Estimates of accuracy of statistical registers based on Anticipated Variance

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

ITSEW 2019 –International Total Survey Error Workshop

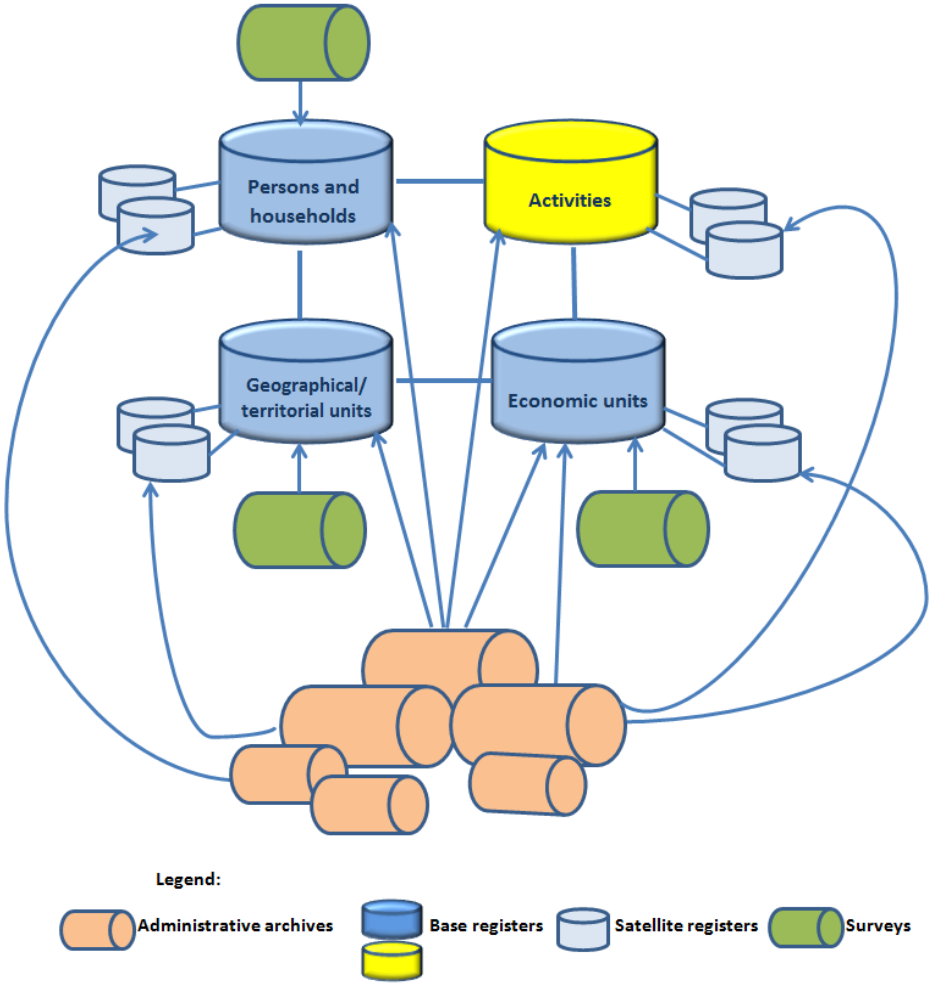
Bergamo, 12 June 2019

*** Istat , ** Università La Sapienza Roma**

Introduction

- Since 2015 ISTAT has been engaged in a deep process of Modernization (Istat, 2016) based on the Vision 2020 (Eurostat, 2015).
- One of the pillars of this process is to produce official statistics by a massive integration of data (Citro, 2016).
- This has been achieved by developing the **Integrated System of Statistical Registers** (**ISSR**) which is a micro data based system built through massive integration of administrative and survey data (Alleva, 2017).

The Italian Integrated System of Statistical Registers (ISSR)



The register values are subject to statistical uncertainty with respect to both units and variables.

The availability of a register enables different users to produce estimates for different domains by summing up the domain values in the register.

Some of these estimates could be highly inaccurate.

Piero Demetrio Falorsi , Giorgio Alleva, Francesca Petrarca, Paolo Righi

Structure of registers (R)

REGISTER				
Unit code in R	True but UNKNOWN Value	PREDICTED Value	Auxiliary Variable With no uncertainty	γ_d Domain membership variable (0,1)
1	y_1	$\hat{y}_1$	$\boldsymbol{x}_1$	1
⋮	⋮	⋮	⋮	⋮
k	y_k	$\hat{y}_k$	$\boldsymbol{x}_k$	0
⋮	⋮	⋮	⋮	⋮
N	y_N	$\hat{y}_N$	$\boldsymbol{x}_N$	1

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

Piero Demetrio Falorsi , Giorgio Alleva, Francesca Petrarca, Paolo Righi

Formalization of the problem

- *R* is a statistical register built at micro-level for the target population U.
- U_d is a statistical domain of interest which is a subset of U.
- R_d is a subset *of R* which represents the target domain U_d.

- The **target parameter** is the total of the variable y in the domain U_d

$$Y_{U_d} = \sum_{k \in U_d} y_k$$

- Let $\hat{y}_k$ be the value in the register that predicts the value y_k of the unit *k*.

- For estimating the target parameter Y_{U_d} the users can simply sum the predicted $\hat{y}_k$ values over R_d.

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

- $\hat{Y}_{R_d}$ is a statistical **estimate** from R (**it is not a true value**).

Piero Demetrio Falorsi , Giorgio Alleva, Francesca Petrarca, Paolo Righi

The measure of accuracy

Different users could compute their own estimates without a previous check of the NSI: but some estimates could be inaccurate.

The main idea here is to define a measure of accuracy that explicitly takes into account all the sources of variability.

Proposed: the **Anticipated Variance (AV)** which has been introduced in literature to deal with different kinds of inference problems (Isaki and Fuller, 1982; Sarndäl *et al*., 1992; Nedyalkova and Tillé, 2008; Nirel, and Glickman, 2009; Falorsi and Righi, 2015):

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

E_M is the Model Expectation and E_P is the Sample Expectation

The measure of accuracy: Anticipated Variance

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

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

$E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] - V_M\left(Y_{U_d}\right)$, if $E_M(\hat{t}) = \vartheta$ and $E_P\left(\hat{Y}_{R_d}|y\right) = Y_{U_d}$

(planned domains*)

$\boldsymbol{\lambda}$ vector of sample membership variables

Sample Design Unbiased with respect to Y_{U_d}

Model Unbiased

ϑ vector of unknown parameters

$E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] + V_M\left(Y_{U_d}\right) - 2Cov_M\left(Y^0_{R_d}, Y_{U_d}\right)$, if $E_M(\hat{t}) = \vartheta$ but $E_P\left(\hat{Y}_{R_d}|y\right) \neq Y_{U_d}$

(un-planned domains)

$$Y^0_{R_d} = E_P\left(\hat{Y}_{R_d}|y\right) \neq Y_{U_d}$$

→ $\boldsymbol{E_P\left[V_M\left(\hat{Y}_{R_d}|\lambda\right)\right]}$ **is the dominant term of the *AV***

(*planned domains: domain membership variables in **the x variable**)

Piero Demetrio Falorsi , Giorgio Alleva, Francesca Petrarca, Paolo Righi

The proposed measure of accuracy

- The AV neutralizes variability due to the pure model variability $V_M\left(Y_{U_d}\right)$ of the **finite population value** Y_{U_d}.

- The *AV* is relatively stable over the time because it does not depend on the specific domain sample size in a given selection.

- It is based only on the first and second moments of the statistical distributions.

- It allows us to easily take into account the different sources of variability resulting from various approaches to inference.

Linearization of $E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right]$ of *AV*

- The calculus of $E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right]$ can be performed by a step by step linearization:

1. first the functions $\hat{y}_k = f(\boldsymbol{x}_k; \hat{\boldsymbol{t}})$ are linearized w.r.t. $\hat{\boldsymbol{t}}$, where $\hat{\boldsymbol{t}}$ represents the estimate of $\boldsymbol{\vartheta}$;

2. then, the estimating equations of $\hat{\boldsymbol{t}}$ are linearized w.r.t. y variables, keeping the λ variables as fixed;

3. finally, the **linearized** vectors of the second step are linearized w.r.t. λ values.

- The final expression is suitable for the computational implementation using **unit level elements** and **plug-in estimate** of the unknown terms.

Linearizing $E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right]$

1. Step: Using the linearized expression for $V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)$

$E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] \cong \boldsymbol{\gamma}_d' \mathbf{F}\{E_P[V_M(\hat{\boldsymbol{t}}|\boldsymbol{\lambda})]\}\mathbf{F}'\boldsymbol{\gamma}_d$ where **F** is the matrix of first derivatives:

$\mathbf{F} = \left\{ f_{ki} = \frac{\partial f(\mathbf{x}_k;\hat{\boldsymbol{t}})}{\partial \vartheta_i}\Big|_{\hat{t}_i=\vartheta_i} : k = 1,\dots,N;\ i = 1,\dots,I \right\}$. For classical simple linear model, $\mathbf{F}=\boldsymbol{X}$.

2. Step:

$V_M(\hat{\boldsymbol{t}}|\boldsymbol{\lambda}) \cong V_M\left(\sum_j \mathbf{u}_{j,y|\lambda}\, y_j|\boldsymbol{\lambda}\right) = \sum_{j\in R}\left[\mathbf{u}_{j,y|\lambda}\mathbf{u}'_{j,y|\lambda}\sigma^2_{yj} + \sum_{\ell\neq j}\mathbf{u}_{j,y|\lambda}\mathbf{u}'_{\ell,y|\lambda}\,\sigma_{yj\ell}\right]$.

For classical simple linear model, the second term is zero because $\boldsymbol{\Sigma}_y = \sigma^2\mathbf{I}$.

3. Step:

Combining 1st+2nd step and linearizing $\mathbf{u}_{j,y|\lambda} \cong \mathbf{u}_{j,y|\pi} + \partial\mathbf{u}_{j,y|\lambda}\left(\lambda_j - \pi_j\right)$ we get $E_P[V_M(\hat{\boldsymbol{t}}|\boldsymbol{\lambda})]$

The vector $\hat{\boldsymbol{t}}$ is obtained as explicit solution of the system of estimating equations:

$\left(\sum_{j\in R}\mathbf{x}_j\,\mathbf{x}_j'\frac{\lambda_j}{\pi_j}\right)^{-1}\sum_{j\in R}\mathbf{x}_j\,y_j\,\frac{\lambda_j}{\pi_j} - \hat{\boldsymbol{t}} = \mathbf{0}_I$. The vectors $\mathbf{u}_{j,y|\lambda}$, $\mathbf{u}_{j,y|\pi}$ and $\partial\mathbf{u}_{j,y|\lambda}$ are given by:

$$\mathbf{u}_{j,y|\lambda} = \left(\sum_{j\in R}\mathbf{x}_j\,\mathbf{x}_j'\frac{\lambda_j}{\pi_j}\right)^{-1}\mathbf{x}_j\frac{\lambda_j}{\pi_j};\ \mathbf{u}_{j,y|\pi} = \left(\sum_{j\in R}\mathbf{x}_j\,\mathbf{x}_j'\right)^{-1}\mathbf{x}_j;$$

$$\partial\mathbf{u}_{j,\lambda|y} = \left(\sum_{j\in R}\mathbf{x}_j\,\mathbf{x}_j'\right)^{-1}\mathbf{x}_j\frac{1}{\pi_j} - \left(\sum_{j\in R}\mathbf{x}_j\,\mathbf{x}_j'\right)^{-1}\mathbf{x}_j\mathbf{x}_j'\frac{1}{\pi_j}\left(\sum_{j\in R}\mathbf{x}_j\,\mathbf{x}_j'\right)^{-1}\mathbf{x}_j.$$

Classical simple linear model

Piero Demetrio Falorsi , Giorgio Alleva, Francesca Petrarca, Paolo Righi

Experimental study

- The experimental study, based on real data, compares the empirical *AV* of a Monte Carlo simulation with the approximate *AV* obtained by a Taylor linearization.
- We consider the projection estimator (Kim and Rao, 2012) assisted by the linear regression model.
- The data set for the empirical study is an administrative dataset that contains information on the population of 21,782 Sapienza University of Rome alumni who graduated between March 1st, 2008 and February 28th, 2009 who signed a job contract in the subordinate or para-subordinate labor markets in the three years following graduation, (Alleva and Petrarca, 2013; Petrarca 2014 a,b)
- We considered only the students who graduated in the disciplinary sectors of *engineering*, *sciences*, *literature*, *economics & statistics*, *psychology*, *chemistry & pharmacy*, and *architecture*. This choice reduced the dataset to 7,085 units.

Experimental study: A simplified statistical framework (1)

1. **No Coverage problems** ➡ *R=U*

2. **Sources of variability**
   - ➢ Model generating the data
   - ➢ Sampling for observing the y *values*

➢ The **target variable** of interest y is the numbers of days worked in the three years after graduation.

Experimental study: A simplified statistical framework (2)

3. Superpopulation model

- We generated 1,000 populations of 7,085 units.
- For each population a vector of the target variables y_k was generated by the sum of two components:

$$y_k = E_M(y_k) + e_k = f(\mathbf{x}_k; \boldsymbol{\vartheta}) + e_k = \tilde{y}_k + e_k$$

$\mathbf{x}_k = (x_{k1}, \cdots, x_{ki}, \cdots, x_{kI})'$ is the vector of I auxiliary variables and

$\boldsymbol{\vartheta} = (\vartheta_1, \cdots, \vartheta_i, \cdots, \vartheta_I)'$ is the vector of I unknown parameters.

- $\tilde{y}_k$ is the vector of the fitted values obtained by a linear regression model tuned on the superpopulation
- e_k is generated by a normal distribution with mean 0 and variance equal to variance of the y_k in the real data set multiplied by a factor 9 ($\sigma^2 = 0.1159733 \cdot 10^7$)

E_M , V_M Model Expectation and Variance

Piero Demetrio Falorsi , Giorgio Alleva, Francesca Petrarca, Paolo Righi

Experimental study: A simplified statistical framework (3)

8 auxiliary variables

- $x_{k1} = 1$;
- x_{k2} dummy variable: gender of a graduate;
- x_{k3} metric variable: age at the time of graduation;
- x_{k4} dummy variable: graduate on time;
- x_{k5} dummy variable: graduate from a second cycle degree;
- x_{k6} metric variable: number of days that a graduate has waited before obtaining a permanent contract;
- x_{k7} metric variable: number of days that a graduate has waited before obtaining a contract with a highly qualified position;
- x_{k8} metric variable: number of days that a graduate has waited before obtaining a contract with actual duration more than or equal to 8 months.

The International Standard Classification of Occupations (ISCO) scale:
https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/@publ/documents/publication/wcms_172572.pdf
The choice of this threshold (at least 8 months) comes from the Italian legislation.

Experimental study: A simplified statistical framework (4)

4. SAMPLING

- For each population, 1,000 samples (S) of n=500 units were selected by a simple random sample design without replacement with inclusion probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k, \ldots, \pi_N)'$.

- $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_k, \ldots, \lambda_N)'$ the *N* column vector of sample membership variables
- $E_P(\boldsymbol{\lambda}) = \boldsymbol{\pi}$
- $$V_P(\boldsymbol{\lambda}\boldsymbol{\lambda}') = \boldsymbol{\Sigma}_\lambda = \begin{bmatrix} \pi_1(1-\pi_1) & & \pi_{1N} \\ \pi_{1k} & \pi_k(1-\pi_k) & \pi_{kN} \\ \pi_{1N} & & \pi_N(1-\pi_N) \end{bmatrix}$$

E_P , V_P Sample Expectation and Variance

Piero Demetrio Falorsi , Giorgio Alleva, Francesca Petrarca, Paolo Righi

Experimental study: A simplified statistical framework (5)

5. PREDICTIONS

- For each sample we get by a simple linear regression model the estimated regression coefficients $\hat{\boldsymbol{t}}$ and

$$\hat{y}_k = f(\boldsymbol{x}_k; \hat{\boldsymbol{t}}) = \boldsymbol{X} \cdot \hat{\boldsymbol{t}}$$

$\hat{\boldsymbol{t}} = (\hat{t}_1, \dots, \hat{t}_i, \dots, \hat{t}_I)'$ represents the estimate of $\boldsymbol{\vartheta}$ based on the observation of the values y_k on the sample S while the values $\mathbf{x}_k$ are available for all the units of R.

- The sum of $\hat{y}_k$ restricted to the domain builds $\hat{Y}_{R_d}$

$$\hat{Y}_{R_d} = \sum_{k \in R_d} \hat{y}_k = \sum_{k \in R_d} f(\mathbf{x}_k; \hat{\boldsymbol{t}})$$

- Through Monte Carlo simulation we calculated all the components needed to estimate AV.
- The evaluation of the Taylor linear approximation in our case requires only to know the matrix of the auxiliary variables, the vectors of sample membership indicators $\boldsymbol{\lambda}$ for each population and for each sample and the vector of the domain membership indicators $\boldsymbol{\gamma_d}$.

Piero Demetrio Falorsi , Giorgio Alleva, Francesca Petrarca, Paolo Righi

Experimental study: Computation of Anticipated Variance

Planned domain: *Gender Female (4,281 units)*

Empirical

The correction w.r.t EpVm is 6%

$$AV\left(\hat{Y}_{R_d}\right) = E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] - V_M\left(Y_{U_d}\right) = \mathbf{7{,}069.9} - \mathbf{426.6} = \mathbf{6{,}649.7}$$

4‰

Linearizated

$$AV\left(\hat{Y}_{R_d}\right) = E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] - V_M\left(Y_{U_d}\right) = \mathbf{7{,}021.6} - \mathbf{496.5} = \mathbf{6{,}623.5}$$

Un-planned domain: *Scientific group (3,368 units)*

Empirical

$$AV\left(\hat{Y}_{R_d}\right) = E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] + V_M\left(Y_{U_d}\right) - 2Cov_M\left(Y^0_{R_d}, Y_{U_d}\right)$$

$$= \mathbf{3{,}045.8} + \mathbf{380.2} - \mathbf{429.5} = \mathbf{2{,}999.4}$$

The correction w.r.t EpVm is 1.6%

2%

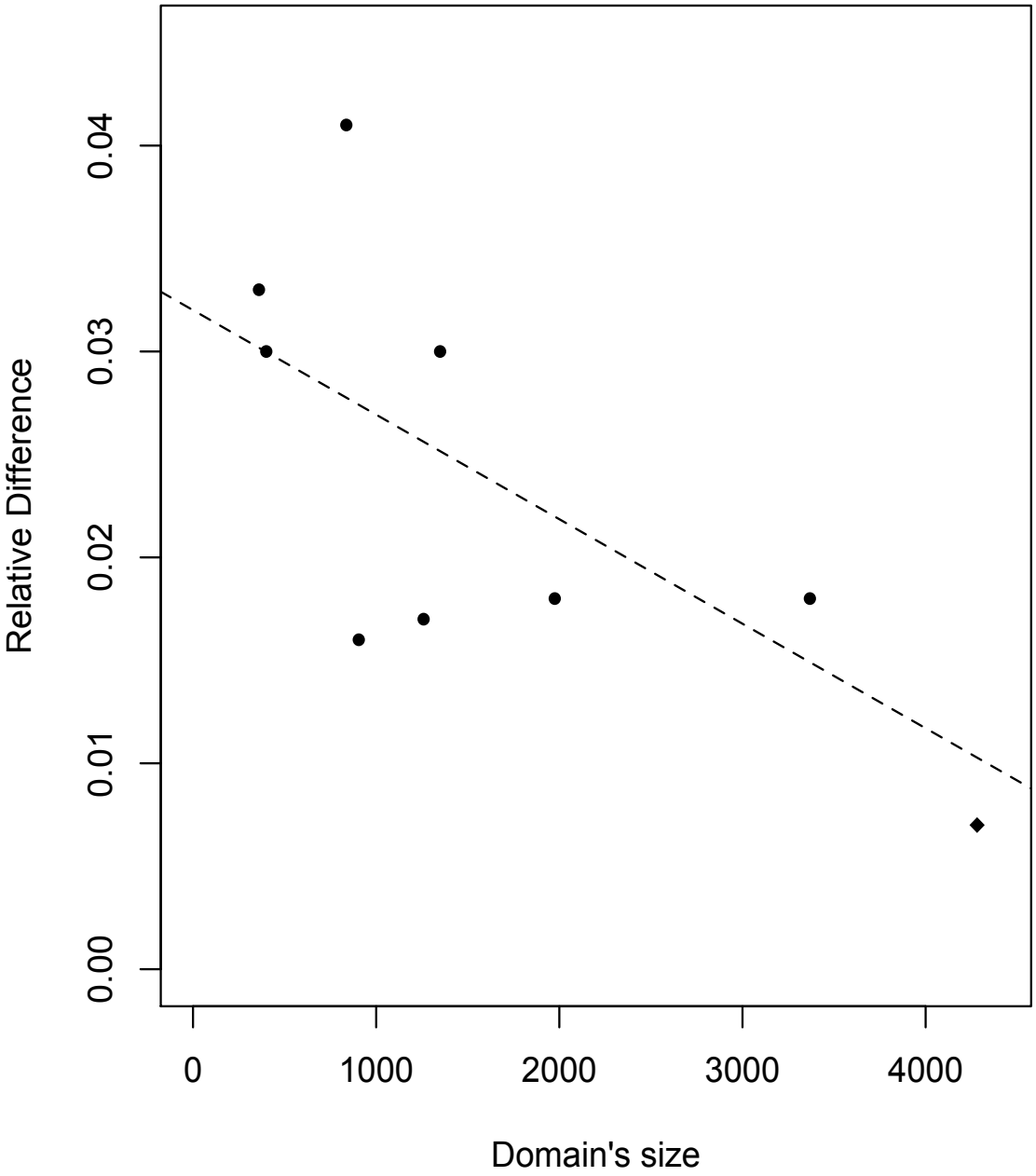
Linearizated

$$AV\left(\hat{Y}_{R_d}\right) = E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] + V_M\left(Y_{U_d}\right) - 2Cov_M\left(Y^0_{R_d}, Y_{U_d}\right) =$$

$$= \mathbf{2{,}990.3} + \mathbf{390.6} - \mathbf{429.5} = \mathbf{2{,}951.4}$$

The number are scaled by a factor of 10^7

Piero Demetrio Falorsi , Giorgio Alleva, Francesca Petrarca, Paolo Righi



Experimental result:

Relative difference of $E_P\left[V_M\left(\widehat{Y}_{R_d}|\lambda\right)\right]$ between simulation and linearization by un-planned domain's size

The un-planned domain is the disciplinary sectors: *sciences, chemistry & pharmacy*, *psychology*, *architecture, engineering*, *economics & statistics*, *literature*, *and scientific_group.*

Fig. shows that the relative difference decreases when the domain sample size increases. These findings confirm that the linearization method produces a downward approximation related to the sample size.

The differences between the two estimates are positive and small, ranging between 16 and 41‰.

Summary

- We have proposed the Anticipated Variance as a suitable measure for evaluating the accuracy of aggregates computed from register.
- We have suggested a strategy for computing the leading components of the *AV* which is based on the linearization of the estimators as a function of random elements y and $\boldsymbol{\lambda}$.
- From the experimental results we see that the method of approximation can be considered as a valid computational strategy. This is based on computation at unit level for each unit in the register.
- The *AV* is simple to use and to communicate to users.
- It explicitly considers the main sources of variability and it may be accepted as a precision measurement.
- It is relatively stable over time.
- Important further steps in the approach to research presented here are those of evaluating the strengths and robustness of the results with further simulation studies.

Main References

Alleva, G. and Petrarca, F. (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,14 giugno 2017, Bologna, https://www.istat.it/it/files//2015/10/Alleva_ITACOSM_14062017.pdf.

Binder D. A. and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. Journal of the American Statistical Association 89, 1035_1043.

Chambers R.L. and Clark R.G. (2015). An Introduction to Model-Based Sampling with Applications. Oxford Statistical Science. 37.

Chen S., and Haziza D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. Biometrika, 102, 439-453.

Citro C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. Survey Methodology, Statistics Canada.

Falorsi P.D. and Righi P. (2015). Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys. Survey Methodology, 41, 215-236.

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

Graf M. (2015). A Simplified Approach to Linearization Variance for Surveys. Technical Report, Institut De Statistique, Université de Neuchâtel.

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

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

Istat, (2016). Istat's Modernisation Programme,https://www.istat.it/en/files/2011/04/IstatsModernistionProgramme_EN.pdf.

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

Lavallée, P., Caron, P. (2001). Estimation Using the Generalised Weight Share Method: The Case of Record Linkage. Survey Methodology, 27, 155-169.

Nedyalkova, D and Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. Biometrika, 95, 521-537.

Nirel, R. and Glickman, H. (2009). Chapter 21 - Sample Surveys and Censuses. In: Rao, C.R. (ed.) Handbook of Statistic, 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 (pp. 159–170). Berlin: Springer. doi: 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. Ph.D. Thesis, Dipartimento di Economia dell'Università degli studi Roma Tre. http://hdl. handle.net/2307/4167.

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

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

Wolter, K. M. (1986). Some Coverage Error Models for Census Data. Journal of the American Statistical Association, 81, 338 - 346.

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

Ziegler, A. (2015). Generalized Estimating Equations, Lecture Notes in statistics. Springer and Verlag.


Thank
You

Alternatives to reflect on

1. **Maximize or minimize** the information from Integrated System of Statistical Registers?

Maximize the Information	▪ **NSI more relevant.** ▪ **More flexible system for the users: allowing different users to produce their own statistics from the ISSR.** ▪ **Need for reducing risk of an erroneous use of the data.**
Minimize the information	Make the use of ISSR more limited and allow the dissemination of only planned outputs (for variables, domains, parameters) having a certified level of accuracy.

2. Should **users be informed** of the accuracy of the estimates or should they **ignore** it?

Making users aware	• **Computationally complex.** • **Reduction of risk of an erroneous use of inferences.** • **Positive for trust and transparency.**
Ignore the problem	▪ High risk of an erroneous use of inferences. ▪ Risk of over-protecting the privacy of data.

- ✓ Its accuracy depends both on the statistical process (including linkage) for the construction of predictions and on the coverage errors of *R.*
- ✓ Different users could compute their own estimates without a previous check of the NSI: but some estimates could be **inaccurate**.
- ✓ **Risk** of an **erroneous use** of inferences for policies and other decisional processes for different **types** of users
- (statistically educated/or not – internal/external – Other,….).

The proposed measure of accuracy

- **Conditional vs unconditional**
- The statisticians who base the inference only on the model assumptions would argue against the use of an unconditional measure of the accuracy, such as the *AV.*
- From a *pure* inferential point of view, we agree with them, but we also observe that :

(i) the conditional model variance $V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)$ is a part of the calculus and its value may be disseminated.

(ii) The *AV* is more stable in time than the conditional model variance $V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)$. This is because the latter measure depends on the specific domain sample size in a given selection.

(iii) This fosters the feasibility for the NSIs of the computational strategy proposed.

Global Variance

- In order to consider explicitly the variability deriving from both the sampling design and the model *M*, Wolter (1986) proposed the concept of the *Global Variance* (GV):

$$GV\left(\hat{Y}_{R_d}\right) = E_P E_M\left(\hat{Y}_{R_d} - \tilde{Y}_{R_d}\right)^2$$

- Under the hypothesis that $E_P E_M\left(\hat{Y}_{R_d}\right) = \tilde{Y}_{R_d}$, the *GV* can be decomposed being conditioned by the realized value of the vector $\boldsymbol{\lambda}$ (Kendall and Stuart, 1976, p.196):

$$GV\left(\hat{Y}_{R_d}\right) = E_P\left[V_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right] + V_P\left[E_M\left(\hat{Y}_{R_d}|\boldsymbol{\lambda}\right)\right].$$

- For the *non-informative sampling designs*, we may apply the operators E_M and E_P in the reverse order and define the *GV* being conditioned by the realized value of the vector $\boldsymbol{y}$:

$$GV\left(\hat{Y}_{R_d}\right) = E_M\left[V_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right)\right] + V_M\left[E_P\left(\hat{Y}_{R_d}|\boldsymbol{y}\right)\right].$$