# MEASURING THE ACCURACY OF AGGREGATES FROM A STATISTICAL REGISTER

*Authors: Piero Demetrio Falorsi, Francesca Petrarca, Paolo Righi,*

*TO BE MODIFIED DO NOT TO QUOTE*

## 1. INTRODUCTION

The Italian National Statistical Institute (Istat) is currently engaged in a modernization programme which includes a significant revision of the methods it has been using for the production of statistics. The principal concept underlying this important change is the use of a system of integrated statistical registers as a basis for the all the production system. This new system will be referred to as the Italian Integrated System of Statistical Registers (ISSR) in the following discussion. ISSR called for a big initial investment in both architecture, methodology and professional competences and continues to require ongoing work.

The ISSR consists of two main elements : the *Base registers* containing the unique identifier of the statistical units and a limited set of some registry "core" variables, *highly identifiable* and *stable in time*, derived from administrative sources, and the *Satellite registers,* containing thematic variables derived from administrative sources or statistical surveys.

The ISSR has been created by a massive integration of administrative archives and survey data. At microdata level, different statistical techniques have been adopted, e.g.: record linkage, statistical matching, projection estimators, model predictions for single units, Hidden Markov Models, etc. These techniques result in defining predictions at the *unit level.* As a consequence, we have an increase in the amount of available information as compared to each source when it is considered individually.

Summarising, the register values are the output of statistical processes subject to statistical uncertainty with respect to both units and variables. We note that the availability of a register (which is a statistical object composed of microdata values) enables different stakeholders to produce estimates for different domains by summing up the domain values in the register. It be could that some of these estimates are highly inaccurate. Therefore, it would be useful not only to make the different stakeholders aware of their level of accuracy but also to adopt a strategy allowing them to compute the accuracy of their estimates autonomously. This a crucial point for NSIs to maintain trust with users, in a responsible and transparent way.

In this paper we deal with this problem. The inferential framework for estimating aggregates from registers is described *Section 2*. In *Section 3* we propose the *Anticipated Variance* (*AV*) (Isaki and Fuller, 1982) as statistical quantity suitable for measuring the accuracy in the context of the production of Official Statistics. The *AV* considers the different source of variability (deriving from sampling design and from statistical model) affecting the accuracy. *Section 4* illustrates some approximated results for defining the explicit expressions of the different components of the *AV*; these are developed considering a simplified statistical setting and some examples are developed in order to illustrate how the *AV* can be computed for some well known statistical models. Finally, *Section 5* develops some considerations for defining a strategy for ensuring the user be aware of the accuracy and in *Section 6* some preliminary conclusions are given.

## 2. ESTIMATES OF AGGREGATES FROM THE REGISTER

Let $y_k$ be the *true unknown* value of a variable of interest $y$ of the $k$th ($k$=1,…, $N$) unit included in the statistical register $R$.

According to a given statistical *working* model, *M*, we can suppose that the $y_k$ value is expressed as the sum of two components:

$$y_k = \tilde{y}_k + e_k, \qquad (2.1a)$$

where $\tilde{y}_k$ is the theoretical value according to which the value of $y$ is generated for the unit $k$ and $e_k$ denotes random noise with model expectations, $E_M(\cdot)$, and model variances, $V_M(\cdot)$, given by

$$E_M(\boldsymbol{e}) = \mathbf{0}_N \;\; ; \;\; V_M(\boldsymbol{e}\boldsymbol{e}') = E_M(\boldsymbol{e}\boldsymbol{e}') = \boldsymbol{\Sigma}_y = \begin{bmatrix} \sigma_{y1}^2 & & \sigma_{y1N} \\ \sigma_{yk1} & \sigma_{yk}^2 & \sigma_{ykN} \\ \sigma_{yN1} & & \sigma_{yN}^2 \end{bmatrix}, \qquad (2.1b)$$

in which: $\boldsymbol{e} = (e_1, \dots, e_k, \dots, e_N)'$, $\mathbf{0}_N$ is the $N$ column vector of zeroes and $\boldsymbol{\Sigma}_y$ is the $N \times N$ model covariance matrix.

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

$$\tilde{y}_k = f(\mathbf{x}_k; \boldsymbol{\vartheta}) \qquad (2.2)$$

in which $\mathbf{x}_k = (x_{k1}, \cdots, x_{ki}, \cdots, x_{kI},)'$ 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.

Suppose furthermore that for observing the $y$ values, a sample *S*, of size *n*, is selected from *R* according to a sample design *P*. Let $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_k, \dots, \lambda_N)'$ denote the $N$ column vector of sample membership indicators being $\lambda_k = 1$ if $k \in S$ and $\lambda_k = 0$, otherwise and let $E_P(\cdot)$ and $V_P(\cdot)$ indicate respectively the operators of sampling expectation and sampling variance, being

$$E_P(\boldsymbol{\lambda}) = \boldsymbol{\pi} \;\; ; \;\; V_P(\boldsymbol{\lambda}\boldsymbol{\lambda}') = \boldsymbol{\Sigma}_\lambda, \qquad (2.3)$$

where, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k, \dots, \pi_N)'$ is the vector of the inclusion probabilities.

Given the model *M* – defined by the relationships (2.1*a*), (2.1*b*) and (2.2) –, and the sample observational setting – defined in (2.3) – the register predictions of the value $\tilde{y}_k$ ($k$=1,...,$N$) can be defined by:

$$\hat{\tilde{y}}_k = f(\mathbf{x}_k; \hat{\boldsymbol{t}}), \qquad (2.4)$$

in which $\hat{\boldsymbol{t}} = \{\hat{t}_i;\ i = 1, \dots, I\}$ represents the estimate of $\boldsymbol{\vartheta}$ based on the observation of the the values $y_k$ on the sample *S*; while the values $\mathbf{x}_k$ are available for all the units of *R*. In many situations, the vector $\hat{\boldsymbol{t}} = \{\hat{t}_i; i = 1, \dots, I\}$ may be obtained as solution of the following system of $I$ equations:

$$\hat{H}_i(\hat{\boldsymbol{t}}) = \sum_{j \in R} g_i\left(y_j; \hat{\boldsymbol{t}};\ \lambda_j; w_j\right) = 0 \quad (i = 1, \dots, I) \qquad (2.5)$$

where $g_i\left(y_j; \hat{\boldsymbol{t}};\ \lambda_j; w_j\right)$ represents the score function of the unit $j$ of the *Generalized Estimating Equation* (GEE) (Ziegler, 2015) for the parameter $\hat{t}_i$ ($i = 1, \dots, I$) and $w_j$ is a generic weight assigned in the estimation phase to the unit $j$; here in the following, for making it simple, we suppose that $w_j$ may be equal either to 1 (solution usual in the classical inferential approach) or to $1/\pi_j$ (solution usual in the model assisted approach). The score function may incorporate in different ways the auxiliary information $\{\mathbf{x}_k; k \in R\}$ in the register.

Let $R_d$ be a *domain* (i.e. a specific subset with $N_d$ units) of *R*, defined on the basis of the variables in *R*, and let

$$Y_d = \sum_{k \in R_d} y_k \qquad (2.6)$$

be the unknown total of the variable $y$ in $R_d$.

The total $Y_d$ may be estimated as the sum over $R_d$ of the predictions $\hat{\hat{y}}_k$

$$\hat{Y}_d = \sum_{k \in R_d} \hat{\hat{y}}_k . \qquad (2.7)$$

**Remark 1**. On the basis of the predictions $\hat{\hat{y}}_k$, it is possible to set up different estimators of $Y_d$. For instance consider

$$\hat{Y}_{d,alt} = \sum_{k \in S \cap R_d} y_k + \sum_{k \in \bar{S} \cap R_d} \hat{\hat{y}}_k . \qquad (2.7a)$$

which is an estimator more usual when using the standard *prediction* approach. However, it is reasonable to assume that (2.7) and (2.7*a*) are roughly equivalent for large domains, (i.e.: $\hat{\hat{Y}}_d \cong \hat{\hat{Y}}_{d,alt}$). For making it simple, from now on will consider only the estimator (2.7).

## 3. THE MEASURE OF ACCURACY

The estimates are obtained summing up the predicted values where the predictions are based on model parameters estimated according to both $M$ and $S$. In our observational setting, the sampling design enables the observation of the sample $S$ and its variability is given by the sample variance/covariance matrix $\boldsymbol{\Sigma}_\lambda$ while the statistical model $M$ generates the variable $y$ and its variability is expressed by the variance/covariance matrix $\boldsymbol{\Sigma}_y$. In order to consider explicitly the variability deriving from both the sampling design and and the model $M$, Wolter (1986) proposes the concept *Global Variance*:

$$GV(\hat{Y}_d) = E_P E_M[\hat{Y}_d - E_P E_M(Y_d)]^2 = E_P[V_M(\hat{Y}_d)|\boldsymbol{\lambda}] + V_P[E_M(\hat{Y}_d)|\boldsymbol{\lambda}].$$

Nevertheless, while we are interested to consider all the sample space and the inducted sampling variability, we want focus on a single determination of the super-population underlying the model $M$. This is achieved by the *Anticipated Variance* ($AV$) which considers the variability of both the sampling design and the statistical model $M$ of the difference $(\hat{Y}_d - Y_d)$. The concept of $AV$, which has been introduced in literature for dealing with different inference problems (Isaki and Fuller, 1982; Sarndäl *et al.*, 1992; Nedyalkova and Tillé, 2008; Nirel, and Glickman, 2009; Falorsi and Righi, 2015), can be defined as:

$$AV(\hat{Y}_d) = E_P E_M(\hat{Y}_d - Y_d)^2 \qquad (3.1)$$

$$= E_P[V_M(\hat{Y}_d)|\boldsymbol{\lambda}] + V_P[E_M(\hat{Y}_d)|\boldsymbol{\lambda}] - V_M(Y_d). \qquad (3.2)$$

As can be seen, the $AV$ neutralizes the variability due to a pure model variability of the parameter $Y_d$ , even if it still considers the model variability of the estimator $\hat{Y}_d$.

In order to derive the explicit expression of the $AV$ we adopt two main approximations: (*i*) we consider the Taylor's series expansion of the function $f(\boldsymbol{x}_k; \hat{\boldsymbol{t}})$ evaluated at the point $f(\boldsymbol{x}_k; \boldsymbol{\vartheta})$ and (*ii*) we approximate the actual sampling design with a *Poisson* sampling design which has the same first order inclusion probabilities as the actual design. This makes for a conservative measure of the sampling variability.

Then we consider the $AV$ of its linear approximation, $\hat{Y}_{d,app}$:

$$\hat{Y}_d \cong \sum_{k \in R_d} f(\mathbf{x}_k; \boldsymbol{\theta}) + \sum_{k \in R_d} \sum_{i=1}^{I} f_{ki} \, (\hat{t}_i - \vartheta_i) + \, O_{kf} \qquad (3.3)$$

where: $f_{ki} = \dfrac{\delta f(\boldsymbol{x_k}; \boldsymbol{\vartheta})}{\delta \vartheta_i}$ and $O_{kf}$ is the rest of minor order.

Therefore, we may assume the following

$$AV(\hat{Y}_d) \cong AV(\hat{Y}_{d,app}) = E_P[V_M(\hat{Y}_{d,app})|\boldsymbol{\lambda}] + V_P[E_M(\hat{Y}_{d,app})|\boldsymbol{\lambda}] - V_M(Y_d) \quad (3.4)$$

in which

$$\hat{Y}_{d,app} = \sum\nolimits_{k \in R_d} \sum\nolimits_{i=1}^{I} f_{ki} \hat{t}_{\boldsymbol{i}} . \quad (3.5)$$

In matrix notation, the above expression may be expressed as:

$\hat{Y}_{d,app} = \boldsymbol{\gamma}_d' \mathbf{F} \hat{\boldsymbol{t}}$.

$\boldsymbol{\gamma}_d = (\gamma_{d1}, \cdots, \gamma_{dk}, \ldots, \gamma_{dN})'$ is the $N$ column vector of the $d$-th domain membership variables being $\gamma_{dk} = 1$, if $k \in R_d$ and $\gamma_{dk} = 0$, otherwise; $\mathbf{F} = \{f_{ki}: k = 1, \ldots, N;\ i = 1, \ldots, I\}$ is an $N \times I$ matrix.

## 4. SOME EXPRESSIONS OF THE COMPONENTS OF THE AV

In this section we will develop some lines of development for defining some explicit expressions of the first two terms on the right hand side of (3.4). Basically we will start from the approximation (3.5).

### 4.1. Term $E_P[V_M(\hat{Y}_{d,app})|\boldsymbol{\lambda}]$

Considering the expression (3.5), we have

$$[V_M(\hat{Y}_{d,app})|\boldsymbol{\lambda}] = \sum\nolimits_{i=1}^{I} v_{\hat{t}_i|\lambda} \sum\nolimits_{k \in R_d} \left( f_{ik}^2 + \sum\nolimits_{k' \neq k} f_{ik} f_{ik'} \right) +$$

$$+ \sum\nolimits_{i=1}^{I} \sum\nolimits_{\ell \neq i} c_{\hat{t}_i, \hat{t}_\ell|\lambda} \sum\nolimits_{k \in R_d} \sum\nolimits_{k' \neq k} f_{ik} f_{\ell k'} , \quad (4.1)$$

where

$v_{\hat{t}_i|\lambda} = V_M(\hat{t}_i|\boldsymbol{\lambda})$ , $c_{\hat{t}_i, \hat{t}_\ell|\lambda} = Cov_M(\hat{t}_i, \hat{t}_\ell|\boldsymbol{\lambda})$ .

In matrix notation we have

$$[V_M(\hat{Y}_{d,app})|\boldsymbol{\lambda}] = \boldsymbol{\gamma}_d' \mathbf{F} [V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}] \mathbf{F}' \boldsymbol{\gamma}_d . \quad (4.2)$$

The $I \times I$ matrix $[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}]$ - with the variances $v_{\hat{t}_i|\lambda}$ and the covariances $c_{\hat{t}_i, \hat{t}_\ell|\lambda}$- may be derived with the usual inferential approaches; however in their specific expressions the vector $\boldsymbol{\lambda}$ should be explicitly introduced in such a way that their formulae are computed on the whole set $R$, instead of considering only units in the sample $S$.

The first component of the $AV$ may be then obtained as

$$E_P[V_M(\hat{Y}_{d,app})|\boldsymbol{\lambda}] = \boldsymbol{\gamma}_d' \mathbf{F} (E_P[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}]) \mathbf{F}' \boldsymbol{\gamma}_d . \quad (4.3)$$

That said, the approach for deriving the explicit formulae of the $I \times I$ matrix $E_P[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}]$ may be different. It is useful to distinguish the case of the linear (or general linear) models from the case in which the score functions $g_i(y_j; \hat{\boldsymbol{t}};\ \lambda_j; w_j)$ are implicit expressions of the vector $\hat{\boldsymbol{t}}$.

**Example. 4.1.** Consider the classical simple linear model, in which $f(\mathbf{x}_k;\boldsymbol{\vartheta}) = \tilde{y}_k = \mathbf{x}_k'\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}_y = \sigma^2\mathbf{I}$, being $\mathbf{I}$ the identity matrix and suppose that $w_j = 1$ for $j \in S$.

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

$$\left(\sum_{j\in R} \mathbf{x}_j\, \mathbf{x}_j'\lambda_j\right)^{-1} \sum_{j\in R} \mathbf{x}_j\, y_j\, \lambda_j - \hat{\boldsymbol{t}} = \mathbf{0}_I.$$

From the above may be easily derived the standard expression for computing the matrix variance

$$[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}] = \sigma^2 \left(\sum_{j\in R} \mathbf{x}_j\, \mathbf{x}_j'\lambda_j\right)^{-1}.$$

### *4.1.1. Linear Models*

We have $\mathbf{F}=\mathbf{X}$ where: $\mathbf{X} = \{\mathbf{x}_j'; j \in R\}$. For the sampling expected values, we consider **the term 0** of the linear approximation of $[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}]$ evaluated at its sampling design expected value.

**Example. 4.2.** For the classical simple linear model given in example 4.1, we have

$$E_P[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}] \cong \sigma^2 \left(\sum_{j\in R} \mathbf{x}_j\, \mathbf{x}_j'\pi_j\right)^{-1}.$$

**Example. 4.3.** Consider an heteroscedastic linear model with no correlation among units in which $\boldsymbol{\Sigma}_y = diag\{\sigma_j^2; j = 1,2,\dots,n\}$. Suppose furthermore that that $w_j = 1$ for $j \in S$.

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

$$\left(\sum_{j\in R} \mathbf{x}_j\, \mathbf{x}_j'\lambda_j/\sigma_j^2\right)^{-1} \sum_{j\in R} \mathbf{x}_j\, \lambda_j y_j/\sigma_j^2 - \hat{\boldsymbol{t}} = \mathbf{0}_I.$$

The standard expression for computing for computing the matrix variance is

$$[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}] = \left(\sum_{j\in R} \mathbf{x}_j\, \mathbf{x}_j'\lambda_j/\sigma_j^2\right)^{-1}.$$

We approximate the sampling expected value of the above matrix as

$$E_P[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}] \cong \left(\sum_{j\in R} \mathbf{x}_j\, \mathbf{x}_j'\pi_j/\sigma_j^2\right)^{-1}.$$

**Example. 4.4.** Consider a general linear model with a general $\boldsymbol{\Sigma}_y$ matrix and suppose that $w_j = 1$ for $j \in S$. Denote with $\boldsymbol{y}_s$the $n$ vector of the $y$ values in the sample $S$ and with $\boldsymbol{\Sigma}_{y,S}$ the model variance covariance matrix of the $y$values in the sample. Let $\mathbf{X}_S$ be the $n \times I$ matrix of the $\mathbf{x}_j'$ values in $S$.

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

$$\left(\mathbf{X}_S'\boldsymbol{\Sigma}_{y,S}^{-1}\mathbf{X}_S\right)^{-1}\mathbf{X}_S'\boldsymbol{\Sigma}_{y,S}^{-1}\boldsymbol{y}_s - \hat{\boldsymbol{t}} = \mathbf{0}_I$$

Using the $\lambda$ values in the matrix $\boldsymbol{D}(\lambda_j) = diag\{\lambda_j;\ j = 1,\dots,N\}$, the above expression may be alternatively defined as:

$$\left(\mathbf{X}'\boldsymbol{D}(\lambda_j)\boldsymbol{\Sigma}_y^{-1}\boldsymbol{D}(\lambda_j)\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{D}(\lambda_j)\boldsymbol{\Sigma}_y^{-1}\boldsymbol{D}(\lambda_j)\boldsymbol{y} - \hat{\boldsymbol{t}} =$$

$$= \left[\sum_{j\in R}\mathbf{x}_j\,\lambda_j\left(\mathbf{x}_j'\nu_{jj} + \sum_{\iota\neq j}\mathbf{x}_\iota'\lambda_\iota\nu_{j\iota}\right)\right]^{-1}\sum_{j\in R}\mathbf{x}_j\,\lambda_j\left(y_j\nu_{jj} + \sum_{\iota\neq j}y_\iota\lambda_\iota\nu_{j\iota}\right) - \hat{\boldsymbol{t}} = \mathbf{0}_I, \qquad (4.4)$$

where $\nu_{j\iota}$ the generic element of the matrix in the position $j\iota$ of the matrix $\boldsymbol{\Sigma}_y^{-1}$.

The model variance of $\hat{\boldsymbol{t}}$ it is .

$$[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}] = \left(\mathbf{X}_S'\boldsymbol{\Sigma}_{y,S}^{-1}\mathbf{X}_S\right)^{-1} = \left(\mathbf{X}'\boldsymbol{D}(\lambda_j)\boldsymbol{\Sigma}_y^{-1}\boldsymbol{D}(\lambda_j)\mathbf{X}\right)^{-1}.$$

Finally approximating the actual sampling design with a *Poisson* sampling design which has the same first order inclusion probabilities of the actual design, we have

$$E_P[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}] \cong \left(\mathbf{X}'\boldsymbol{D}(\pi_j)\boldsymbol{\Sigma}_y^{-1}\boldsymbol{D}(\pi_j)\mathbf{X}\right)^{-1} = \left[\sum_{j\in R}\mathbf{x}_j\,\pi_j\left(\mathbf{x}_j'\nu_{jj} + \sum_{\iota\neq j}\mathbf{x}_\iota'\pi_\iota\nu_{j\iota}\right)\right]^{-1}$$

where $\boldsymbol{D}(\pi_j) = diag\{\pi_j;\ j = 1,\dots,N\}$.

### *4.1.2. Generalized Estimating Equations*

For the GEE estimation system (2.5), using the linear approximation proposed by Chambers (2012, p.124) we have

$$\hat{\boldsymbol{t}} \cong \boldsymbol{\vartheta} - [\boldsymbol{A}(\boldsymbol{\vartheta})|\boldsymbol{\lambda}]^{-1}\widehat{\boldsymbol{H}}(\boldsymbol{\vartheta}), \qquad (4.5)$$

where

$$[\boldsymbol{A}(\boldsymbol{\vartheta})|\boldsymbol{\lambda}] = \left(\frac{\delta\sum_{j\in R}g_i\left(y_j;\boldsymbol{\vartheta};\ \lambda_j;w_j\right)}{\delta\boldsymbol{\vartheta}}\right), \qquad (4.6)$$

$$\widehat{\boldsymbol{H}}(\boldsymbol{\vartheta}) = \left\{\widehat{H}_i(\boldsymbol{\vartheta}) = \sum_{j\in R}g_i\left(y_j;\boldsymbol{\vartheta};\ \lambda_j;w_j\right) = 0;\ (i = 1,\dots,I)\right\} \qquad (4.7)$$

Thus, it is

$$[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}] = [\boldsymbol{A}(\boldsymbol{\vartheta})|\boldsymbol{\lambda}]^{-1}V_M\left[\widehat{\boldsymbol{H}}(\boldsymbol{\vartheta})\left(\widehat{\boldsymbol{H}}(\boldsymbol{\vartheta})\right)'\middle|\boldsymbol{\lambda}\right][\boldsymbol{A}(\boldsymbol{\vartheta})|\boldsymbol{\lambda}]^{-1}. \qquad (4.8)$$

The Sampling design expected value of the above may by the term 0 of its linear approximation:

$$E_P[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}] \cong [\boldsymbol{A}(\boldsymbol{\vartheta})|\boldsymbol{\pi}]^{-1}V_M\left[\widehat{\boldsymbol{H}}(\boldsymbol{\vartheta})\left(\widehat{\boldsymbol{H}}(\boldsymbol{\vartheta})\right)'\middle|\boldsymbol{\pi}\right][\boldsymbol{A}(\boldsymbol{\vartheta})|\boldsymbol{\pi}]^{-1} \qquad (4.9)$$

where

$$[\boldsymbol{A}(\boldsymbol{\vartheta})|\boldsymbol{\pi}] \cong \left(\frac{\delta\sum_{j\in R}g_i\left(y_j;\boldsymbol{\vartheta};\ \pi_j;w_j\right)}{\delta\boldsymbol{\vartheta}}\right),$$

in which the expression of $V_M\left[\widehat{\boldsymbol{H}}(\boldsymbol{\vartheta})\left(\widehat{\boldsymbol{H}}(\boldsymbol{\vartheta})\right)'\middle|\boldsymbol{\pi}\right]$ is obtained by that of $V_M\left[\widehat{\boldsymbol{H}}(\boldsymbol{\vartheta})\left(\widehat{\boldsymbol{H}}(\boldsymbol{\vartheta})\right)'\middle|\boldsymbol{\lambda}\right]$ with the replacement of the values $\lambda_j$ with their expected values $\pi_j$.

**Example. 4.5.** Consider a generalized linear model (GLM) and suppose that the weights $w_j = 1$ for $j \in S$. According to Ziegler (2015, p.121), the estimating equations $\widehat{\boldsymbol{H}}(\boldsymbol{\vartheta})$, can be defined as

$$\widehat{\boldsymbol{H}}(\hat{\boldsymbol{t}}) = \mathbf{F}_S'\boldsymbol{\Sigma}_{y,S}^{-1}\left[\widehat{\tilde{\boldsymbol{y}}}_S(\hat{\boldsymbol{t}}) - \boldsymbol{y}_S\right] = \mathbf{0}_I, \tag{4.10}$$

where $\widehat{\tilde{\boldsymbol{y}}}_S(\hat{\boldsymbol{t}}) = (\hat{\tilde{y}}_1, \dots, \hat{\tilde{y}}_n)'$ denotes the $n$ vector of the $\hat{\tilde{y}}$ values in the sample $S$, each of which is an implicit function of the parameter $\hat{\boldsymbol{t}}$. Adopting the same computational trick proposed in 4.4, the (4.10), may be expressed as

$$\widehat{\boldsymbol{H}}(\hat{\boldsymbol{t}}) = \mathbf{F}'\boldsymbol{D}(\lambda_j)\boldsymbol{\Sigma}_y^{-1}\boldsymbol{D}(\lambda_j)\left[\widehat{\tilde{\boldsymbol{y}}}(\hat{\boldsymbol{t}}) - \boldsymbol{y}\right] = \mathbf{0}_I \tag{4.11}$$

Therefore, we have:

$$V_M\left[\widehat{\boldsymbol{H}}(\boldsymbol{\vartheta})\left(\widehat{\boldsymbol{H}}(\boldsymbol{\vartheta})\right)'\middle|\boldsymbol{\lambda}\right] =$$

$$= \mathbf{F}'\boldsymbol{D}(\lambda_j)\boldsymbol{\Sigma}_y^{-1}\boldsymbol{D}(\lambda_j)\left[V_M\left(\widehat{\tilde{\boldsymbol{y}}}(\hat{\boldsymbol{t}})\middle|\boldsymbol{\lambda}\right) + \boldsymbol{\Sigma}_y - 2Cov_M\left(\widehat{\tilde{\boldsymbol{y}}}(\hat{\boldsymbol{t}}), \boldsymbol{y}|\boldsymbol{\lambda}\right)\right]\boldsymbol{D}(\lambda_j)\boldsymbol{\Sigma}_y^{-1}\boldsymbol{D}(\lambda_j)\mathbf{F}. \tag{4.12}$$

Therefore, considering the (4.8), it is

$$[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}] = [\boldsymbol{A}(\boldsymbol{\vartheta})|\boldsymbol{\lambda}]^{-1}\mathbf{F}'\boldsymbol{D}(\lambda_j)\boldsymbol{\Sigma}_y^{-1}\boldsymbol{D}(\lambda_j)\left[V_M\left(\hat{\tilde{y}}(\hat{t})\middle|\lambda\right) + \Sigma_y - 2Cov_M\left(\hat{\tilde{y}}(\hat{t}), y|\lambda\right)\right]\cdot$$

$$\cdot\, \boldsymbol{D}(\lambda_j)\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-\mathbf{1}}\boldsymbol{D}(\lambda_j)\mathbf{F}\,[\boldsymbol{A}(\boldsymbol{\vartheta})|\boldsymbol{\lambda}]^{-\mathbf{1}} \tag{4.13}$$

Finally, Applying the (4.9) and approximating the actual sampling design with a *Poisson* sampling design which has the same first order inclusion probabilities as the actual design, we have:

$$E_P[V_M(\hat{\boldsymbol{t}})|\boldsymbol{\lambda}] = [\boldsymbol{A}(\boldsymbol{\vartheta})|\boldsymbol{\pi}]^{-1}\mathbf{F}'\boldsymbol{D}(\pi_j)\boldsymbol{\Sigma}_y^{-1}\boldsymbol{D}(\pi_j)\left[V_M\left(\widehat{\tilde{\boldsymbol{y}}}(\hat{\boldsymbol{t}})\middle|\boldsymbol{\pi}\right) + \boldsymbol{\Sigma}_y - 2Cov_M\left(\widehat{\tilde{\boldsymbol{y}}}(\hat{\boldsymbol{t}}), \boldsymbol{y}|\boldsymbol{\pi}\right)\right]\cdot$$

$$\cdot\, \boldsymbol{D}(\pi_j)\,\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{D}(\pi_j)\mathbf{F}\,[\boldsymbol{A}(\boldsymbol{\vartheta})|\boldsymbol{\pi}]^{-1}. \tag{4.14}$$

**Remark**. In order to determine an explicit expression of the model variance $V_M\left(\hat{\tilde{y}}(\hat{t})\middle|\lambda\right)$ and $Cov_M\left(\hat{\tilde{y}}(\hat{t}), y|\lambda\right)$, we can consider a linear approximation of the estimating equation (given in 2.5) as

$$\widehat{H}_i(\hat{\boldsymbol{t}}) = \sum\nolimits_{j\in R} g_i\left(\tilde{y}_j; \vartheta;\ \lambda_j; w_j\right) + \sum\nolimits_{\mathrm{j}\in\mathrm{R}} g_{ji(y)}\left(y_j - \tilde{y}_j\right) \qquad (\mathrm{i} = 1, \dots, \mathrm{I})$$

where

$$g_{ji(y)} = \left.\frac{\delta g_i\left(y_j; \boldsymbol{\vartheta};\ \lambda_j; w_j\right)}{\delta y_j}\right|_{y_j=\tilde{y}_j}.$$

Then, adopting a matrix notation, we have

$$\hat{\tilde{\mathbf{y}}}(\hat{\boldsymbol{t}}) \cong D(\lambda_j)\mathbf{F}\mathbf{G}_y'D(\lambda_j)\boldsymbol{y}$$

Where $\mathbf{G}_y' = \left\{g_{ji(y)}: j = 1, \dots, N; i = 1, \dots, I\right\}$. Therefore, we have:

$$V_M\left(\widehat{\tilde{\boldsymbol{y}}}(\hat{\boldsymbol{t}})\middle|\boldsymbol{\lambda}\right)$$

$$\cong \boldsymbol{D}(\lambda_j)\mathbf{F}\mathbf{G}_y'\boldsymbol{D}(\lambda_j)\boldsymbol{\Sigma}_y\boldsymbol{D}(\lambda_j)\mathbf{G}_y\mathbf{F}'\boldsymbol{D}(\lambda_j), \text{ and } Cov_M\left(\widehat{\tilde{\boldsymbol{y}}}(\hat{\boldsymbol{t}}), \boldsymbol{y}|\boldsymbol{\lambda}\right) = \boldsymbol{D}(\lambda_j)\mathbf{F}\mathbf{G}_y'\boldsymbol{D}(\lambda_j)\boldsymbol{\Sigma}_y. \tag{4.15}$$

### 4.2 Term $V_P\left[E_M\left(\hat{Y}_{d,app}\right)|\boldsymbol{\lambda}\right]$

Considering, the first order linear approximation, we have:

$$\left[E_M\left(\hat{Y}_{d,app}\right)|\boldsymbol{\lambda}\right] = \sum_{k\in R_d}\sum_{i=1}^{I} f_{ki}\hat{\tilde{t}}_{\boldsymbol{i}}\,, \tag{4.16}$$

where $\hat{\tilde{t}}_{\boldsymbol{i}}$ is obtained as solution of the following systems of $I$ estimating equations:

$$\widehat{H}_i\left(\hat{\tilde{\boldsymbol{t}}}\right) = \sum_{j\in R} g_i\left(\tilde{y}_j; \hat{\tilde{\boldsymbol{t}}};\ \lambda_j; w_j\right) = 0 \quad (i = 1, \dots, I). \tag{4.17}$$

In matrix notation, the (4.16) expression may be expressed as:

$$\left[E_M\left(\hat{Y}_{d,app}\right)|\boldsymbol{\lambda}\right] \cong \boldsymbol{\gamma}_d'\mathbf{F}\hat{\tilde{\boldsymbol{t}}}. \tag{4.18}$$

In which $\hat{\tilde{\boldsymbol{t}}}$ denotes the vector solution of (4.17).

**Example. 4.6.** Consider the simple linear model illustrated in the example 4.1.

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

$$\left(\sum_{j\in R}\mathbf{x}_j\,\mathbf{x}_j'\lambda_j\right)^{-1}\sum_{j\in R}\mathbf{x}_j\,\tilde{y}_j\,\lambda_j - \hat{\tilde{\boldsymbol{t}}} = \mathbf{0}_I.$$

Then we have:

$$\left[E_M\left(\hat{Y}_{d,app}\right)|\boldsymbol{\lambda}\right] \cong \boldsymbol{\gamma}_d'\mathbf{X}\left(\sum_{j\in R}\mathbf{x}_j\,\mathbf{x}_j'\lambda_j\right)^{-1}\sum_{j\in R}\mathbf{x}_j\,\tilde{y}_j\,\lambda_j,$$

Therefore, it is

$$V_P\left[E_M\left(\hat{Y}_{d,app}\right)|\boldsymbol{\lambda}\right] \cong \boldsymbol{\gamma}_d'\mathbf{F}V_P\left(\hat{\tilde{\boldsymbol{t}}}\right)\mathbf{F}'.$$

Even in simple linear models, $\hat{\tilde{\boldsymbol{t}}}$ is an implicit expression of the $\lambda$ values. Hence, for developing an explicit expression $V_P\left(\hat{\tilde{\boldsymbol{t}}}\right)$, it is necessary, to consider its linear approximation

$$\hat{\tilde{\boldsymbol{t}}} = E_P\left(\hat{\tilde{\boldsymbol{t}}}\right) + \mathbf{G}_\lambda'(\boldsymbol{\lambda} - \boldsymbol{\pi}) \tag{4.19}$$

where $\mathbf{G}_\lambda$ is an $N \times I$ which generic element $g_{ji(y)}$ is

$$g_{ji(y)} = \left.\frac{g_i\left(\tilde{y}_j; \hat{\tilde{\boldsymbol{t}}};\ \lambda_j; w_j\right)}{\delta\lambda_j}\right|_{\lambda_j=\pi_j} \qquad (j = 1, \dots, N; i = 1, \dots, I).$$

According to the above, it is

$$V_P\left(\hat{\tilde{\boldsymbol{t}}}\right) \cong \mathbf{G}_\lambda'V_P(\boldsymbol{\lambda})\mathbf{G}_\lambda \leq \mathbf{G}_\lambda'\boldsymbol{D}\left[\pi_j\left(1-\pi_j\right)\right]\mathbf{G}_\lambda \tag{4.20}$$

where $\boldsymbol{D}\left[\pi_j\left(1-\pi_j\right)\right] = diag\left\{\pi_j\left(1-\pi_j\right); j = 1, \dots, N\right\}$ is the diagonal matrix of the variance under a *Poisson sampling* of the $\boldsymbol{\lambda}$ values.

**Example. 4.7.** Consider the simple linear model, illustrated in the example (4.1). The matrix $\mathbf{G}'_\lambda$ is given by

$$\mathbf{G}'_\lambda = \left(\sum_{j\in R} \mathbf{x}_j \mathbf{x}_j^{'}\pi_j\right)^{-1}\left[\sum_{j\in R} \mathbf{x}_j \tilde{y}_j - \left(\sum_{j\in R} \mathbf{x}_j \mathbf{x}_j^{'}\right)\left(\sum_{j\in R} \mathbf{x}_j \mathbf{x}_j^{'}\pi_j\right)^{-1}\sum_{j\in R} \mathbf{x}_j \pi_j \tilde{y}_j\right]$$

$$= \left(\sum_{j\in R} \mathbf{x}_j \mathbf{x}_j'\pi_j\right)^{-1}\left[\boldsymbol{I} - \left(\sum_{j\in R} \mathbf{x}_j \mathbf{x}_j'\right)\left(\sum_{j\in R} \mathbf{x}_j \mathbf{x}_j'\pi_j\right)^{-1}\right]\mathbf{X}'\boldsymbol{D}\left(\tilde{y}_j\right)\boldsymbol{D}(1-\pi_j)$$

Where $\boldsymbol{D}\left(\tilde{y}_j\right) = diag\left\{\tilde{y}_j; j = 1, \dots, N\right\}$ and $\boldsymbol{D}(1-\pi_j) = diag\left\{1-\pi_j; j = 1, \dots, N\right\}$.

**Example. 4.8.** Consider the simple linear model, illustrated in the example (4.1), but suppose that the weights $w_j = 1/\pi_j$ (as in the model assisted approach). In this case the matrix $\mathbf{G}'_\lambda$ equals to $\mathbf{0}_{I\times N}$ (an $I \times N$ matrix of zeroes . Indeed, it is

$$\mathbf{G}'_\lambda = \left(\sum_{j\in R} \mathbf{x}_j \mathbf{x}_j^{'}\right)^{-1}\left[\sum_{j\in R} \mathbf{x}_j \tilde{y}_j - \left(\sum_{j\in R} \mathbf{x}_j \mathbf{x}_j^{'}\right)\left(\sum_{j\in R} \mathbf{x}_j \mathbf{x}_j^{'}\right)^{-1}\sum_{j\in R} \mathbf{x}_j \tilde{y}_j\right] = \mathbf{0}_{I\times N},$$

and thus we have the *Result* that if we use in the estimation phase the weights $w_j = 1/\pi_j$, we obtain:

$V_P\left[E_M\left(\hat{Y}_{d,app}\right)\middle|\boldsymbol{\lambda}\right] = 0.$

The same result may be obtained, if we consider the heteroscedastic linear model of the example (4.3). Indeed, using the $w_j = 1/\pi_j$ it is

$$\mathbf{G}'_\lambda = \left(\sum_{j\in R} \mathbf{x}_j \mathbf{x}_j^{'}/\sigma_j^2\right)^{-1}\left[\sum_{j\in R} \mathbf{x}_j \tilde{y}_j/\sigma_j^2 - \left(\sum_{j\in R} \mathbf{x}_j \mathbf{x}_j^{'}/\sigma_j^2\right)\left(\sum_{j\in R} \mathbf{x}_j \mathbf{x}_j^{'}/\sigma_j^2\right)^{-1}\sum_{j\in R} \mathbf{x}_j \tilde{y}_j/\sigma_j^2\right]$$

**Remark**. 1. The above examples make explicit the more general result according to which, if the parameter $\hat{\boldsymbol{t}}$ represents an unbiased estimate of the corresponding population parameter $\boldsymbol{t}$ then the component $V_P\left[E_M\left(\hat{Y}_{d,app}\right)\middle|\boldsymbol{\lambda}\right]$ of the $AV$ vanishes.

**Remark**. 2. The above result confirms the *Result* 5.10.1 in Sarndal *et al.* 1992.

**Example. 4.9.** Consider the GLM model introduced in example 4.5 and suppose that the weights $w_j = 1$ for $j \in S$. According to the example (4.5), the estimating equations for the parameter $\hat{\tilde{\boldsymbol{t}}}$ can be defined as

$\widehat{\boldsymbol{H}}(\hat{\boldsymbol{t}}) = \mathbf{F}'\boldsymbol{D}(\lambda_j)\boldsymbol{\Sigma}_y^{-1}\boldsymbol{D}(\lambda_j)\left[\hat{\tilde{\boldsymbol{y}}}(\hat{\tilde{\boldsymbol{t}}}) - \tilde{\boldsymbol{y}}\right] = \mathbf{0}_I.$

where $\hat{\tilde{\boldsymbol{y}}}_S(\hat{\tilde{\boldsymbol{t}}})$ denotes the $N$ vector of the predictions $\hat{\tilde{y}}$ estimated with the sample parameter $\hat{\tilde{\boldsymbol{t}}}$.

## 5. STRATEGIES FOR MAKING THE STAKEHOLDERS AWARE OF THE ACCURACY

Two main strategies are suggested in the paper for making the user aware of the accuracy:

1. The first is based on the development of a software applications that together with the production of the aggregates $\hat{Y}_d$ will provide the user the estimates of the corresponding *AV.*
2. The second is based on that developed for the *synthetic presentation of the sampling errors* in social sample surveys.

*Software application*

The plug-in estimate of the $AV$ may be computed by replacing the estimates $\hat{\boldsymbol{t}}$ , $\hat{\tilde{\boldsymbol{y}}}$ and $\hat{\boldsymbol{\Sigma}}_y$ instead of the

unknown parameters $\boldsymbol{\vartheta}$ , $\tilde{\boldsymbol{y}}$ and $\boldsymbol{\Sigma}_y$ in the expressions of the different components of the *AV*. According to *Ziegler* (2015, point 5, pp.121) these plug-in estimates are strongly consistent estimator of the different components of the variance. Other approaches may be based on the Bootstrap methods (Scholtus, 2018).

As far as concerns the computational feasibility of the proposal, let us consider that $\left[V_M\left(\hat{Y}_{d,app}\right)\middle|\boldsymbol{\lambda}\right]$ can be computed according to the matrix expression according to formula (4.2).

Nevertheless, from the register perspective, the variance $\left[V_M\left(\hat{Y}_{d,app}\right)\middle|\boldsymbol{\lambda}\right]$ could be computed according to (4.1) applying the sum function. In particular, for each record *k*, the $\left[V_M\left(\hat{Y}_{d,app}\right)\middle|\boldsymbol{\lambda}\right]$ needs the knowledge of the following variables:

- ✓ $I$ column variables such as: $v_{\hat{t}_i|\lambda} f_{ik}^2$;
- ✓ $I$ column variables such as: $v_{\hat{t}_i|\lambda}[f_{ik}(F_i - f_{ik})]$;
- ✓ $I(I\text{-}1)$ column variables such as: $c_{\hat{t}_i,\hat{t}_\ell|\lambda}[f_{ik}(F_\ell - f_{\ell k})]$,

being $F_i = \sum_k f_{ik}$ *and* $F_\ell = \sum_k f_{\ell k}$.

The above approach requires the storage of the above quantities, but it will be less computational demanding when the user asks for specific statistics because much of the computation is moved up. The sum of the columns for the records belong to the domain (the only parameter of the computation) must be carried out.

Finally we note, that the parameters to stored for a given *Y* are [2*I*+ $I(I\text{-}1)$].

*Synthetic presentation of the sampling error*

This approach largely used for the presentation of social surveys, is based on the hypothesis that the estimates $\hat{Y}_d$ and their $AV$ are linked by means of a functional models, as for instance:

$$\frac{AV\left(\hat{Y}_d\right)}{\hat{Y}_d^2} = \epsilon^2\left(\hat{Y}_d\right) = f\left(\hat{Y}_d, \alpha_1, \dots, \alpha_q\right) + u_d,$$

where $\alpha_1, \dots, \alpha_q$ are unknown parameters and $u_d$ is an error term.

From a practical point of view the estimates $\hat{\alpha}_1, \dots, \hat{\alpha}_q$ of the parameters $\alpha_1, \dots, \alpha_q$ are obtained with the standard regression techniques fitting the models on a cloud of different couple of points

$$\left[\frac{\widehat{AV}\left(\hat{Y}_d\right)}{\hat{Y}_d^2}, \hat{Y}_d\right],$$

where $\widehat{AV}\left(\hat{Y}_d\right)$ is the estimate of the $AV$. Then, on the basis of the parameters $\hat{\alpha}_1, \dots, \hat{\alpha}_q$, the users can compute the accuracy of their estimates by mean of

$$\widehat{\widehat{AV}}\left(\hat{Y}_d\right) = f\left(\hat{Y}_d, \hat{\alpha}_1, \dots, \hat{\alpha}_q\right)\hat{Y}_d^2$$

In practice this could be done with a software application, which should have stored the functional for of the function $f$ and the parameters $\hat{\alpha}_1, \dots, \hat{\alpha}_q$.

A form of $f$ useful in practice for frequencies hypothesizes a decreasing relationship between the $\hat{Y}_d$ value and the $\epsilon^2\left(\hat{Y}_d\right)$ :

$$\epsilon^2\left(\hat{Y}_d\right) = \alpha_1 \hat{Y}_d^{\alpha_2} u_d.$$

## 6. PRELIMINARY CONCLUSIONS

In this document we have proposed different strategies which allow the different users of a

statistical register to be of aware of the accuracy of their estimates.

The main result of this research line is the identification of the $AV$ as a suitable for evaluating the accuracy of register aggregates and the definition of the computational formulae for defining its values, considering a simplified statistical setting.

The main further steps in this research line are those of evaluating the strengths and robustness of the results with some simulation studies.

Another aspect to be considered is that of the computational feasibility of the proposal. This is strictly related on the definition of the best strategy for presenting the accuracy to the users.

## MAIN REFERENCES


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

Falorsi, P.D., 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.

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

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

Nirel, R. , Glickman, H. (2009). Chapter 21 - Sample Surveys and Censuses. In: Rao, C.R. (ed.) *Handbook of Statistic*, Elsevier.

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

Scholtus S., (2019), *A bootstrap method for estimators based on combined administrative and survey data.* NTTS Conference 2019, to be published.

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

Ziegler, A. (2011), *Generalized Estimating Equations*, Lecture Notes in statistics.