

Generalized estimation procedure for non planned domains based on modified-GREG estimator

Claudia De Vitiis, Paolo Righi

Italian National Statistical Institute (e-mail: devitiis@istat.it, parighi@istat.it)

Abstract: The paper proofs some relevant properties of *modified-GREG* estimator useful when new estimation for unplanned domain are required for a sample survey for which the estimates for some other planned domains have been already published. An application on real survey data is shown.

Keywords: small area estimation, unplanned domain.

1. Introduction

The paper illustrates a procedure to calculate estimates for unplanned domains, functional when the estimates have to be obtained on the basis of sample data deriving from a survey planned to produce estimates for other different domains. The theoretical work context assume that the estimates of the unplanned domains are required after the dissemination of the planned domain estimates. The proposed procedure implements the *modified-GREG* (MGREG) estimator and defines a weighting system of the sampling units with the following relevant properties: (a) for each unplanned domain the estimate is obtainable even if no sampling unit belongs to the domain; (b) the estimates are calibrated with respect to archive known totals for the unplanned domains; (c) for each variable of interest the estimates calculated for the planned domains coincide with the estimate already disseminated.

The paper is organised as follows. Section 2 describes the MGREG estimator and gives proofs of some theoretical properties. Section 3 shows an application on real survey data. Section 4 gives some brief conclusions.

2. Modified-Greg estimator and sample survey application context

Let U be the reference population and $U_1, \dots, U_d, \dots, U_D$ the partition in domains of interest. Let us assume that a sampling strategy (design and estimator), planning the sizes of the sample s_d ($d=1, \dots, D$) for each domain and using the generalized regression estimator, or GREG estimator (Särnadal et al, 1992), has been implemented and the estimates have been published. Given this sampling context, new estimates for some unplanned domains are required. Let $U_{(1)}, \dots, U_{(a)}, \dots, U_{(A)}$ be the new domain partition and $t_{(a)} = \sum_{k \in U_{(a)}} y_k$ the parameters to be estimated, being y_k the value of the interest variable. A general expression of the MGREG estimator (Rao, 2003) is given by

$${}_{mg}\hat{t}_{(a)} = \sum_{k \in U_{(a)}} \tilde{y}_{r,k} + \sum_{k \in s_{(a)}} a_k (y_{r,k} - \tilde{y}_{r,k}) , \quad (2.1)$$

Where: $s_{(a)}$ is the observed unplanned sample for domain $U_{(a)}$, a_k is the base weight and $\tilde{y}_k = \hat{\mathbf{B}}_d' \mathbf{x}_k$ is the value predicted by means of the superpopulation model underling the estimator, $\hat{\mathbf{B}}_d = [\sum_{k \in s_d} \mathbf{x}_k \mathbf{x}_k' a_k]^{-1} \sum_{k \in s_d} \mathbf{x}_k y_k a_k$, where $\mathbf{x}_k$ is the vector of the auxiliary variables known for the sampling units. As $\hat{\mathbf{B}}_d$ depends on the sample s_d , the (2.1) may be decomposed as follows: $mg \hat{t}_{(a)} = \sum_{d \in D} [\sum_{k \in U_{d(a)}} \tilde{y}_k + \sum_{k \in s_{d(a)}} a_k (y_k - \tilde{y}_k)]$, being $U_{d(a)} = U_d \cap U_{(a)}$ and $s_{d(a)} = s_d \cap s_{(a)}$. Finally, with some calculation we obtain an operative formulation of the MGREG estimator $mg \hat{t}_{d(a)}$ for the domain $U_{d(a)}$

$$mg \hat{t}_{(a)} = \sum_{d \in D} \left[\sum_{k \in s_d} w_{k,d(a)} y_k \right] = \sum_{d \in D} mg \hat{t}_{d(a)}, \quad (2.2)$$

being $w_{k,d(a)} = a_k [\delta_{k,(a)} + (t_{\mathbf{x},d(a)} - \hat{t}_{\mathbf{x},d(a)})' (\sum_{k \in s_d} \mathbf{x}_k \mathbf{x}_k' a_k)^{-1} \mathbf{x}_k]$ and $\delta_{k,(a)} = 1$ if unit $k \in U_{(a)}$ and equal to zero otherwise; $\hat{t}_{\mathbf{x},d(a)}$ is the Horvitz- Thompson estimate of the known totals of the auxiliary variables $t_{\mathbf{x},d(a)}$ on the population $U_{d(a)}$. The (2.2) shows that that MGREG estimator assigns to each unit k belonging to the domain d a set of weights $\{w_{k,d(a)}\}$ of cardinality equal to the number of domains $U_{d(a)} \neq \emptyset$.

For the MGREG estimator some relevant properties hold: (i) it produces estimates even if $s_{(a)} = \emptyset$; (ii) the estimates on the auxiliary variables reproduce the known totals on the new domains, $\sum_{d \in D} [\sum_{k \in s_d} w_{k,d(a)} \mathbf{x}_k'] = t_{\mathbf{x}(a)}$, but, for our aims, the most important property is that (iii) it allows to reproduce the same estimates on variables of interest in the planned domains U_d ($d=1, \dots, D$), as shows the following proof:

$$\begin{aligned} mg \hat{t}_d &= \sum_{a \in A} mg \hat{t}_{d(a)} = \\ &= \sum_{a \in A} \left\{ \sum_{k \in s_d} y_k a_k \left[\delta_{k,(a)} + (t_{\mathbf{x},d(a)} - \hat{t}_{\mathbf{x},d(a)})' \left(\sum_{k \in s_d} \mathbf{x}_k \mathbf{x}_k' a_k \right)^{-1} \mathbf{x}_k \right] \right\} = \\ &= \sum_{k \in s_d} y_k a_k \left\{ \sum_{a \in A} \left[\delta_{k,(a)} + (t_{\mathbf{x},d(a)} - \hat{t}_{\mathbf{x},d(a)})' \left(\sum_{k \in s_d} \mathbf{x}_k \mathbf{x}_k' a_k \right)^{-1} \mathbf{x}_k \right] \right\} = \\ &= \sum_{k \in s_d} \left\{ y_k a_k \left[1 + (t_{\mathbf{x},d} - \hat{t}_{\mathbf{x},d})' \left(\sum_{k \in s_d} \mathbf{x}_k \mathbf{x}_k' a_k \right)^{-1} \mathbf{x}_k \right] \right\} = g \hat{t}_d, \end{aligned} \quad (2.3)$$

being ${}_g\hat{t}_d$ the GREG estimator on the domain U_d .

The MGREG estimator is computer intensive as it requires a calibration process for each domain $U_{d(a)}$. This is one of the major drawbacks for its utilization especially in large scale surveys. As a consequence, given a set of auxiliary variables used in the GREG estimator, it could result too large for the MGREG estimator. In this case, it could be necessary a reduction of the calibration constraints and property (2.3) no more holds. Nevertheless, if the excluded auxiliary variables are weakly predictive, the MGREG still reproduces the GREG estimates for the planned domain with good approximation.

3. Application

The MGREG estimator was implemented for the 2005 Small & Medium Enterprises (SME) survey carried out by the Italian Statistical Institute (ISTAT) for producing estimates according to the NACE Rev. 2 classification. This classification of economic activities replaces the classification NACE Rev. 1.1. The application has been performed for an ISTAT research project aiming to produce the estimates according to the new classification, requested by the EU Short Term Statistics Regulation n° 1165/98. Current SME domains are based on Nace Rev. 1.1 according to the Structural Business Survey Regulation n° 58/97 and the sampling strategy has been projected for these different domain partitions (Siesto, 2009). The domains are: the 4-digit economic activity class; the 3-digits economic activity group by size class (defined in terms of number of persons employed); the economic activity division (2-digits) by region. NACE Rev. 2 defines different domains from Nace Rev.1, because enterprises belonging to the same category in Nace Rev.1.1 can be classified in more than one category for the Nace Rev.2. Therefore, domains defined on Nace Rev. 2 are unplanned domains and it could occurs that some of them have no sample units. The MGREG estimator has been usefully implemented and the estimates obtained have been used for the purpose of the EU Regulation.

Here we show the results of the comparison between the MGREG estimator and the Ratio Raking (RR) estimator (Deming and Stephan, 1940). The second estimator is a calibration estimator and is used to compute a set of final weights satisfying the known total of the auxiliary variables (number of enterprises, number of employed persons and turnover) at the same time for the 2-digit Nace Rev. 1.1 domains and the 4-digit Nace Rev. 2 domains (Casciano, 2009). This estimates have been considered only as benchmarking values because the procedure is highly time spending and it doesn't allow to produce estimates for Nace Rev. 2 domains empty of sample units. As far as the MGREG estimator is concerned, the final weights are computed to satisfy the known total of number of enterprises and the number of persons employed for each domain $U_{d(a)}$ (2-digit Nace Rev 1.1 and 4-digit Nace Rev. 2).

Table 1 shows the relative difference between RR e MGREG estimates of turnover and value added at sector level of the Nace Rev. 2. The last line of the table points out that the overall change in the values of the two estimates is very low, 0.3%. The greatest differences emerge for sector D, where MGREG overestimates the turnover of 4% and the value added of 3.2%, and for sector I being the MGREG estimates smaller than RR estimates respectively of 2.5% and 4.2%. However, we may observe that they are residual sector in terms of the two variables. According to the RR estimates, turnover of sector D represents 3.1% of the overall turnover, while the value added is the 1.3%. Sector I produces the 2.7% of the overall turnover and the 3.8% of the value added. Analysing the largest sectors (C, F, G, H-J, L-M-N) producing 88.9%

of the turnover and an 86.4% of the overall value added, RR and MGREG estimators do not present great differences. Finally, the table points out that has not been possible to assign a new 4-digit Nace Rev. 2 for each sample enterprises, and therefore, these units are not included in any sector. According to the RR estimates the non classified units produce the 0.5% of the turnover and the 0.8% of value added.

Table 1: *Relative differences between Ratio Raking and Modified GREG estimates of the turnover and value added at sector level based on Nace Rev. 2*

Sector	Description	Turnover		Value added	
		% value according to the RR estimator	(RR estimate / Megreg estimate)%-100	% value according to the RR estimator	(RR estimate / Megreg estimate)%-100
B	MINING AND QUARRYING	0.3	3.4	0.4	3.4
C	MANUFACTURING	24.2	1.0	26.9	0.8
D	ELECTRICITY, GAS, STEAM AND AIR CONDITIONING SUPPLY	3.1	-4.0	1.3	-3.2
F	CONSTRUCTION	11.6	0.9	13.8	0.5
G	WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	38.4	-0.3	22.4	0.3
I	ACCOMMODATION AND FOOD SERVICE ACTIVITIES	2.7	2.5	3.8	4.2
H-J	TRANSPORTATION AND STORAGE - INFORMATION AND COMMUNICATION	5.8	1.9	7.5	0.2
K	FINANCIAL AND INSURANCE ACTIVITIES	1.5	-0.3	1.2	0.4
L-M-N	REAL ESTATE ACT. - PROFESSIONAL, SCIENTIFIC AND TECHNICAL ACT. -ADMINISTRATIVE AND SUPPORT SERVICE ACT.	8.9	-0.8	15.8	-1.6
P	EDUCATION	0.2	-3.2	0.4	0.2
Q	HUMAN HEALTH AND SOCIAL WORK ACTIVITIES	1.7	-0.5	4.0	-0.6
R-S	ARTS, ENTERTAINMENT AND RECREATION - OTHER SERVICE ACTIVITIES	1.0	3.3	1.7	3.2
*	NO CLASSIFIED SAMPLE	0.5	18.2	0.8	12.3
TOTAL		100.0	0.3	100.0	0.3

4. Concluding remarks

The MGREG estimator discussed in this paper has turned out to be very helpful for dealing with complex estimation contexts related with unplanned domains and classification change affecting the domain definition. The MGREG allows to obtain estimates for the unplanned domains through the construction of a weighting system, without resorting to small area methods which generally do not permit to provide the sample units with weights useful to achieve all the desired estimates respecting simultaneously constraints on auxiliary variables and coherence among estimates.

References

- Casciano C. (2009) Appendice 2. Metodo Basato su Calibrazioni Iterative (Metodo R.R) in: *Rapporto conclusivo delle attività del “Gruppo di Lavoro per la definizione delle metodologie necessarie a gestire le problematiche connesse all’adozione della ATECO 2007”*, to appear.
- Deming, W.E., Stephan, F.F. (1940) On a least squares Adjustment Procedures of a sample frequency table when the expected marginal totals are known, *Annals of Mathematical Statistics*, 11, 427-444.
- Rao J. N. K. (2003) *Small Area Estimation*, Wiley, New York.
- Siesto G. (2009) Stime derivanti dalla rilevazione PMI e impatto della nuova classificazione Nace Rev.2 con riferimento ai regolamenti SBS e STS in: *Rapporto conclusivo delle attività del “Gruppo di Lavoro per la definizione delle metodologie necessarie a gestire le problematiche connesse all’adozione della ATECO 2007”*, to appear.
- Särndal C. E., Swensson B., Wretman, J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.