On bias correction in small area estimation: An M-quantile approach

Bias correction in stima per piccole aree: un approccio M-quantile

Gaia Bertarelli, Francesco Schirripa Spagnolo, Raymond Chambers and David Haziza

Abstract In this paper we propose two bias correction approaches in order to reduce the prediction bias of the robust M-quantile predictors in small area estimation in the presence of representative outliers. A bootstrap procedure is considered for the estimation of the mean squared error. A Monte-Carlo simulation study is conducted. Results confirm that our approaches improve the efficiency and reduce the prediction bias of M-quantile predictors when the population contains units that may be influential if selected in the sample.

Abstract *L'obiettivo del lavoro è quello di proporre due approcci per ridurre l'errore di predizione degli stimatori basati modello di regressione M-quantile nella stima per piccole aree in presenza di outliers rappresentativi . Per valutare la variabilità degli stimatori è utilizzato un approccio bootstrap. Uno studio di simulazione è stato implementato ed i risultati hanno evidenziato che gli approcci proposti migliorano l'efficienza e riducono l'errore di predizione quando la popolazione contiene unità che possono essere influenti se selezionate nel campione.*

Key words: Robust methods; Small Area Estimation; M-quantile

Gaia Bertarelli
Istituto di Management Scuola Universitaria Superiore Sant'Anna, Pisa, Italy, e-mail: gaia.bertarelli@santannapisa.it

Francesco Schirripa Spagnolo
Dipartimento di Economia e Management, Università di Pisa, Pisa, Italy, e-mail: francesco.schirripa@ec.unipi.it

Raymond Chambers
National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, Australia, e-mail: ray@uow.edu.au

David Haziza
Département de mathématiques et de statistique, Université de Montréal, Canada, e-mail: haziza@dms.umontreal.ca

1 Introduction

Outliers occur frequently in sample surveys when the data distribution is highly skewed. Accordingly, to the terminology of [3] sample outliers can be classified into two categories. The first type is the 'non-representative outliers', which are sample elements whose data values are incorrect or they are unique. In this case, they can be identified and removed or corrected before estimation. However, in other cases, sample values associated with the outliers have been correctly recorded and they cannot be considered as unique. These are called 'representative outliers' because they are representative of the non-sampled part of the population; in other words, there is no reason to assume that there are no more similar outliers in the non-sampled part of the population. Such outliers values can seriously affect the survey estimates. Consequently, several methods have been developed in order to mitigate the effects of outliers on survey estimates.

Representative outliers are even more concerning in the small area estimation (SAE) context, where sample sizes are very small and the estimation is often model-based [5]. [4] addressed the issue of outlier robustness in SAE by proposing an M-quantile approach aiming at overcoming the issue of outliers by avoiding the normal assumption.[7] addressed the same issue from the perspective of linear mixed models. Both these approaches use plug-in robust prediction replacing parameter estimates in optimal but outlier-sensitive predictors by outlier robust versions. These predictors are efficient under the correct model but may be sensitive to the presence of outliers because they use plug-in robust prediction which usually leads to a low prediction variance and a considerable prediction bias. [6] and [5] proposed a bias correction method for models with continuous response variables. The main aim of this work is to propose new M-quantile predictors in SAE with correction terms for the bias. Two approaches are studied. The first estimator is a unified approach to M-quantile predictors based on a full bias correction and it could be viewed as a generalization of [3]. The second proposal is developed following the conditional bias approach by [1] and [6].

2 Bias corrected M-quantile-based estimator

Let θ_i be a finite population parameter for area i. That is, θ_i is a well-defined function of the values of a random variable Y associated with the N_i elements of such a small area finite population of interest. For ease of notation, we assume that both Y and θ_i are scalar, and we denote

$$\theta_i = f(\mathbf{y}_{U_i}),$$

where $\mathbf{y}_{U_i}$ denotes the vector of population values of Y for small area i and f is a known function. A basic sample survey inference problem is then one of predicting the value of θ_i give a sample of $n < N$ values from $\mathbf{y}_U$. Without loss of generality we

put $\mathbf{y}_s$ equal to the population sub-vector defined by these values, where s denotes the set of sampled population units. We define (i) $\mathbf{y}_{U_i}$ vector of population values of Y for area i with $U = \bigcup_{i=1}^{m} U_i$ with m is the number of small areas; (ii) $\mathbf{y}_{s_i}$ vector of sampled population values in small area i with $s = \bigcup_{i=1}^{m} s_i$. Suppose that, given $\mathbf{y}_{s_i}$ we can impute the remaining values $\hat{\mathbf{y}}_{U_i}$ denote this imputed vector. A popular method of predicting the unobserved value of θ_i is via the Plug-In Predictor (PIP)

$$\hat{\theta}_i = f(\hat{\mathbf{y}}_{U_i}). \tag{1}$$

Adopting a model-based approach, the empirical PIP for θ_i based on this plug-in approximation is

$$\hat{\theta}_i = f(\mathbf{y}_{s_i}, \{\hat{y}_{ij}^{opt}; j \in U_i - s_i\}) \tag{2}$$

where the set $U_i - s_i$ contains the $N_i - n_i$ indices of the non-sampled units, $\hat{y}_{ij}^{opt} = E[y_{ij}|\mathbf{y}_s; \delta = \hat{\delta}]$ is the plug-in approximation of the minimum mean squared error predictor (MMSEP) of y_{ij}^{opt} for a non-sampled population unit j for area i, and δ is a vector of unkown parameters. The above PIP (2) for small area can be also computed using the M-quantile approach. It can be obtained by using the estimated regression coefficients by M-quantile approach, $\hat{\beta}_\tau$, leading to

$$\hat{\theta}_i^{MQ} = f\Big(\mathbf{y}_{s_i}, \{g^{-1}(\mathbf{x}_{ij}^T \hat{\beta}_{\tau_i}); j \in U_i - s_i\}\Big), \tag{3}$$

where τ_i represents the order of M-quantile for area i. Its computation varies depending on the type of the data.

We propose two small area estimators based on Generalised version of M-quantile regression models.

The first estimator is a unified approach to M-quantile predictors based on a full bias correction. Following [3], the first order approximation to the prediction bias of $\hat{\theta}_i^{MQ}$ is

$$E[\hat{\theta}_i^{MQ} - \theta_i] \simeq \sum_{j \notin s_i} \Big(\frac{\partial f}{\partial y_{ij}}\Big)_{\mathbf{y}_{U_i} = \mathbf{m}_{U_i}} E[\hat{y}_{ij} - y_{ij}] \simeq \sum_{i \in r_j} \Big(\frac{\partial f}{\partial y_{ij}}\Big)_{\mathbf{y}_U = \hat{\mathbf{m}}_{U\bar{q}_j}} \Big(\frac{\partial g^{-1}}{\partial \eta}\Big)_{\eta = \mathbf{x}_{ij}^T \hat{\beta}_{\bar{q}_j}} \mathbf{x}_{ij}^T E[\hat{\beta}_q - \beta_q],$$

The bias corrected robust predictor MQC for the population average of Y in the ith area will be:

$$\theta_i^{MQC} = N_i^{-1} \left(\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{\mu}_{ij} + \sum_{j \in r_i} \Big(\frac{\partial f}{\partial y_{ij}}\Big)_{\mathbf{y}_U = \hat{\mathbf{m}}_{U\bar{q}_j}} \Big(\frac{\partial g^{-1}}{\partial \eta}\Big)_{\eta = \mathbf{x}_{ij}^T \hat{\beta}_{\bar{q}_j}} \mathbf{x}_{ij}^T \hat{\mathbf{B}}_i \right) \tag{4}$$

where $d_{jh\bar{q}_j} = 2\left\{\bar{q}_j I(r_{hj} > 0) + (1 - \bar{q}_j) I(r_{hj} \leq 0)\right\}$ and $\hat{\mathbf{B}}_i$ has to be computed depending of the type of the response variable. If y_{ij} is continuous

$$\hat{\mathbf{B}}_i = \left(\sum_{h=1}^{m} \sum_{j \in s_h} \mathbf{x}_{hj} \hat{d}_{hj} \mathbf{x}_{hj}^T \right)^{-1} \sum_{h=1}^{m} \sum_{j \in s_h} \mathbf{x}_{hj} \hat{d}_{hj} \hat{\sigma}_{hj} \phi \left\{ \frac{y_{hj} - \mathbf{x}_{ij}^T \hat{\beta}_{\tau_i}}{\hat{\sigma}_{hj}} \right\}. \tag{5}$$

The second proposal is developed following the conditional bias approach by [1] and [6]. In a model based approach, the conditional bias attached to unit ij is

$$B_{ij} = E[\hat{\theta} - \theta | s; Y_{ij} = y_{ij}].$$

The prediction error $\hat{\theta}_i - \theta_i$ can be approximated as:

$$\hat{\theta}_i - \theta_i \simeq \sum_{j \in r_i} B_{ij}(I_{ij} = 0) + \sum_{j \in s_i} B_{ij}(I_{ij} = 1). \tag{6}$$

To determine the conditional bias, we need to distinguish two cases, whether the unit belongs to the sample or not. The main problem is that the conditional bias of a non-sampled unit can't be estimated since it depends on the Y-values on the non-sample units, which are not observed. A robust predictor of the mean in the ith area can be expressed as

$$N_i^{-1}\left(\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} g^{-1}(\mathbf{x}_{ij}^T \beta) - \sum_{j \in s_i} B_{ij}(I_{ij} = 1) + \phi\left\{\sum_{j \in s_i} B_{ij}(I_{ij} = 1)\right\}\right)$$

where ϕ is the Huber function. Translating the idea for MQ we have:

$$\theta_i^{MQD} = N_i^{-1}\left(\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} g^{-1}(\mathbf{x}_{ij}^T \beta_{\bar{q}_j}) - \sum_{h=1}^{m} \sum_{j \in s_h} \hat{B}_{jh}(I_{jh} = 1) + \phi\left\{\sum_{h=1}^{m} \sum_{j \in s_h} \hat{B}_{jh}(I_{jh} = 1)\right\}\right). \tag{7}$$

The ϕ-function in MQD depends on a tuning constant c. Using min-max method to compute the optimal tuning constant we obtain

$$\theta_i^{MQD} = N_i^{-1}\left(\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} g^{-1}(\mathbf{x}_{ij}^T \beta_{\bar{q}_j}) - \frac{1}{2}\left(min\left\{B_{jh}(I_{jh} = 1)\right\} + max\left\{B_{jh}(I_{jh} = 1)\right\}\right)\right) \tag{8}$$

where the conditional bias for unit j has to be computed depending of the type of the response variable. If y_{ij} is a continuous

$$\hat{B}_{hj}(I_{hj} = 1) = \sum_{i \notin s_i} \mathbf{x}_{ij}^T \left\{\sum_{h=1}^{m} \sum_{j \in s_h} \mathbf{x}_{hj} \hat{d}_{hj} \mathbf{x}_{hj}^T\right\}^{-1} \hat{d}_{hj} \mathbf{x}_{hj}(y_{hj} - \mathbf{x}_{hj}^T \hat{\beta}_{\tau_i}). \tag{9}$$

3 Model-based simulations

In this section, we provide results regarding model-based simulation scenarios for continuous variables. Following [5], population data are generated from $m = 40$ small areas with samples selected by a simple random sampling without replacement within each area. The population and sample size are the same for all areas and are fixed at $N_i = 100$ and $n_i = 5$. Values for x are generated as i.i.d. from a lognormal distribution with a mean of 1 and a standard deviation of 0.5 on the log scale. Values for Y are generated as $y_{ij} = 100 + 5x_{ij} + u_i + \varepsilon_{ij}$, where i refers to

the areas and j to the population units. The random area and individual effects are independently generated according to the following scenarios:

a) [0,0,0] - no outliers, $u \sim N(0,3)$ and $e \sim N(0,6)$;
b) [e,0,0] - individual outliers only, $u \sim N(0,3)$ and $e \sim \delta N(0,6)+(1-\delta)N(20,150)$; $\delta \sim Ber(0.03)$;
c) [e,u,0] - outliers in both area (fixed) and individual effects , $u \sim N(0,3)$ for areas 1–36, $u \sim N(9,20)$ for areas 37–40 and $e \sim \delta N(0,6)+(1-\delta)N(20,150)$.

Each scenario is independently simulated 1000 times. For each simulation the population values are generated according to the underlying scenario, a sample is selected in each area and the sample data are then used to compute estimates of each of the actual area means for y. Nine different estimators are used for this purpose: the M-quantile estimator MQ by [4] which serves as a reference for the MQ regression based estimators, the bias corrected M-quantile estimator MQBC by [5], the M-quantile estimator based on full bias correction MQC (see equation (4)), the M-quantile estimator based on conditional bias correction MQD (see equation (8)),the standard EBLUP which serves as a reference for all the considered estimators, the robust eblup REBPLUP by [7] and its robust bias corrected version REBLUP–BC by [5], the CBEBLUP and CEBLUP predictorS by [6]. The influence function ϕ that is used in MQBC, MQC, REBLUP BC, CBEBLUP and CEBLUP is a Huber proposal 2 type. For each estimator, we test three different tuning constant for the bias correction part equal to 3, 6 and 9. The performance of the proposed indicators is evaluated according to min-max plots (Figure 1). The values on the x-axis and $y-$axis on plots are:

$$AbsRBias = \frac{\text{Median}\big[\text{AbsB}(\theta_{ki})\big] - \min\big\{\text{Median}\big[\text{AbsB}(\Theta_i)\big]\big\}}{\max\big\{\text{Median}\big[\text{AbsB}(\Theta_i)\big]\big\} - \min\big\{\text{Median}\big[\text{AbsB}(\Theta_i)\big]\big\}}$$

and

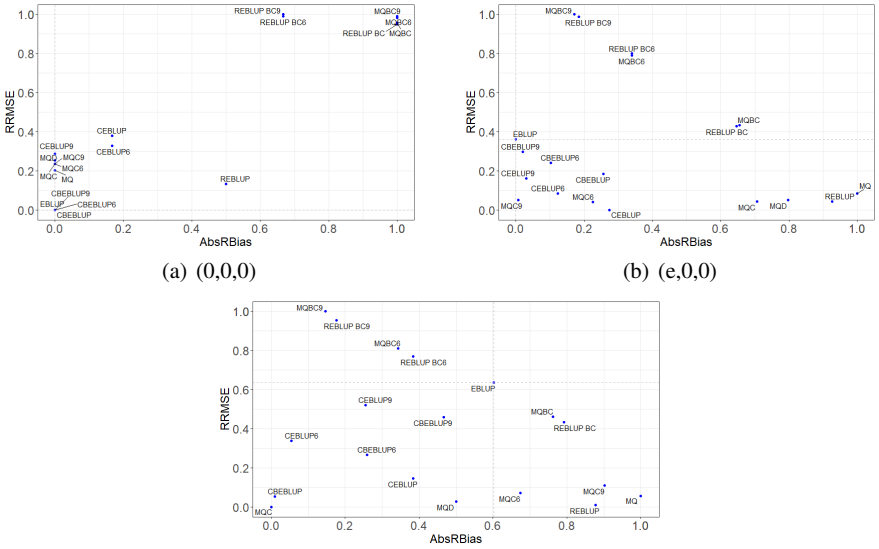
$$RRMSE = \frac{\text{Median}\big[\text{RRMSE}(\theta_{ki})\big] - \min\big\{\text{Median}\big[\text{RRMSE}(\Theta_i)\big]\big\}}{\max\big\{\text{Median}\big[\text{RRMSE}(\Theta_i)\big]\big\} - \min\big\{\text{Median}\big[\text{RRMSE}(\Theta_i)\big]\big\}},$$

where θ_{ki} is the kth estimator in the ith area and Θ_i is the vector all K predictors in area i.

Results confirm our expectations regarding the behaviour of the MQC and MQD estimators. With respect to MQ estimator, the new proposed estimators reduce the bias in the presence of outliers and the variance with only unit-outliers.

We now examine the performance of the MSE estimators. We use the bounded-block-bootstrap [2] for MQC and MQD, with a constant equal to 3 for scenarios a and b and equal to 1.345 for c. Results are reported in table 3.

Scenario	Estimator							
	MQ	MQBC	MQBC6	MQBC9	MQC	MQC6	MQC9	MQD
[0,0,0]	-3.38	-6.85	-5.54	-5.38	2.16	2.12	2.14	2.24
[e,0,0]	-18.01	-8.90	-5.16	-3.77	-2.32	-1.52	-2.26	-3.26
[e,u,0]	-11.09	-8.96	-5.22	-3.83	4.15	-0.53	-2.66	3.51



(a) (0,0,0) (b) (e,0,0)

(c) (e,u,0)

Fig. 1 Min-Max plots for MQ, MQBC, MQC, MQD, EBLUP, REBLUP, REBLUP BC, CBEBLUP and CEBLUP under selected simulation scenarios.

References

[1] Beaumont J.F., Haziza D., Ruiz-Gazen A.: A unified approach to robust estimation in finite population sampling. Biometrika **100(3)**, 555–569 (2013)

[2] Bertarelli G., Chambers, R., Salvati, N.: Outlier robust small domain estimation via bias correction and robust bootstrapping. Stat. Methods Appl. (2020) doi:10.1007/s10260-020-00514-w-

[3] Chambers, R.: Outlier robust finite population estimation. JASA **81(396)**, 1063–1069 (1986)

[4] Chambers, R., Tzavidis, N.: M-quantile models for small area estimation. Biometrika **93(2)**, 255–268 (2006)

[5] Chambers, R., Chandra, H., Salvati, N., Tzavidis, N.: Outlier Robust Small Area Estimation. J. Roy. Stat. Soc. B **76(1)**, 47–69 (2014)

[6] Dongmo-Jiongo, V., Haziza, D., Duchesne, P.: Controlling the bias of robust small area estimators. Biometrika **100(4)**, 843–858 (2013)

[7] Sinha, S. K., Rao, J.: Robust small area estimation. Canadian Journal of Statistics **37(3)**, 381–399 (2009)

[8] Chambers, R., Salvati, N. and Tzavidis, N.: Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the uk. Journal of the Royal Statistical Society: Series A **179 (2)**, 453-479 (2016)

The address component of the Statistical Base Register of Territorial Entities

La componente indirizzi del Registro Statistico Base dei Luoghi (RSBL)

D. Fardelli, E. Orsini, A. Pagano[1]

Abstract The new Statistical Base Register of Territorial and geographical entities of ISTAT (RSBL) is a multidimensional register integrating several components addresses with geographic coordinates, micro-zones and census blocks, buildings and housing units, administrative zones, statistical and functional zones. All components will be integrated with other components according hierarchical and geographical principles. The RSBL, with the other Registers of Institute, will provide a bridge between the statistical units, such as individuals, families and economic. In this paper, we present the component of addresses of RSBL, illustrating his structure, his process and some preliminary results on the data contained in it about the geocoding and georeferencing.

Abstract *Il Registro Statistico di Base dei Luoghi (RSBL) è il pilastro di tutte le attività che prevedono la georeferenziazione delle informazioni statistiche contenute negli altri registri o raccolte attraverso le indagini. La componente indirizzi di RSBL acquisisce le informazioni relative agli indirizzi derivanti dal progetto ANNCSU e attraverso il processo di integrazione, implementa al suo interno gli indirizzi di diversi archivi amministrativi (Lista Anagrafica, Anagrafe Tributaria, Catasto, Asia). Ogni indirizzo sarà corredato, oltre ad indicatori di qualità dedicati, di coordinata geografica e di sezione di censimento. Tale infrastruttura permetterà la georeferenziazione e la geocodifica alla sezione di censimento e alla griglia regolare delle unità statistiche (individui, famiglie, unità locali, etc.).*

Key words: Register, Integration, Georeferencing, Geocoding, Addresses, Geographic Coordinates, Harmonization, Geospatial integration.

[1] Davide Fardelli, ISTAT; fardelli@istat.it:
Enrico Orsini, ISTAT; eorsini@istat.it:
Andrea Pagano, ISTAT; andreapagano@istat.it.

1 Introduction

One of the pillars of NSIs modernization program is the system of integrated statistical registers as the basis for surveys and statistical production; this system has been denominated Integrated System of Statistical Registers (ISSR). The ISSR integrates information relating to: (i) individuals, families and cohabitation; (ii) economic units; (iii) places; and (iv) activities [4] [9]. In this system, every unit referred to places will be inside the Statistical Base Register of Territorial Entities (RSBL). The base registers are connected by codes and are maintained updated over time using mainly administrative sources [8]. According to Global Statistical Geospatial Framework (GSCF) [7] [3], the geospatial information is an important data sources for statistics. Therefore, RSBL will assume a dual role: (i) georeferencing and/or geocoding the statistical units (demographic/economic) [2] and (ii) spatial data production (e.g. surfaces, altitudes, distances, contiguities, statistics on buildings, population grid, etc.). The RSBL has been designed like a multidimensional register integrating several components with heterogeneous nature: addresses with geographic coordinates, micro-zones (old census block), buildings and dwellings units, administrative zones and statistical and functional zones. All components will be integrated with other components according hierarchical and geographical principles. The RSBL, with the other Registers, also will provide a bridge between the statistical units, such as individuals, families and economic. They will be geocoded at census block or regular grid. In this paper, we will illustrate the addresses component of RSBL. It has been released in a prototypal form in 2018 and a new release in 2020. The RSBL has been used like sample frame for the permanent census of population. The aim is to build the register only once and to keep it updated in time.

2 The structure of RSBL-Addresses

The addresses component of RSBL should include all the addresses existence on national territory. Every address will be admitted and identified in RSBL with a unified address code (CUI). The attribution of a code will simplify the integration with other registers, and the code will avoid errors of linkage, due at several form of strings of addresses. Every CUI will have a geographic coordinate and/or census block geocoded. The geographic information is always accompanied with quality indicators both of coordinate and of geocoding.

The innovation of RSBL is the integration of addresses from several administrative and geographical archives. The statistic unit in this component is the address interpreted like the direct or indirect access, from a street to a housing unit or other units like economic activities.

The National Archive of Addresses of Urban Streets (ANNCSU), born in 2012, is the first archive to populate RSBL. It is the primary source of register and it is considered like a benchmark, because is provided straight by municipality. It represents the administrative archive, which contains streets and addresses for the entire country