} \tilde{\theta}_{j'k}^{MT} = 1\right)$, or globally standardized $\left(\sum_{i' \in U_F} \sum_{j' \in U_M} \tilde{\theta}_{i'j'}^{FM}\tilde{\theta}_{j'k}^{MT} = 1\right)$. 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.

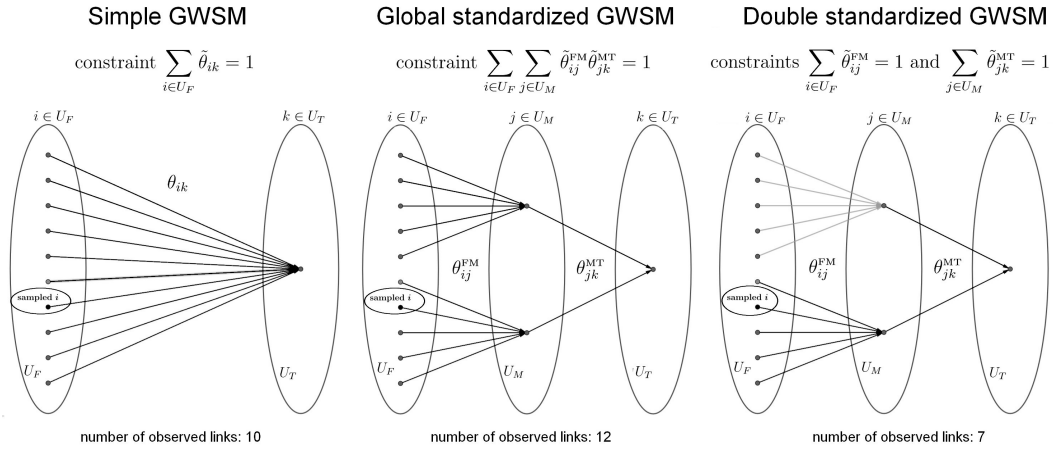

Figure 3: Number of observed links for each standardization

With double standardization, we can derive $\tilde{\theta}_{ij}^{FM}$ and $\tilde{\theta}_{jk}^{MT}$ separately. To compute $\tilde{\theta}_{ij}^{FM}$, 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 $N_{Fj} = \sum_{i' \in U_F} l_{i'j}^{FM}$. Similarly, to compute $\tilde{\theta}_{jk}^{MT}$, the links between U_M and the indirectly sampled $k \in U_T$ must be observed which leads to $N_{Mk} = \sum_{j' \in U_M} l_{j'k}^{MT}$ 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. 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. 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 (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. 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

$$N_{Fj} + N_{Mk} < N_{Fj} N_{Mk} = N_{Fk}, \tag{9}$$

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,

$$N_{Fk} = \sum_{j \in U_M} l^{MT}_{jk} N_{Fj} = C_k \sum_{j \in U_M} l^{MT}_{jk} = C_k N_{Mk} = N_{Fj} N_{Mk}. \tag{10}$$

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 4).

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

$$\sum_{j \in U_M} \tilde{\theta}^{FM}_{ij} \tilde{\theta}^{MT}_{jk} = \sum_{j \in U_M} \frac{l^{FM}_{ij}}{N_{Fj}} \frac{l^{MT}_{jk}}{N_{Mk}} = \frac{l_{ik}}{N_{Fk}} = \tilde{\theta}_{ik}. \tag{11}$$

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 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. 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.

4 Simulation study

We have shown in Subsection 3.1 that, in the case of MtO links and sampling designs with the $\mathbf{\Delta}$-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.

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 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 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 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 3.3 (see

(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 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 (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.

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 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 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

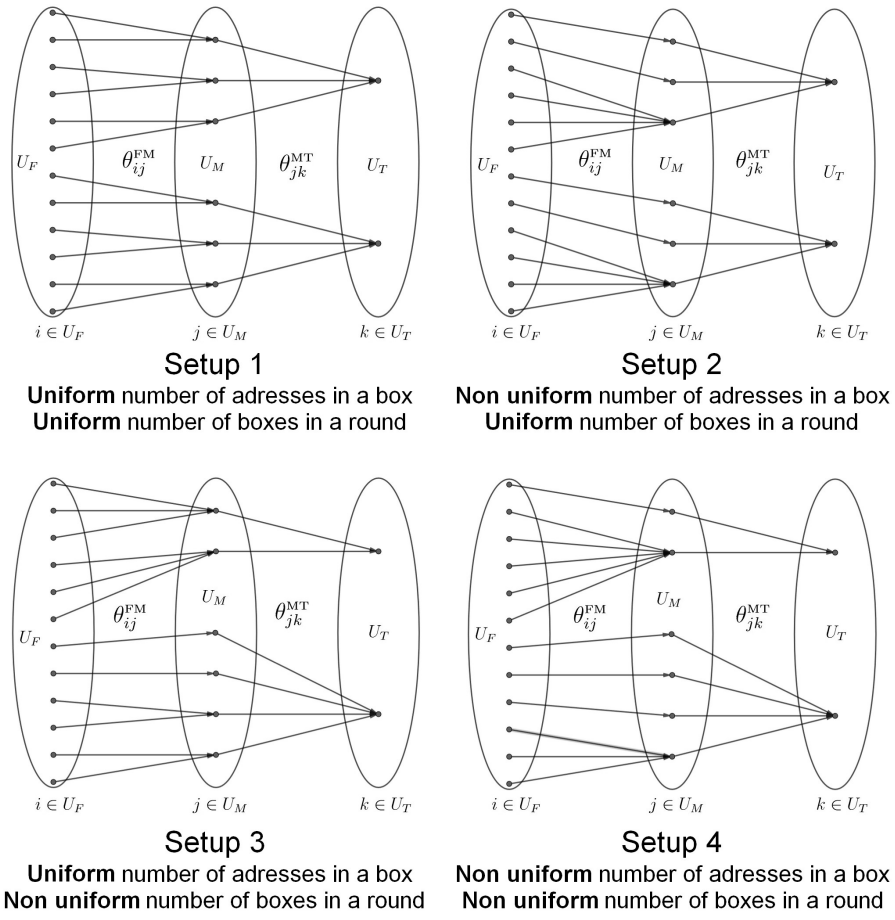

Figure 4: The four setups

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

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 1: Relative bias and efficiency of the DGWSM estimate under different links setups.

not depend on the simulation setup, the simple GWSM does not vary between the setups.

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

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

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

Table 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 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 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 (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 (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.

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 and 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, the $\mathbf{\Delta}$-property holds for this sampling design (see details in Subsection 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 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 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

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: 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.

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 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). 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). 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 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 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, 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.

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.

Acknowledgements

This work has been partly supported by the French *Agence Nationale de la Recherche* through CIFRE contract 2019/1966 and through the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010.

7 Appendix

Proof of Proposition 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:

$$\tilde{\boldsymbol{\Theta}}_k = \begin{pmatrix} \mathbf{0} \\ \tilde{\boldsymbol{\theta}}_k \\ \mathbf{0} \end{pmatrix}, \quad k \in U_F, \tag{12}$$

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) can be written as follows:

$$\begin{aligned} \mathrm{Var}(\hat{t}_{y1}) &= \sum_{k \in U_T} \sum_{k' \in U_T} y_k y_{k'} \mathrm{Cov}(\hat{t}_{\tilde{\theta}_k}, \hat{t}_{\tilde{\theta}_{k'}}) \\ &= \sum_{k \in U_T} \sum_{k' \in U_T} y_k y_{k'} \tilde{\boldsymbol{\Theta}}_k^t \boldsymbol{\Delta} \tilde{\boldsymbol{\Theta}}_{k'} \\ &= \sum_{k \in U_T} \sum_{k' \in U_T} y_k y_{k'} \tilde{\boldsymbol{\theta}}_k^t \boldsymbol{\Delta}_{kk'} \tilde{\boldsymbol{\theta}}_{k'}, \end{aligned} \tag{13}$$

where $\boldsymbol{\Delta} = \left(\dfrac{\pi_{ii'} - \pi_i \pi_{i'}}{\pi_i \pi_{i'}} \right)_{i,i' \in U_F}$ 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 (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:

$$\mathbb{1}_k^t \tilde{\boldsymbol{\theta}}_k = 1 \quad \text{for all} \quad k \in U_F, \tag{14}$$

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

$$\begin{aligned} \mathrm{Var}(\hat{t}_{y1}) &= \sum_{k \in U_T} y_k^2 \tilde{\boldsymbol{\theta}}_k^t \boldsymbol{\Delta}_k \tilde{\boldsymbol{\theta}}_k + \sum_{k \in U_T} \sum_{k' \neq k \in U_T} y_k y_{k'} \tilde{\boldsymbol{\theta}}_k^t \boldsymbol{\Delta}_{kk'} \tilde{\boldsymbol{\theta}}_{k'} \\ &= \sum_{k \in U_T} y_k^2 \tilde{\boldsymbol{\theta}}_k^t \boldsymbol{\Delta}_k \tilde{\boldsymbol{\theta}}_k + \sum_{k \in U_T} \sum_{k' \neq k \in U_T} y_k y_{k'} c_{kk'} \tilde{\boldsymbol{\theta}}_k^t \mathbb{1}_k \mathbb{1}_{k'}^t \tilde{\boldsymbol{\theta}}_{k'} \\ &= \sum_{k \in U_T} y_k^2 \mathrm{Var}(\hat{t}_{\tilde{\theta}_k}) + \sum_{k \in U_T} \sum_{k' \neq k \in U_T} y_k y_{k'} c_{kk'}. \end{aligned} \tag{15}$$

This is exactly (6) and finishes the proof of the first part of Proposition 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 $\tilde{\boldsymbol{\theta}}_k^{opt}, k \in U_F$ that minimize

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

under the standardization constraint (14). The variance $\mathrm{Var}(\hat{t}_{y1})$ is minimized for vectors $\tilde{\boldsymbol{\theta}}_k$ verifying the following equation (see Deville and Lavallée, 2006, equation 6.4):

$$y_k \sum_{k' \in U_T} \boldsymbol{\Delta}_{kk'} \tilde{\boldsymbol{\theta}}_{k'} y_{k'} = \lambda_k \mathbb{1}_k, \quad k \in U_F, \tag{16}$$

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

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

where $\boldsymbol{\Delta}_k = \boldsymbol{\Delta}_{kk}$. Equation (16) can be rewritten as:

$$y_k^2 \boldsymbol{\Delta}_k \tilde{\boldsymbol{\theta}}_k + y_k \sum_{k' \neq k \in U_T} \boldsymbol{\Delta}_{kk'} \tilde{\boldsymbol{\theta}}_{k'} y_{k'} = \lambda_k \mathbb{1}_k, \quad k \in U_F,$$

which implies

$$\tilde{\boldsymbol{\theta}}_k = y_k^{-2} \boldsymbol{\Delta}_k^{-1} \left(\lambda_k \mathbb{1}_k - y_k \sum_{k' \neq k \in U_T} \boldsymbol{\Delta}_{kk'} \tilde{\boldsymbol{\theta}}_{k'} y_{k'} \right), \quad k \in U_F. \tag{17}$$

By using the fact that $\boldsymbol{\Delta}_{k,k' \neq k} = c_{kk'} \mathbb{1}_k \mathbb{1}_{k'}^t$ and the standardization constraints (14), we get:

$$\tilde{\boldsymbol{\theta}}_k = y_k^{-2} \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k \left(\lambda_k - y_k \sum_{k' \neq k \in U_T} c_{kk'} y_{k'} \right), \quad k \in U_F. \tag{18}$$

Multiplying by $\mathbb{1}_k^t$ the equation (18) and using again the standardization constraints (14), we obtain the following expression for the Lagrange multipliers:

$$\lambda_k = y_k^2 \left(\mathbb{1}_k^t \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k\right)^{-1} + y_k \sum_{k' \neq k \in U_T} c_{kk'} y_{k'}, \quad k \in U_F. \tag{19}$$

Finally, by plugging the expression of λ_k given in (19) in the expression of $\tilde{\boldsymbol{\theta}}_k$ from (18), we obtain:

$$\begin{aligned}
\tilde{\boldsymbol{\theta}}_k^{opt} &= y_k^{-2} \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k \left(\lambda_k - y_k \sum_{k' \neq k \in U_T} c_{kk'} y_{k'} \right) \\
&= y_k^{-2} \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k \left(y_k^2 \left(\mathbb{1}_k^t \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k\right)^{-1} + y_k \sum_{k' \neq k} c_{kk'} y_{k'} - y_k \sum_{k' \neq k} c_{kk'} y_{k'} \right) \\
&= \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k \left(\mathbb{1}_k^t \boldsymbol{\Delta}_k^{-1} \mathbb{1}_k\right)^{-1}.
\end{aligned}$$

This finishes the proof of 2.1. □

Proof of Proposition 2.2

Result (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:

$$\text{Var}(\hat{t}_{y1}^{opt}) = \sum_{k \in U_T} y_k^2 \text{Var}(\hat{t}_{\tilde{\theta}_k^{opt}}) + \sum_{k \in U_T} \sum_{k' \neq k \in U_T} y_k y_{k'} c_{kk'}. \tag{20}$$

Using equations (15) and (20), we get:

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

Let us show that for the optimal weights derived in Proposition 2.1, we have:

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

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

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

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 (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

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

which ends the proof of Proposition 2.2. □

Proof of Proposition 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 $\text{Var}(\hat{t}_y) < \text{Var}(\hat{t}_{y1}^{opt})$ 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 $k \in U_F$, 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 2.3. □

References

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 statistica ufficiale*, 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*, pages 219–227. Dunod, Paris, France.

Deville, J. and Maumy-Bertrand, M. (2006). Extension of the indirect sampling method and its application to tourism. *Survey Methodology*, 32(2):177–185.

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

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.

Haziza, D. and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32(2):206–226.

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.

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, New York.

Rendtel, U. and Harms, T. (2009). Weighting and calibration for household panels. *Methodology of longitudinal surveys*, pages 265–286.

Xu, X. and Lavallée, P. (2009). Treatments for link nonresponse in indirect sampling. *Survey Methodology*, 35(2):153–164.