The anticipated variance as a measure for the accuracy of complex multisource statistics

Piero Demetrio Falorsi, Paolo Righi

Abstract

Budget constraints, declining response rate, coverage errors and the need to have wider, deeper, quicker and better statistics foster the National Statistical Offices to investigate new data sources, paradigms and tools for producing data. Since 2014, Italian National Statistical Institute relies business statistics on an extensive use of administrative data ([1], [2]). Households' statistics, instead, are still based on standard surveys. The last 2011 population census has been carried out according to the traditional approach using multiple modes to facilitate more cost-effective response. Nevertheless for the next 2021 round a register-based census has been planned. Here, a multisource approach should be applied. In particular, multiple frames such as administrative population register, tax register, social security data, etc. will be used to face coverage concerns.

Several statistical procedures such as record linkage or statistical matching enable the definition of the Statistical Register (SR). The administrative data gives also the main contribution for observing the target variables of the SR or for imputing missing values for specific sub-populations according to a predictive approach. At this step the SR still suffers from over-coverage and under-coverage concerns.

Surveys will support census process to estimate the under/over coverage of the register (population size).

The accuracy of multi-sources statistics has to be taken into account three potential sources of uncertainty, the model variance of the prediction approach, the variance of the random sampling, the variance of the Capture/Recapture (CR) model for coverage errors.

In particular, CR model deals with uncertainty of the population size using the conditional variance decomposition which takes into account randomization and CR model uncertainty ([3], [4]). Here, we propose to use the anticipated variance (AV) approach ([5]). The AV makes inference on the population size given the two sources of variability conditionally to the current observed SR. This approach sounds more suitable for official statistic product with respect to the unconditional variance in which the inference considers all the potential realization of SR.

1. Introduction

Istat according to the modernization programme is going to shift the data production from a traditional survey based statistics to a register-based statistics. Register-based statistics means that, given a set of registers, the statistics are obtained by adding up the values of the target variables at unit level (micro level). Each register is built up using a multiple frames procedure. Administrative population register, tax register, social security data, etc. will be used to face coverage concerns. Usual techniques at microdata level (record linkage, statistical matching, prediction process or imputation) and macrodata level (calibration estimator, model based estimator, bayesian models) enable the creation of the register, hereinafter denoted as Statistical Register (SR). More than one SR is defined. Mainly, the modernization programme assumes a Base Register (BR), identifying statistical units and the main demographic variables, and several Satellite Registers (RS) containing thematic variables mainly derived from administrative sources or surveys. All these registers define the Integrated System of Statistical Registers (ISSRs), underlining the coherence constraints among each SR.

The BR and SRs are statistical products. They could suffer from under/over-coverage problem and, linkage errors and the variables can be imputed or predicted and not directly observed in administrative archives. Therefore, is crucial to define an uncertainty measure of the register-based statistics.

The document delineates the quality frameworks to deal with the effects of under/over-coverage with respect to the population size estimate. Basically, the uncertainty of the population size ignoring the linkage process. The framework should be easily extended to other sources of uncertainty such as, linkage or imputation. Three different frameworks are shown. Two of them are quite standard in the literature of CR problem. The third, based on the concept of Anticipated Variance, is original if applied in this context.

The document does not show results or quantitative evidence. It does the groundwork to develop the further research on this topic.

2. Basic Notation

Let us introduce the basic notation. Other notation will be defined throughout the document.

- U_L: target population (L, indicates living);
- N_L: target parameter – people living (U_L size);
- R: the BR (during the document we use them as synonymous);
- N_R: size of R – people registered (size of R).
- $U_{g,L}$ and R_g (with g=1, ..., G): sub- set of U_L and R, in which units have homogeneity behavior with respect to under-coverage and over-coverage phenomenon (same probability to be under-covered and over-covered) of sub-population $U_{g,L}$. In practice notation g identifies a **g** vector of covariate generally known R;
- $U_{ga,L}$ and R_{ga} (with a=1, ..., A): is a further partition $U_{g,L}$ and R_g being a the geographic area defining the finest partition of the Country. The population size estimates will be consistent at aggregate levels of these areas;
- $N_{ga,L}$ and $N_{ga,R}$ e$N_{ga,RL}$: the size respectively of $U_{ga,L}$, R_{ga} and $U_{ga,L} \cap R_{ga}$.
- $P_{gaL|R} = N_{ga,RL}/N_{ga,R}$: proportion living people with respect to the total of personsR_{ga}, where 1-$P_{gaL|R}$ is the over-coverage proportion of the BR.
- $P_{gaR|L} = N_{ga,RL}/N_{ga,L}$: proportion of persons in R_{ga} with respect to the living people, where 1-$P_{gaR|L}$ is the under-coverage proportion of BR;

3. Living population statistics

We assume the goal of the register-based statistics is to use the sum operator to achieve the estimate of the target parameter. Let BR suffer from over / under-coverage problems, we want to assign to each unit of the BR a d_k weight ($k = 1, \dots N_R$) that takes into account the over / under-coverage conditions. In practice N_L depends on N_R according to these weights.

The weights are defined by the **g** vector of auxiliary known variable in BR. We do not introduce the case where the covariates in **g** are partially observed in R.

The general expression of weight is $d_k = d_{ga} \;\; \forall k \in R_{ga}$ being

$$d_{ga} = \frac{P_{ga,L|R}}{P_{ga,R|L}}.$$

The weights d_k will be used for every statistics related to the living population in a given area a.

The living population (size of $U_{ga,L}$) is given by

$$N_{ga,L} = \textstyle\sum_{R_{ga}} d_k = N_{ga,R} d_{ga}.$$

The population of domain D (cutting-across the sub-population define by the couple ga) is

$$N_{D,L} = \textstyle\sum_{R_D} d_k$$

The overall size population is

$$N_L = \textstyle\sum_{ga} \sum_{R_{ga}} d_k.$$

The expression is based on standard equation given, for example, by [6]

$$N_{ga,L} \times P_{ga,R|L} = N_{ga,R} \times P_{ga,L|R} \Leftrightarrow N_{ga,L} = N_{ga,R} \times \frac{P_{ga,L|R}}{P_{ga,R|L}} = \frac{N^*_{ga,R}}{P_{ga,R|L}}. \qquad (3.1)$$

The last equality introduces$N^*_{ga,R}$, that is the size of R removing the over-coverage.

The terms$N_{ga,L}$, $P_{ga,L|R}$ e $P_{ga,R|L}$ have to be estimated.

As final remark, at this stage of the research the formulation of d_k weights tackle the coverage issue at unit level (persons), while they are not suitable to deal with coverage for households [3]

3.1 Estimation of living population

The estimate for domain D is given by

$$\widehat{N}_{D,L} = \sum_k \hat{d}_k$$

with $\hat{d}_k = \hat{d}_{ga}\ \forall k \in R_{ga}$. The estimate is based on the Extended Dual System Estimator (EDSE – [7]). The EDSE produces correct estimates when some assumptions hold on the capture / recapture process. We do not further investigate the assumptions ([8];[7]). Moreover, EDSE assumes homogeneity in ga also for over-coverage probability.
The EDSE consider two list of objects defined independently: the B list of clusters of units (such as census sections/enumeration areas or dwellings) not affected by over-coverage concerns.
The capture process is identified by R, while the recapture is identified by a survey on B.
As far the over-coverage is concerned we introduce two survey strategies:
(a) A list survey from R, counting the ineligible people (over-counted);
(b) A follow-up of units not observed in B but included in R. The follow-up verifies, if there is an over-coverage problem in R for unmatched units (it is important in R can be identified the clusters of B). Further details in Nirel and Glickman (2009, section 2.2.1 and 2.2.2).
We assume the d_k is observed if the survey on B is a census and the survey for the over-coverage is a census as well.

You should note we do not consider uncertainty on d_k given by the super-population models generating under-coverage and over-coverage ([8];[7]). In other terms, the authors assume that if we repeat the procedure for counting the population we obtain different totals even though the procedure makes a census on R and B because of the underlining models. Here, we propose to make inference conditionally to the given realization of the super-population model.

We come back on this topic in section 4.
In practice, a sample from B and/or from R (for the over-coverage) is carried out. We obtain an estimate

$$\hat{d}_k = \widehat{P}_{gaL|R} / \widehat{P}_{gaR|L}$$

where $\widehat{P}_{gaL|R}$ and $\widehat{P}_{gaR|L}$ are the sampling estimates of $P_{gaL|R}$ and $P_{gaR|L}$.
According to the survey strategy (a) we estimate $P_{gaL|R}$ as follows:

- draw a sample s_o from R;
- compute the sampling weight w_k;
- collect the variable $e_{ga,k}$ being equal to 1 if $k \in U_{ga,L} \cap R_{ga}$ and 0 otherwise (eligible unit);
- compute$\widehat{N}_{ga,RL} = \sum_{k \in s_o} e_{ga,k}\, w_k$ and $\widehat{N}_{ga,R} = \sum_{k \in s_o} w_k$;
- compute $\widehat{P}_{gaL|R} = \widehat{N}_{ga,RL} / \widehat{N}_{ga,R}$.

According to the survey strategy (b) we estimate $P_{gaL|R}$ as follows:

- draw a cluster sample $\mathrm{s_u}$ from B;
- compute the sampling weight $\mathrm{w_j}$ for cluster j;
- define $\mathrm{N_{ga,jR}}$: number of units in $\mathrm{R_{ga}}$ of cluster j (value known);
- collect the variable $\mathrm{e_{ga,j}}$: number of elegible people in cluster j (people matched in the area sampling and in the follow-up);
- compute $\widehat{\mathrm{N}}_{\mathrm{ga,RL}} = \sum_{\mathrm{j \in s}} \mathrm{e_{ga,j}}\, \mathrm{w_j}$ and $\widehat{\mathrm{N}}_{\mathrm{ga,R}} = \sum_{\mathrm{j \in s}} \mathrm{N_{ga,jR} w_j}$;
- compute $\widehat{\mathrm{P}}_{\mathrm{gaL|R}} = \widehat{\mathrm{N}}_{\mathrm{ga,RL}} / \widehat{\mathrm{N}}_{\mathrm{ga,R}}$.

The estimation of $P_{gaR|L}$ follows the ordinary process of the Dual System Estimator (Wolter, 1986) from the sample s_u,

- collect the variable $u_{ga,jB}$ number of individuals of cluster j in $U_{ga,L}$ (number of living people observed in the sample);
- collect the variable $\mathrm{c_{ga,jRB}}$ number of individuals of cluster j in $\mathrm{U_{ga,L}} \cap \mathrm{R_{ga}}$ (number of living people observed in the sample and matched in the BR);
- compute $\widehat{\mathrm{N}}_{\mathrm{ga,B}} = \sum_{\mathrm{j \in S}} \mathrm{u_{ga,jB}}\, \mathrm{w_j}$ and $\widehat{\mathrm{N}}_{\mathrm{ga,RB}} = \sum_{\mathrm{j \in S}} \mathrm{c_{ga,jRB} w_j}$;
- $\widehat{\mathrm{P}}_{\mathrm{gaR|L}} = \widehat{\mathrm{N}}_{\mathrm{ga,RB}} / \widehat{\mathrm{N}}_{\mathrm{ga,B}}$ (unbiased estimate under the DSE assumptions).

The estimation of (3.1) is given by

$$\widehat{N}_{ga,L} = N_{ga,R} \times \frac{\widehat{P}_{ga,L|R}}{\widehat{P}_{ga,R|L}} = \frac{N_{ga,R}\ \widehat{P}_{ga,L|R}}{\widehat{P}_{ga,R|L}} = \frac{\widehat{N}^*_{ga,R}}{\widehat{P}_{ga,R|L}} = \left[\frac{\widehat{N}^*_{ga,R}\ \widehat{N}_{ga,B}}{\widehat{N}_{ga,RB}}\right],$$

and for $N_{D,L}$

$$\widehat{N}_{D,L} = \sum\nolimits_{k\in R_D} \hat{d}_k = \sum\nolimits_{g=1}^{G}\sum\nolimits_{a=1}^{A} N_{gaD,R}\hat{d}_{ga} = \sum\nolimits_{g=1}^{G}\sum\nolimits_{a=1}^{A} \widehat{N}_{gaD,L}$$

Note that survey strategy (b) uses a unique set of weights for over-coverage and under-coverage. The $\widehat{N}_{D,L} = \sum_{R_D}\hat{d}_k$ follows a standard estimation approach of nonlinear parameter. In the simple case of $U_{ga,L}$ we have

$$\widehat{N}_{ga,L} = N_{ga,R}\left[\left(\frac{\sum_{j\in S} w_j\ e_{ga,j}}{\sum_{j\in S} w_j\ N_{ga,jR}}\right)\left(\frac{\sum_{j\in S}\ w_j u_{ga,j}}{\sum_{j\in S}\ w_j c_{ga,j}}\right)\right].$$

4. Uncertainty on estimation of living population

Let us introduce the topic examining the DSE first. The DSE assumes multinomial model

$$L(\mathcal{N}, p_{R|L}, p_{B|L}) = \binom{\mathcal{N}}{N_{RB}, N_{R\bar{B}}, N_{\bar{R}B}} (p_{R|L})^{N_R} (p_{B|L})^{N_B} \\ \times (1-p_{R|L})^{\mathcal{N}-N_R} (1-p_{B|L})^{\mathcal{N}-N_B}$$

generating the under-coverage of *R*, being the table 1 the results of capture/recapture process.

Table 1. Capture /Recapture results (*ga* notation omitted for simplicity)

		Recapture B (Area list)		
		In	Out	
Capture R	In	N_{RB}	$N_{R\bar{B}}$	N_R
	Out	$N_{\bar{R}B}$	$N_{\overline{RB}}$	$N_{\bar{R}}$
		N_B	$N_{\bar{B}}$	N

According to table 2 we have: the super-population parameters, the parameters of the finite observed population and the estimates of both.

Table 2. Parameters and estimator in capture/recapture process (*ga* notation omitted for simplicity)

Multinomial super population parameter	Realization under DS estimator	Estimation using a sample for recapture
$p_{R\vert L} \equiv E_M(P_{R\vert L})$	$P_{R\vert L} = \frac{N_{RB}}{N_B}$	$\widehat{P}_{R\vert L} = \frac{\widehat{N}_{RB}}{\widehat{N}_B}$
$p_{B\vert L} \equiv E_M(P_{B\vert L})$	$P_{B\vert L} = \frac{N_{RB}}{N_R}$	$\widehat{P}_{B\vert L} = \frac{\widehat{N}_{RB}}{\widehat{N}_R}$
$p_{RB\vert L} = p_{R\vert L}p_{B\vert L}$	$P_{RB\vert L} = \frac{N_{RB}}{N_B}\frac{N_{RB}}{N_R}$	$\widehat{P}_{RB\vert L} = \frac{\widehat{N}_{RB}}{\widehat{N}_B}\frac{\widehat{N}_{RB}}{\widehat{N}_R}$

Wolter ([8]) states the estimator is subject to two sources of variability: sampling variability and model variability associated with the coverage error model,

$$V(\widehat{N}_{ga,L}) = E_P E_M\left[\widehat{N}_{ga,L} - E_P E_M(\widehat{N}_{ga,L})\right]^2 \qquad (4.1)$$

in which the $E_M(.)$ operator is the expectation with respect to the multinomial coverage error model and $E_P(.)$ is the expected value with respect to the sampling design. In case of unbiased estimator we have $E_P E_M(\hat{N}_{ga,L}) = \mathcal{N}_{ga}$.

The $V(.)$ operator measures the uncertainty on the unknown parameters $p_{R|L}$, $p_{B|L}$and $p_{RB|L}$. $V(\hat{N}_{ga,L}) \neq 0$ when $\hat{N}_{ga,B} = N_{ga,B}$ and $\hat{N}_{ga,RB} = N_{ga,RB}$. That is $N_{ga,L}$ is an observation of a random variable generated by the multinomial model.

The aim is to make inference on the parameters of the super-population model. Nirel and Glickman ([7]) extend the expression (4.1) taking into account a *Poisson* model for over-coverage condition.

We argue the official statistics aim to make inference on $N_{ga,L}$ conditionally to the observed register R and the observed area sample. That means the task is to make inference on living people without considering every possible registers we could observe. Target parameters are N_L, $P_{R|L}$, $P_{B|L}$ and $P_{RB|L}$ proportions. The $P_{L|R}$ proportion is added due to over-coverage.

According to this approach, we propose to use

$$AV(\hat{N}_{ga,L}) = V(\hat{N}_{ga,L}|N_{ga,R}; N_{ga,B}; N_{ga,RB}) = E_P E_M[\hat{N}_{ga,L} - E_P(\hat{N}_{ga,L})]^2 \quad (4.2)$$

introduced in [5]. Finally, Pfeffermann ([6]) suggests the use of the sampling design variance

$$V_P(\hat{N}_{ga,L}) = V_P(\hat{N}_{ga,L}|N_{ga,R}; N_{ga,B},; N_{ga,RB}) = E_P[\hat{N}_{ga,L} - E_P(\hat{N}_{ga,L})]^2. \quad (4.3)$$

The expressions (4.2) and (4.3) are null when $\hat{N}_{ga,B} = N_{ga,B}$ and $\hat{N}_{ga,RB} = N_{ga,RB}$.
Let us investigate the expression (4.2) when $E_P(\hat{N}_{ga,L}) = N_{ga,L}$. The (4.2) is replaced by with

$$AV(\hat{N}_{ga,L}) = E_P E_M[\hat{N}_{ga,L} - N_{ga,L}].^2 \quad (4.4)$$

According to [9] the (4.4) is equal to

$$AV(\hat{N}_{ga,L}) = E_M V_P(\hat{N}_{ga,L}). \quad (4.5)$$

Since $\hat{N}_{ga,L} = N_{ga,R}\left[\left(\frac{\sum_{j\in S} w_j\, e_{ga,j}}{\sum_{j\in S} w_j\, N_{ga,jR}}\right)\left(\frac{\sum_{j\in S}\, w_j u_{ga,j}}{\sum_{j\in S}\, w_j c_{ga,j}}\right)\right]$ then we apply the first order Taylor series approximation method in the point $[E_P E_M(\hat{N}_{ga,L}|\mathcal{N}, p_{R|L}, p_{B|L}, p_{RB|L})]$ for computing $V_P(\hat{N}_{ga,L})$.
Now, for sake of simplicity, we express the *AV* in case of *Poisson* sampling (cluster sampling introduce complexity in the formula but it does not give new elements of discussion).
In the general case of the Horvitz-Thompson estimator, $\hat{Y} = \sum_s y_j/\pi_j$ of a total Y where π_j is the inclusion probability, the design variance is $V_P(\hat{Y}) = \sum_U y_j^2\left(\frac{1}{\pi_j} - 1\right)$. In the context the EDSE we approximate $V_P(\hat{N}_{ga,L})$ with

$$V_P(\hat{N}_{ga,L}) = N_{ga,R}^2\, V_P\left(\frac{\hat{P}_{ga,L|R}}{\hat{P}_{ga,R|L}}\right) \cong \sum_{j\in U_J}\left[(z_{ga,j})^2\left(\frac{1}{\pi_j} - 1\right)\right], \quad (4.6)$$

where U_J is the cluster population, being

$$z_{ga,j} = a_{ga,e}\, e_{ga,j} + a_{ga,x}\, N_{ga,Rj} + a_{ga,u}\, u_{ga,j} + a_{ga,c}\, c_{ga,j}$$

the Woodruff transformation ([10]) based on Taylor linearization method in which

$$a_{ga,e} = N_{ga,R}\left[\left(\frac{w_j}{\sum_{j\in S} w_j\, N_{ga,jR}}\right)\left(\frac{\sum_{j\in S}\, w_j u_{ga,j}}{\sum_{j\in S}\, w_j c_{ga,j}}\right)\right],$$

$$a_{ga,N} = -N_{ga,R}\left[w_j\left(\frac{\sum_{j\in S} w_j\ e_{ga,j}}{\left(\sum_{j\in S} w_j\ N_{ga,jR}\right)^2}\right)\left(\frac{\sum_{j\in S}\ w_j u_{ga,j}}{\sum_{j\in S}\ w_j c_{ga,j}}\right)\right],$$

$$a_{ga,u} = N_{ga,R}\left[\left(\frac{\sum_{j\in S} w_j\ e_{ga,j}}{\sum_{j\in S} w_j\ N_{ga,jR}}\right)\left(\frac{w_j}{\sum_{j\in S}\ w_j c_{ga,j}}\right)\right],$$

$$a_{ga,c} = -N_{ga,R}\left[w_j\left(\frac{\sum_{j\in S} w_j\ e_{ga,j}}{\sum_{j\in S} w_j\ N_{ga,jR}}\right)\left(\frac{\sum_{j\in S}\ w_j u_{ga,j}}{\left(\sum_{j\in S}\ w_j c_{ga,j}\right)^2}\right)\right].$$

In the linearization process it should be noted that $N_{ga,R}$ is treated as a fixed value.
By the definition of $z_{ga,j}$ we can express the (4.6) as

$$V_P\left(\hat{N}_{ga,L}\right) \cong N_{ga,R}^2\left[\frac{V_P(\hat{P}_{ga,L|R})}{\left[E_P(\hat{P}_{ga,R|L})\right]^2} + \frac{\left[E_P(\hat{P}_{ga,L|R})\right]^2}{\left[E_P(\hat{P}_{ga,R|L})\right]^4}V_P(\hat{P}_{ga,R|L})\right]. \quad (4.7)$$

in accordance with [6]. In formula (4.7) no model uncertainty is taken into account.
We must apply the $E_M(.)$ operator for defining the AV. That means we treat $z_{ga,j}$ as a random value generated by the super- population model,

$$AV\left(\hat{N}_{ga,L}\right) \cong N_{ga,R}^2\ E_M\left\{\sum_{j\in U_J}\left[\left(z_{ga,j}\right)^2\right]\left(\frac{1}{\pi_j}-1\right)\right\}.$$

For independence assumptions on CR and over-coverage, we achieve

$$AV\left(\frac{\hat{P}_{ga,L|R}}{\hat{P}_{ga,R|L}}\right) \cong \sum_{j\in U_J}\left[\left(E_M\left(z_{ga,j}\right)\right)^2 + V_M\left(z_{ga,j}\right)\right]\left(\frac{1}{\pi_j}-1\right). \quad (4.8)$$

Finally the (4.8) may be reformulated as

$$AV\left(\frac{\hat{P}_{ga,L|R}}{\hat{P}_{ga,R|L}}\right) \cong \sum_{j\in U_J}\left[E_M\left(z_{ga,j}\right)\right]^2\left(\frac{1}{\pi_j}-1\right) + \sum_{j\in U_J}\frac{V_M\left(z_{ga,j}\right)}{\pi_j} - \sum_{j\in U_J}V_M\left(z_{ga,j}\right). \quad (4.9)$$

The (4.9) highlights the AV is based on three components

$$AV\left(\frac{\hat{P}_{ga,L|R}}{\hat{P}_{ga,R|L}}\right) \cong V_P E_M\left(\frac{\hat{P}_{ga,L|R}}{\hat{P}_{ga,R|L}}\right) + E_P V_M\left(\frac{\hat{P}_{ga,L|R}}{\hat{P}_{ga,R|L}}\right) - V_M\left(\frac{P_{ga,L|R}}{P_{ga,R|L}}\right) \quad (4.10)$$

The (4.10) evidences the difference with the variance decomposition formula

$$E_P E_M\left[\frac{\hat{P}_{ga,L|R}}{\hat{P}_{ga,R|L}} - E_P E_M\left(\frac{\hat{P}_{ga,L|R}}{\hat{P}_{ga,R|L}}\right)\right]^2 = V_P E_M\left(\frac{\hat{P}_{ga,L|R}}{\hat{P}_{ga,R|L}}\right) + E_P V_M\left(\frac{\hat{P}_{ga,L|R}}{\hat{P}_{ga,R|L}}\right).$$

5. Discussion

The document deals with the variability of the population size estimation by means of register-based approach. The proposed quality framework starts from the idea that: if the quality survey of the BR includes with certainty every cluster (enumeration areas, for example) of the country; if this survey is not affected by non-sampling errors; if the super-population model for coverage errors is correct (no bias), we should expect to estimate the population size with certainty.
The proposed quality framework, based on Anticipated Variance, is not completely model free as design variance. As formula (4.10) states, the Anticipated Variance differs from the unconditional variance given formula (4.1), by a subtracting term that identify the model variance.
The quality framework for the population size estimation could be extended to include all the processes involved in the definition of the BR (such linkage, matching, etc.). As far SR is concerned the

predictive models have to be included. The role of these new elements should be the same of the sampling design in the Anticipated Variance.
The research requires further investigation and an experimental phase to understand the feasibility of a concrete use of the framework.

References

[1] Luzi O., R. Monducci. (2016). The new statistical register "Frame SBS": overview and perspectives. *Rivista di Statistica Ufficiale*, 1/2016, pp. 5-14.
[2] Righi P. Estimation procedure and inference for component totals of the economic aggregates in the "Frame SBS" (2016). *Rivista di Statistica Ufficiale*, 1/2016, pp. 84-97.

[3] Zhang, L.-C. (2012). Topics of Statistical Theory for Register-Based Statistics and Data Integration. *Statistica Neerlandica*, 66: 41–63.

[4] Zhang, L.-C. (2012). On the Accuracy of Register-Based Census Employment Statistics." Paper presented at the European Conference on Quality in Official Statistics, May 30–June 1 2012, Athens. http://www.q2012.gr/articlefiles/sessions/23.4_Zhang_AaccuracyRegisterStatistics.pdf (accessed April 2017).

[5]Isaki, C.T., and Fuller, W.A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

[6] Pfeffermann, D. (2013). Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture. *Journal of Survey Statistics and Methodology*, 3, 425-483.

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

[8] Wolter

[9] Kendall, M.G. and Stuart, A. (1976) *The Advanced Theory of Statistics*. Volume 3. Charles Griffin & Co. Ltd., London & High Wycombe.

[10]Woodruff

Questions

1. Traditionally, the register considers the inclusion of units without uncertainty (i.e. no weights or weights equal to 1). We propose to associate to each unit a $\hat{d}_k$ weight (homogeneous at level g), taking into account over-coverage/under-coverge of the register.
   Is the use of these weights manageable for a complex organization as an NSO?
2. In the register-based statistics, with weights equal to 1, the longitudinal estimates are straightforward. In the proposed approach, for each unit the $\hat{d}_k$ could vary between two occasions. Can we use the $\hat{d}_k$ computed in the last occasion for longitudinal estimates?

   2.a. The use of $\hat{d}_k$ at unit level (person) could be the base to estimate the population size of composite units (households)?
3. . In our approach we propose to use the anticipated variance

$$AV(\widehat{N}_{gL}) = V(\widehat{N}_{gL}|N_R) = E_p E_M\left[\widehat{N}_{gL} - E_P(\widehat{N}_{gL})\right]^2$$

instead of the unconditional variance as proposed by Wolter and or Nirel and Glickman

$$V(\widehat{N}_{gL}) = E_p E_M\left[\widehat{N}_{gL} - E_P E_M(\widehat{N}_{gL})\right]^2$$

Is it correct to follow this measure of variability?