3rd Workshop on Methodologies for Official Statistics
Rome, Istat, 4-5 December 2024

Linearization approach for measuring the accuracy of multinomial outcomes from a statistical register

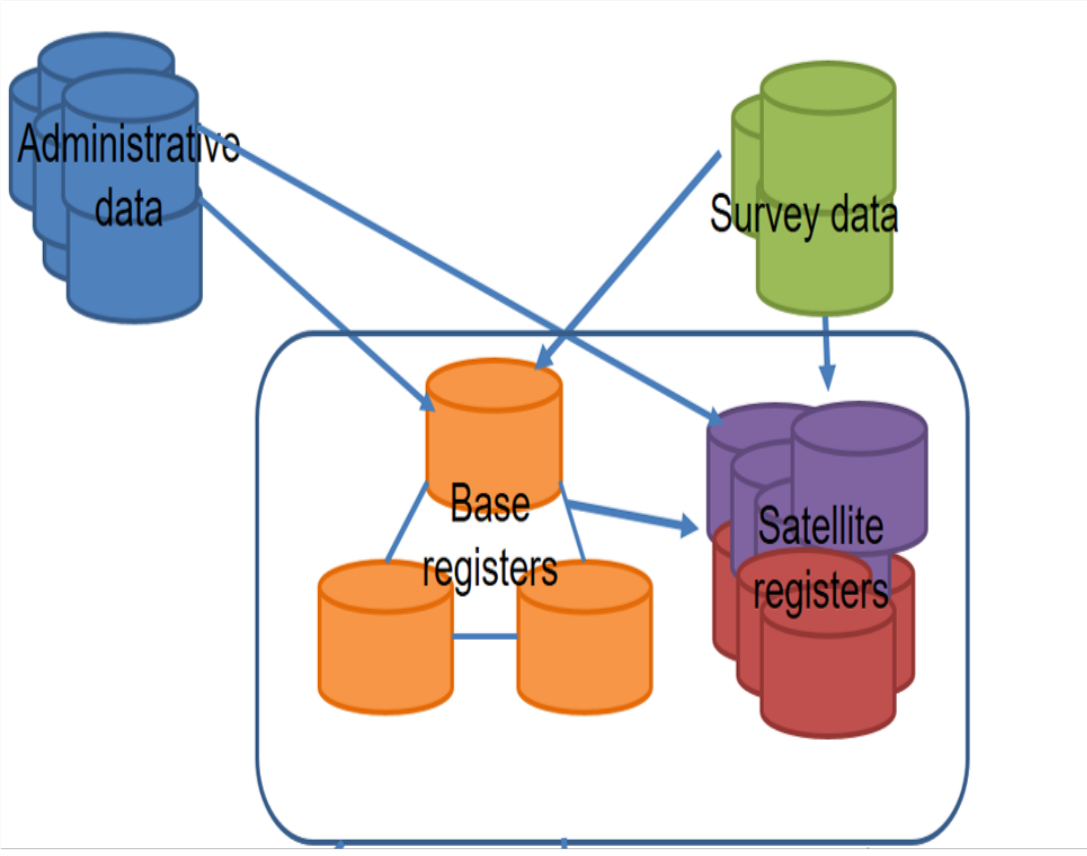
S.Falorsi, D. Chianella, R. Filippini, S. Toti
(ISTAT - Italy)

G. Alleva, N. Deliu, P.D. Falorsi
(Sapienza Univ. of Rome)

Contents

- One of the pillars of the modernization is the Integrated System of Statistical Registers (ISSR)
- ISSR, is composed of many variables many of which are obtained through multi-source statistical processes
- The Generalized Mean Squared Error, GMSE, is the proposed accuracy measure for planned and unplanned (possibly computed directly by users) estimates from variables of the ISSR
- An application on the Attained Level of Education by means of Mass Imputation based on Multinomial-distribution as part of the Base Register of Individuals (BRI) + Other Census Variables like non employment status
- Comparative analysis of empirical properties of GMSE vs Bootstrap MSE estimation by means of a Monte Carlo Simulation based on synthetic data generation from Attained Level of Education

The Istat Integrated System of Statistical Registers (ISSR)



- ISSR consists in a number of coherent registers to produce several types of statistical outputs
- The release of each statistical register is based on a multi-source process, that mainly integrates administrative data, but also surveys or other registers results
- This innovation has lead to create new processes that can vary a lot in complexity
- The need of a new proper quality system to assess and monitor the new processes and their results has been expressed from several points of view

Quality of estimates from ISSR

- The estimates obtained from the statical registers should be associated by a measure of their uncertainty.
- If it is difficult for traditional sample surveys to produce accuracy measures that take into account different error sources, this is even more challenging for estimates from statistical registers:
  - Because the process includes different type of data sources and different statistical methods (such as record linkage, statistical matching, or imputation/prediction).
  - Because a great potentiality of ISSR is to produce estimates on unplanned domains, thus a way to calculate on-the-fly uncertainty measures should be provided to the user.

Examples on Italian Registers

Topic	Register	Statistical Analysis
Living population (with weights for over/undercoverage)	Base Population	Overcoverage/ Undercoverage Models
Level of education	Base Activity	GLM
Employment status	Base Activity	HMM
Census Microdata Database		SAE Projections
Local units (main variables)	Base Economic units	Regression
Economic variables	Extend Register of economic units	Model assisted projection
Main cultivar	Farm Register	Model assisted projection

Population, Statical Register and estimation domains

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	Overcoverage
			⋮	⋮	⋮	⋮	
1	y_1	1	⋮			0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
k	y_k	0	k	$\hat{y}_k$	$\mathbf{x}_k$	1	
⋮	⋮	⋮		⋮	⋮	1	
⋮	⋮	⋮	$N_{(R)}$	$\hat{y}_{N_{(R)}}$	$\mathbf{x}_{N_{(R)}}$	⋮	
⋮	⋮	⋮					Undercoverage
$N_{(U)}$	$y_{N_{(U)}}$	1					

$\hat{y}_k =$ Value Built by an explicit or implicit statistical model or algorithm.

Example: The Attained Level of Education

(?)

- The attained level of education is part of the Base Register of Individuals (BRI).
- Sources of data:
  - (i) administrative information (MIUR) $\rightarrow$ data from 2011 onwards with 16 modalities;
  - (ii) 2011 Census information $\rightarrow$ educational status at 2011 with 12 modalities;
  - (iii) sample survey to cope with delay and informative gaps $\rightarrow$ permanent Italian census from 2018 with 17 modalities (5% of the population);
  - (iv) other auxiliary information (transfer of residence; 2012-2017 with 4 modalities)

$$\Downarrow$$

Reconciliation: $K = 8$ categories

Example: The Attained Level of Education

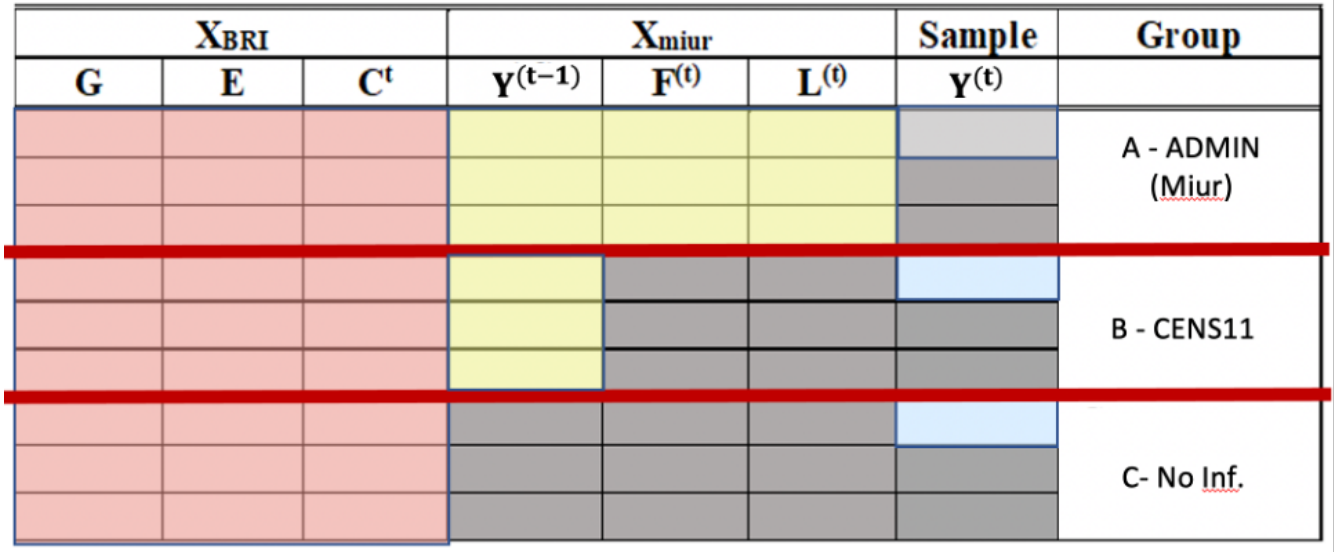


Figure: Informative context for mass imputation of $Y^{(t)}$ (**?**)

- The informative context is quite heterogeneus:
  - Red: Individual characteristics known for all the population of interest: G = gender; E = age classes; C = citizenship (It / not It)
  - Yellow / lightblue: partially available
  - Grey: Missing data
- Mass imputation procedure to estimate the attained level of education $Y \doteq Y^{(t)}$, by means of *multinomial* log-linear models.

(One of the) Challenges

- The ISSR is thus the result of statistical processes subject to different sources of statistical uncertainty (sampling uncertainty, model uncertainty, etc.)

How to deal with uncertainty?

- Evaluation of the sources of uncertainty and errors
- Evaluation of the accuracy of the (imputed) data
- Responsibility and transparency on the quality of data
- Inform the end users (unplanned, on the fly)

Our Goal(s)

Istat-Sapienza Project

- Come up with feasible measures to calculate estimates' accuracy, accounting for different sources of uncertainty
- Focusing on the context of the attained level of education:
  - we based it on a previously introduced (generic) measure of global uncertainty (GMSE) see Alleva et al. (2021: J Off Stat, 37(2), 481-503)
  - adapt it to a multinomial setting

Setup

Our target parameter is the *unknown* population totals $\boldsymbol{\theta}_k^{(d)}, \quad k = 1, \ldots, K$, for a given *domain* d (e.g., number of individuals having a PhD in the province of Bologna):

$$\boldsymbol{\theta}_k^{(d)} = \sum_{i=1}^{N} \gamma_i^{(d)} Y_{ik},$$

- The response variable $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$ is modelled as

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iK}) \sim Mult(1, \boldsymbol{p_i}), \quad i = 1, \ldots N,$$

  with $Y_{ik} \in \{0, 1\}$ and s.t. $\sum_{k=1}^{K} Y_{ik} = 1$, and $\boldsymbol{p_i} = \{p_{i1}, \ldots, p_{iK}\}$, where $p_{ik} = P(Y_{ik} = 1 | X_i)$, being the unknown parameter vector.
- $\boldsymbol{\gamma}^{(d)}$ is the binary domain membership vector.

Setup

Let now $\hat{\boldsymbol{\theta}}_k^{(d)}$ be a consistent estimator of $\boldsymbol{\theta}_k^{(d)}$. We consider:

$$\hat{\boldsymbol{\theta}}_k^{(d)} = \sum_{i=1}^{N} \gamma_i^{(d)} f_k(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}),$$

where

$$f_k(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}) = \hat{p}_{ik} = \frac{\exp \boldsymbol{x}_i{}^T \hat{\boldsymbol{\beta}}_k}{1 + \sum_{k=1}^{K-1} \exp \boldsymbol{x}_i{}^T \hat{\boldsymbol{\beta}}_k}, \quad k = 1, \ldots, K-1,$$

with $\boldsymbol{x}_i \in \mathbb{R}^J$ a set of covariates and $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kJ})$ the unknown model coefficients, $k = 1, \ldots, K$.

The goal is to estimate the accuracy of $\hat{\boldsymbol{\theta}}_k^{(d)}$, so that they can be provided *on-the-fly* by register users.

The Global Mean Squared Error

- Idea: take into consideration all the random components $(R_1, \ldots, R_p)$ involved in the inferential process.
- We focus on two sources of uncertainty: (1) M = model variability; (2) Π = sampling variability.
- is the sample's indicator variable vector

Definition (GMSE)

Given $\boldsymbol{\theta}^{(d)}$ and $\hat{\boldsymbol{\theta}}^{(d)}$, the GMSE is defined as:

$$GMSE(\hat{\boldsymbol{\theta}}^{(d)}) = \mathbb{E}_{(R_1,\ldots,R_p)}(\hat{\boldsymbol{\theta}}^{(d)} - \boldsymbol{\theta}^{(d)})^2$$

$$= \mathbb{E}_{\Pi}\mathbb{E}_{M}\left(\sum_{i=1}^{N}\gamma_i^{(d)} f(\boldsymbol{x}_i;\hat{\boldsymbol{\beta}}) - \sum_{i=1}^{N}\gamma_i^{(d)} Y_i \Big| \boldsymbol{\lambda}\right)^2.$$

The Global Mean Squared Error

Definition (Upper bound on GMSE)

Incorporating both sampling and model uncertainty, under the assumption that the estimator is design and model unbiased, the GMSE can be expressed as:

$$\begin{aligned} GMSE(\hat{\boldsymbol{\theta}}^{(d)}) &= \mathbb{E}_{\Pi}\mathbb{E}_{M}(\hat{\boldsymbol{\theta}}^{(d)} \pm \tilde{\theta}^{(d)} - \boldsymbol{\theta}^{(d)}|\boldsymbol{\lambda})^2 \\ &= \mathbb{E}_{\Pi}\mathbb{V}\mathrm{ar}_{M}(\hat{\boldsymbol{\theta}}^{(d)}|\boldsymbol{\lambda}) - \mathbb{V}\mathrm{ar}_{M}(\boldsymbol{\theta}^{(d)}) \\ &\leq \mathbb{E}_{\Pi}\mathbb{V}\mathrm{ar}_{M}(\hat{\boldsymbol{\theta}}^{(d)}|\boldsymbol{\lambda}) \end{aligned}$$

with $\tilde{\theta}^{(d)} \doteq \mathbb{E}(\hat{\boldsymbol{\theta}}^{(d)}) = \sum_{i=1}^{N} \gamma_i^{(d)} f(\boldsymbol{x}_i;\boldsymbol{\beta}) = \sum_{i=1}^{N} \gamma_i^{(d)} p_i$

The GMSE for multinomial data

Under K response categories, the (scalar) GMSE is generalized to a GMSE matrix, with the dominant component $\mathbb{E}_{\Pi}\mathbb{V}\mathrm{ar}_M(\hat{\boldsymbol{\theta}}^{(d)}|\boldsymbol{\lambda})$ involving:

$$\mathbb{V}\mathrm{ar}_M(\hat{\boldsymbol{\theta}}^{(d)}|\boldsymbol{\lambda}) = \begin{bmatrix} \mathbb{V}\mathrm{ar}_M(\hat{\theta}_1^{(d)}|\boldsymbol{\lambda}) & \mathbb{C}\mathrm{ov}_M(\hat{\theta}_1^{(d)}, \hat{\theta}_2^{(d)}|\boldsymbol{\lambda}) & \dots & \mathbb{C}\mathrm{ov}_M(\hat{\theta}_1^{(d)}, \hat{\theta}_K^{(d)}|\boldsymbol{\lambda}) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{C}\mathrm{ov}_M(\hat{\theta}_K^{(d)}, \hat{\theta}_1^{(d)}|\boldsymbol{\lambda}) & \mathbb{C}\mathrm{ov}_M(\hat{\theta}_K^{(d)}, \hat{\theta}_2^{(d)}|\boldsymbol{\lambda}) & \dots & \mathbb{V}\mathrm{ar}_M(\hat{\theta}_K^{(d)}|\boldsymbol{\lambda}) \end{bmatrix}.$$

Computation of the GMSE

Two steps of linear approximation

(1) The estimator $\hat{\theta}_k^{(d)} = \sum_{i=1}^{N} \gamma_i^{(d)} f_k(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}})$ is linearized at $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$:

$$\mathbb{V}\mathrm{ar}_M(\hat{\theta}_k^{(d)}|\boldsymbol{\lambda}) = \boldsymbol{\gamma}^{(d)T} \boldsymbol{F}_k \mathbb{V}\mathrm{ar}_M\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\boldsymbol{\lambda}\right) \boldsymbol{F}_k^T \boldsymbol{\gamma}^{(d)}, \qquad k = 1, \ldots, K,$$

(2) We then use the result originally proposed in Chambers and Clark (2015) to linearize $\hat{\boldsymbol{\beta}}$ around their expected value $\mathbb{E}_\Pi \mathbb{E}_M(\hat{\boldsymbol{\beta}})$:

$$\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \approx -\bar{\boldsymbol{A}}_{\boldsymbol{\beta}}^{-1} \sum_{i=1}^{N} \lambda_i \boldsymbol{g}_i(\boldsymbol{\beta}|y),$$

and we thus get:

$$\mathbb{V}\mathrm{ar}_M(\hat{\theta}_k^{(d)}|\boldsymbol{\lambda}) \approx \boldsymbol{\gamma}^{(d)T} \boldsymbol{F}_k \left(\sum_{i=1}^{N} \boldsymbol{U}_i \Sigma_{y_i} \boldsymbol{U}_i^T\right) \boldsymbol{F}_k^T \boldsymbol{\gamma}^{(d)},$$

with $\boldsymbol{U}_i = \lambda_i \bar{\boldsymbol{A}}_{\boldsymbol{\beta}}^{-1} \dot{\boldsymbol{X}}_i$ and

$$A = \frac{\partial^2 l(\beta \mid y)}{\partial \beta_{kj} \partial \beta_{k'j'}}$$

Computation of the GMSE

In summary...

The $\boldsymbol{U}_i$ terms are linear wrt the design randomness in λ_is and the expectation $\mathbb{E}_\Pi$ can be computed directly. An approximation for the GMSE is thus obtained as:

$$
\begin{aligned}
GMSE(\hat{\theta}_k^{(d)}) &= \mathbb{E}_\Pi \mathbb{V}\mathrm{ar}_M(\hat{\theta}_k^{(d)} \mid \boldsymbol{\lambda}) \\
&\approx \mathbb{E}_{\lambda \sim \Pi} \left[\boldsymbol{\gamma}^{(d)T} \boldsymbol{F}_k \left(\sum_{i=1}^{N} \boldsymbol{U}_i \Sigma_{y_i} \boldsymbol{U}_i^T \right) \boldsymbol{F}_k^T \boldsymbol{\gamma}^{(d)} \right] \\
&= \boldsymbol{\gamma}^{(d)T} \boldsymbol{F}_k \sum_{i=1}^{N} \pi_i \bar{\boldsymbol{A}}_{\boldsymbol{\beta}}^{-1} \dot{\boldsymbol{X}}_i \Sigma_{y_i} (\bar{\boldsymbol{A}}_{\boldsymbol{\beta}}^{-1} \dot{\boldsymbol{X}}_i)^T \boldsymbol{F}_k^T \boldsymbol{\gamma}^{(d)},
\end{aligned}
$$

since $\mathbb{E}(\lambda_i^2) = \pi_i$ in a simple balanced design, where $\lambda_i \sim Bern(\pi_i)$.

A Kronecker based formulation

- In order to get a more computationally efficient estimation process an alternative formulation of Gmse has been studied, based on kronecker matrix algebra
- Compared to the basic formulation which works for a single record (and then based on loop on the N units of population), the new formulation is more efficient from a computational point of view and allows us to consider blocks of matrices
- We have developed a code R generalized for the different data structures at istat

Kronecker formulation details

$$GMSE^{\text{Lin}}\left(\hat{\theta}\right) = \mathbf{\Gamma}^{T}\left(\dot{\pi}\#\dot{\mathbf{U}}\dot{\mathbf{\Sigma}}_{\mathbf{Y}}\dot{\mathbf{U}}\right)\mathbf{\Gamma}$$

$$\dot{\mathbf{U}} = -[\dot{\mathbf{X}}\#\dot{\pi}]\mathbf{\Lambda}_{\beta}^{-\mathbf{1}},$$

$$\dot{\mathbf{\Sigma}}_{\mathbf{Y}} = \mathbf{1}_{T} \otimes \mathbf{\Sigma}_{\mathbf{Y}}$$

$$\mathbf{\Sigma}_{\mathbf{Y}} = \mathbf{I}_{K;N} - \mathbf{p}\#[\mathbf{P} \otimes \mathbf{1}_{J}].$$

Simulation comparison

Population sample

- N = 100.000, 300.000. 500.000
- n=5000, 15.000, 25.000
- K=8 and J=14

Simulations

- G = 100 independent replicas of the sampling design
- M = 100 replicas of the response variable for each sample replicate

$$\widehat{\mathrm{GMSE}}^{\mathrm{MC}}\left(\hat{\theta}_k^{(d)},\theta_k^{(d)}\right)-\frac{1}{G}\sum_{g=1}^{G}\widehat{\mathrm{MSE}}_g^{\mathrm{MC}}\left(\hat{\theta}_k^{(d)},\theta_k^{(d)}\right)-\frac{1}{G}\sum_{g=1}^{G}\left(\frac{1}{M}\sum_{m=1}^{M}\left(\sum_{i=1}^{N}\gamma_i^{(d)}\hat{p}_{ik}^{(m,g)}-\sum_{i=1}^{N}\gamma_i^{(d)}p_{ik}\right)^2\right),$$

$$\widehat{\mathrm{CV}}^{\mathrm{MC}}\left(\hat{\theta}_k^{(d)},\theta_k^{(d)}\right)=\frac{\sqrt{\widehat{\mathrm{GMSE}}^{\mathrm{MC}}\left(\hat{\theta}_k^{(d)},\theta_k^{(d)}\right)}}{\mathbf{E}\left(\theta_k^{(d)}\right)},\quad k=1,\ldots,K.$$

Setup

Table 1: Register data (covariates, sample and domain membership) with their nature and marginal properties for Subgroup B of the register.

Covariates	Variable type	Marginal frequency $\boldsymbol{p}_{X_j}$
X_1 : `age`	Categorical: $\{1,\ldots,5\}$	$(0.012, 0.115, 0.208, 0.395, 0.270)$
X_2 : `gender` (`Female`)	Binary: $\{0, 1\}$	$(0.477, 0.523)$
X_3 : `citizenship` (`Italy`)	Binary: $\{0, 1\}$	$(0.071, 0.929)$
X_4 : `2011 Education` $\boldsymbol{Y}^{(2011)}$	Categorical: $\{1,\ldots,8\}$	$(0.005, 0.024, 0.171, 0.294, 0.363, 0.026, 0.113, 0.004)$
Sample membership		
λ: Sample indicator (`Yes`)	Binary: $\{0, 1\}$	$(0.950, 0.050)$
Domain membership (internal)		
$\gamma^{(d)}, d - X_2$: `gender`	Binary: $\{0, 1\}$	$(0.477, 0.523)$
Domain membership (external)		
$\gamma^{(d)}, d - X_5$: `municipality`	Categorical: $\{1,\ldots,9\}$	$(0.075, 0.065, 0.09, 0.1, 0.09, 0.12, 0.08, 0.16, 0.23)$

Setup

Table 2: Model coefficients $\beta_k, k = 1, \ldots, K - 8$ for Subgroup B based on the sample survey data. Coefficients for $k - 8$ are all set to zero since this is used as baseline category.

	Response Category k							
Model covariate (dummy)	$k-1$	$k-2$	$k-3$	$k-4$	$k-5$	$k-6$	$k-7$	$k-8$
X_0: Intercept	8.897	-0.543	6.021	8.730	8.018	0.028	0.574	0.000
X_1 :: `[29, 39]`	-1.344	6.426	-0.130	-1.586	-2.059	-1.687	-2.862	0.000
X_1 :: `[40, 49]`	-0.894	7.927	0.577	-0.852	-1.427	-1.638	-2.270	0.000
X_1 :: `[50, 69]`	0.966	8.214	2.157	-0.267	-1.068	-1.392	-1.825	0.000
X_1 :: `[70, )`	2.640	11.259	4.524	0.755	-0.085	-1.151	-0.867	0.000
X_2 :: `Female`	0.821	0.947	0.286	-0.046	0.079	0.311	0.233	0.000
X_3 :: `Italy`	-2.048	-1.250	-0.411	0.056	0.032	-0.291	0.311	0.000
X_4 :: $k-2$	0.251	3.542	3.475	2.110	2.391	8.305	1.772	0.000
X_4 :: $k-3$	-0.999	1.997	4.431	2.827	2.548	7.951	8.471	0.000
X_4 :: $k-4$	-6.377	-4.379	-2.595	-0.143	-0.492	3.409	3.863	0.000
X_4 :: $k-5$	0.016	3.170	4.512	5.732	9.772	13.889	13.409	0.000
X_4 :: $k-6$	-16.979	-14.753	-7.478	-5.416	-2.491	7.184	5.581	0.000
X_4 :: $k-7$	-10.654	-9.324	-7.827	-6.885	-4.805	2.696	5.800	0.000
X_4 :: $k-8$	-14.816	-14.945	-15.966	-10.870	-9.100	-8.896	-0.020	0.000

Final Results 1

Table 3: Estimates of totals $\hat{\theta}_k^{(d)} = \sum_{i=1}^{N} \gamma_i^{(d)} \hat{Y}_{ik}^{(t)}, k = 1, \ldots, 8$ for the full register and for domain $d \in X_2$: Gender, with their estimated GMSE. The sample fraction $n_k^{(d)}/\hat{\theta}_k^{(d)}$ is between 3.9% and 5.2% across all cases. Boostrap estimates are based on $B = 1000$ boostrap resamples (with replacement).

			Linearized Estimator		Bootstrap Estimator	
Category k	$\hat{\theta}_k^{(d)}$	Sample size $n_k^{(d)}$	$\widehat{\text{GMSE}}_k^{\text{Lin}}$	$\widehat{\text{CV}}_k^{\text{Lin}}$	$\widehat{\text{GMSE}}_k^{\text{Boot}}$	$\widehat{\text{CV}}_k^{\text{Boot}}$
Full register: $\gamma_i = 1, \quad i = 1, \ldots, N$						
1 Illiterate	1039	49	15195	11.86%	25938	15.38%
2 Literate but no education	6649	340	97462	4.70%	141041	5.65%
3 Primary	49886	2572	288343	1.08%	764254	1.75%
4 Lower secondary	84174	4285	530144	0.87%	2557366	1.90%
5 Upper secondary	113719	5682	497936	0.62%	545618	0.65%
6 Bachelor degree	7234	364	91337	4.18%	256850	7.00%
7 Master degree	32810	1524	171777	1.26%	706125	2.56%
8 PhD level	1054	44	16074	12.02%	35973	18.11%
Internal domain: $\gamma^{(d)}, d =$ Male (47.7%)						
1 Illiterate	300	14	4435	22.22%	12137	36.49%
2 Literate but no education	1569	81	24444	9.97%	28522	10.74%
3 Primary	19631	1015	114420	1.72%	300830	2.79%
4 Lower secondary	45853	2306	261562	1.12%	1142715	2.33%
5 Upper secondary	56374	2775	243509	0.88%	209053	0.81%
6 Bachelor degree	2701	132	36359	7.06%	153317	14.50%
7 Master degree	14443	656	76429	1.91%	353488	4.12%
8 PhD level	510	20	7813	17.33%	14960	24.13%
Internal domain: $\gamma^{(d)}, d =$ Female (52.3%)						
1 Illiterate	739	35	10721	14.01%	12041	14.73%
2 Literate but no education	5080	259	72826	5.31%	167424	8.06%
3 Primary	30255	1557	172843	1.37%	219218	1.55%
4 Lower secondary	38321	1979	265811	1.35%	461937	1.77%
5 Upper secondary	57345	2907	251083	0.87%	309445	0.97%
6 Bachelor degree	4533	232	54303	5.14%	53537	5.10%
7 Master degree	18367	868	93330	1.66%	113597	1.84%
8 PhD level	545	24	8069	16.5%	9646	18.18%

Final Results 2

Table 4: Estimates of the CV with respect to the register totals estimates $\hat{\theta}_k^{(d)} = \sum_{i=1}^{N} \gamma_i^{(d)} \hat{Y}_{ik}^{(t)}, k = 1, \ldots, 8$, with $\gamma_i = 1, i = 1, \ldots, N$ with the three different estimators, and for $N \in \{100.000, 300.000, 500.000\}$.

Category k	$\hat{\theta}_k^{(d)}$	Sample size $n_k^{(d)}$	$\widehat{\mathrm{CV}}_k^{\mathrm{Lin}}$	$\widehat{\mathrm{CV}}_k^{\mathrm{Boot}}$	$\widehat{\mathrm{CV}}_k^{\mathrm{MC}}$
$N = 100,000; \quad n = 5,000$					
1 Illiterate	308	15	19.73%	25.77%	19.95%
2 Literate but no education	1886	81	9.64%	13.20%	9.47%
3 Primary	14537	724	2.42%	3.18%	2.39%
4 Lower secondary	30854	1517	1.55%	1.96%	1.53%
5 Upper secondary	38667	1929	1.10%	1.36%	1.08%
6 Bachelor degree	2183	107	7.69%	9.78%	7.50%
7 Master degree	11247	609	1.92%	2.59%	1.87%
8 PhD level	317	12	15.52%	21.43%	15.53%
$N = 300,000; \quad n = 15,000$					
1 Illiterate	949	58	14.21%	23.34%	13.90%
2 Literate but no education	5613	292	5.48%	5.94%	5.22%
3 Primary	41411	2216	1.43%	2.10%	1.39%
4 Lower secondary	92610	4626	0.87%	1.18%	0.84%
5 Upper secondary	116326	5892	0.63%	0.73%	0.61%
6 Bachelor degree	6464	296	4.37%	5.18%	4.40%
7 Master degree	33711	1704	1.15%	1.3%	1.14%
8 PhD level	915	42	11.41%	17.47%	11.13%
$N = 500,000; \quad n = 25,000$					
1 Illiterate	1581	80	10.86%	14.92%	9.91%
2 Literate but no education	9375	469	4.22%	4.51%	4.10%
3 Primary	72608	3656	1.10%	1.30%	1.08%
4 Lower secondary	154465	7763	0.68%	0.76%	0.66%
5 Upper secondary	193322	9680	0.49%	0.52%	0.49%
6 Bachelor degree	10936	552	3.38%	4.01%	3.35%
7 Master degree	56129	2880	0.89%	1.16%	0.87%
8 PhD level	1583	79	8.77%	10.48%	8.38%

Final Results 3

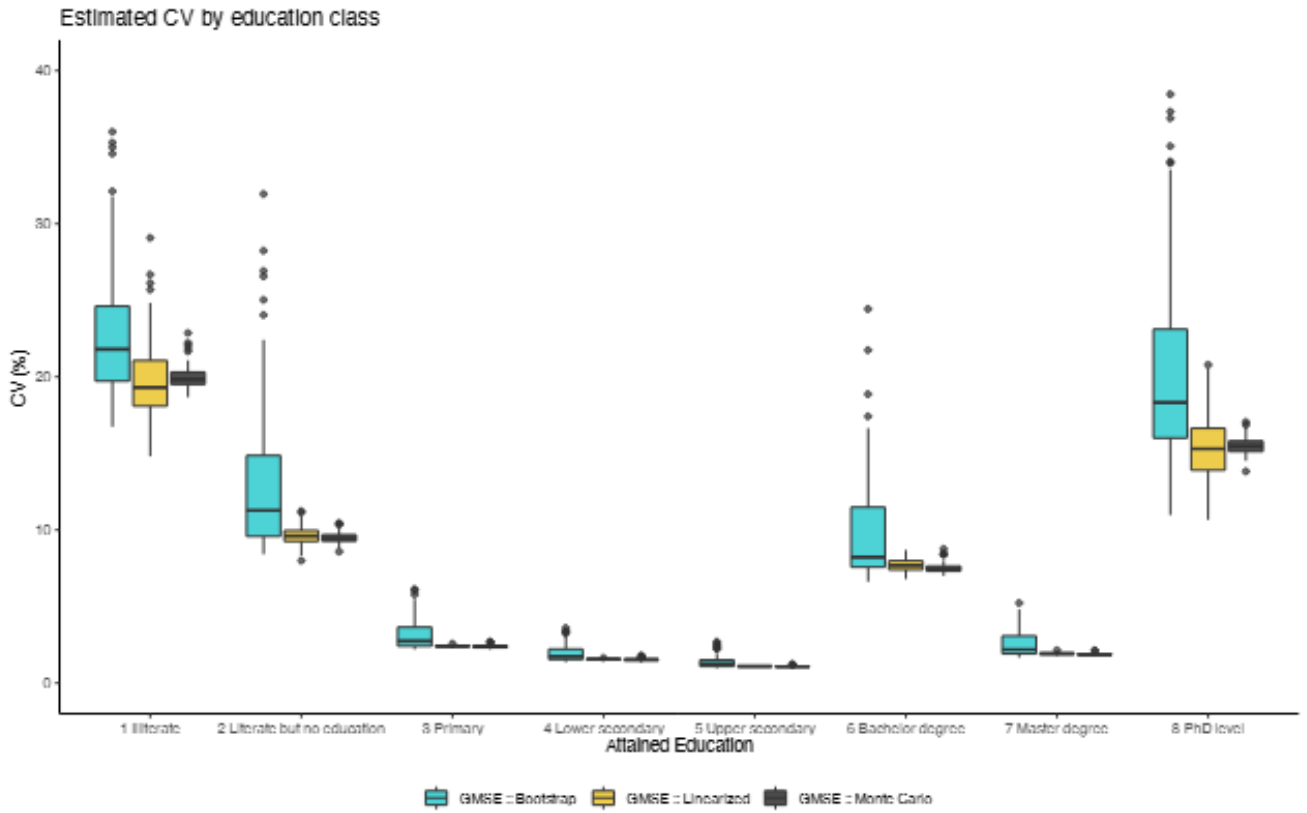


Figure 3: Estimates of the CV with the three different estimators, and for $N = 100.000$, with $n = 0.05 \times N = 5000$. All results are based on $S = 100$ replications of the three evaluation procedures.

Conclusions

In this work, motivated by ISTAT's programme, we took our first steps to evaluate the feasibility / validity of a global measure of accuracy: GMSE.

Key advantages:

(i) Computationally and memory efficient

(ii) Allows on-the-fly estimation

Ongoing and Future Work:

(i) Extend and evaluate the GMSE to other structures of data / models.

(ii) Consider other methodological developments, such as: the case of latent class models, the Bayesian framework, to incorporate the additional uncertainty arising from the a priori distribution of model parameters used for prediction.

(iii) Provide implementation support.

References 1

References

Alleva, G. (2017a). Emerging challenges in official statistics: new sources, methods and skills. Technical report, SIS2017 Statistical Conference. Statistics and Data Science: new challenges, new generations.

Alleva, G. (2017b). The new role of sample surveys in official statistics. Technical report, ITACOSM 2017: The 5th Italian Conference on Survey Methodology.

Alleva, G., Falorsi, P. D., Petrarca, F., and Righi, P. (2021). Measuring the accuracy of aggregates computed from a statistical register. *Journal of Official Statistics*, 37(2):481–503.

Ascari, G., Blix, K., Brancato, G., Burg, T., McCourt, A., van Delden, A., Krapavickaitė, D., Ploug, N., Soltus, S., Stoltze, P., et al. (2020). Quality of multisource statistics–the KOMUSO project. *The survey statistician*, 81:36–51.

Biemer, P., Trewin, D., Bergdahl, H., and Japec, L. (2014). A system for managing the quality of official statistics. *Journal of Official Statistics*, 30(3):381–415.

Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, 74(5):817–848.

References 2

Bruengger, H. (2008). How should a modern national system of official statistics look? *UNECE, Statistical Division*.

Bycroft C, M.-D. N. (2020). Use of administrative records for non-response in the new zealand 2018 census. *Statistical Journal of the IAOS*, 36(1):107–115.

Chambers, R. L. and Clark, R. (2012). *An introduction to model-based survey sampling with applications*. Oxford University Press.

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40(2):137–162.

Daalmans, J. (2017). *Mass imputation for census estimation*. Statistics Netherlands.

Daas, P., Ossen, S., Tennekes, M., Zhang, L.-C., Hendriks, C., Foldal Haugen, K., Cerroni, F., Di Bella, G., Laitila, T., Wallgren, A., et al. (2011). Report on methods preferred for the quality indicators of administrative data sources. Technical report, Blue - ETS Project, Deliverable 4.2.

Daas, P. J., Puts, M. J., Buelens, B., and Hurk, P. A. v. d. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, 31(2):249–262.

De Broe, S., Struijs, P., Daas, P., van Delden, A., Burger, J., van den Brakel, J., ten Bosch, O., Zeelenberg, K., and Ypma, W. (2021). Updating the paradigm of official statistics: New quality criteria for integrating new

References 3

De Waal, T., van Delden, A., and Scholtus, S. (2020). Multi-source statistics: basic situations and methods. *International Statistical Review*, 88(1):203–228.

Di Zio, M., Filippini, R., and Rocchetti, G. (2019a). An imputation procedure for the italian attained level of education in the register of individuals based on administrative and survey data. *Measuring well-being at local level using remote sensing and official statistics data*, page 143.

Di Zio, M., Filippini, R., and Rocchetti, G. (2019b). An imputation procedure for the italian attained level of education in the register of individuals based on administrative and survey data. *Rivista di Statistica Ufficiale*, Issue 2-3(2019):143–174.

Eurostat (2014). ESS handbook for quality reports. Technical report, European Statistical System.

Eurostat (2020). European Statistical System handbook for quality and metadata reports. Technical report, Luxembourg: Publications Office of the European Union.

Fabrizi, E., Montanari, G. E., and Giovanna Ranalli, M. (2016). A hierarchical latent class model for predicting disability small area counts from survey data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 179(1):103–131.

Graf, E. and Tillé, Y. (2014). Variance estimation using linearization for poverty and social exclusion indicators. *Survey Methodology*, 40(1):61–80.

Groves, R. M. (2011). Three eras of survey research. *Public opinion quarterly*, 75(5):861–871.

References 4

Holt, D. T. (2007). The official statistics olympic challenge: Wider, deeper, quicker, better, cheaper. *The American Statistician*, 61(1):1–8.

Isaki, C. T. and Fuller, W. A. (1982). Survey Design under the Regression Superpopulation Model. *Journal of the American Statistical Association*, 77(377):89–96.

Istat (2016). ISTAT'S MODERNISATION PROGRAMME. Technical report, ISTAT: Istituto Nazionale di Statistica.

Kim, J. K. and Rao, J. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96(4):917–932.

Kuhn, T. S. (1997). *The structure of scientific revolutions*, volume 962. University of Chicago press Chicago.

Lohr, S. L. (2021). *Sampling: design and analysis*. Chapman and Hall/CRC.

Lohr, S. L. and Raghunathan, T. E. (2017). Combining Survey Data with Other Data Sources. *Statistical Science*, 32(2):293–312.

López-Vizcaíno, E., Lombardía, M. J., and Morales, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 178(3):535–565.

References 5

Lundy, E. R. (2022). Predicting the quality and evaluating the use of administrative data for the 2021 canadian census of population. *Statistical Journal of the IAOS*, 38(4):1177–1183.

Marino, M. F., Ranalli, M. G., Salvati, N., and Alfo, M. (2019). Semiparametric empirical best prediction for small area estimation of unemployment indicators. *The Annals of Applied Statistics*, 13(2):1166–1197.

Mashreghi, Z., Haziza, D., and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10:1–52.

Molina, I., Saei, A., and José Lombardía, M. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 170(4):975–1000.

Nedyalkova, D. and Tille, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95(3):521–537.

Radermacher, W. J. (2018). Official statistics in the era of big data opportunities and threats. *International Journal of Data Science and Analytics*, 6(3):225–231.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

References 6

Scholtus, S. and Daalmans, J. (2021). Variance estimation after mass imputation based on combined administrative and survey data. *Journal of Official Statistics*, 37(2):433–459.

Shao, J. (2003). Impact of the bootstrap on sample surveys. *Statistical Science*, 18(2):191–198.

Vallée, A.-A. and Tillé, Y. (2019). Linearisation for variance estimation by means of sampling indicators: Application to non-response. *International Statistical Review*, 87(2):347–367.

Valliant, R. (2009). Model-based prediction of finite population totals. In *Handbook of Statistics*, volume 29, pages 11–31. Elsevier.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Wallgren, A. and Wallgren, B. (2014). *Register-based Statistics: statistical methods for administrative data*. John Wiley & Sons.

Wolter, K. M. (1986). Some Coverage Error Models for Census Data. *Journal of the American Statistical Association*, 81(394):338.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66(1):41–63.

Ziegler, A. (2011). *Generalized estimating equations*, volume 204. Springer Science & Business Media.

Final

This work is the result of the collaboration of many colleagues

Thanks to all of them

Thank you for the attention!