A two-step selective editing procedure based on contamination models

Marco Di Zio[1] *Ugo Guarnera*[1]

Sommario

Ilves and Laitila (2009) propongono una procedura in due fasi per l'editing selettivo. Il loro approccio prevede, oltre alla selezione delle unità affette da possibili errori influenti, anche l'estrazione di un campione dalle rimanenti unità al fine di rimuovere l'eventuale distorsione residua. In questo articolo viene studiato l'uso del modello di contaminazione implementato in SeleMix (Buglielli and Guarnera, 2011) nella suddetta procedura a due fasi. Viene effettuato uno studio di valutazione sui dati dell'indagine Istat sulle piccole e medie imprese del 2008, con errori simulati in base ad alcuni meccanismi frequentemente incontratri nel contesto delle indagini negli Istituti di Statistica.

Parole Chiave: Controllo e correzione dei dati, Errori influenti, Modelli mistura, Modelli a classi latenti, Funzioni punteggio

Abstract

Ilves and Laitila (2009) propose a two-step procedure for selective editing. According to their approach, in addition to the units selected as affected by influential errors, a sample from the remaining observations is drawn in order to remove the possible residual bias. In this paper, the use of a contamination model as implemented in SeleMix (Buglielli and Guarnera, 2011) in the two-step procedure is studied. An evaluation study is performed by using data from 2008 Istat survey on small and medium enterprises and by simulating errors based on some mechanisms frequently met in NSI surveys.

Keywords: Data editing, Influential errors, Mixture models, Latent class models, Score function

1. INTRODUCTION

In the last years, it has been accepted the idea that only a small subset of observations is affected by errors having a high impact on the estimates, while the rest of

[1] Istat, Integration, Quality, Research and Production Networks Development Department. email: dizio@istat.it, guarnera@istat.it. The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat

the observations are not contaminated or contain errors having small impact on the estimates. This assumption and the fact that interactive editing procedures, like for instance recontact of respondents, are resource demanding, have motivated the idea at the basis of selective editing, that is to look for important errors (errors with an harmful impact on estimates) in order to focus the expensive interactive treatments (follow up, recontact) only on this subset of units. This should reduce the cost of the editing phase maintaining at the same time an acceptable level of quality of estimates.

Although the connection of selective editing with the estimation phase is evident, most of the papers deepen the editing aspect of such a procedure disregarding the inferential feature of selective editing. Some exception can be found in literature. Ilves and Laitila (2009) and Ilves (2010) see selective editing as a part of an estimation process aimed to reduce the impact of measurement error on the final estimates. They propose a two-step procedure for selective editing. Their proposal is motivated by the fact that the non-selected observations may still be affected by errors resulting in a biased target parameter estimator. To obtain an unbiased estimator, a sub-sample is drawn from the unedited observations, follow-up activities with recontacts are carried through and the bias due to remaining errors is estimated. The bias estimate is used to make the target parameter estimator unbiased.

Other papers focusing on the inferential aspect of selective editing are those by Buglielli *et al*. (2011) and Di Zio and Guarnera (2011). In these papers a model-based approach is proposed: log-normal data are supposed to be affected by errors according to a contamination model. With this approach it is possible to estimate the expected error affecting data and consequently it is allowed to estimate the impact of the residual error on the target estimates after editing the subset of selected units. In this setting the score function is based on the estimated expected error. The procedure is implemented in the R package *SeleMix*.

In this paper we study the joint use of the two approaches, that is to apply the two-step procedure proposed by Ilves and Laitila and to draw the sample in the second phase for removing the bias according to a sampling design with inclusion probabilities proportional to the scores (expected error) computed by using the contamination model proposed by Buglielli *et al*. (2011). The use of expected errors for sampling may give a more efficient strategy for removing the bias from the final estimates. An evaluation study is performed by using data from the 2008 Istat survey on small and medium enterprises and by simulating errors based on some mechanisms frequently encountered in the NSI surveys.

The paper is structured as follows. Section 2. shortly describes selective editing. The two-step selective editing approach is illustrated in Section 3.. Section 4. illustrates the contamination model used for selective editing as implemented in SeleMix. The experiments are described in Section 5., while results and conlusions are discussed in Section 6..

2. Selective editing

Selective editing is based on the idea of looking for important errors in order to focus the treatment on the corresponding subset of units to reduce the cost of the

editing phase, while maintaining the desired level of quality of estimates. In practice, observations are ranked according to the values of a score function expressing the impact of their potential errors on the target estimates (Latouche and Berthelot 1992), and all the units with a score above a given threshold are selected.

The score function is a tool to prioritise observations according to the expected benefit of their correction on the target estimates. According to this definition, it is natural to think of the score function as an estimate of the error affecting data. The estimate is generally based on comparing observed values with predictions (sometimes called *anticipated values*) obtained from some explicit or implicit model for the data. In the case of sample surveys, the comparison should also include sampling weights in order to properly take into account the error impact on the estimates.

An additional element often considered in the context of selective editing, is the degree of suspiciousness, that is an indicator measuring, loosely speaking, the probability of being in error. The necessity of this element arises from the implicit assumption of the intermittent nature of the error in survey data, i.e., the assumption that only a certain proportion of the data are affected by error, or, from a probabilistic perspective, that each measured value has a certain probability of being erroneous. Some authors do not introduce this element, others implicitly use it in their proposals. Norberg *et al.* (2010) state that several case studies indicate that procedures based only on the comparison of observed and predicted values without the use of a degree of suspiciousness tend to generate a large proportion of false alarm. Several score functions are proposed in literature, the difference being mainly given by the kind of prediction and the use of *degree of suspiciousness*. Among the different methods used to obtain predictions it is worthwhile to mention the use of information gathered in a previous occasion of the survey (Latouche and Berthelot 1992), regression models (Norberg *et al.*, 2010), contamination models (Buglielli *et al*. 2011). A detailed review can be found in De Waal (2011).

As far as the degree of suspiciousness is concerned, a common drastic approach consists in introducing it in the score function through a zero-one indicator that multiplies the difference between observed and predicted values, where zero and one correspond to consistency or inconsistency respectively with respect to some edit rules. In this case it is assumed that errors appear only as edit failures and observations that pass the edits are considered error-free without uncertainty (Latouche and Berthelot 1992). More refined methods to estimate the probability of being in error can be found in Norberg *et al.* (2010) and Buglielli *et al*. (2011).

Prediction and suspiciousness can be combined to form a score for a single variable, named *local score*. A local score frequently used for the unit i with respect to the variable Y_j is

$$S_{ij} = \frac{p_i w_i |y_{ij} - \hat{y}_{ij}|}{\hat{T}_j}$$

where p_i is the degree of suspiciousness, y_{ij} is the observed value of the variable Y_j on the ith unit, $\hat{y}_{ij}$ is the corresponding prediction, w_i is the sampling weight, and $\hat{T}_j$ is an estimate of the target parameter. Once the local scores for the variables of interest are computed, a global score to prioritise observations is needed. Several

functions can be used to obtain the global score, see Hedlin (2008), for instance the maximum of the local scores $GS_i^{\infty} = max_j S_{ij}$.

Once the observations have been ordered according to their global score, a threshold should be chosen in order to select the subset of units to be edited such that the impact on the target estimates of the errors remaining in the unedited observations is negligible.

3. Probabilistic selective editing under a two-phase sampling approach

Ilves and Laitila (2009) incorporate the selective editing in a two-phase sampling approach in order to obtain an unbiased estimator also with respect to measurement error. More in detail, let $U = 1, 2, \ldots, N$ be a finite population from which a sample s_a of size n_a is drawn according to a sample design $p_a(\cdot)$. Let us assume that the observed variable Y in the sample s_a is possibly affected by a measurement error. The target is the estimation of the population total $T_{y^*} = \sum_{k \in U} y_k^*$, where y^* are the true values corresponding to the observed and possibly contaminated y values. The Horvitz-Thompson (HT) estimator computed on the observed values y_k, for $k \in s_a$, is

$$\hat{t}_y = \sum_{k \in s_a} \frac{y_k}{\pi_{ak}}$$

where π_{ak} are the first order inclusion probabilities. The HT estimator is unbiased for the total, however because of measurement errors, $\hat{t}_y$ can be a biased estimator of the target total T_{y^*}.

When selective editing is performed, n_{ed} units of the sample s_a are recontacted and for them the true value is supposed to be recovered and finally considered instead of the corresponding observed values for the estimate. This task is carried out in order to limit the impact of measurement errors on the accuracy of the final estimates. Nevertheless, the selective editing procedure may not be perfect and some errors biasing the estimates may still remain in data. The idea is to estimate the residual bias still present in the unedited units and to remove it from the HT estimator computed on the final data, say $\tilde{y}$, composed of n_{ed} edited observations and $n_a - n_{ed}$ unedited observations. To this aim, a subsample s_b of size n_b is drawn from the unedited observations of the sample s_a, and all the units in s_b are edited in order to compute the error $e_k = (y_k - y_k^*)$ for each unit. An unbiased estimator for T_{y^*} is obtained by subtracting the estimated bias from the biased total estimate

$$\hat{t}_{\tilde{y}} = \sum_{k \in s_a} \frac{\tilde{y}_k}{\pi_{ak}} - \sum_{k \in s_b} \frac{e_k}{\pi_{ak}\pi_{bk}} \tag{1}$$

where π_{bk} is the first order inclusion probability for the unit k in the second phase sample s_b.

4. Selective editing through contamination models

The key elements for selective editing via contamination models are:

1. specification of a parametric model for the true (non-contaminated) data,
2. specification of an error model.

We assume that two sets of variables are observed: the variables of the first group, say X-variables, are assumed to be correctly measured while the second set of variables, say Y-variables, corresponds to items possibly affected by measurement errors. In this set-up, which can be useful when some variables are available from administrative sources or are measured with high accuracy, it is quite natural to treat the variables that are observed with error as response variables and the reliable variables as covariates. In the following we model true data through a normal probability distribution. They allow to derive the distribution of the true data conditional on the observed data. This distribution is central in the proposed selective editing method. We remark that, a model is also studied for the case when no auxiliary variables X are available, details can be found in Buglielli *et al.* (2011).

An important point is that the model specification reflects the intermittent nature of the error mechanism. This means that errors are assumed to affect only a subset of data, or in other words, each unit in the dataset is affected by an error with an (unknown) a priori probability p. The assumption of intermittent error, which is very common in the context of survey data treatment, naturally leads to the model specification of the error model in terms of a mixture of probability distributions. As a consequence, the observed data distribution is also a mixture whose components correspond to error-free and contaminated data respectively. Such models are often referred to as contamination models and are commonly applied in the context of outlier identification. In the following, the model is described in some detail.

4.1 True data model

True data corresponding to possible contaminated items are represented as a $n \times m$ matrix Y^* of n independent realizations from a random m-vector assumed to follow a Gaussian distribution whose parameters may depend on some set of q covariates not affected by error. Thus, we have the regression model:

$$Y^* = XB + U \tag{2}$$

where Y^* is the $n \times p$ true data matrix, X is a $n \times q$ matrix whose rows are the measures of the q covariates on the n units, B is the $q \times m$ matrix of the coefficients, and U is the $n \times p$ matrix of normal residuals:

$$U \sim N(0, \Sigma). \tag{3}$$

4.2 Error model

In order to model the intermittent nature of the error mechanism we introduce a Bernoulli r.v. I with parameter p, where $I = 1$ if an error occurs and $I = 0$ otherwise. In the sequel, Y will denote possible contaminated variables. Thus, given that $I = 0$, it must hold $Y = Y^*$. Furthermore, given that $I = 1$, errors affect data through an additive mechanism represented by a Gaussian r.v. with zero mean and covariance matrix Σ_ϵ proportional to Σ, i.e., given $I = 1$:

$$Y = Y^* + \epsilon, \quad \epsilon = N(0, \Sigma_\epsilon), \quad \Sigma_\epsilon = (\lambda - 1)\Sigma, \quad \lambda > 1.$$

It is convenient to represent the error model through the conditional distribution:

$$f_{Y|Y^*}(y|y^*) = (1 - p)\delta(y - y^*) + pN(y; y^*, \Sigma_\epsilon) \tag{4}$$

where p (mixing weight) is the *a priori* probability of contamination and $\delta(t' - t)$ is the delta-function with mass at t. In case that the set of X-variates is empty, the variables Y_i $(i = 1, \ldots, n)$ are normally distributed with common mean vector μ. It is worthwhile noting that, due to the intermittent error assumption, it is conceptually possible to think of data as partitioned into correct and erroneous, and to estimate, for each observation, the probability of being correct or corrupted. The distribution of the observed data is easily derived multiplying the normal density for the true data implied by (2) and (3) and the error density (4), and integrating over Y^*:

$$f_Y(y) = (1 - p)N(y; B'X, \Sigma) + pN(y; B'X, \lambda\Sigma) \tag{5}$$

The distribution (5) refers to observed data and can be easily estimated by maximizing the likelihood based on n sample units via an ECM algorithm.

4.3 Score function and threshold

In order to define the score function for selective editing we derive the distribution of the error-free data Y^* conditional on observed data (including covariates X). A straightforward application of the Bayes formula provides:

$$f(y_i^*|y_i) = \tau_1(y_i)\delta(y_i^* - y_i) + \tau_2(y_i)N(y_i^*; \tilde{\mu}_i, \tilde{\Sigma}) \tag{6}$$

where

$$\tilde{\mu}_i = \frac{(y_i + (\lambda - 1)\mu_i)}{\lambda}; \quad \tilde{\Sigma} = \left(1 - \frac{1}{\lambda}\right)\Sigma,$$

$\delta(y_i^* - y_i)$ is the delta function with mass at y_i, and $\tau_1(y_i)$, $\tau_2(y_i)$ are the posterior probabilities that a unit with observed values y_i belongs to correct and erroneous data group respectively:

$$\begin{aligned}\tau_1(y_i) &= Pr(y_i = y_i^*|y_i) = \frac{(1-p)N(y_i;\mu_i,\Sigma)}{(1-p)N(y_i;\mu_i,\Sigma)+pN(y_i;\mu_i,\lambda\Sigma)},\\ \tau_2(y_i) &= Pr(y_i \neq y_i^*|y_i) = 1-\tau_1(y_i),\\ i &= 1,\ldots,n.\end{aligned}$$

It is natural to define predictions $\hat{y}_i$ as estimates of the expected errors $E(y_i^*|y_i)$. From (6) it follows:

$$E(y_i^*|y_i) = \tau_1(y_i)y_i + \tau_2(y_i)\tilde{\mu}_i, \quad i < 0, 1, \ldots, n. \tag{7}$$

Predictions can be obtained by replacing the parameters in formula (7) with their corresponding estimates.

It is worthwhile to remark that in the context of economic surveys, when positive variables are to be analyzed, logarithms of data, instead of data in their original scale, are often modeled through a Gaussian distribution. The previous methodology can be easily adapted to the lognormal case.

Given the predictions for each unit of a dataset, an appropriate score function can be defined in terms of the expected error: $y_i - \hat{y}_i = \tau_2(y_i)(y_i - \hat{\tilde{\mu}}_i)$, where $\hat{\tilde{\mu}}_i$ is an estimate of $\tilde{\mu}_i$. We provide details for the univariate case. Let us suppose the target aggregate to estimate is the total T_{y^*} of the variable $Y*$, and let $\hat{t}_{y^*} = \sum_{i=1}^n w_i y_i^*$ be the corresponding estimator based on true values. Let us define the relative individual error for the ith unit with respect to the variable Y^* as the ratio between the (weighted) expected error and an estimate $\hat{T}_{y^*}$ of the target parameter, that is

$$r_i = \frac{w_i(y_i - \hat{y}_i)}{\hat{T}_{y^*}}. \tag{8}$$

The score function is simply defined as $S_i = |r_i|$. Moreover, based on error predictions, the expected residual error in the unedited data can also be computed. More precisely, we define the residual error remaining in data after editing the i units with the highest score as:

$$R_i = \Big|\sum_{k>i}^{n} r_k\Big|.$$

The previous definitions allow to relate the number of units to select for interactive editing to the desired level of accuracy for the target estimates. In fact, once an accuracy level (threshold) η is chosen, the selective editing procedure consists of:

1. sorting the observations in descending order according to the value of S_i;
2. find $n_{ed} \equiv n_{ed}(\eta)$ such that $n_{ed} = min\{k^* \in (0, 1, \ldots, n) \mid R_k < \eta, \ \forall k \geq k^*\}$, i.e., select the first n_{ed} units such that, all the residual errors R_k computed from the $(n_{ed}+1)$th to the last observation are below η. .

The algorithm so far described is easily extended to the multivariate case by defining a global score function in terms of the local score functions for the different variables, see Di Zio and Guarnera (2011).

The parameters involved in the computation of (8) are estimated through the ECM algorithm, while a robust estimate of T^* can be obtained by using the preditions $\hat{y}_i$,

$$\hat{T}_{y^*} = \sum_i w_i \hat{y}_i.$$

5. Experiments

In this section we describe an experimental application where selective editing based on SeleMix is jointly used with the two-step estimation procedure proposed by Ilves and Laitila. According to their approach, units that have not been selected for interactive editing are subsampled and the second phase sample is used to estimate the bias associated with measurement errors remaining in data. Selective edititing is based on the contamination model approach described in (Buglielli *et al.* 2011) and implemented in the R-package *SeleMix*. Moreover, as described in the following, the score function in SeleMix is also used in some of the analysed estimation methods for the second phase sampling.

We have conducted the experiments on data from the 2008 Istat *survey on small and medium enterprises*. In particular we have considered enterprises in the Nace Rev2 sections B, C, D and E corresponding to aggregation of economic activities in *Manifacturing, mining and quarrying and other industry*. This group of units ($N = 8723$) has been used in the experiment as reference population (U) and for this population the variables *turnover* (X) and *labour cost* (Y) have been used assuming that the available data are error-free. Errors are artificially introduced in the Y variable according to error mechanisms frequently encountered in the context of NSI surveys, they are explicitely described in the next paragraphs. We suppose that the population parameter to be estimated is the total of the variable Y. The variable turnover is used as a covariate in the contamination model to obtain predictions for (Y).

A Monte Carlo study based on 2000 iterations has been carried out in order to study the impact of the use of a contamination model in the two-step procedure. We study the situations where the number of recontacts cannot exceed a certain amount n_{rec} determined by budget constraints. Hence, in the following, n_{rec} is kept fixed.

Each iteration of the Monte Carlo experiment consists of the following steps:

1. ***Sampling***
   a simple random sample without replacement (srswor) s_a of $n_a = 1000$ observations is extracted from the target population U

2. ***Data contamination***
   errors on the variable Y are artificially introduced according to the following mechanisms:

- Multiply Y values by 10, $(err.10)$,
- Multiply Y values by 100, $(err.100)$,
- Multiply Y values by 1000, $(err.1000)$,
- inversion of the first two digits, $(inv.first)$,
- inversion of the last two digits, $(inv.last)$,
- replacement of the reported value with the value "1", $(err.one)$.

3. ***Model estimation and score computation***
   SeleMix is used to estimate a contamination model and to assign scores according to (8) to each unit. Records are accordingly ordered.

4. ***Selective editing***
   The observed values of the first n_{ed} observations are replaced by the correspondig true values. Three cases are analysed:
   - $n_{ed} = n_{rec}$, all units are edited;
   - $n_{ed} = 0$, no units are edited;
   - $n_{ed} = n_{th}$ where n_{th} is the number of units selected by SeleMix corresponding to a level of accuracy parameter equal to 0.01.

5. ***Second-phase sampling***
   Two subsamples $s_b^{(1)}$, $s_b^{(2)}$ of $n_b = n_{rec} - n_{ed}$ units are extracted from the $n_a - n_{rec}$ unedited data using 1) srswor and 2) sampling with inclusion probabilities proportional to the scores (8). For each sampled unit the difference y_k and y_k^* between the observed and the true value of the variable Y is computed.

6. ***Estimation***
   Different estimators are used to estimate the total of variable Y, and the corresponding errors are computed by comparing the estimates with the true population value of the total. The estimators are described below.

As benchmark estimator the Horwitz-Thompson estimator based on the true values of Y^* ($\hat{t}_{y^*}$) in the sample s_a is used:

$$\hat{t}_{y^*} = \frac{N}{n_a} \sum_{k \in s_a} y_k^*.$$

The corresponding HT estimator $\hat{t}_y$ based on observed unedited data is defined analogously:

$$\hat{t}_y = \frac{N}{n_a} \sum_{k \in s_a} y_k.$$

Estimators based on both edited and sampled data are also computed. According to the cases introduced in step (4), three situations are analysed:

1. all the n_{rec} units are edited and no unit is subsampled (estimator $\hat{t}^{SE}$)
2. no unit is edited and all the n_{rec} observations are subsampled and used for bias correction, ($\hat{t}^{SP1}$ and $\hat{t}^{SP2}$; corresponding to the SRSWOR and PPS sampling respectively);
3. n_{th} units selected by SeleMix at a level of accuracy equal to 0.01 are edited, while $n_{rec} - n_{th}$ observations are subsampled ($\hat{t}^{SM1}$ and $\hat{t}^{SM2}$ corresponding to the SRSWOR and PPS sampling respectively).

We remark that, the estimator $\hat{t}^{SE}$ does not include the bias correction term and is defined as:

$$\hat{t}^{SE} = \frac{N}{n_a} \sum_{k \in s_a} \tilde{y}_k = \frac{N}{n_a} \sum_{k \in E} y_k^* + \frac{N}{n_a} \sum_{k \in s_a \setminus E} y_k, \tag{9}$$

where E is the set composed of the n_{ed} edited units.

The other estimators can be expressed according to formula (1) by using the appropriate inclusion probabilities.

We remark that for $\hat{t}^{SP1}$ and $\hat{t}^{SP2}$ the first term in (1) is computed on the observed unedited data

$$\sum_{k \in s_a} \frac{\tilde{y}_k}{\pi_{ak}} = \frac{N}{n_a} \sum_{k \in s_a} y_k,$$

while for $\hat{t}^{SM1}$ and $\hat{t}^{SM2}$ the first term in (1) is analogous to the one in formula (9)

$$\sum_{k \in s_a} \frac{\tilde{y}_k}{\pi_{ak}} = \frac{N}{n_a} \sum_{k \in E} y_k^* + \frac{N}{n_a} \sum_{k \in s_a \setminus E} y_k$$

but the set E is composed of the n_{th} units selected by SeleMix.

6. Results and conclusions

The results of two experiments ($Ex1, Ex2$) are reported in Table (6.). Estimators are evaluated through the empirical relative root mean squared error (RRMSE) and the empirical relative bias (RB).

The incidence of errors is the same in the two experiments for the following error mechanisms: $err.1000$ (0.5%), $err.100$ (1%), $inv.first$ (1%), $inv.last$ (2%), $err.one$ (1%).

The error parameter varying in the two experiments is only $err.10$ that in $Ex1$ is not introduced at all, while in $Ex2$ is $err.10$ (0.15%). These different settings are introduced to reproduce the following situations:

- target estimates are mainly affected by errors caused by outliers, $Ex1$.
- target estimates are due to errors caused by both outliers and inliers, $Ex2$.

These two situations are analyzed at different number of recontacted units (n_{rec}) to assess the behaviour of the different estimators when a low number of units can be recontacted ($n_{rec} = 30$ for $Ex1$), and when a higher number of recontacts is

allowed ($n_{rec} = 150$ for $Ex2$). We remark that for $Ex1$ the estimators based on a combination of selective editing and the two-phase sampling strategy ($\hat{t}^{SM1}$, $\hat{t}^{SM2}$) are not evaluated because of the low number of edited units.

Table 1 - RRMSE and RB of the anlysed estimators based on selective editing and a two-phase sampling

Experiment		$\hat{t}_{y^*}$	$\hat{t}_y$	$\hat{t}^{SE}$	$\hat{t}^{SP1}$	$\hat{t}^{SP2}$	$\hat{t}^{SM1}$	$\hat{t}^{SM2}$
Ex1	RRMSE%	4	611	4	1649	5	-	-
	RB%	0	545	0	0	0	-	-
Ex2	RRMSE%	4	837	13	999	10	33	10
	RB%	0	743	11	1	0	1	0

The first comment concerns the sampling design for bias correction. In both the experiments the estimator based on a PPS sampling, where the inclusion probabilities are proportional to the scores provided by SeleMix, is much more efficient than the estimator based on SRSWOR.

When the accuracy of estimates is mainly affected by outliers ($Ex1$), the selective editing procedure is able to remove the bias, and the RRMSE is almost the same than that obtained by using true data. In this situation the estimator $\hat{t}^{SE}$ overperforms the other estimator $\hat{t}^{SP2}$ whose RRMSE is dominated by a high variability.

When the accuracy of the estimates is also affected by inliers ($Ex2$), the estimator $\hat{t}^{SE}$ is strongly biased (the main component of the RRMSE). The estimators based on sub-sampling are all able to remove the bias, even though they are characterised by a strong variability that makes the RRMSE close to the one obtained with $\hat{t}^{SE}$.

The results emphasize that an optimal strategy should be based on an accurate analysis of the trade-off between variance and bias of estimators. In fact, although the estimators based only on selective editing can be seriously biased, at level of MSE they are still comparable to the estimators based on a two-phase sampling, in fact the advantage due to the bias reduction is less appreciable because of the increase of the variance.

References

Buglielli, T., and Di Zio, M., and Guarnera, U., and Pogelli, F.R., (2011). “Selective Editing of Business Survey Data Based on Contamination Models: an Experimental Application." *Proceedings of NTTS 2011 New Techniques and Technologies for Statistics, Bruxelles, 22-24 February 2011.*

Buglielli, T., Guarnera, U., (2011). *S*eleMix: Selective Editing via Mixture models. R package version 0.8.1. http:/CRAN.R-project.orgpackage=SeleMix

De Waal, T., and Pannekoek, J., and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*, Wiley.

Di Zio, M., and Guarnera, U., (2011). “SeleMix: an R Package for Selective Editing via Contamination Models", *Proceedings of the 2011 International Methodology Symposium, Statistics Canada. November 1-4, 2011, Ottawa, Canada*.

Hedlin, D. (2008). “Local and global score functions in selective editing". *UN/ECE Work Session on Statistical Data Editing, Wien*.

Ilves, M. and Laitila, T. (2009). “Probability-Sampling approach to Editing." *Austrian Journal of Statistics*, Vol. 38, no. 3, 171-182.

Ilves, M. (2010). “Probabilistic approach to editing." *Workshop on Survey Sampling Theory and Methodology Vilnius, Lithuania, August 23-27, 2010.*

Latouche, M., and Berthelot, J.M. (1992). “Use of a Score Function To Prioritise and Limit Recontacts in Business Surveys", *Journal of Official Statistics*, Vol. 8, no. 3, 389-400.

Norberg, A., and Adolfsson, C., and Arvidson, G., and Gidlund, P., and Nordberg, L. (2008). “A General Methodology for Selective Data Editing". *Statistics Sweden*.

Norme redazionali

La Rivista di statistica ufficiale pubblica contributi originali nella sezione "Temi trattati" ed eventuali discussioni a largo spettro nella sezione "Interventi". Possono essere pubblicati articoli oggetto di comunicazioni a convegni, riportandone il riferimento specifico. Gli articoli devono essere fatti pervenire al Comitato di redazione delle pubblicazioni scientifiche corredati da una nota informativa dell'autore contenente attività, qualifica, indirizzo, recapiti e autorizzazione alla pubblicazione. Ogni articolo prima della pubblicazione dovrà ricevere il parere favorevole di due referenti scelti tra gli esperti dei diversi temi affrontati.

Per l'impaginazione dei lavori gli autori sono tenuti a conformarsi rigorosamente agli standard editoriali fissati dal Comitato di redazione e contenuti nel file RSU stili o nella classe LaTex, entrambi disponibili on line. La lunghezza dei contributi originali per entrambe le sezioni dovrà essere limitata entro le 35 pagine. Una volta che il lavoro abbia superato il vaglio per la pubblicazione, gli autori sono tenuti ad allegare in formato originale tavole e grafici presenti nel contributo, al fine di facilitare l'iter di impaginazione e stampa. Per gli standard da adottare nella stesura della bibliografia si rimanda alle indicazioni presenti nel file on line.

Tutti i lavori devono essere corredati di un sommario nella lingua in cui sono redatti (non più di 120 parole); quelli in italiano dovranno prevedere anche un abstract in inglese.

Nel testo dovrà essere di norma utilizzato il corsivo per quei termini o locuzioni che si vogliano porre in particolare evidenza (non vanno adoperati, per tali scopi, il maiuscolo, la sottolineatura o altro).

Gli articoli pubblicati impegnano esclusivamente gli autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

Per contattare la redazione o per inviare lavori: rivista@istat.it. Oppure scrivere a:
Segreteria del Comitato di redazione delle pubblicazioni scientifiche
all'attenzione di Gilda Sonetti
Istat
Via Cesare Balbo, 16
00184 Roma