sciendo

Journal of Official Statistics, Vol. 37, No. 2, 2021, pp. 481–503, http://dx.doi.org/10.2478/JOS-2021-0021

Measuring the Accuracy of Aggregates Computed from a Statistical Register

Giorgio Alleva[1], *Piero Demetrio Falorsi*[2], *Francesca Petrarca*[1], *and Paolo Righi*[2]

The Italian National Statistical Institute (Istat) is currently engaged in a modernization programme that foresees a significant revision of the methods traditionally used for the production of official statistics. The main concept behind this transformation is the use of the Integrated System Statistical Registers, created by a massive integration of administrative archives and survey data. In this article, we focus on how to measure the accuracy of register estimates of a population total from measurements calculated at the unit level. We propose the global mean squared error (GMSE) as a statistical quantity suitable for measuring accuracy in the context of the production of official statistics. It can be defined to explicitly consider the main sources of uncertainty that may affect registers. The article suggests a feasible calculation strategy for the GMSE that allows National Statistical Institutes to build algorithms that can promptly be applied for each user request, thus improving the relevance, transparency and confidence of official statistics. Through a simulation study, we verified the efficacy of the proposed strategy.

Key words: Integration; anticipated variance; linearization; mean square error; total survey error.

1. Background and Scope of the Article

The article focuses on how to measure the accuracy of population totals obtained from register data starting from uncertainty measures computed at the unit level. We consider the context where the users may autonomously define their totals, having direct access to the register microdata. In general, their statistics are unplanned and unpredictable in advance. The estimates of population totals are fundamental for knowing the dimension of quantitative variables or the level of diffusion of qualitative variables in a population. They represent the dominant part of the output that the different users produce from data and the common target parameters of the National Statistical Institute (NSIs). Standard linearization techniques (Särndal et al. 1992; Wolter 1986) allow extending the approach proposed herein for measuring the accuracy of non-linear statistics (such as correlations or regression parameters, and so on of a distribution) computable from the register microdata.

The background of the research activity described in this article is the modernization programme that the Italian National Statistical Institute (Istat 2016) launched some years

[1] Sapienza University of Rome, Via del Castro Laurenziano 9, 00161 Rome, Italy. Emails: giorgio.alleva@uniroma1.it and francesca.petrarca@uniroma1.it
[2] Italian National Institute of Statistics (Istat), Via Cesare Balbo, 16 – 00184 Rome, Italy. Emails: piero.falorsi@gmail.com and parighi@istat.it

ago. The main concept underlying this transformation is the use of the Integrated System of Statistical Registers (ISSR) as the basis for the production of all official statistics. This transformation represents a strategic challenge: it proposes abandoning the paradigm of statistical inference based on sample surveys that has been used for the past 75 years and moving on to a mixed data source paradigm for the future (Citro 2014; Alleva 2017). The ISSR is the result of a massive integration of administrative archives with survey data.

According to the statistical quality framework followed by Statistics Canada (2009), also reported by Wallgren and Wallgren (2014), the term survey includes the following components: (1) a census, which attempts to collect data from all members of a population; (2) a sample survey, in which data are collected from a (usually random) sample of population members; (3) a collection of data from administrative records, in which data are derived from 2 records originally kept for non-statistical purposes; and (4) a derived statistical activity, in which data are estimated, modelled, or otherwise derived from existing statistical data sources. Each of the previous components introduces different sources of uncertainty that should be considered both for predicting the target variables at the individual level and for the register aggregate. For instance, component (1) introduces the possibility of coverage errors, which we can address with specific statistical models. Component (2) includes sampling errors, and components (3) and (4) comprise the uncertainty derived by models adopted for building predictions at the individual level.

To construct ISSRs as the single informative infrastructure for the production of official statistics starting from a microdata level, different statistical techniques have been adopted. Many of these techniques result in computing predictions at the unit level. The register values remain the output of statistical processes subject to statistical uncertainty for both units and variables. The main strategic choice is whether to make the use of ISSR limited and to allow the dissemination of only planned outputs with a certified accuracy or to make the system more flexible, allowing different users, to produce their own statistics from the ISSR. We propose here to opt for the second option, which makes the Institute more relevant for its users, but exposes the NSI to the threat of inappropriate use of the register data by unaware users. Indeed, users who have access to the microdata could conceivably produce their estimates fully unaware of any problems associated with the quality of their register statistics. The European Statistical System (ESS) is aware of the importance of producing new measures of accuracy for multi-source statistics such as those produced by statistical registers (Eurostat 2019). In this article, we suggest a computational strategy for facilitating flexible and correct use of register data by enabling users to quickly estimate global mean squared error (GMSE) on their own. In Section 2, we give the notation and introduce the measure of accuracy we propose to adopt for the register aggregates. Then, in Section 3, to facilitate comprehension, we introduce a simplified statistical framework in which the register is not affected by coverage errors. Section 4 describes the calculus of the GMSE for the simplified statistical framework. Section 5 illustrates the main computational challenges. Section 6 deals with coverage errors. Sections 7 and 8 show the first results of a simulation study and provide preliminary conclusions with some initial reflections on how to develop a feasible validation approach. The derivations of the main results are available in Appendix (Section 9) and in Appendices 2, 3 and 4 published as online supplementary materials for this article.

2. Notation and Proposed Measure of Accuracy

Let U be the unknown target population of interest, including $N_{(U)}$ statistical units. Let U_d be a statistical *domain of interest*, which is a subset of U with $N(_{U_d})$ units. The target parameter of interest, Y_{U_d}, is the total of the variable y within the domain U_d:

$$Y_{U_d} = \sum_{k \in U_d} y_k, \tag{1}$$

where y_k is the true value of the variable y for unit k.

Let R be a statistical register, including $N_{(R)}$ statistical units: ideally, each statistical unit in U should be represented by a *corresponding* unit in R.

Furthermore, let R_d be a subset of R of size $N(_{R_d})$, which represents the target domain U_d. Let $\hat{y}_k$ be the value recorded in the register that predicts the value y_k. These *predicted* values can be computed according to different statistical models or algorithms. For estimating Y_{U_d}, the users can simply sum the predicted $\hat{y}_k$ values over R_d:

$$\hat{Y}_{R_d} = \sum_{k \in R_d} \hat{y}_k. \tag{2}$$

$\hat{Y}_{R_d}$ is a register-based statistic as in Wallgren and Wallgren (2014) and is the result of an estimation process. We may define the accuracy of $\hat{Y}_{R_d}$ based on the difference between this statistic and the actual value, Y_{R_d}. The accuracy depends on various factors, such as the coverage error of the register and the measurement errors of predictions.

The data structure of the population U and the statistical register, R, are illustrated in Table 1, where the right part represents the population and the left part represents the statistical register. In our table $\mathbf{x}_k$ denotes a vector of I, auxiliary variables available in R for each unit k. Note that the true y_k values are rarely available in the register. The last columns on both parts of the table (the right and the left) are dichotomous membership variables indicating whether the unit is included in domain d. The true values of these

Table 1. Data structure in population U and in statistical register R.

Population U			Statistical register R				
Identifier of the population unit true unknown	True y Value	True membership variable (0,1) of the domain d	Code in R	Predicted value	Auxiliary variables	Register membership variable of the domain d	
			1	$\hat{y}_1$	$\mathbf{x}_1$	1	Over-coverage
1	y_1	1				0	
k	y_k	0	k	$\hat{y}_k$	$\mathbf{x}_k$	1	
						1	
			$N_{(R)}$	$\hat{y}_{N_{(R)}}$	$\mathbf{x}_{N_{(R)}}$		
$N_{(U)}$	$y_{N_{(U)}}$	1					Under-coverage

variables in the population may coincide with what is reported in the register. The areas with a grey background in the table highlight the over/under-covered units in R.

Depending on the specific objective, the target variable, denoted with the symbol y, and the auxiliary variables, indicated with the symbol x, can be represented by information provided by each of the survey components illustrated in Section 1. To better clarify the different roles, the auxiliary variables are known, and the target variables are those predicted at the unit level. The auxiliary variables are, for example, the sex and age of people in the population register or tax paid by firms in the business register.

In a total survey error approach (Biemer 2010), *the mean squared error* (MSE) represents the simplest way to measure the accuracy of registered-based statistics. It is expressed as the expected squared difference between the estimator and the true unknown population value:

$$MSE(\hat{Y}_{R_d}) = E(\hat{Y}_{R_d} - Y_{U_d})^2 = V(\hat{Y}_{R_d}) + [Bias(\hat{Y}_{R_d})]^2, \tag{3}$$

where $E(\cdot)$ denotes the operator of expectation and $V(\cdot)$ indicates the operator of the variance. Each specific approach to inference focuses on different sources of variability and bias in the definition of the MSE; these are related to what is treated as fixed or random in the specific inferential approach. For instance, *design-based* (Cochran 1977) or model-assisted approaches (Särndal et al. 1992) treat the population values y_k as unknown constants, and the sample selected, with the sample design P, is the only source of randomness; therefore, they develop their expectations considering only the variability of the sampling design. The *model- based* approach (Vaillant 2009; Chambers and Clark 2015) considers the sample as *fixed* and the y_k values as random variables generated according to the model, M; thus, they develop the MSE considering only the variability embedded in the model. The expectation developed from the model generating the data will be denoted as E_M, and the expectation calculated from the variability of the sampling design will be indicated as E_P. The same notation will be adopted for the operator of variance, thus defining the operators V_M and V_P.

Here, we propose to develop the MSE, taking into consideration all the random components involved in the inferential process for building the predictions. We can simply do this by defining the operator of expectation in Equation (3) as a concatenation of elementary expectation operators, each of which considers a specific random component. Following the proposal of Wolter (1986), who introduces the concept of *global variance*, the measure we propose could be denominated as a *global MSE* (GMSE). The GMSE can be expressed as

$$GMSE(\hat{Y}_{R_d}) = E_P E_M(\hat{Y}_{R_d} - Y_{U_d})^2. \tag{4}$$

When planning the sampling design, GMSE is also known as anticipated variance (Isaki and Fuller 1982; Särndal et al. 1992; Nedyalkova and Tillé 2008; Nirel and Glickman 2009; Falorsi and Righi 2015). Here, the measure is not limited to the sampling context and incorporates additional sources of variability and bias. For instance, the nonresponse by defining GMSE as:

$$GMSE(\hat{Y}_{R_d}) = E_P E_M E_{NR}(\hat{Y}_{R_d} - Y_{U_d})^2, \tag{5}$$

in which E_{NR} indicates the expectation under the models adopted for imputing the nonresponse in survey data. Continuing the illustration, let us consider the case in which we collect the y variable from a census affected by nonresponse. In this case, we can define the GMSE *as*

$$\text{GMSE}(\hat{Y}_{R_d}) = E_{NR}(\hat{Y}_{R_d} - Y_{U_d})^2.$$

The GMSE could be accepted as a measure of precision by the main professional families of methodologists within the NSIs: at least, those who base their inference only on statistical models and those who use the statistical models as a support for inference that continues to be based essentially on sampling design. The global measure has a number of advantageous qualities, including the following: generality, stability over time and robustness in the case of model failures. GMSE is simple to use and communicate to users. It is based on the first and second moments of the random distributions of the specific source of uncertainty. Its calculus does not imply full knowledge of the underlying distributions.

We observe that well-known approaches for estimating GMSE are based on replication methods (Scholtus 2019). However, these techniques are highly time-consuming: the replicates have to produce the whole process generating $\hat{Y}_{R_d}$. Therefore, in the context of the massive and continuous production of official statistics by NSIs, replication methods do not seem to be a feasible solution.

To facilitate informed and correct use of a registry, once the user defines the target total and the specific domain of interest (R_d), it would be useful to build a dynamic data vector, for example, $\hat{\sigma}^2_{dy,k}$ for $k = 1, \ldots, N_{(R_d)}$, so that the GMSE estimate is:

$$\widehat{\text{GMSE}}\left(\hat{Y}_{R_d}\right) = \sum_{k \in R_d} \hat{\sigma}^2_{dy,k}. \tag{6}$$

The quantities $\hat{\sigma}^2_{dy,k}$ are unit and domain specific. However, we will see in Section 5 that the amount of information to be stored for their calculation is limited for each unpredictable user request. The dependency on the domain is limited to a few useful variables in the register. Therefore, in our proposal, we do not suggest storing the $\hat{\sigma}^2_{dy,k}$ values, but we do recommend stockpiling some intermediate values, not domain specific, from which the $\hat{\sigma}^2_{dy,k}$ may be easily calculated, thus making the proposed solution applicable in the real contexts of NSI informative infrastructures (this will become clearer in Section 5 below). We also note that the definition of the computational formulae for determining quantities $\hat{\sigma}^2_{dy,k}$ represents a relevant result since it enables the NSI to build algorithms that can be applied promptly for each user request even those not planned in advance. To produce the proposed assurance of accuracy at the micro-level (given in Equation 6), we propose a computational strategy based on the following two primary approximations:

1. the linearization of estimator$\hat{Y}_{R_d}$ with respect to each specific source of variability considered in GMSE. The validity of this assumption will be proven as true in the typical asymptotic contexts that are used in these cases (see Appendix A3 in online supplementary materials);

2. the adoption of a form of calculus of the sampling variances based only on first inclusion probabilities.

The computational strategy is made simpler by the well-known result of Kendall and Stuart (1976, 196), according to which we can express the GMSE as a sum of conditional values that is more manageable for the calculus. We repeatedly use this result in Section 4 and in Appendix (Section 9). Finally, we note that although some believe that the use of approximations could weaken the methodological proposal, we find ourselves today in a context where NSIs do not calculate accuracy at all. Furthermore, approximate solutions are usual, for example, in the case of sample variances where closed forms are not available.

3. A Simplified Statistical Framework

Here, we suppose that *coverage error* is negligible, which implies $R_d \equiv U_d$. We also assume that the $\mathbf{x}_k = (x_{k1}, \cdots, x_{ki}, \cdots, x_{kI})'$ values are not subject to the measurement error. The main sources of variability considered here are the model, M, generating the data y and the sampling design, P. The model and the sampling design are always developed under the non-informative assumption of the current survey sampling activity. The model formulation and fitting are independent of the sampling design and vice versa.

3.1. Model Uncertainty

Regarding the first source, we can suppose that each y_k value is to be expressed as the sum of two components:

$$y_k = \tilde{y}_k + e_k, \tag{7}$$

where $\tilde{y}_k = E_M(y_k)$ is the theoretical value according to which the value of y is generated from a given statistical model, M, for unit k, and e_k denotes the random error with model expectations given by

$$E_M(e_k) = 0,\ E_M(e_k^2) = V_M(e_k) = \sigma_{y,k}^2,\ E_M(e_k e_\ell) = \sigma_{y,k,\ell}, \tag{8}$$

where e_ℓ indicates the random error of unit ℓ.

The structure of a model expectation is quite general and may be easily applied to different statistical models. For instance, consider a well-known model with domain random effects, adopted as a small area estimation technique, $y_k = \tilde{y}_k + \varepsilon_k + z_d$ for $k \in U_d$, in which ε_k is random noise and z_d is a random domain effect; we may then reformulate the model expectation structure of this model, according to Equation 8, by defining $e_k = \varepsilon_k + z_d$ for $k \in U_d$.

3.2. Sampling Uncertainty

A generalized framework for defining sampling designs, illustrated in detail in Falorsi and Righi (2015), assumes a sample S of fixed size n selected from R, in accordance with sample design P with inclusion probabilities $\pi = (\pi_1, \ldots, \pi_k, \ldots, \pi_{N(R)})'$. Many practical sampling designs define domains that are planned sub-populations so that the

sample sizes have been fixed before selecting the sample. Denote by $R_{[h]} (h = 1, \ldots, H)$ the planned domain of size $N_{(R_{[h]})} = \sum_{k \in R} d_{k(h)}$ where $d_{k(h)} = 1$ if $k \in R_h$ and $d_{k(h)} = 0$ otherwise. Fixed size sampling designs are those satisfying $\sum_{k \in S} \boldsymbol{d}_k = \boldsymbol{n}$, where $\boldsymbol{d}_k = (d_{k(1)}, \ldots, d_{k(h)}, \ldots d_{k(H)})'$ and $\boldsymbol{n} = (n_1, \ldots, n_h, \ldots, n_H)'$ is the vector of integer numbers defining the sample sizes fixed at the design stage, with $\sum_{k \in S} d_{k(h)} \pi_k = \boldsymbol{n}$. In our setting, planned domains can overlap; therefore, unit k may have more than one value $d_{k(h)} = 1$ (for $h = 1, \ldots, H$). Several customary fixed size sampling designs invite particular consideration. A well-known example is the stratified simple random sampling without replacement (SSRSWOR) design, where strata are the planned domains and each $\boldsymbol{d}_k$ vector has $H - 1$ elements equal to zero and one element equal to 1, which implies that each unit j belongs to one and only one planned domain. The total Y estimated with the Horvitz-Thompson estimator is $\hat{Y}_{HT} = \sum_{k \in S} y_k (1/\pi_k)$. We suppose that the $N_{(R)} \times H$ matrix $\mathbf{D} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_j, \ldots, \boldsymbol{d}_{N_{(R)}})'$ is non-singular. According to this general sampling design framework, Deville and Tillé (2005) proposed an approximated expression of the variance for $\hat{Y}_{HT}$ based on the Poisson sampling theory given by

$$V_P\left(\hat{Y}_{HT}\right) \cong \left[\frac{N_{(R)}}{(N_{(R)} - H)}\right] \sum_{k \in R} \left(\frac{1}{\pi_k} - 1\right) \eta_k^2 \tag{9}$$

$$\text{where } \eta_k = y_k - \pi_k \boldsymbol{d}_k' \left[\sum_{J \in R} \boldsymbol{d}_j \boldsymbol{d}_j' \pi_j (1 - \pi_j)\right]^{-1} \sum_{j \in R} \pi_j \left(\frac{1}{\pi_j} - 1\right) \boldsymbol{d}_j y_j. \tag{10}$$

Equation 9 resembles the variance expression of the Horvitz-Thompson estimator under a Poisson sampling design, but it uses the residuals, η_k, instead of the original value, y_k. In practice, when $H = 1$, this is the variance approximation of the conditional Poisson sampling (Deville and Tillé 2005). The above approximation works well when the number of domains H remains small compared to the overall population size $N_{(R)}$. A conservative approximation of Equation (10) may be obtained by substituting η_k with the y_k values.

3.3. Predictions

Let us suppose that $\tilde{y}_k$ can be expressed as a function $f(\cdot)$

$$\tilde{y}_k = f(\mathbf{x}_k; \boldsymbol{\vartheta}), \tag{11}$$

in which $\boldsymbol{\vartheta} = (\vartheta_1, \ldots \vartheta_i, \ldots, \vartheta_I)'$ is a vector of I unknown parameters.
Let

$$\hat{y}_k = f(\mathbf{x}_k; \hat{\boldsymbol{t}}), \tag{12}$$

be the register predictions, where $\hat{\boldsymbol{t}} = (\hat{t}_1, \ldots, \hat{t}_i, \ldots, \hat{t}_I)'$ represents the estimate of $\boldsymbol{\vartheta}$ based on observation of the values y_k on the sample S. The resulting estimator $\hat{Y}_{R_d}$, given by the Equation 2 belongs to the class of *projection estimators* that can be developed either under the *model-assisted approach* (Kim and Rao 2012) or the *model-based approach* (Chambers and Clark 2015; FAO, part.5, 2014).

Remark. The predictions $\hat{y}_k$ can be built in different ways, thus defining different forms of Equation 2, as follows:

$$\hat{y}_k = \begin{cases} y_k \text{ if } k \in S \cap R \\ f(\mathbf{x}_k;\hat{\boldsymbol{t}}) \text{ if } k \in \overline{S} \cap R \end{cases} \quad (a), \qquad \hat{y}_k = \begin{cases} y_k \text{ if } k \in S \cap R \\ f(\mathbf{x}_k;\hat{\boldsymbol{t}}) + \hat{e}_k \text{ if } k \in \overline{S} \cap R \text{`} \end{cases} \quad (b)$$

$\hat{e}_k$ is a residual that can be selected either from the residuals estimated in sample S or from the estimated distribution of the y *values* (Chen and Haziza 2017). With predictions built as in expression (a), the use of the resulting estimator is more common when using the standard *prediction* approach for inference. Form (b) is appropriate in cases where register values are used for calculating indicators, such as quantiles or correlations in which the variability at the unit level is relevant. Furthermore, when y is categorical, each specific value of the $\hat{y}_k$ values in R can be set equal to one of the standard modalities of the y variable.

4. The Calculus or the GMSE in the Simplified Statistical Framework

4.1. Decomposition of GMSE

In the observational setting described in Section 3, there are two random vectors of $N_{(R)}$ units: $\boldsymbol{y} = (y_1, \ldots, y_k, \ldots, y_{N_{(R)}})'$ and $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_k, \ldots, \lambda_{N_{(R)}})'$, which is the vector of sample membership indicators with $\lambda_k = 1$ if $k \in S$ and $\lambda_k = 0$ otherwise. We suppose that the estimate $\hat{\boldsymbol{t}}$ of $\boldsymbol{\vartheta}$ is *model unbiased*, which means $E_M(\hat{\boldsymbol{t}}|\boldsymbol{\lambda}) = \boldsymbol{\vartheta}$, where $E_M(\hat{\boldsymbol{t}}|\boldsymbol{\lambda})$ denotes the *model expectation* conditioned on the sample realized value of the vector $\boldsymbol{\lambda}$. Thus, GMSE may be expressed as (see Appendix, Section 9)

$$GMSE\left(\hat{Y}_{R_d}\right) =$$

$$\begin{cases} E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] - V_M\left(Y_{U_d}\right) & \text{if } E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right) \neq Y_{U_d} \quad (13a) \\ E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] - V_M\left(Y_{U_d}\right) - 2Cov_M\left[E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right), Y_{U_d}\right] & \text{if } E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right) \neq Y_{U_d} \quad (13b) \end{cases},$$

where $E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right)$ denotes the sampling expectation conditioned on the realized value of the vector $\boldsymbol{y}$. As seen from Equation (13a), if the estimator $\hat{Y}_{R_d}$ is design unbiased for the aggregateY_{U_d}, then GMSE will neutralize variability attributing it to the pure model variability of population parameter Y_{U_d}. The conditions for fulfilling design unbiasedness are given in Section 3 of Kim and Rao (2012).

4.2. Calculus of the Dominant Component of GMSE

The dominant component of $GMSE(\hat{Y}_{R_d}))$ is $E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right]$. Indeed, the term $V_M\left(Y_{U_d}\right)$ contributes negatively to the expression Equation (13a), whereas it enters a positive component of the difference $\left(V_M\left(Y_{U_d}\right) - 2Cov_M\left[E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right), Y_{U_d}\right]\right)$ in the Equation (13b). We also note that the term $-2Cov_M\left[E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right), Y_{U_d}\right]$ is of the same order of magnitude as the component $V_M\left(Y_{U_d}\right)$ and represents the model covariance between the two population totals Y_{U_d} and $E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right)$. The overall difference $\left(V_M\left(Y_{U_d}\right) - 2Cov_M\left[E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right), Y_{U_d}\right]\right)$ becomes negative when $E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right) > \frac{1}{2}Y_{U_d}$ with $Y_{U_d} > 0$. In most of the empirical situations that are encountered, the difference tends to be negligible or negative.

In the following discussion, we will present some asymptotic behaviors of the variables of interest. We refer to the results shown in the literature on the subject (see, for example, Isaki and Fuller 1982; Wolter 1985; Särndal et al. 1992; Deville 1999; Breidt and Opsomer 2017). Here, we limit ourselves to recalling the general framework of our assumptions (as according to Deville 1999): assuming that (1) the size N of the population and the size n of the sample tend to infinity; (2) $N^{-1}Y$ has a finite limit, where Y is the total of the variable y, N is the size of a sequence of populations of increasing size; (3) $N^{-1}(\hat{Y} - Y)$ with $\hat{Y}$ the estimator of the total converges in probability to zero; and (4) $n^{-1/2}N^{-1}(\hat{Y} - Y)$ tends to a multi-dimensional normal distribution, observing the central limit theorem. A consequence of these assumptions is that the terms that are $O_p(n^{-1/2})$ in the decomposition of $E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right]$ can be considered small, and the product of two such small quantities can be deemed negligible.

We focus now on the predominant component $E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right]$.First, we note that the estimate $\hat{Y}_{R_d}$ may be seen as an overall function of three random components, $\boldsymbol{y}$, $\boldsymbol{\lambda}$ and $\hat{\boldsymbol{t}}$, in which the quantities $\mathbf{x}_k$ and $\gamma_{d,k}$ (for $k \in R$) are considered known and not random, with $\gamma_{d,k} = 1$ if $k \in R_d$ and $\gamma_{d,k} = 0$ otherwise. We can express the register-based statistic

$$\hat{Y}_{R_d} = \hat{Y}_{R_d}\left(\hat{\boldsymbol{t}}; \boldsymbol{y}; \boldsymbol{\lambda}|\boldsymbol{X}, \boldsymbol{\gamma}_d, \vartheta\right)$$

as a function of the three random components, given as the fixed auxiliary variables, $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_{N_{(R)}})'$, domain membership variables, $\boldsymbol{\gamma}_d = (\gamma_{d,1}, \ldots, \gamma_{d,k}, \ldots, \gamma_{d,N_{(R)}})'$, and super population parameter, $\boldsymbol{\vartheta}$.

We arrive at the final computable expression of $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$ through a three-step linearization (see Appendix A4 in online supplementary materials). We carry out the first linearization on the random quantities $\hat{\boldsymbol{t}}$ and the second and third linearization steps on the random vectors $\boldsymbol{y}$, and $\boldsymbol{\lambda}$. We calculate each that are first derived, taking into consideration the specific source of variability for the given conditional set-up.

1. *First linearization step*. The estimator $\hat{Y}_{R_d}$ is linearized with respect to the vector $\hat{\boldsymbol{t}}$ where the derivatives are computed at the model expected value $\boldsymbol{\vartheta}$, thus obtaining:

$$\hat{Y}_{R_d} = E_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}, \boldsymbol{\vartheta}\right) + \sum_{k \in R} \sum\nolimits_{i=1}^{I} \gamma_{d,k} f_{ki}(\hat{t}_i - \vartheta_i) + r_1, \tag{14}$$

where

$$f_{ki} = \left.\frac{\partial f(\mathbf{x}_k; \hat{\boldsymbol{t}})}{\partial \hat{t}_i}\right|_{\hat{t}_i = \vartheta_i} : k = 1, \ldots, N_{(R_d)};\ i = 1, \ldots, I, \text{ being } r_1 = 0_p\left(1/\sqrt{n}\right), \tag{15}$$

is a remainder of minor order. Discarding the remainder, the model variance, $V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})$, which represents the core part of $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$, is given by:

$$\begin{aligned} V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right) &\cong V_M\left[\sum_{k \in R} \sum\nolimits_{i=1}^{I} \gamma_{d,k} f_{ki}\left(\hat{t}_i - \vartheta_i\right)|\boldsymbol{\lambda}\right] = V_M\left[\boldsymbol{\gamma}_d' \mathbf{F}\left(\hat{\boldsymbol{t}} - \boldsymbol{\vartheta}\right)|\boldsymbol{\lambda}\right] \\ &= \boldsymbol{\gamma}_d' \mathbf{F}\left[V_M\left(\hat{\boldsymbol{t}}|\boldsymbol{\lambda}\right)\right]\mathbf{F}'\boldsymbol{\gamma}_d, \end{aligned} \tag{16}$$

where $\mathbf{F} = [f_{ki}]$ is a $N_{(R)} \times I$ matrix.

2. *Second linearization step.* The term

$$\sum_{k\in R}\sum\nolimits_{i=1}^{I}\gamma_{d,k}f_{ki}\left(\hat{t}_i-\vartheta_i\right)=\boldsymbol{\gamma}_d^{'}\mathbf{F}\left(\hat{\boldsymbol{t}}-\boldsymbol{\vartheta}\right)$$

in Equation (16) is linearized with respect to the y variables, where the derivatives are computed at $\hat{\boldsymbol{t}}=\boldsymbol{\vartheta}$ and $\boldsymbol{y}=\tilde{\boldsymbol{y}}=(\tilde{y}_1,\ \ldots,\tilde{y}_k,\ \ldots,\tilde{y}_{N(R)})^{'}$ keeping the $\boldsymbol{\lambda}$ vector fixed. The mathematical explanations are detailed in Appendix A2 in the online supplementary materials. Here, in this section, we limit ourselves to providing and describing the main and essential results. Let

$$\sum_{j\in R}\boldsymbol{g}_j(\hat{\boldsymbol{t}};\boldsymbol{y};\boldsymbol{\lambda})=\mathbf{0}_I, \tag{17}$$

be the system of GEE, generalized estimating Equations (Ziegler 2015) for estimating the vector $\hat{\boldsymbol{t}}$ in which

$$\boldsymbol{g}_j(\hat{\boldsymbol{t}};\boldsymbol{y};\boldsymbol{\lambda})=[g_{j1}(\hat{\boldsymbol{t}};\boldsymbol{y};\boldsymbol{\lambda}),\ \ldots,g_{ji}(\hat{\mathbf{t}};\boldsymbol{y};\boldsymbol{\lambda}),\ \ldots,g_{jI}(\hat{\boldsymbol{t}};\boldsymbol{y};\boldsymbol{\lambda})]', \tag{17b}$$

is the I vector of the score functions $g(\hat{\boldsymbol{t}};\boldsymbol{y};\boldsymbol{\lambda})$ for the parameter $\hat{t}_i$ $(i=1,\ldots,I)$ of unit j, where $\mathbf{0}_I$ is a vector of I zeroes. Adopting the linear approximation first proposed by Binder and Patak (1994) and then, among others, by Chambers and Clark (2015, 123–125), we have:

$$(\hat{\boldsymbol{t}}-\boldsymbol{\vartheta})\cong\boldsymbol{A}_{\vartheta}^{-1}\sum_{j\in R}\boldsymbol{g}_j(\boldsymbol{\vartheta};\boldsymbol{y};\boldsymbol{\lambda}), \tag{18}$$

being

$$\boldsymbol{A}_{\vartheta}=\left[a_{i\ell|\vartheta}=\frac{\partial\sum_{j\in R}g_{ji}\left(\hat{\boldsymbol{t}};\boldsymbol{y};\boldsymbol{\lambda}\right)}{\partial\hat{t}_\ell}\bigg|_{\hat{\boldsymbol{t}}=\boldsymbol{\vartheta}}\right],$$

a $(I\times I)$ matrix (in which $i,\ell=1,\ldots,I$), and where $\boldsymbol{g}_j(\boldsymbol{\vartheta};\ \boldsymbol{y};\ \boldsymbol{\lambda})=\{g_{ji}(\boldsymbol{\vartheta};\ \boldsymbol{y};\ \boldsymbol{\lambda});\ i=1,\ \ldots,\ I\}$ is defined in Equation (17*b*) by the substitution of $\hat{\boldsymbol{t}}$ with $\boldsymbol{\vartheta}$. Thus, according to the Equation (A3) of Appendix A2 in online supplementary materials, we have the following approximation, which holds for $n\gg I$:

$$\begin{aligned}V_M\left(\hat{Y}_{R_d}\middle|\boldsymbol{\lambda}\right)&\cong V_M\left[\sum_{k\in R}\sum_{j\in R}\sum\nolimits_{i=1}^{I}\gamma_{d,k}f_{ki}u_{ji|\lambda}y_j\middle|\boldsymbol{\lambda}\right]\\&=V_M\left(\boldsymbol{\gamma}_d^{'}\mathbf{F}\sum_{j\in R}\mathbf{u}_{j|\lambda}y_j\middle|\boldsymbol{\lambda}\right)\\&=\boldsymbol{\gamma}_d^{'}\mathbf{F}\sum_{j\in R}\left[\mathbf{u}_{j|\lambda}\mathbf{u}_{j|\lambda}^{'}\sigma_{y,j}^2+\sum_{\ell\neq j}\mathbf{u}_{j|\lambda}\mathbf{u}_{\ell|\lambda}^{'}\sigma_{y,j\ell}\right]\mathbf{F}'\boldsymbol{\gamma}_d,\end{aligned} \tag{19}$$

where $\mathbf{u}_{j|\lambda}=\frac{-\partial[\boldsymbol{A}_{\vartheta}^{-1}\boldsymbol{g}_j(\boldsymbol{\vartheta};\boldsymbol{y};\boldsymbol{\lambda})]}{\partial y_j}\Big|_{\boldsymbol{y}=\tilde{\boldsymbol{y}}}(u_{j1|\lambda},\ \ldots,u_{j1|\lambda},\ldots u_{j1|\lambda})'$.

For the general linear model in which $\tilde{y}_k = \mathbf{x}_k^{'}\boldsymbol{\vartheta}$, the matrix $\boldsymbol{A}_{\boldsymbol{\vartheta}}$ is an $I \times I$ identity matrix.

3. *Third linearization step*. According to the approach proposed by Graf (2015) and by Vallée and Tillé (2019), the terms $\mathbf{u}_{j|\lambda}$, included in Equation (19), are linearized with respect to the sampling indicators $\boldsymbol{\lambda}$ around the sample design expected value $\boldsymbol{\pi}$. We have,

$$\mathbf{u}_{j|\lambda} \cong \mathbf{u}_{j,\pi} + \partial \mathbf{u}_j(\lambda_j - \pi_j), \tag{20}$$

where $\mathbf{u}_{j,\pi}$ is obtained as $\mathbf{u}_{j|\lambda}$ by substituting the values of λ_j with the corresponding expected values π_j and

$$\partial \mathbf{u}_j = \frac{\partial \mathbf{u}_{j|\lambda}}{\partial \lambda_j}\bigg|_{\lambda=\pi} = \left(\partial u_{j1}, \ldots, \partial u_{ji}, \ldots, \partial u_{jI}\right)^{'}.$$

Then, we have

$$E_P\left(\mathbf{u}_{j|\lambda}\mathbf{u}'_{j|\lambda}\right) = \mathbf{u}_{j,\pi}\mathbf{u}'_{j,\pi} + \partial \mathbf{u}_j \partial \mathbf{u}'_j \pi_j(1 - \pi_j). \tag{21}$$

$$E_P\left(\mathbf{u}_{j|\lambda}\mathbf{u}'_{\ell|\lambda}\right) = \mathbf{u}_{j,\pi}\mathbf{u}'_{\ell,\pi} + \partial \mathbf{u}_j \partial \mathbf{u}'_\ell (\pi_{j\ell} - \pi_j \pi_\ell), \tag{22}$$

where $\pi_{j\ell}$ is the joint inclusion probability of units j and ℓ.
Then, considering the sampling expected values Equations (21) and (22) into Equation (19), we have

$$\begin{aligned} E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] &\cong \boldsymbol{\gamma}'_d \mathbf{F}\left\{E_P\left[V_M\left(\hat{\mathbf{t}}|\boldsymbol{\lambda}\right)\right]\right\}\mathbf{F}'\boldsymbol{\gamma}_d \\ &= \boldsymbol{\gamma}'_d \mathbf{F}\left\{\sum_{j\in R}\left[\left(\mathbf{u}_{j,\pi}\mathbf{u}'_{j,\pi} + \partial \mathbf{u}_j \partial \mathbf{u}'_j \pi_j(1-\pi_j)\right)\sigma^2_{y,j} + \right.\right. \\ &\left.\left. \sum_{\ell \neq j}\left(\boldsymbol{u}_{j,\pi}\mathbf{u}'_{\ell,\pi} + \partial \mathbf{u}_j \partial \mathbf{u}'_\ell (\pi_{j\ell} - \pi_j \pi_\ell)\right)\sigma_{y,j\ell}\right]\right\}\mathbf{F}'\boldsymbol{\gamma}_d. \end{aligned} \tag{23}$$

The above expression cannot be computed for many usual sampling designs since the joint inclusion probabilities, $\pi_{j\ell}$, are unknown. Starting from result of Equation (9), we propose an upward approximation of Equation (23) based on the first-order inclusion probabilities, which we recommend for the calculus of $E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right]$:

$$\begin{aligned} &E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] \\ &\leq \boldsymbol{\gamma}'_d \mathbf{F}\sum_{j\in R}\left[\left(\mathbf{u}_{j,\pi}\mathbf{u}'_{j,\pi} + \partial \mathbf{u}_j \partial \mathbf{u}'_j \pi_j(1-\pi_j)\right)\sigma^2_{y,j} + \left(\sum_{\ell \neq j}\mathbf{u}_{j,\pi}\mathbf{u}'_{\ell,\pi}\sigma_{y,j\ell}\right)\right]\mathbf{F}'\boldsymbol{\gamma}_d. \end{aligned} \tag{24}$$

In Appendix A3 in the online supplementary materials, we give a lower bound of $E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right]$. Equation (24) is defined using unit-level elements: $\mathbf{u}_{j,\pi}$, $\partial \mathbf{u}_j$, π_j, $\sigma^2_{y,j}\sigma_{y,j\ell}$ which are not domain specific.

4.3. Calculus of the Other Components of GMSE

According to the model setting (3.2), the component $V_M(Y_{U_d})$ is given by

$$V_M\left(Y_{U_d}\right) = \sum_{j\in R_d}\left(\sigma^2_{y,j} + \sum_{\ell\neq j}\sigma_{y,j\ell}\right). \tag{25}$$

For the *calculus* of $Cov_M[E_P(\hat{Y}_{R_d}|\mathbf{y}),Y_{U_d}]$, we adopted a linearization consisting of two steps, where in the first step, the estimator $\hat{Y}_{R_d}$ is linearized with respect to the vector $\hat{\mathbf{t}}$ and the first step is computed at the sampling design expected value $\boldsymbol{t} = E_P(\hat{\boldsymbol{t}}|\boldsymbol{\lambda})$. In the second step, adopting the same approach as Binder and Patak (1994), the estimating expressions of $\boldsymbol{t}$ are linearized around $\tilde{\mathbf{y}}$ and $\boldsymbol{\vartheta}$, thus obtaining:

$$Cov_M\left[E_P\left(\hat{Y}_{R_d}|\mathbf{y}\right), Y_{U_d}\right] \cong \sum_{k\in R}\sum_{j\in R}\sum_{i=1}^{I}\gamma_{d,k}f_{ki}\,u_{ji,\pi}\left(\sigma^2_{y,j} + \sum_{(\ell\neq j)\cap(\ell\in R_d)}\gamma_{d,\ell}\,\sigma_{y,j\ell}\right). \tag{26}$$

4.4. Plug-In Estimate of the GMSE

The plug-in estimate of GMSE may be computed by inserting the estimates of $\hat{\mathbf{t}}$, $\hat{y}_k$ $(k = 1,\ldots,N_{(R)})$, $\hat{\sigma}^2_{y,j}$ and $\hat{\sigma}_{y,j\ell}$ $(j,\ell = 1,\ldots,N_{(R)})$ in place of the unknown parameters $\boldsymbol{\vartheta}$, $\tilde{y}_k$ $(k = 1,\ldots,N_{(R)})$, $\sigma^2_{y,j}$ and $\sigma_{y,j\ell}$ $(j,\ell = 1,\ldots,N_{(R)})$ in the expressions of the different components of the GMSE. According to Ziegler (2015, point 5, 121), these plug-in estimates are strongly consistent estimators of the different components of the variance.

4.5. Example with the Logistic Model

Consider a logistic model where $y_k = 1$ with probability $\tilde{y}_k = f(\mathbf{x}_k;\boldsymbol{\vartheta})$ and $y_k = 0$ with probability $1-\tilde{y}_k$, where $f(\mathbf{x}_k;\boldsymbol{\vartheta}) = e^{\mathbf{x}'_k\boldsymbol{\vartheta}}/\left(1+e^{\mathbf{x}'_k\boldsymbol{\vartheta}}\right)$. Under the *model-based approach*, the estimating equations for the calculus of the GMSE using the first-order linear approximations, derived by the log-likelihood, are

$$\sum_{j\in R}\boldsymbol{g}_j\left(\hat{\boldsymbol{t}};\mathbf{y};\boldsymbol{\lambda}\right) = \sum_{j\in R}\mathbf{x}_j\left[y_j - \frac{e^{\mathbf{x}'_j\hat{\boldsymbol{t}}}}{1+e^{\mathbf{x}'_j\hat{\boldsymbol{t}}}}\right]\lambda_j.$$

The matrix $\boldsymbol{A}_\vartheta$ is given by:

$$\boldsymbol{A}_\vartheta = -\sum_{j\in R}\left[\frac{\mathbf{x}_j\mathbf{x}'_j e^{\mathbf{x}'_j\boldsymbol{\vartheta}}}{\left(1+e^{\mathbf{x}'_j\boldsymbol{\vartheta}}\right)^2}\right]\lambda_j$$

and the vectors $\mathbf{u}_{j|\lambda}$, $\mathbf{u}_{j,\pi}$ and $\partial\mathbf{u}_j$ are expressed as $\mathbf{u}_{j|\lambda} = -\boldsymbol{A}_\vartheta^{-1}\mathbf{x}_j\lambda_j, \mathbf{u}_{j,\pi} = -\boldsymbol{A}_\vartheta^{-1}\mathbf{x}_j\pi_j$, $\partial\mathbf{u}_j = -\boldsymbol{A}_\vartheta^{-1}\mathbf{x}_j$.

5. Tips on Computational Aspects

GMSE may be expressed as the sum of elementary unit variances, $\sigma^2_{dy,k}$, over the register domain units:

$$\mathrm{GMSE}\left(\hat{Y}_{R_d}\right) \cong \sum_{k \in R_d} \sigma^2_{dy,k}. \tag{27}$$

Considering jointly the Equations (13*a*) or (13*b*), (24), (25) and (26), we have that the quantities $\sigma^2_{dy,k}$ are given by

$$\sigma^2_{dy,k} = \begin{cases} \sigma^2_{Ady,k} - \sigma^2_{Bdy,k} & \text{if } E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right) = Y_{U_d} \quad (28a) \\ \sigma^2_{Ady,k} + \sigma^2_{Bdy,k} - 2\sigma^2_{Cdy,k} & \text{if } E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right) \neq Y_{U_d} \quad (28b) \end{cases}$$

where $\sigma^2_{Ady,k}$, $\sigma^2_{Bdy,k}$ and $\sigma^2_{Cdy,k}$ are the elementary unit variances from the sum of which(over the register domain units) the three components of GMSE are obtained (namely, $E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right]$, $V_M\left(Y_{U_d}\right)$ and $Cov_M\left[E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right), Y_{U_d}\right]$).

With simple algebra from Equations (24), (25) and (26), we have:

$$\sigma^2_{Ady,k} = \sum_{(k' \neq k) \cap (k' \in R_d)} \sum\nolimits_{i=1}^{I} \sum\nolimits_{i=1}^{I} f_{ki} f_{k't'} V_{y,ii'}, \tag{29}$$

$$\sigma^2_{Bdy,k} = \sigma^2_{yk} + \sum_{(k' \neq k) \cap (k' \in R_d)} \sigma_{y,kk'}, \tag{30}$$

$$\sigma^2_{Cdy,k} = \sum_{j \in R} \sum\nolimits_{i=1}^{I} f_{ki}\, u_{ki,\pi} \left(\sigma^2_{yj} + \sum_{(k' \neq k) \cap (k' \in R_d)} \sigma_{y,kk'} \right), \tag{31}$$

in which

$$v_{y,ii'} = \sum_{j \in R} \left(\sigma^2_{y,} u_{ji,\pi} u_{ji',\pi} + \sum_{\ell \neq j} u_{ji,\pi} u_{\ell i',\pi} \sigma_{y,j\ell} \right) + \partial u_{ji} \partial u_{ji'} \sigma^2_{y,j} \pi_j \left(1 - \pi_j\right). \tag{32}$$

From Equation (24), we see that the sum over the domain units of $\sigma^2_{Ady,k}$ is an upward approximation of the component of $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$

$$E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})] \leq \sum_{k \in R_d} \sigma^2_{Ady,k} = \sum_{k \in R_d} \sum_{k' \in R_d} \sum\nolimits_{i=1}^{I} \sum\nolimits_{i=1}^{I} f_{ki} f_{k'i'} v_{y,ii'}. \tag{33}$$

The estimates $\hat{\sigma}^2_{dy,k}$ (introduced in Equation 6) of the elementary unit variances $\sigma^2_{dy,k}$ may be estimated using the usual plug-in technique

$$\hat{\sigma}^2_{dy,k} = \begin{cases} \hat{\sigma}^2_{Ady,k} - \hat{\sigma}^2_{Bdy,k} & \text{if } E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right) = Y_{U_d} \quad (34a) \\ \hat{\sigma}^2_{Ady,k} + \hat{\sigma}^2_{Bdy,k} - 2\sigma^2_{Cdy,k} & \text{if } E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right) \neq Y_{U_d} \quad (34b) \end{cases}$$

by substituting the expressions of $\hat{\sigma}^2_{Ady,k}$, $\hat{\sigma}^2_{Bdy,k}$ and $\hat{\sigma}^2_{Cdy,k}$ the estimates $\hat{\boldsymbol{t}}$, $\hat{y}_k$ $(k = 1, \ldots, N_{(R)})$, $\hat{\sigma}^2_{y,j}$ and $\hat{\sigma}_{y,j\ell}\left(j, \ell = 1, \ldots, N_{(R)}\right)$, in place of the unknown parameters $\boldsymbol{\vartheta}$, $\tilde{y}_k$ $(k = 1, \ldots, N_{(R)})$, $\sigma^2_{y,j}$ and $\sigma_{y,j\ell}$ $\left(j, \ell = 1, \ldots, N_{(R)}\right)$.

From the above expressions, the main results are as follows.

1. *Domain characterization.* The quantities $v_{yii'}$, f_{ki}, $u_{ji,\pi}$, ∂u_{ji}, σ^2_{yj}, and $\sigma_{y,j\ell}$ (and their corresponding plug-in estimates $\hat{v}_{y,ii'}$, $\hat{f}_{ki}$, $\hat{u}_{ji,\pi}$, $\hat{\sigma}^2_{yj}$ and $\hat{\sigma}_{y,j\ell}$) are not domain specific. $\sigma^2_{Ady,k}$, $\sigma^2_{Bdy,k}$ and $\sigma^2_{Cdy,k}$ are domain specific since they are defined as a sum over R_d.
2. *Space for the storage*. A small amount of space is needed for storing the $(I \times I)$ matrix $\hat{\mathbf{v}}_y = \left\{\hat{v}_{yii'}; i, i' = 1, \ldots, I\right\}$ whereas storing the matrix $\hat{\mathbf{F}} = \left\{\hat{f}_{ki}; k = 1, \ldots, N_{(R)}; i = 1, \ldots, I\right\}$ requires a large volume of space. The quantities $\hat{u}_{ji,\pi}$, $\hat{\sigma}^2_{y,j}$, and $\hat{\sigma}_{y,j\ell}$ require a large volume of space. They are directly involved only in the calculus of the subdominant parts of GMSE and generally result in providing a negative contribution to this quantity.
3. *Computational complexity*. Regarding the predominant component, the calculus of the matrix $\hat{\mathbf{v}}_y$ involves the estimation of the parameters $\hat{\boldsymbol{t}}$, $\hat{\sigma}^2_{y,j}$ and$\hat{\sigma}^2_{y,j\ell}$ and the calculus of the vectors $\hat{\mathbf{u}}_{j,\pi}$ and $\partial\hat{\mathbf{u}}_j$ including different steps of linearization. On the other hand, the calculus of $\hat{\mathbf{F}}$ will indeed seem simple once the estimate of the parameter $\hat{\boldsymbol{t}}$ has beenobtained, and the calculus of the values $\hat{f}_{ki}$ may be obtained directly by just applying the specific *analytical expression* to the $\mathbf{x}_k$ vector. On the fly, the calculus of the sub dominant components of GMSE may be cumbersome mainly because of the domain dependency on the sum $\sum_{(k'\neq k)\cap(k'\in R_d)} \sigma_{y,kk'}$; nevertheless, if we neglect these terms, we introduce only as light upward approximation of the GMSE.
4. *Stability over time.* The quantities $\hat{\mathbf{v}}_y$ and the functional form of $\hat{\mathbf{F}}$ are relatively table over time. The functional form of $\hat{\mathbf{F}}$ does not change unless the method of prediction is modified. The matrix $\hat{\mathbf{v}}_y$ is essentially a function of (A) the sample design properties(and does not depend on the specific sample selection),which change only rarely when ever the survey is restructured, and (B) the structure of the model variances and covariances $\hat{\sigma}^2_{y,j}$ and $\hat{\sigma}^2_{y,j\ell}$, which change rarely and only in the case where there is a structural break in the y value sand in the model for their generation.

The above expressions and the results of the empirical experiment (in Section 7) offer some suggestions on how to develop a feasible and robust computational strategy. First, we note that the subdominant components give a negative contribution to GMSE and that this tends to be negligible for unplanned domains in which $E_P\left(\hat{Y}_{R_d}|\mathbf{y}\right) \neq Y_{U_d}$. This finding invites the examination of two different alternatives: one for the planned domain (for which $E_P\left(\hat{Y}_{R_d}|\mathbf{y}\right) = Y_{U_d}$) and the other for unplanned domains.

For the **planned case** where domains are well known in advance and limited in number, the Equation (13a) is used for the computation. In particular:

1. The matrix $\hat{\mathbf{v}}_y$ is computed and stored.
2. The matrix $\hat{\mathbf{F}}$ is not stored. The values $\hat{f}_{ki}$ are computed on the fly on the basis of its functional form, which links these values directly to the auxiliary variables $\mathbf{x}_k$ and to the parameters $\hat{\boldsymbol{t}}$.
3. With regard to the computation of matrix $\hat{\mathbf{F}}$, the only two objects that are permanently stored are the vector of the parameter $\hat{\boldsymbol{t}}$ and the functional forms that permit the computation of the $\hat{f}_{ki}$ values.
4. The elementary unit variances $\hat{\sigma}^2_{Bdy,k}$ are computed and stored.

For the **unplanned domains** (Equation (13*b*)), generally unpredictable in their number, the quantity $E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right]$ is used as an upward approximation of the GMSE. This can be easily computed on the fly on the basis of the stored material as defined in points 1 and 3 listed above.

6. The Coverage Errors

The register may be affected by coverage errors, which implies $N_{(R_d)} \neq N_{(U_d)}$, where $N_{(R_d)}$ is the number of units in domain d listed in the register, and $N_{(U_d)}$ is the domain population size.

An unbiased estimate $\hat{N}_{unb,(U_d)}$ may be obtained directly from the register with the Extended Dual System Estimator (Nirel and Glickman 2009; Pfeffermann 2015) by summing up the predicted values of a particular w variable over the register units:

$$\hat{N}_{unb,(U_d)} = \sum_{k \in R_d} \hat{w}_k, \tag{35}$$

in which

$$\hat{w}_k = \frac{\hat{P}\left(k \in U | k \in R\right)}{\hat{P}\left(k \in R | k \in U\right)} \tag{36}$$

represents the model (roughly) unbiased prediction of

$$w_k = \frac{P\left(k \in U | k \in R\right)}{P\left(k \in R | k \in U\right)} \tag{37}$$

where $\hat{P}(k \in U|k \in R)$ is the model's unbiased estimate of the conditional probability $P(k \in U|k \in R)$ such that unit k included in the register belongs to the population and $\hat{P}(k \in R|k \in U)$ is the model's unbiased estimate of the conditional probability $P(k \in R|k \in U)$ such that unit k belonging to the population is included in the register. This approach has been adopted for the Italian Population Base Register by integrating the register data with the Census *Population Coverage Survey* carried out each year as a component of the Italian Permanent Census Survey System (Righi et al. 2021). The GMSE of $\hat{N}_{unb,(U_d)}$ may be expressed as and specified as described in Subsection 4.2 by defining the predictions and model's expected values of the w variables (instead of the y variables).

Moreover, we note that in the case in which the register is affected by coverage errors, a weighted estimator of the total of a generic y variable in the domain U_d, Y_{U_d} can conveniently be expressed as:

$$\hat{Y}_{U_d} = \sum_{k \in R_d} \hat{y}_k \hat{w}_k.$$

The GMSE of this estimator can be obtained by considering its linear approximation:

$$\mathrm{GMSE}\left(\hat{Y}_{U_d}\right) \cong \mathrm{GMSE}\left(\sum_{k \in R_d} \hat{y}_k w_k + \sum_{k \in R_d} \tilde{y}_k \hat{w}_k\right). \tag{38}$$

We omit further technical developments, which can be easily derived according to the procedure given in Section 4.

7. Experimental Study

This experimental study, which is based on real data, compares the empirical GMSE of a Monte Carlo simulation with the approximate GMSE obtained from Taylor approximations.

The data set for the empirical study is an administrative archive that contains information regarding the population of 21,782 Sapienza University of Rome (Italy) alumni who graduated between March 1, 2008 and February 28, 2009 and who signed a job contract in the three years following graduation (Alleva and Petrarca 2013; Gruppo UNI. CO 2015; Petrarca 2014, a, b). The study focuses on the disciplinary sectors of engineering, sciences, literature, economics and statistics, psychology, chemistry and pharmacy, and architecture. The data set has 7,085 units. However, the simulations conducted on other subsets of disciplinary sectors confirm the results shown here.

The target y variable is the number of days worked during the three years after graduation. The vector of auxiliary variables for a unit is $\mathbf{x}_k = (x_{k1}, x_{k2}, x_{k3}, x_{k4}, x_{k5}, x_{k6}, x_{k7}, x_{k8})'$, where $x_{k1} = 1$; x_{k2}: gender of a graduate; x_{k3}: age at the time of graduation; x_{k4}: graduation on time (yes/no); x_{k5}: graduation from a second-cycle programme (yes/no); x_{k6}: number of days that a graduate has waited before obtaining a permanent contract; x_{k7}: number of days that a graduate has waited before obtaining a contract with a highly qualified position (ISCO 1-ISCO 2); and x_{k8}: number of days that a graduate has waited before obtaining a contract with an actual duration of more than or equal to eight months.

7.1. Standard Simulation and Linear Approximation

We generated 1,000 populations of 7,085 units. For each population, a vector of the target variables y_k was generated as described in Section 3 by taking the sum of two components: $y_k = E_M(y_k) + e_k = \tilde{y}_k + e_k$, where $\tilde{y}_k$ is the vector of the fitted values obtained from a linear regression model attuned to the super-population, and e_k is generated with a normal distribution with mean 0 and variance equal to the variance of the y_k in the real data set ($\sigma^2 = 0.1159733 \cdot 10^7$). For each population, 1,000 samples of $n = 500$ units were selected utilizing a simple random sample design without replacement. The two processes that generate populations and samples allow us to simulate the model and sampling uncertainty. For each sample we obtained from a simple linear regression model, the estimated regression coefficients $\hat{\boldsymbol{t}}$ formulate the $\hat{\boldsymbol{y}} = \mathbf{X}\hat{\mathbf{t}}$ vector to be utilized for constructing predictions. The sum of these values restricted to the domain built the $\hat{Y}_{R_d}$ estimates.

The evaluation of our information by linear approximation only requires us to know the matrix of the auxiliary variables and then implement the calculation of the formulae given in the text. We had two types of domains: the internal domains for which the domain membership variable belongs to the vector of auxiliary variables and the external domains otherwise. Note that in the case of internal domains, we apply a generalized regression estimator (Särndal et al. 1992). The size of each domain is presented in Table 2.

7.2. Results

First, we discuss the simulation results concerning *GMSE for* the two large domains: *gender* and *scientific group* (obtained by summing *sciences*, *chemistry and pharmacy*,

Table 2. Internal and external domains with their size.

Internal domain	Population size	External domain	Population size
Gender_female	4,281	Scientific group	3,368
Gender_male	2,804	Others	3,717

economics and statistics and *engineering*) and the domain *others* (containing *architecture*, *literature* and *psychology*).

In Tables 3 and 4, we report the values of GMSE and its components for the large internal and external domains, respectively. For the sake of brevity, the tables display only the results *for gender female* and *scientific group*. Similar results were produced for the other two domains. Part A of Table 3 reports the experimental results of the $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$ in the case of an internal domain (*gender female*), whereas part B of Table 3 shows the results for an external domain (*scientific group*) achieved using the Monte Carlo simulation and linear approximations, hereinafter referred to as "empirical" and "linearized" expectations, respectively.

For the empirical expectations of the GMSE, we considered Equations (4) and (13*a*) for the internal domains. In the case of the external domain, the evaluation of GMSE is based on Equation (13b) because in this case, the projection estimator is biased, $E_P[(\hat{Y}_{R_d}|\boldsymbol{y})] \neq Y_{U_d}$. Equations (3.2) and (3.7) of Table 3 show that the dominant part of GMSE comes from $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$.

For the linearized Equations (3.3–7.5 of Table 3-part A; 3.8–3.10 of Table 3-part B), we give the estimates for the dominant contribution of GMSE computed from the definitions of Equation (23) applied to the linear regression model. As expected, these evaluations are sufficiently in agreement with one another.

Table 3. GMSE and its components.

A: Case of an internal domain (*Gender_female*)

Empirical

$V_M(Y_{U_d}) = 426.6$

$\text{GMSE}(\hat{Y}_{R_d}) = E_P E_M(\hat{Y}_{R_d} - Y_{U_d})^2 = 6{,}649.0$ (3.1)

$\text{GMSE}(\hat{Y}_{R_d}) \cong E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})] - V_M(Y_{U_d}) = 7{,}069.9\text{-}426.6 = 6{,}643.3$ (3.2)

Linearized

$V_M(Y_{U_d}) = \boldsymbol{\gamma}'_d \sum_y \boldsymbol{\gamma}_d = 496.5$ (3.3)

$E_P[V_M(\hat{\mathbf{t}}|\boldsymbol{\lambda})] = \sum_{j\in R}[E_P(\boldsymbol{u}_{j|\lambda}\boldsymbol{u}'_{j|\lambda})\sigma^2_{yj}] = 7,021.6$ (3.4)

$\text{GMSE}(\hat{Y}_{R_d}) = E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})] - V_M(Y_{U_d}) = 7{,}021.6\text{-}496.5 = 6{,}525.1$ (3.5)

A: Case of an external domain (*Scientific group*)

Empirical

$V_M(Y_{U_d}) = 380.2$

$\text{GMSE}(\hat{Y}_{R_d}) = E_P E_M(\hat{Y}_{R_d} - Y_{U_d})^2 = 2{,}996.4$ (3.6)

$\text{GMSE}(\hat{Y}_{R_d}) \cong E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})] + V_M(Y_{U_d}) - 2Cov_M[E_P(\hat{Y}_{R_d}|\boldsymbol{y}), Y_{U_d}]$

$= 3{,}045.8 + 380.2 - 429.5 = 2{,}996.5$ (3.7)

Linearized

$V_M(Y_{U_d}) = \boldsymbol{\gamma}'_d \sum_y \boldsymbol{\gamma}_d = 390.6$ (3.8)

$E_P[V_M(\hat{\mathbf{t}}|\boldsymbol{\lambda})] = \sum_{j\in R}[E_P(\boldsymbol{u}_{j|\lambda}\boldsymbol{u}'_{j|\lambda})\sigma^2_{yj}] = 2{,}990.3$ (3.9)

$\text{GMSE}(\hat{Y}_{R_d}) = E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})] + V_M(Y_{U_d}) - 2Cov_M[E_P(\hat{Y}_{R_d}|\boldsymbol{y}), Y_{U_d}]$

$= 2{,}990.3 + 390.6\text{-}429.5 = 2{,}951.4$ (3.10)

The numbers are scaled by a factor of 10^7.

Table 4. Values of $E_P[V_M(\hat{Y}_{R_d}|\lambda)]$ in the case of simulation and linearization.

Domains	Domain population size	Expected sampling fraction (n_d/N_d)	$E_P[V_M(\hat{Y}_{R_d}\|\boldsymbol{\lambda})]$		Difference* (A-B)	Relative difference (A-B)/A
			Empirical* (A)	Linearized* (B)		
Architecture	905	0.128	214.2	210.8	3.36	0.016
Chemistry and Pharmacy	400	0.056	85.8	83.2	2.56	0.030
Economics and Statistics	1,349	0.190	472.6	458.3	14.23	0.030
Engineering	1,259	0.178	590.6	580.5	10.13	0.017
Literature	1,975	0.279	1,039.4	1,020.7	18.70	0.018
Psychology	837	0.118	248.0	237.9	10.15	0.041
Sciences	360	0.051	38.8	37.5	1.26	0.033
Scientific group	3,368	0.475	3,045.8	2,990.3	55.50	0.018

*The numbers are scaled by a factor of 10^7.

Our attention now focuses on $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$. Experiment serves to investigate the external disciplinary sector domains, allowing us to compute the expectation $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$ for small and large domains. Table 4 reports the values of $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$ in the case of simulation (column A) and linearization (column B). The differences between the two estimates are positive and small, ranging between 16 and 41‰.

Figures 1a and 1b show the data of Table 4, where the circled points correspond to the values of external domains, whereas the diamond points correspond to the values for the

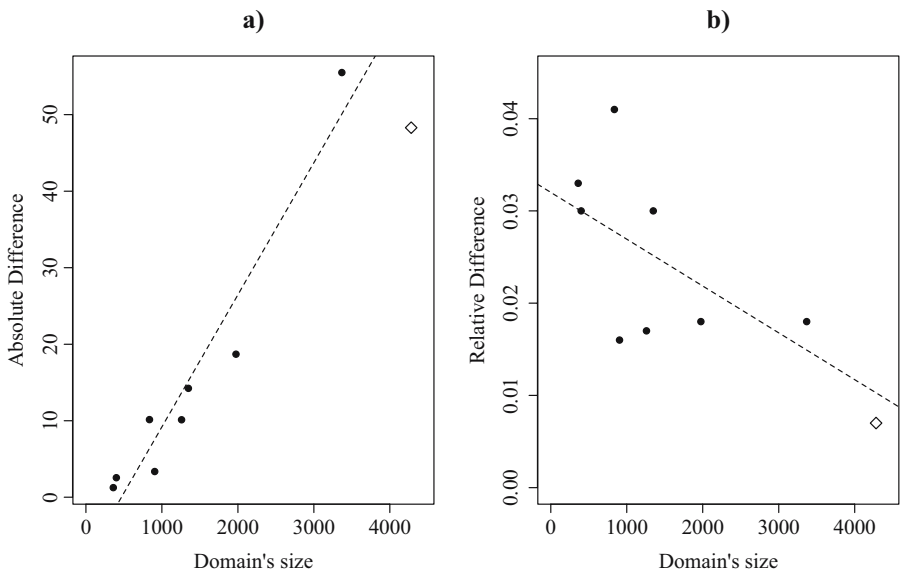

Fig. 1. Absolute (a) and relative (b) differences in $E_P[V_M(\hat{Y}_{R_d}|\lambda)]$ between simulation and linearization according to the size of the domain.
**The point ◇ refers to the internal domain gender_female*

internal domain *gender female*, also shown here for completeness. The trend lines were drawn without considering the values of the internal domain.

As shown in Figure 1a, the difference between the empirical and linearized expectations indicates a positive relationship with the population size of the external domains. On the other hand, Figure 1b, shows that the relative difference increases when the domain sample size decreases. These findings confirm that the linearization method produces a downward approximation directly related to the sample size.

8. Conclusions

In this article, we have proposed the *global mean square error* as an appropriate measure to assess the accuracy of register aggregates. This measure has some relevant qualities: generality, stability over time and robustness in the case of model failure. It is easy to use and communicate to users and could be accepted as a measure of accuracy by the main professional families of methodologists within the National Statistical Institutes.

Our approach is based on only the first and second moments of the random distributions of the specific source of uncertainty. Its calculus does not imply full knowledge of the underlying distributions.

In addition, we suggested an immediate GMSE calculation strategy for any unexpected user request by simply aggregating domain-dependent variances estimated at the unit level. The amount of information to store for this calculation is limited, and the domain dependency is limited to a few useful variables. The calculation strategy suggested here is a powerful advantage of our proposal, as it allows NSIs to build algorithms that can be applied *instantly* to any user request, thus improving the relevance, transparency and confidence of official statistics.

The simulation conducted confirms the accuracy of the different GMSE decompositions proposed in Section 4 as model-assisted projection estimators, whether they are design-unbiased or biased. Furthermore, the very small discrepancies between empirical and linearized expectations suggest that the proposed approximation method can be undertaken as a valid computational strategy. We emphasize once again that the linearized variance is calculated using unit-level elements: $\mathbf{u}_{j,\pi}$, $\partial\mathbf{u}_j$, π_j, $\sigma^2_{y,j}$ and is suitable for the calculation of an accuracy measurement of the statistics based on registers.

In addition to the more extensive validation studies to be launched to confirm the benefits and robustness of the empirical results, the main further steps to be taken in the research outlined here are the definition of a validation strategy and targeted extensions with regard to both other sources of uncertainty and parameters other than means and totals that cannot be expressed as simple linear functions of the register predictions.

Although significant validation may not be feasible at this stage, we suggest that NSIs plan, on a regular (e.g., annual) basis, experimental studies conducted for specific domains where the GMSE values obtained as proposed in this article are compared with corresponding values obtained from other approaches. We suggest considering replication methods that repeat the whole process of calculating the register predictions. These experiments may show inconsistencies in the experimental evidence on which further empirical and theoretical investigations should be considered. Furthermore, they could validate whether the asymptotic properties adopted here hold.

Beyond the coverage error considered here, other extensions need to be developed. In our view, those on which research should be prioritized are linkage errors and predictions based on machine learning algorithms. For both cases, it should be analyzed whether the main tools adopted here, the linearization and decomposition of the GMSE into simpler conditional components, could be successfully applied. Other simple and straight forward extensions could be obtained by considering standard linearization techniques to measure the accuracy of non- linear parameters derived from register microdata, such as correlations, regression parameters or quantiles.

Finally, we stress that another aspect to consider in facilitating the strategy feasibility and its wider applicability is to implement software tools that make it easy to calculate GMSE from the microdata of the register and from the functional form used to build the forecasts.

9. Appendix

9.1. *Demonstration of the Equations (13a) and (13b)*

To derive Equation (13*a*), we add and subtract the overall mean, $E_P E_M(\hat{Y}_{R_d}) = \tilde{Y}_{R_d}$ in the expression of GMSE. We have

$$\begin{aligned} GMSE(\hat{Y}_{R_d}) &= E_P E_M(\hat{Y}_{R_d} - E_P E_M(\hat{Y}_{R_d}) + E_P E_M(\hat{Y}_{R_d}) - Y_{U_d})^2 \\ &= E_P E_M(\hat{Y}_{R_d} - E_P E_M(\hat{Y}_{R_d}))^2 + E_P E_M(E_P E_M(\hat{Y}_{R_d}) - Y_{U_d})^2 \\ &\quad + 2E_P E_M\left[\left(\hat{Y}_{R_d} - E_P E_M(\hat{Y}_{R_d})\right)\left(E_P E_M(\hat{Y}_{R_d}) - Y_{U_d}\right)\right] \\ &= E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})] - V_M(Y_{U_d}), \end{aligned}$$

since, from Kendall and Stuart (1976, 196), it is

$$E_P E_M\left(\hat{Y}_{R_d} - E_P E_M\left(\hat{Y}_{R_d}\right)\right)^2 = E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] + V_P\left[E_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right],$$

and $E_M(\hat{\mathbf{t}}|\boldsymbol{\lambda}) = \boldsymbol{\vartheta}$, then

$$V_P\left[E_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] \cong V_P\left[\sum_{k\in R_d} f[\mathbf{x}_k; E_M(\hat{\mathbf{t}}|\boldsymbol{\lambda})]\right] = V_P\left[\sum_{k\in R_d} f(\mathbf{x}_k; \vartheta)\right] = 0$$

$$E_P E_M\left(\tilde{Y}_{R_d} - Y_{U_d}\right)^2 = E_M\left(\tilde{Y}_{R_d} - Y_{U_d}\right)^2 = V_M\left(Y_{U_d}\right) \text{ and}$$

$$E_P E_M\left[\left(\hat{Y}_{R_d} - E_P E_M\left(\hat{Y}_{R_d}\right)\right)\left(E_P E_M\left(\hat{Y}_{R_d}\right) - Y_{U_d}\right)\right]$$

$$= \left[\left(E_P E_M\left(\hat{Y}_{R_d}\right)\right)^2 - E_M\left(Y_{U_d}^2\right)\right] = -V_M\left(Y_{U_d}\right).$$

To demonstrate Equation (13*b*), consider that

$$\begin{aligned} &E_M E_P\left(\hat{Y}_{R_d} - E_P E_M\left(\hat{Y}_{R_d}\right)\right)\left(E_P E_M\left(\hat{Y}_{R_d}\right) - Y_{U_d}\right) \\ &= -E_M\left[\left(E_P\left(\hat{Y}_{R_d}|\mathbf{y}\right)Y_{U_d}\right)\right] + E_P E_M\left(\hat{Y}_{R_d}\right) = -Cov_M\left(E_P\left(\hat{Y}_{R_d}|\mathbf{y}\right), Y_{U_d}\right). \end{aligned}$$

10. References

Alleva, G., and F. Petrarca. 2013. "New indicators for investigating the Integration of Sapienza graduates into the labor market." Working papers n. 120/2013 del Dipartimento Memotef, ISSN 2239-608X.

Alleva, G. 2017. "The new role of sample surveys in official statistics." ITACOSM 2017, The 5th Italian Conference on Survey Methodology, June 14, 2017, Bologna, Italy. Available at: https://www.istat.it/it/files//2015/10/Alleva_ITACOSM_14062017.pdf (accesssed May 2021).

Biemer, P.P. 2010. "Total Survey Error Design, implementation, and evaluation." *Public Opinion Quarterly* 4(5) : 817–848. DOI: https://doi.org/10.1093/poq/nfq058.

Binder, D.A., and Z. Patak. 1994. "Use of estimating functions for estimation from complex surveys." *Journal of the American Statistical Association* 89: 1035–1043. DOI: https://doi.org/10.1080/01621459.1994.10476839.

Breidt, F.J., and J.D. Opsomer. 2017. "Model-Assisted Survey Estimation with Modern Prediction Techniques." *Statistical Science* 32(2) : 190–205. DOI: https://doi.org/10.1214/16-STS589.

Chambers, R.L., and R.G. Clark. 2015. "An Introduction to Model-Based Sampling with Applications." *Oxford Statistical Science*. 37. DOI: https://doi.org/10.1093/acprof:oso/9780198566625.001.0001.

Chen, S., and D. Haziza. 2017. "Multiply robust imputation procedures for the treatment of item nonresponse in surveys." *Biometrika* 102: 439–453. DOI: https://doi.org/10.1007/s40300-017-0128-9.

Citro, C.F. 2014. "From multiple modes for surveys to multiple data sources for estimates." *Survey Methodology*. Statistics Canada. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14128-eng.pdf?st=emZzAE9_(accessed May 2021).

Cochran, W.G. 1977. *Sampling techniques*, (Third edition). New York: Wiley. Available at: https://glad.geog.umd.edu/Potapov/_Library/Cochran_1977_Sampling_Techniques_ Third_Edition.pdf (accessed May 2021).

Deville, J.-C. 1999. "Variance estimation for complex statistics and estimators: Linearization and residual techniques." *Survey Methodology* 25(2) : 193–203. Statistics Canada. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/ 1999002/article/4882-eng.pdf?st=qNANeGtP (accessed May 2021).

Deville, J.-C., and Y. Tillè. 2005. "Variance approximation under balanced sampling." *Journal of Statistical Planning and Inference* 128(2) : 569–591. Available at: https://core.ac.uk/download/pdf/43673958.pdf (accessed May 2021).

Eurostat. 2019. Available at: https://ec.europa.eu/eurostat/cros/content/essnet-quality-multisource-statistics-komuso_en (accessed May 2021).

Falorsi, P.D., P. Lavallée, and P. Righi. 2019. "Cost Optimal Sampling for the Integrated Observation of Different Populations." *Survey Methodology*. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019003/article/00004-eng.pdf?st=qZIAruhQ (accessed May 2021).

FAO. 2014. *Technical Report on the Integrated Survey Framework*, Technical Report Series GO-02- 2014. Available at: http://gsars.org/wp-content/uploads/2014/07/Technical_report_on-ISF-Final.pdf (accessed May 2021).

Graf, M. 2015. *A Simplified Approach to Linearization Variance for Surveys*. Technical Report, Institut De Statistique, Université de Neuchâtel. Available at: https://doc.rero.ch/record/324723/files/00002754.pdf (accessed May 20121).

Gruppo UNI.CO. 2015. *La Domanda di Lavoro per i laureati*. I risultati dell'integrazione tra gli archivi amministrativi dell'Universita' Sapienza di Roma e del Ministero del Lavoro e delle Politiche Sociali, Edizioni Nuova Cultura- Roma. ISBN 9788868124816. DOI: https://doi.org/10.4458/4816.

Isaki, C., and W.A. Fuller. 1982. "Survey design under the regression superpopulation model." *Journal of the American Statistical Association* 77: 89–96.

ISCO 1-ISCO 2. Available at: https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/@publ/documents/publication/wcms_172572.pdf

Istat. 2016. *Istat's Modernisation Programme*. Available at: https://www.istat.it/en/files/2011/04/IstatsModernistionProgramme_EN.pdf (accessed May 2021).

Kendall, M.G, and A. Stuart. 1976. *The Advanced Theory of Statistics: Design and analysis, and time- series*. Hafner.

Kim, J.K., and J.N.K. Rao. 2012. "Combining data from two independent surveys: a model-assisted approach." *Biometrika* 99(1) : 85–100. DOI: https://doi.org/10.1093/-biomet/asr063.

Nedyalkova, D., and Y. Tillé. 2008. "Optimal sampling and estimation strategies under the linear model." *Biometrika* 95: 521–537. DOI: https://doi.org/10.1093/biomet/asn027.

Nirel, R., and H. Glickman. 2009. "Chapter 21 – Sample Surveys and Censuses." In *Handbook of Statistics*, edited by C.R. Rao.: Elsevier.

Petrarca, F. 2014a. "Non-metric PLS path modeling: Integration into the labour market of Sapienza graduates." In *Advances in latent variables. Studies in theoretical and applied statistics*: 159–170. Berlin: Springer. DOI: https://doi.org/10.1007/10104_2014_16

Petrarca, F. 2014b. Assessing Sapienza University alumni job careers: Enhanced partial least squares latent variable path models for the analysis of the UNI.CO administrative archive. PhD diss., Dipartimento di Economia dell'Universita' degli studi Roma Tre. Available at: http://hdl. handle.net/2307/4167 (accessed May 2021).

Pfeffermann, D. 2015. "Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture." *Journal of Survey Statistics and Methodology* 3(4) : 425–483. DOI: https://doi.org/10.1093/jssam/smv035.

Righi, P., P.D. Falorsi, S. Daddi, E. Fiorello, P. Massoli, and M.D. Terribili. 2021. "Optimal sampling for the Population Coverage Survey of the new Italian Register Based Census." Journal of Official Statistics (September 2021)

Särndal, C.E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer-Verlag.

Scholtus, S. 2019. "A bootstrap method for estimators based on combined administrative and survey data." In NTTS Conference 2019. Brussels, Belgium. Available at: https://slidetodoc. com/download.php?id=4386.

Statistics Canada. 2009. *Statistics Canada Quality Guidelines*, (6th edition). Available at: https://www150.statcan.gc.ca/n1/en/pub/12-539-x/12-539-x2019001-eng.pdf?st=y2AqFuiY.

Wallgren, A., and B. Wallgren. 2014. *Register-based Statistics: statistical methods for administrative data*, (2nd Edition). Chichester: Wiley.

Wolter, K.M. 1985. *Introduction to Variance Estimation*. New York: Springer.

Wolter, K.M. 1986. "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association* 81: 338–346. DOI: https://doi.org/10.2307/2289222.

Vaillant, R. 2009. "Model based predictions of finite population totals." In *Chapter 23 in Handbook of statistics 29: Design, Methods and Applications*, edited by P. Pfefferman and C.R. Rao. Amsterdam: North Holland.

Vallée, A.A., and Y. Tillé. 2019. "Linearisation for Variance Estimation by Means of Sampling Indicators: Application to Non-response." *International Statistical Review* 0(0) : 1–21. DOI: https://doi.org/10.1111/insr.12313.

Ziegler, A. 2015. *Generalized Estimating Expressions*, Springer and Verlag. Lecture Notes in statistics.

Received May 2019
Revised August 2020
Accepted January 2021