WORKING PAPERS

N° 1344

June 2022

“QR Prediction for Statistical Data Integration”

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

Toulouse School of Economics

QR Prediction for Statistical Data Integration

Estelle Medous [*,†,‡], Camelia Goga[†], Anne Ruiz-Gazen[*],

Jean-François Beaumont[§], Alain Dessertaine[‡], Pauline Puech[‡]

Abstract

In this paper, we investigate how a big non-probability database can be used to improve estimates 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, proposed by Särndal and Wright (1984) to handle the case where the non-probability database contains auxiliary variables but no study variable. We provide conditions ensuring that the QR predictor is asymptotically design-unbiased. Assuming the probability sampling design is not informative, the QR predictor is also model-unbiased regardless of the validity of those conditions. We compare the design properties of different predictors, in the class of QR predictors, through a simulation study. They include 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. As expected, the model-based predictor did not perform well when the underlying model was misspecified.

Keyword: cosmetic estimator, dual-frame, GREG estimator, non-probability sample, probability sample.

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

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

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

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

[§]Statistique Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada. Email: jean-francois.beaumont@statcan.gc.ca

through monthly probability surveys. Controlling the sampling design allows design-based inference without resorting to modeling of the study variables, and therefore is particularly attractive to 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), and Kim (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 (the different types of mails sent) are only available in the probability sample whereas auxiliary information is only available in the non-probability database. The aim of the present paper is to study this particular context thoroughly.

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 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 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 condition such that these predictors are asymptotically design-unbiased. We also show that they are unbiased under the model and sampling design. In Section 4, we use Monte Carlo simulations to compare several QR predictors and show that the cosmetic estimator is a good compromise for several setups. Finally we conclude and give perspectives in Section 5.

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 the sample 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 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. Let us denote 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 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 (rather) 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:

$$\hat{T}_{DI} = T_{NP} + \sum_{k\in s_P} d_k(1-\delta_k)y_k, \tag{1}$$

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 already proposed in 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 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 $CV_{NP} = \sqrt{S^2_{Y,NP}}/\bar{Y}_{NP}$ the coefficient of variation of Y in s_{NP}, with $S^2_{Y,NP} = \frac{1}{N_{NP}-1}\sum_{k\in U}\delta_k(y_k-\bar{Y}_{NP})^2$.

The proof of Proposition 2.1 is given in the appendix. Intuitively, the result of Proposition 2.1 (ii) 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:

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

where

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

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 auxiliary variables available on s_{NP} in the calibration equation.

Following Kim and Tam (2021), it is possible to use the linearization approach and derive the approximate variance of $\hat{T}_{PDI}$, denoted as $\text{AVar}(\hat{T}_{PDI})$. 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 $T_C = \sum_U (1-\delta_k)y_k$. The gain in efficiency when using Hájek is not true in general (see, e.g., Särndal et al. (1992)) but it can be substantial in some contexts as illustrated in the simulation setups of Section 4 when comparing $\hat{T}_{HT}$ and $\hat{T}_{PDI}$ for Poisson sampling. 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 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 precise expression of the difference between the variances.

Proposition 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 concerns 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 $k \in s_{NP} - s_P$.

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 2 by the Hájek-type estimator $\hat{T}_C^{(\mathrm{Ha})}$. In the present section, y_k in unknown for $k \in s_{NP}$, and contrarily to Section 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 study bias properties of the QR predictor under the design as well as under the joint distribution induced by the model and the sampling design. 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 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 1: Data characteristics in the data integration context of Section 3.

3.1 QR predictors

The variable Y is not available in s_{NP} and we cannot use anymore $\hat{T}_{PDI}$ since the sum $T_{NP} = \sum_{k\in U} \delta_k y_k$ 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 $\boldsymbol{x}_k$:

$$y_k = \boldsymbol{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \quad k \in s_{NP}, \tag{2}$$

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 (2) conditionally on observed auxiliary variables $\boldsymbol{x}_k$, $k \in s_{NP}$. Note that model (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

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k \in s_P} q_k \delta_k \boldsymbol{x}_k \boldsymbol{x}_k^\top\right)^{-1} \left(\sum_{k \in s_P} q_k \delta_k \boldsymbol{x}_k y_k\right), \tag{3}$$

where q_k are known positive constants for $k \in s_{NP}$. We assume that the $p \times p$ dimensional matrix $\sum_{k \in s_P} q_k \delta_k \boldsymbol{x}_k \boldsymbol{x}_k^\top$ and $\sum_{k \in U} \pi_k q_k \delta_k \boldsymbol{x}_k \boldsymbol{x}_k^\top$ are 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):

$$\begin{aligned} \hat{T}_{NP}^{(\mathrm{QR})} &= \sum_{k \in U} \delta_k \hat{y}_k + \sum_{k \in s_P} r_k \delta_k (y_k - \hat{y}_k) \\ &= \sum_{k \in U} \delta_k \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}} + \sum_{k \in s_P} r_k \delta_k (y_k - \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}}), \end{aligned} \tag{4}$$

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

$$\hat{T}^{(\mathrm{QR})} = \hat{T}_{NP}^{(\mathrm{QR})} + \hat{T}_C^{(\mathrm{Ha})}. \tag{5}$$

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 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 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 3.1. *(projection form) Consider the QR predictor $\hat{T}_{NP}^{(\mathrm{QR})}$ given by (4). Under the condition that there exists a vector $\boldsymbol{\mu} \in \mathbf{R}^p$ such that*

$$(Proj): \quad \boldsymbol{\mu}^\top \boldsymbol{x}_k q_k = r_k \quad \textit{for all} \quad k \in s_{NP} \cap s_P, \tag{6}$$

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 $k \in s_{NP} \cap s_P$. This condition is satisfied when v_k is one of the auxiliary variables in the model. If $v_k = 1$, it is satisfied provided an 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 $k \in s_{NP} \cap s_P$. A consequence of Proposition 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 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*

$$(QR): \quad 1 - \pi_k r_k = \pi_k q_k \boldsymbol{x}_k^\top \boldsymbol{\lambda} \quad \textit{for all} \quad k \in s_{NP}. \tag{7}$$

Then:

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

where

$$\hat{T}_{NP}^{(\mathrm{Q}\pi)} = \sum_{k\in U} \delta_k \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}} + \sum_{k\in s_P} d_k \delta_k (y_k - \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}}) \tag{8}$$

is the model-assisted type predictor of T_{NP} with $\hat{\boldsymbol{\beta}}$ given by (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.

3.2 Bias properties

Let us consider the QR class of predictors that satisfy the (QR) condition given by (7). For this class of predictors, the final estimator of T is

$$\hat{T}^{(\mathrm{Q}\pi)} = \hat{T}_{NP}^{(\mathrm{Q}\pi)} + \hat{T}_{C}^{(\mathrm{Ha})}.$$

The total error is given by:

$$\hat{T}^{(\mathrm{Q}\pi)} - T = (\hat{T}_{NP}^{(\mathrm{Q}\pi)} - T_{NP}) + (\hat{T}_{C}^{(\mathrm{Ha})} - T_C).$$

The estimator $\hat{T}^{(\mathrm{Q}\pi)}$ is not exactly design-unbiased because of the nonlinearity of $\hat{\boldsymbol{\beta}}$ and of the Hajek estimator $\hat{T}_{C}^{(\mathrm{Ha})}$. Following Särndal (1980), we rather look at the asymptotic design-unbiasedness of the estimators.

Let us consider the asymptotic framework from Isaki and Fuller (1982), which allows for the population and the sample sizes to grow to infinity. A predictor $\hat{T}$ 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 E_p is the expectation under the design. Wright (1983) proved that the (QR) condition given in proposition 3.2 is a sufficient condition for $\hat{T}_{NP}^{(\mathrm{Q}\pi)}$ to be asymptotically design-unbiased for T_{NP}, provided that

$$\lim_{N\to\infty} \frac{1}{N} \mathrm{E}_p \left[\left(\sum_{k\in U} \delta_k \boldsymbol{x}_k - \sum_{k\in s_P} d_k \delta_k \boldsymbol{x}_k \right)^\top (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \right] = 0, \tag{9}$$

where $\tilde{\boldsymbol{\beta}} = (\sum_{k\in U} \pi_k q_k \delta_k \boldsymbol{x}_k \boldsymbol{x}_k^\top)^{-1} \sum_{k\in U} \pi_k q_k \delta_k \boldsymbol{x}_k y_k$. Following Breidt and Opsomer (2000), if the sampling fraction n/N converges to a constant different from 0, assuming mild conditions on the 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 (9) follows by assuming that the regression coefficient estimator satisfies $\lim_{N\to\infty} \mathrm{E}_p ||\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}||^2 = 0$. The estimator $\hat{T}_{C}^{(\mathrm{Ha})}$ is a Hájek-type estimator of T_C. Assuming that the probability of the intersection set $s_P \cap s_{NP}$ to be empty is negligible, then $\hat{T}_{C}^{(\mathrm{Ha})}$ is asymptotically design-unbiased for T_C. From the above, we conclude that the (QR) predictor is asymptotically design-unbiased for T.

Assuming the sampling design is not informative with respect to model (2), we can prove that the the QR predictor $\hat{T}_{NP}^{(\mathrm{QR})}$ is model-unbiased for T_{NP}. The model bias of $\hat{T}_{NP}^{(\mathrm{QR})}$ is given by:

$$\mathrm{E}_m(\hat{T}_{NP}^{(\mathrm{QR})} - T_{NP}) \quad = \sum_{k\in U} \delta_k \mathrm{E}_m(\boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}} - y_k) + \sum_{k\in s_P} r_k \delta_k \mathrm{E}_m(y_k - \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}}), \tag{10}$$

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 model (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 $\mathrm{E}_m(\boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}} - y_k) = 0$. Thus, $\hat{T}_{NP}^{(\mathrm{QR})}$ is model-unbiased for T_{NP} without requiring the QR condition. As a result, under non-informative sampling and conditioning on y_k, $k \in U - s_{NP}$, $\hat{T}^{(\mathrm{QR})}$ is asymptotically mp-unbiased for T.

4 Simulations

In this section, we conduct a Monte-Carlo study to compare the efficiency of some of the QR predictors, $\hat{T}^{(\mathrm{QR})} = \hat{T}_{NP}^{(\mathrm{QR})} + \hat{T}_C^{(\mathrm{Ha})}$ from Section 3, namely the model-assisted, the model-based and the cosmetic estimators, assuming that $v_k = 1$ in model (2). We are also interested in comparing these estimators with the expansion estimator and the PDI estimator defined in Section 2. To illustrate that the superiority of some estimators compared to others depend on the data, we define three different setups based on different artificial populations. As mentioned in Subsection 3.1, 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.

4.1 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, the conditional variable $Y|X_1, X_2$ 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:

$$\mu_{Y|X_1,X_2} = a_0 + a_1 X_1 + a_2 X_2.$$

2. For Model 2, $\mu_{Y|X_1,X_2}$ is a quadratic function of X_1 and a linear function of X_2:

$$\mu_{Y|X_1,X_2} = b_0 + b_1(X_1 - \bar{X}_1)^2 + b_2 X_2 \text{ with } \bar{X}_1 \text{ the mean of } X_1 \text{ over } U.$$

3. For Model 3, $\mu_{Y|X_1,X_2}$ is a linear function of X_2:

$$\mu_{Y|X_1,X_2} = c_0 + c_2 X_2.$$

To compare the results 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:

$$\gamma = \mathrm{Var}(a_1 X_1)/\mathrm{Var}(a_2 X_2) = \mathrm{Var}(b_1(X_1 - \bar{X}_1)^2)/\mathrm{Var}(b_2 X_2).$$

  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 Subsection 4.2, 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 2. A non-probability sample of size 900 is drawn

Model	Mean of (X_1, X_2)	Std of (X_1, X_2)	Mean of $Y\|X_1, X_2$	R^2
1			$\mu_Y = a_0 + a_1 X_1 + a_2 X_2$	equal
2	(20,30)	(15,20)	$\mu_Y = b_0 + b_1(X_1 - \bar{X}_1)^2 + b_2 X_2$	between
3			$\mu_Y = c_0 + c_2 X_2$	populations

Table 2: Population models with $\mu = 100$, σ^2=100, and $\gamma = 0.5$.

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 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 3: Three studied setups.

4.2 Results

Let us consider the three setups defined above and compare the following estimators:

- $\hat{T}_{HT} = \sum_{k \in s_P} d_k y_k,$
- $\hat{T}_{PDI} = T_{NP} + (N - N_{NP}) \frac{\sum_{k \in s_P} d_k (1 - \delta_k) y_k}{\sum_{k \in s_P} d_k (1 - \delta_k)},$
- $\hat{T}^{(\text{MB})} = \sum_{k \in U} \delta_k \hat{y}_k^{(\text{MB})} + \sum_{k \in s_P} \delta_k (y_k - \hat{y}_k^{(\text{MB})}) + (N - N_{NP}) \frac{\sum_{s_P} d_k (1 - \delta_k) y_k}{\sum_{k \in s_P} d_k (1 - \delta_k)},$
- $\hat{T}^{(\text{MA})} = \sum_{k \in U} \delta_k \hat{y}_k^{(\text{MA})} + \sum_{k \in s_P} \delta_k d_k (y_k - \hat{y}_k^{(\text{MA})}) + (N - N_{NP}) \frac{\sum_{k \in s_P} d_k (1 - \delta_k) y_k}{\sum_{k \in s_P} d_k (1 - \delta_k)},$
- $\hat{T}^{(\text{Cos})} = \sum_{k \in U} \delta_k \hat{y}_k^{(\text{Cos})} + \sum_{k \in s_P} \delta_k (y_k - \hat{y}_k^{(\text{Cos})}) + (N - N_{NP}) \frac{\sum_{k \in s_P} d_k (1 - \delta_k) y_k}{\sum_{k \in s_P} d_k (1 - \delta_k)}.$

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 the estimators:

$$\text{RB}_{MC}(\hat{R}) = 100 \times \frac{1}{L} \sum_{l=1}^{L} \frac{\hat{T}^{(l)} - T}{T}$$

where $\hat{T}^{(l)}$ is an estimate of T ($\hat{T}_{HT}$, $\hat{T}^{(\text{MB})}$, $\hat{T}^{(\text{MA})}$, $\hat{T}^{(\text{Cos})}$ or $\hat{T}_{PDI}$), 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 $\hat{T}^{(\text{Cos})}$):

$$\text{RMSE}_{MC}(\hat{T}^{(\text{Cos})}) = 100 \times \frac{\text{MSE}_{MC}(\hat{T})}{\text{MSE}_{MC}(\hat{T}^{(\text{Cos})})},$$

where

$$\text{MSE}_{MC}(\hat{T}) = L^{-1} \sum_{l=1}^{L} \left(\hat{T}^{(l)} - T \right)^2.$$

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
		$\mathrm{RVar}_{MC}(\hat{T}^{(\mathrm{Cos})})$	23566.93	55.62	114.06	100.00	20.97
		$\mathrm{RMSE}_{MC}(\hat{T}^{(\mathrm{Cos})})$	22897.58	2715.21	113.91	100.00	20.65
	Setup 2	RB_{MC}	-0.07	-1.65	-0.06	-0.05	0.02
		$\mathrm{RVar}_{MC}(\hat{T}^{(\mathrm{Cos})})$	36947.99	84.94	118.21	100.00	23.17
		$\mathrm{RMSE}_{MC}(\hat{T}^{(\mathrm{Cos})})$	36638.27	1056.44	118.42	100.00	23.15
	Setup 3	RB_{MC}	0.03	-0.01	0.01	0.01	0.01
		$\mathrm{RVar}_{MC}(\hat{T}^{(\mathrm{Cos})})$	41088.93	58.38	100.49	100.00	33.47
		$\mathrm{RMSE}_{MC}(\hat{T}^{(\mathrm{Cos})})$	41080.51	58.39	100.48	100.00	33.51

Table 4: Relative bias (in %), relative variance and MSE compared to the Cosmetic estimator (in %) of the different estimators for the 3 different setups. Expected size of the probability sample: 200. Size of the non-probability sample: 900.

We also compute the Monte Carlo relative variance (RVar) of an estimator $\hat{T}$ (relative to $\hat{T}^{(\mathrm{Cos})}$):

$$\mathrm{RVar}_{MC}(\hat{T}^{(\mathrm{Cos})}) = 100 \times \frac{\mathrm{Var}_{MC}(\hat{T})}{\mathrm{Var}_{MC}(\hat{T}^{(\mathrm{Cos})})},$$

where

$$\mathrm{Var}_{MC}(\hat{T}) = L^{-1} \sum_{l=1}^{L} \left(\hat{T}^{(l)}\right)^2 - \left(L^{-1} \sum_{l=1}^{L} \hat{T}^{(l)}\right)^2.$$

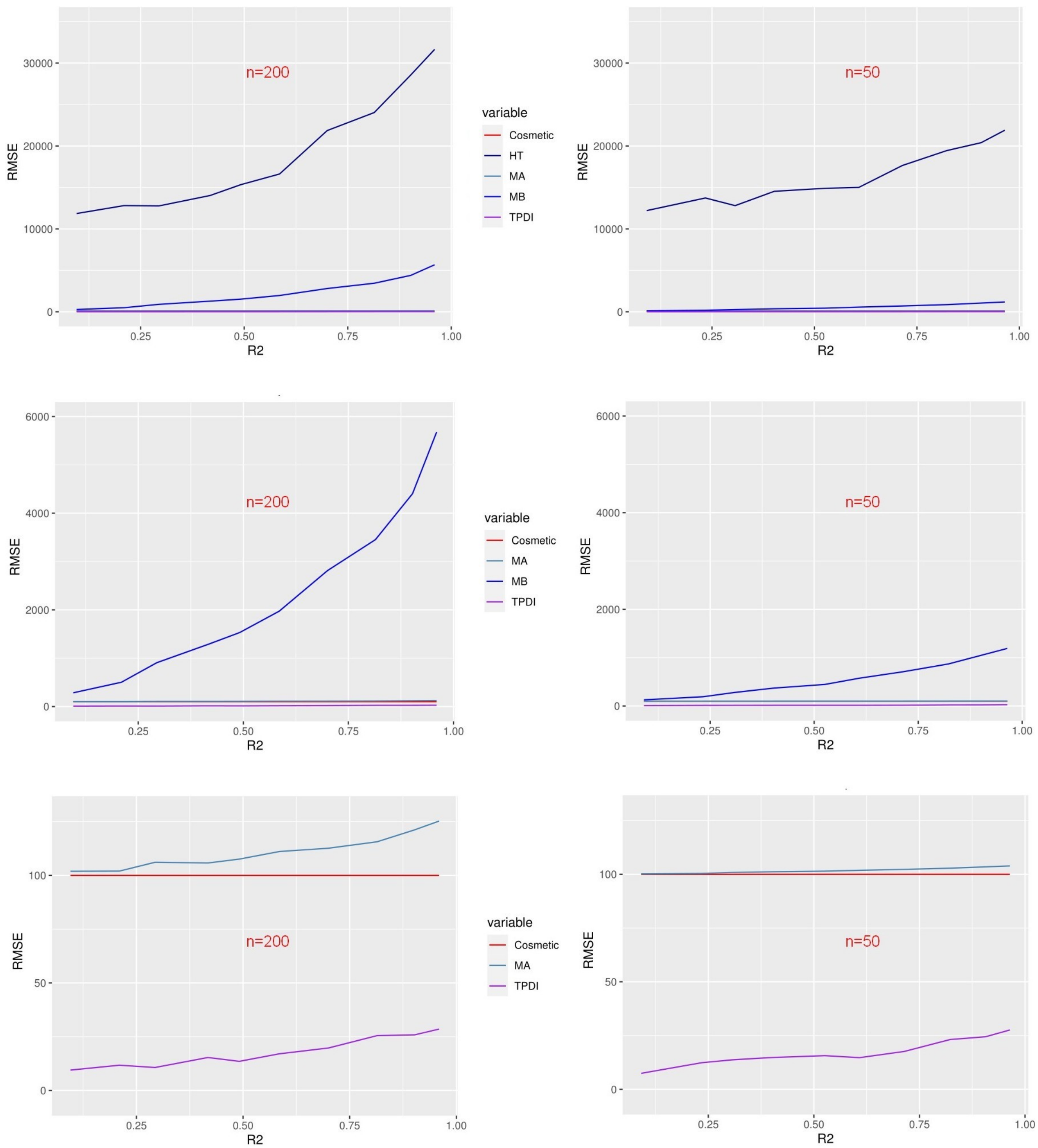
Table 4 contains the simulation results for the three setup when $R^2 = 0.8$. In all setups, we confirm that both $\hat{T}_{PDI}$ and $\hat{T}_{HT}$ have a small Monte Carlo bias, as expected. In terms of MSE, $\hat{T}_{PDI}$ is the most precise estimator, while $\hat{T}_{HT}$ is the least precise estimator 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 while $\hat{T}^{(\mathrm{MB})}$ 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 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 $\hat{T}^{(\mathrm{Cos})}$). The main difference lies in the use of a non weighted sum of residuals for $\hat{T}^{(\mathrm{Cos})}$ and of a weighted sum of residuals for $\hat{T}^{(\mathrm{MA})}$. When weighted regression methods are used to predict $y_k, k \in s_{NP}$, an unweighted sum of the residuals is recommended in the definition of the estimator when the model is misspecified.

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 $\hat{T}^{(\mathrm{MA})}$ because the sum of residuals in the Cosmetic estimator is unweighted.

To better understand the impact of the R^2 on the results, we also plot, on the y-axis of Figures 1, 2 and 3, the $\mathrm{RMSE}_{MC}(\hat{T}^{(\mathrm{Cos})})$ 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 1 (resp. Figure 2 and 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 is 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 1 and 2), $\hat{T}_{HT}$ (and $\hat{T}^{(\mathrm{MB})}$ for Figures 1 and 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 compared to $\hat{T}^{(\mathrm{Cos})}$ 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 1 (resp. 2) shows the evolution of $\mathrm{RMSE}_{MC}(\hat{T}_{Cos})$ with respect to the 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, 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 setups, a smaller size reduces the difference between $\mathrm{RMSE}_{MC}(\hat{T}^{(\mathrm{Cos})})$ of QR predictors. Figure 3 shows the evolution of $\mathrm{RMSE}_{MC}(\hat{T}^{(\mathrm{Cos})})$ with respect to the R^2 in the non-informative setup. This time, $\hat{T}^{(\mathrm{MB})}$ 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 $\hat{T}^{(\mathrm{Cos})}$ and $\hat{T}^{(\mathrm{MA})}$ 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



i

Figure 1: Relative MSE (in %), with the Cosmetic estimator as the baseline, versus R^2 in the informative setup

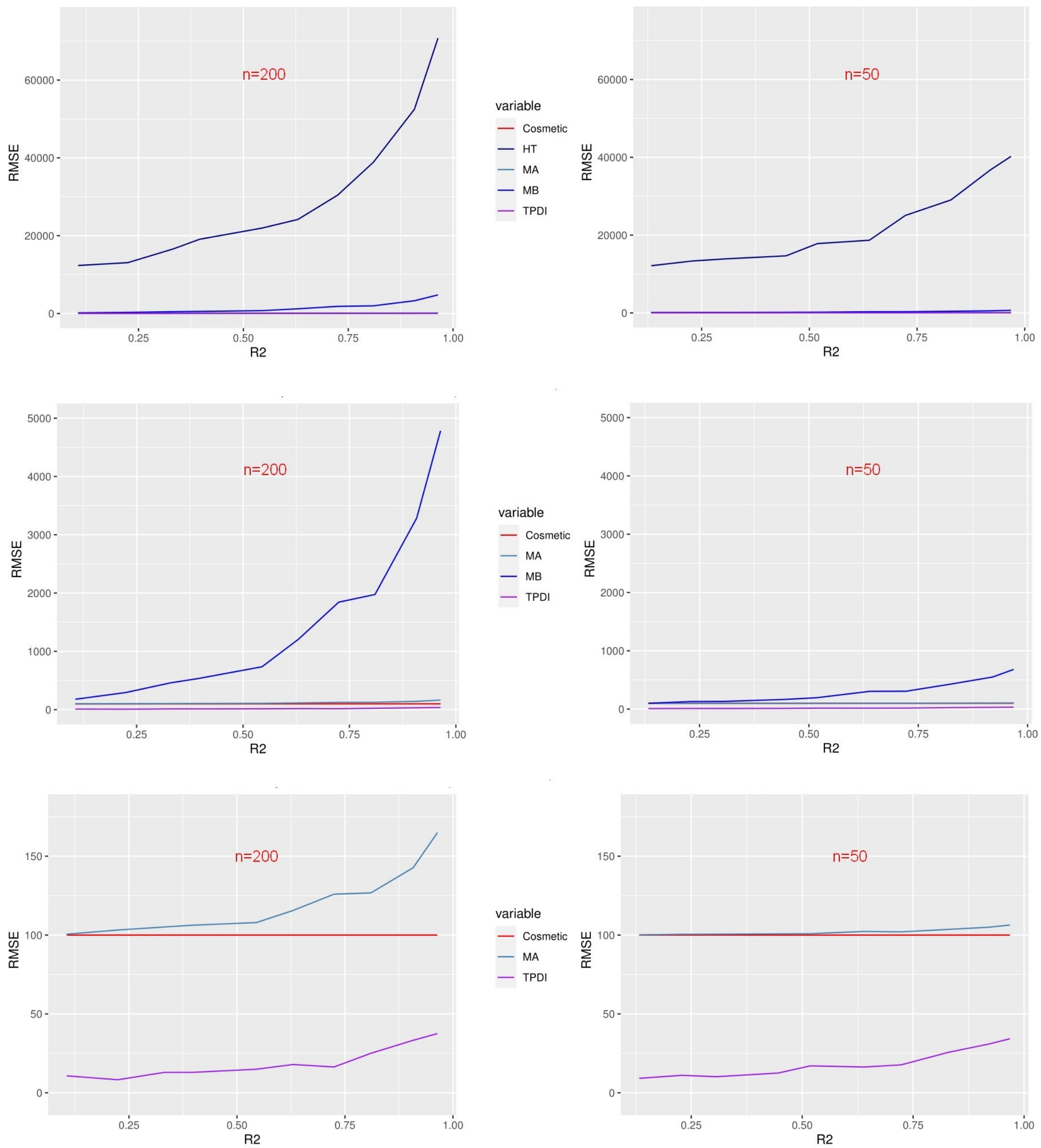


Figure 2: Relative MSE (in %), with the Cosmetic estimator as the baseline, versus R^2 in the quadratic setup

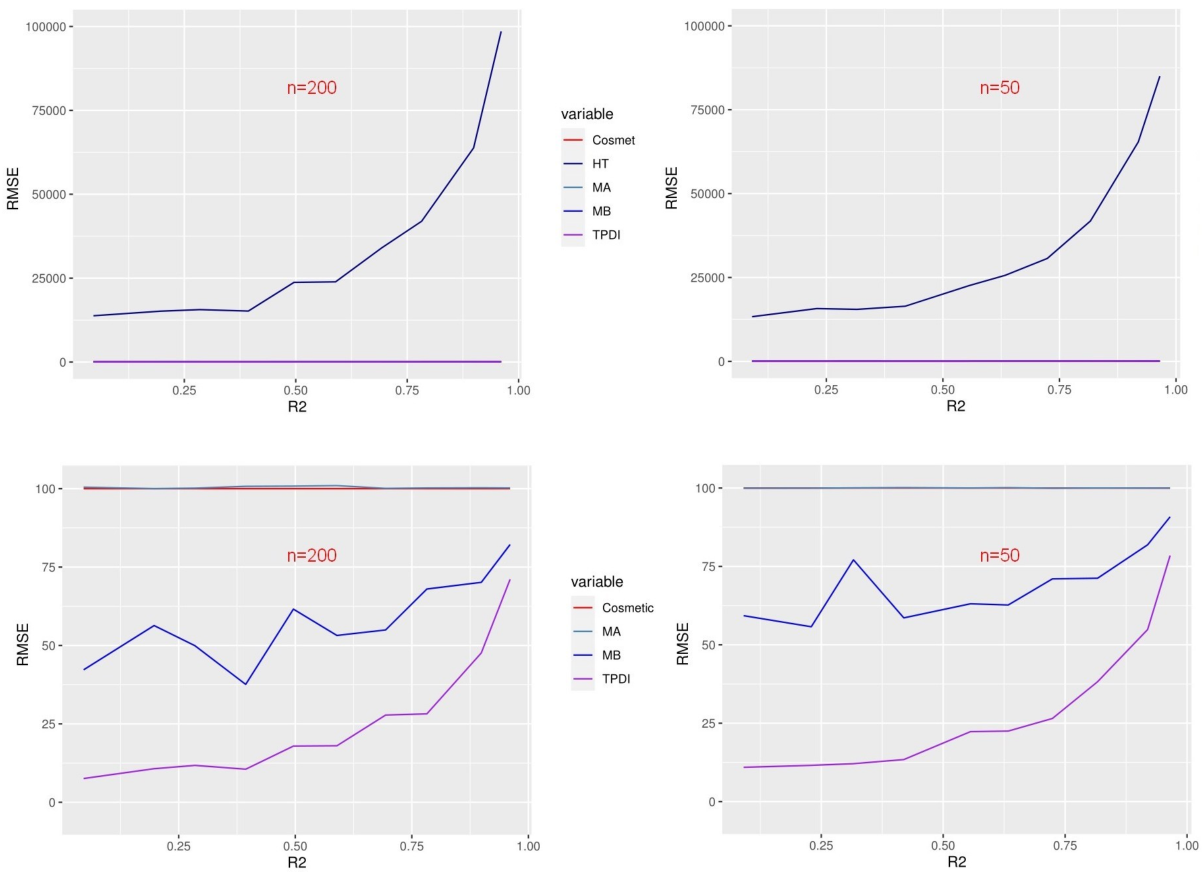


Figure 3: Relative MSE (in %), with the Cosmetic estimator as the baseline, versus R^2 in the non-informative setup

increases faster with R^2 than the other estimators. The advantage of $\hat{T}^{(\mathrm{Cos})}$ over $\hat{T}^{(\mathrm{MA})}$ might disappear in a scenario where the probability sample size would be a smaller fraction of the population size. $\hat{T}^{(\mathrm{MB})}$ 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, $\hat{T}^{(\mathrm{MB})}$ 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.

5 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 in presence of auxiliary information. 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 properties. We have also 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.

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.

References

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.

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.

Kim, J. K. (2022). A gentle introduction to data integration in survey sampling. *The survey statistician*, 85:19–29.

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.

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.

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. and Wright, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian J. of Statistics*, 11:146–156.

Wright, R. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78(384):879–884.

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.

Appendix

Proof of Proposition 2.1

We recall that $\hat{T}_{DI} = \sum_{k\in U} \delta_k y_k + \sum_{k\in s}(1-\delta_k)d_k y_k$ and $\hat{T}_{HT} = \sum_{k\in s} d_k y_k = \sum_{k\in s} \delta_k d_k y_k + \sum_{k\in s}(1-\delta_k)d_k y_k$. Thus, we have:

$$\text{Var}(\hat{T}_{HT}) - \text{Var}(\hat{T}_{DI}) = \text{Var}\left(\sum_{k\in s} \delta_k d_k y_k\right) + 2\,\text{Cov}\left(\sum_{k\in s} \delta_k d_k y_k, \sum_{k\in s}(1-\delta_k)d_k y_k\right).$$

(i) For Poisson sampling, we have:

$$\text{Cov}\left(\sum_{k\in s} \delta_k d_k y_k, \sum_{k\in s}(1-\delta_k)d_k y_k\right) = \sum_{k\in s} \delta_k(1-\delta_k)d_k y_k^2 = 0$$

and

$$\text{Var}(\hat{T}_{HT}) - \text{Var}(\hat{T}_{DI}) = \text{Var}\left(\sum_{k\in s} \delta_k d_k y_k\right) = \sum_{k\in U} \delta_k(d_k-1)y_k^2 \geq 0,$$

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:

$$\text{Var}\left(\sum_{k\in s} \delta_k d_k y_k\right) = \frac{N}{n}\frac{N-n}{N(N-1)}\left(N(N_{NP}-1)S^2_{y,NP} + N_{NP}\bar{Y}^2_{NP}\,(N-N_{NP})\right),$$

$$\text{Cov}\left(\sum_{k\in s} \delta_k d_k y_k, \sum_{k\in s}(1-\delta_k)d_k y_k\right) = -\frac{N}{n}\frac{N-n}{N(N-1)}N_{NP}\bar{Y}_{NP}\left(N\bar{Y}_U - N_{NP}\bar{Y}_{NP}\right),$$

and thus

$$\text{Var}(\hat{T}_{HT}) - \text{Var}(\hat{T}_{DI}) = \frac{N}{n}\frac{N-n}{N(N-1)}\left(N(N_{NP}-1)S^2_{y,NP} + N_{NP}\bar{Y}_{NP}\left((N+N_{NP})\bar{Y}_{NP} - 2N\bar{Y}_U\right)\right).$$

We conclude that $\text{Var}(\hat{T}_{HT})$ is larger than or equal to $\text{Var}(\hat{T}_{DI})$ 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 2.2

We have:

$$\mathrm{Var}(\hat{T}_{HT}) = \mathrm{Var}\left(\sum_{k\in s} d_k y_k\right) = N^2(1-f)\frac{S_y^2}{n},$$

$$\mathrm{AVar}(\hat{T}_{PDI}) = \mathrm{Var}\left(\sum_{k\in s}(1-\delta_k)d_k(y_k-\bar{Y}_C)\right) = \mathrm{Var}\left(\sum_{k\in s} d_k\tilde{y}_k\right) = N^2(1-f)\frac{S_{\tilde{y}}^2}{n}$$

where

$$\begin{aligned} S_y^2 &= \frac{1}{N-1}\sum_{k\in U}(y_k-\bar{Y}_U)^2,\\ \tilde{y}_k &= (1-\delta_k)(y_k-\bar{Y}_C),\\ S_{\tilde{y}}^2 &= \frac{1}{N-1}\sum_{k\in U}(\tilde{y}_k-\bar{\tilde{Y}}_C)^2 = \frac{1}{N-1}\sum_{k\in U}\tilde{y}_k^2. \end{aligned}$$

Using some basic but tedious calculus, we obtain:

$$\begin{aligned} \mathrm{Var}(\hat{T}_{HT}) - \mathrm{AVar}(\hat{T}_{PDI}) &= N^2(1-f)\frac{S_y^2-S_{\tilde{y}}^2}{n}\\ &= N^2(1-f)\frac{1}{n}\frac{1}{N-1}\left(\sum_{k\in U}\delta_k(y_k-\bar{Y}_U)^2+(N-N_{NP})(\bar{Y}_C-\bar{Y}_U)^2\right)\\ &= N^2(1-f)\frac{1}{n}\frac{1}{N-1}\left(S_{NP}^2(N_{NP}-1)+N_{NP}\frac{N}{N-N_{NP}}(\bar{Y}_{NP}-\bar{Y}_U)^2\right) \end{aligned}.$$

□

Proof of Proposition 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 $\mathbf{Q}_{\mathrm{xsp}}^\top = \mathbf{X}_{s_P}^\top\mathrm{diag}(q_k\delta_k)_{k\in s_P}$. Then $\hat{\boldsymbol{\beta}} = (\mathbf{Q}_{\mathrm{xsp}}^\top\mathbf{X}_s)^{-1}\mathbf{Q}_{\mathrm{xsp}}^\top\mathbf{y}_{s_P}$. 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:

$$\sum_{k\in s_P} r_k\delta_k(y_k-\hat{y}_k) = \mathbf{1}_{s_P}^\top\mathbf{R}_{s_P}(\mathbf{y}_{s_P}-\mathbf{X}_s\hat{\boldsymbol{\beta}}),$$

where $\mathbf{1}_{s_P}$ is a vector of ones with dimension the size of s_P. Then, $\mathbf{1}_{s_P}^\top\mathbf{R}_{s_P}(\mathbf{y}_s-\mathbf{X}_{s_P}\hat{\boldsymbol{\beta}}) = 0$ when $\mathbf{1}_{s_P}^\top\mathbf{R}_{s_P}$ spans the row space of $\mathbf{Q}_{\mathrm{xsp}}^\top$, namely if there exists $\boldsymbol{\mu}\in\mathbf{R}^n$ such that $\boldsymbol{\mu}^\top\mathbf{Q}_{\mathrm{xsp}}^\top = \mathbf{1}_{s_P}^\top\mathbf{R}_{s_P}$, which is equivalent to $\boldsymbol{\mu}^\top\boldsymbol{x}_k q_k - r_k = 0 \quad \text{for all} \quad k\in s_{NP}\cap s_P$. □

Proof of Proposition 3.2

We have

$$\begin{aligned} \hat{T}_{NP}^{(\mathrm{QR})} - \hat{T}_{NP}^{(\mathrm{Q}\pi)} &= \sum_{k\in s_P}(r_k-d_k)\delta_k(y_k-\boldsymbol{x}_k^\top\hat{\boldsymbol{\beta}})\\ &= \boldsymbol{\lambda}^\top\sum_{k\in s_P} q_k\delta_k\boldsymbol{x}_k(y_k-\boldsymbol{x}_k^\top\hat{\boldsymbol{\beta}}) = 0. \end{aligned}$$