WP.22
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (iv): New and emerging methods

CONTAMINATION MODELS FOR THE DETECTION OF OUTLIERS AND INFLUENTIAL ERRORS IN CONTINUOUS MULTIVARIATE DATA

Submitted by ISTAT[1]

I. INTRODUCTION

Particularly in economic survey data, it is not necessary to identify all the errors affecting data in order to provide statistical information of high quality since in general a large proportion of response errors tend to have little collective impact on the target output when corrected. Efforts should be concentrated on *influential errors*, i.e., those having a significant influence on publication statistics, thus allowing to re-direct the saved resources to survey phases and data processing activities having a higher pay-off in terms of improvement of data quality.
More in general, the need of reducing costs due to micro editing has to be extended to the problem of identification and treatment of influential observations and outliers, that generally characterize economic survey data. Influential observations can be defined as population units for which their inclusion or exclusion from an estimator may lead to substantial changes in the values and properties of the target estimate. Outliers are generally defined as observations that deviate from a specified data model (for an overview, see among others, Barnett and Lewis, 1994; Lee, 1995). In other words, their influence is usually not explicitly taken into account in their definition. Outliers are frequently caused by defects in some phases of the survey process such as misreporting or misrecording errors (in this case, they belong to the class of *non representative outliers*, see Chambers, 1986). Therefore, these observations are important not only because their presence may highlight not (fully) appropriate data models, but also because they may correspond to errors which may be influential, in the sense that they can seriously affect estimation and inference. Identification of outliers actually due to errors is often complicated by the fact that in observed data this type of data is mixed with extreme but correct values.
At National Statistical Institutes (NSIs) it is common practice to deal with the identification of outliers corresponding to gross measurement errors and errors that are influential with respect to some basic estimates (e.g., totals) at the editing and imputation phase. In complex surveys, the choice of the methods for the detection of these types of data depends, among others, on the target estimates and the model possibly assumed for the population.

[1]Prepared by Marco Di Zio (dizio@istat.it), Ugo Guarnera (guarnera@istat.it), Orietta Luzi (luzi@istat.it).

Concerning outliers, in most NSIs their detection is often performed in practice through univariate or bivariate methods, i.e. methods which are based on the (sometimes graphical) analysis of either marginal distributions of the target variables, or ratios between strictly related items. However, in economic data outliers have usually a multivariate nature. Multivariate outlier detection methods generally use the concept of "distance" among each observation and the "center" of the data. The definition of the distance often relies on some model assumption and requires the estimation of the model parameters. Distances that account for the covariance structure of the data and which are robust against the presence of outliers (Hampel et al., 1986) are usually preferred. A wide class of multivariate outlier detection methods assumes a multinormal data distribution. This class includes methods that use the Mahalanobis distance, and the *contaminated normal model* (see Little, 1988), where both data and error mechanisms are explicitly modelled.
As for influential errors, *selective editing* (Latouche and Berthelot, 1992) is generally adopted for their detection. In selective editing, *score functions* are used to prioritize the units that are in error and have a non negligible potential impact on the target parameter's estimate. All the data having the associated score higher than a pre-defined threshold are considered as *critical*, hence requiring an accurate revision. Selective editing poses a number of problems in the definition of the various elements to be estimated to build score functions. An important issue is the choice of a cut-off for the score function giving the threshold that classifies observations in units to be re-contacted (or not) (Hedlin, 2003). There is not a rule for the cut-off determination, it must be determined ad-hoc for each survey, using available information, for instance previous data. Conversely, it is clear that it is easy to determine the cut-off value if only a fixed number of re-contacts is allowed. Another critical issue in selective editing is the estimation of the error, including both the probability to be in error (also called *risk component*) and the magnitude (*influence component*). These elements are implicitly used by the score functions introduced in literature, and they are also mentioned explicitly in Jäder and Norberg (2005). Considering the influence component, both the prediction and the target estimate are to be determined for estimating it. As for target estimate, it is generally accepted to use a first guess (for instance obtained through robust methods or historical values). Examples of score functions using such estimates are those proposed in Latouche and Berthelot, 1992. Considering the probability of an observation to be in error, it is generally estimated through either the use of logical/mathematical constraints (either cross-sectional or longitudinal), and/or through the degree of outlyingness of the observation (see Lawrence and McKenzie, 1994; Lawrence and McDavitt, 1994; Costa and Cardno, 2005; Jäder and Norberg, 2005). Sometimes, in practice, the risk component is assumed equal to 1 leaving all the quantification of degree of errors to the estimated influence component. In general, the interpretation of the scores as expected values of errors is in general difficult unless data and error mechanism are explicitly modelled.
In this paper we propose a multivariate error model for continuous variables which allows to estimate both the error probability and the error magnitude, based on the assumption that the distribution of the erroneous data can be obtained by the distribution of the error free data by inflating the variance. Following the idea of Ghosh-Dastidar and Shafer (2006), we assume a contaminated multivariate normal distribution for the data, where covariance matrices of acceptable and erroneous observations are proportional by a constant factor to be estimated. The resulting model is a mixture of two Gaussian distributions having the same mean vector, where the components represent good and bad data, respectively. The use of a latent class model allows to estimate the distribution of the good data conditional on the observed data (undistinguished good and bad data), unlike other approaches which require their simultaneous availability and knowledge of the state (again good or bad) of the observation. In this way it is possible to define the scores as expectations, with respect to the estimated conditional distribution, of the difference between the observed values and the corresponding "true" unobserved ones. This explicitly formalizes the idea of a score function as product of an influence component and a risk component.

The model has been developed in the context of studies aiming at developing alternative approaches for the detection of outliers and influential errors for the ISTAT survey on *Balances of Agricultural Firms*(RICA-REA). In particular, the variables taken into account are those relating to firms labour costs (number of employees, wages and salaries, labour cost).
The paper is structured as follows. In Section II, the proposed score function is introduced. Section III contains details on the adopted error contamination model. Section IV contains the discussion of the results of an experimental application of the proposed method on the ISTAT data from the RICA-REA survey. Section V contains final considerations and future work.

II. SCORE FUNCTIONS AND THRESHOLD

The general idea of selective editing is to find the most important errors and to leave the ones with negligible impact on the final estimates. The strategy of selecting editing to balance costs and benefits, is that of ranking the observations according to a score function and reviewing those observations which have a score higher than a specified threshold. The score function should estimate the error probability and the impact of this error on the estimates, and the threshold should be built according to a required degree of reliability of estimates (that is in fact affected by the errors remaining in the data).
Let us denote with $\tilde{T}_Y$ the estimate of the total for the variable Y computed after the selective editing, and let T^*_Y be the estimate computed on data free of error, according to a sample S. Roughly speaking, conditionally on the sample S, the goal is to have $|\frac{\tilde{T}-T^*}{T^*}| < \epsilon$ (for simplicity we have disregarded the subscript Y). In this comparison, there are two main different random mechanisms that should be taken into account:

(1) the error mechanism;
(2) the probability of recovering the true value (error free) for the unit selected for revision.

One way of taking into account those components is to compute either $E|\frac{\tilde{T}-T^*}{T^*}|$ or $|E(\frac{\tilde{T}-T^*}{T^*})|$ where the expected value is with respect to the random mechanism previously introduced. However, up to now the probability of recovering the true value is considered high and it is not considered. In other word we suppose that, once an erroneous value is identified in a unit, the true value is recovered. In our approach, we suppose that units are i.i.d., this is not always verified especially when a complex survey design is used, however in a stratified sample design, this can be approximated by assuming the i.i.d. hypothesis within the strata. With these assumptions, the expected value may be computed according only to the error random mechanism. Once it is clear what is the objective of the selective editing, that is the expected difference between the two estimates, $\tilde{T}$ and T^*, we need to pass from the evaluation at an aggregated level to one at a level of single unit (observation), which means to build a score function.
Let us denote with $\tilde{Y}$ the r.v. of the values obtained after the selective editing, with Y the r.v. corresponding to the observed values, and with Y^* the r.v. corresponding to the true values. We note that, under the assumption that the true values is always recovered for the selected observations, i.e., $\tilde{Y}_i = Y^*_i$ for $i \in M$ and $\tilde{Y}_i = Y_i$ for $i \notin M$, where M is the set of selected observations. It is not clear whether to use $E|\frac{\tilde{T}-T^*}{T^*}|$ or $|E(\frac{\tilde{T}-T^*}{T^*})|$. The function $E|\frac{\tilde{T}-T^*}{T^*}|$ requires the computation of an expected value of an absolute difference, this is generally not simple to obtain, however $E|\frac{\tilde{T}-T^*}{T^*}| = E|\sum_{i\notin M}\frac{Y_i-Y^*_i}{T^*}| \leq \sum_{i\notin M} E|\frac{Y_i-Y^*_i}{T^*}|$, hence it is natural to consider $E|\frac{Y_i-Y^*_i}{T^*}|$ as a score function, and the sum of the ordered (decreasing) values of the score functions ensures an upper bound for the (estimated) difference of estimates.
The quantity $|E(\frac{\tilde{T}-T^*}{T^*})|$ is easier to manage because we need to compute the expected value of a difference. Since $|E(\frac{\tilde{T}-T^*}{T^*})| = |\sum_{i\notin M} E(\frac{Y_i-Y^*_i}{T^*})| \leq \sum_{i\notin M} |E(\frac{Y_i-Y^*_i}{T^*})|$, a natural choice for the score

function is $|E(\frac{Y_i - Y_i^*}{T^*})|$. Also in this case we have an approximation for the total error left in data (an upper bound), and moreover there is a further pitfall, in fact the errors may compensate each other. Hence, it can happen that the difference between the estimates is acceptable, but some important errors (at individual level) are still left in data, and this can be dangerous for instance if we release data and the user makes analysis at a finer level of aggregation.
In order to solve these problems, we propose to use the following way of selecting the observations to be recontacted.

(1) order the observations with respect to the $|E(\frac{Y_i - Y_i^*}{T^*})|$ (in a decreasing order);
(2) select for reviewing all the (ordered) observations until the $\bar{k}$th, where $\bar{k}$ is such that for any $k > \bar{k}$, $|\sum_{i=k}^{n} E(\frac{Y_i - Y_i^*}{T^*})| < \epsilon$.

The first step is concerned with the score function and the second with the threshold. These two steps overcome the problem just mentioned of having left errors in data higher than 2ϵ, in fact it is easy to show that it implies that $|E(\frac{Y_i - Y_i^*}{T^*})| < 2\epsilon, \forall i > \bar{k}$. Moreover the two steps so far described for the selection of units allows to avoid too high upper bounds for the definition of the threshold. In fact, the two previously mentioned strategies, based on summing the expectations of the absolute values of single errors, or the absolute values of the expectations of single errors, lead to a strongly conservative approach, resulting in a number of re-contacted observations much higher than actually needed to fulfill the requirement. Using the usual terminology, we use as a score function $|E(\frac{Y_i - Y_i^*}{T^*})|$ and $|\sum_{i=k}^{n} E(\frac{Y_i - Y_i^*}{T^*})| < \epsilon$ to compute the threshold.
The key problem now is the computation of $E(\frac{Y_i - Y_i^*}{T^*})$. In general this expected value is $0 \times P(Y_i = Y_i^*) + E(\frac{Y_i - Y_i^*}{T^*})P(Y_i \neq Y_i^*)$. It can be noticed that two quantities are involved: the probability of being in error (risk component), and the magnitude of the error (influence component). In the next section we will explicitly introduce a model for the error, making it possible to compute coherently all the elements involved in the score function.
We remark that the same algorithm for the selection of critical units is valid also in the case the target estimate is the mean.

III. CONTAMINATION MODEL

Outliers in multivariate data are often described through contamination models. These models describe contaminated data as realizations from a probability distribution which is obtained by the "good" data distribution by shifting (additive error) or multiplying (multiplicative error) the corresponding random variable, by some (generally random) value. Thus, models seem to be particularly useful to describe the situation of the selective editing. In fact, in this case we assume that some data are affected by an error, and we want to find the most important ones. When dealing with quantitative variables, a frequent approach is to take logarithms of the original data in order to symmetrize the data distribution and to apply linear models. In our approach we assume a Gaussian model for the log-transformed error free data, and suppose that data are corrupted by an additive Gaussian error with zero mean and with the covariance matrix proportional to the covariance covariance matrix of the error free data. This is the model also used by Ghosh-Dastidar and Schafer (2006) for the identification of gross errors. According to this framework we assume that data free of errors are represented by a random p-vector $\mathbf{Y}^*$, and that $\mathbf{X}^* = log(\mathbf{Y}^*)$ is distributed according to a probability distribution $f_{\mathbf{X}^*}(\cdot)$, that is a p-multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Our model assumes that the observed data are obtained by corrupting, with a certain probability, the r.v. $\mathbf{X}^*$ with an error that inflates the variance of the Gaussian distribution. We can write the model as

$$\mathbf{x} = \mathbf{x}^* + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{x}^* \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma}_\epsilon)$. As already mentioned, we suppose that the error covariance matrix $\boldsymbol{\Sigma}_\epsilon$ is proportional to the covariance matrix $\boldsymbol{\Sigma}$, i.e., we suppose that $\boldsymbol{\Sigma}_\epsilon = (\alpha - 1)\boldsymbol{\Sigma}$ for some $\alpha > 1$. This restriction simplifies the analysis, but at the same time it allows to handle the situation characterised by a modest number of gross errors (see Ghosh-Dastidar and Schafer, 2006). The important feature of the model is that errors do not affect all the data but only a portion of them. This intermittent error mechanism implies that the distribution of the observed data $\mathbf{X}$, conditional on the error free data $\mathbf{X}^*$, can be represented as a mixture of two probability distributions corresponding to erroneous and not erroneous data respectively

$$f_{\mathbf{x}|\mathbf{x}^*} = p\delta(\mathbf{x} - \mathbf{x}^*) + (1 - p)N(\mathbf{x}^*, \boldsymbol{\Sigma}_\epsilon) \tag{2}$$

where $p \in [0, 1]$ can be interpreted as the proportion of non erroneous data, $\delta(\mathbf{x} - \mathbf{x}^*)$ is the Dirac delta, assigning all the mass probability to the point $\mathbf{x}$ when $\mathbf{x} = \mathbf{x}^*$, and $N(\mathbf{x}^*, \boldsymbol{\Sigma}_\epsilon)$ represents the error distribution that is centered on the non erroneous data $\mathbf{x}^*$ with a covariance matrix $\boldsymbol{\Sigma}_\epsilon$. Based on this conditional distribution and the marginal distribution of $\mathbf{x}^*$ the density function $f_{\mathbf{x}}(\mathbf{x})$ of the observed data $\mathbf{X}$ is easily obtained:

$$f_{\mathbf{x}}(\mathbf{x}) = pf_{\mathbf{x}^*}(\mathbf{x}) + (1 - p)N(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_\epsilon) \tag{3}$$

The real goal of the model is to compute the probability density function (pdf) of the free of error data given that we have observed some values, in other words the conditional pdf $f_{\mathbf{x}^*|\mathbf{x}}(\mathbf{x}^*|\mathbf{x})$.
According to the probability distribution of the error free data, and by considering (3), we have

$$f_{\mathbf{x}^*|\mathbf{x}}(\mathbf{x}^*|\mathbf{x}) = \frac{pf_{\mathbf{x}^*}\delta(\mathbf{x}^* - \mathbf{x}) + (1 - p)N(\mathbf{x}; \boldsymbol{\mu}, \alpha\boldsymbol{\Sigma})N(\mathbf{x}^*; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})}{pf_{\mathbf{x}^*}(\mathbf{x}) + (1 - p)N(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_\epsilon)} \tag{4}$$

where

$$\tilde{\boldsymbol{\mu}} = (\boldsymbol{\Sigma}_\epsilon^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Sigma}_\epsilon^{-1}\mathbf{x} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) = \frac{1}{\alpha}[\mathbf{x} + (\alpha - 1)\boldsymbol{\mu}]$$

and

$$\tilde{\boldsymbol{\Sigma}} = (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_\epsilon^{-1})^{-1} = \left(1 - \frac{1}{\alpha}\right)\boldsymbol{\Sigma}$$

Note that (4) can be written as:

$$f_{\mathbf{x}^*|\mathbf{x}}(\mathbf{x}^*|\mathbf{x}) = \tau_1(\mathbf{x})\delta(\mathbf{x}^* - \mathbf{x}) + \tau_2(\mathbf{x})N(\mathbf{x}^*; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \tag{5}$$

where $\tau_1(\mathbf{x})$ is the posterior probability, given the observed values $\mathbf{x}$, of being free of error, and $\tau_2(\mathbf{x})$ is the posterior probability of being in error. In fact, from formulas (1) and (2) it follows that, for each unit i $(i = 1, \ldots, n\ )$, $\tau_1(\mathbf{x}_i) = Pr(\mathbf{x}_i = \mathbf{x}_i^*|\mathbf{x}_i)$ and $\tau_2(\mathbf{x}_i) = Pr(\mathbf{x}_i \neq \mathbf{x}_i^*|\mathbf{x}_i)$.
The error model (1) implies that in the original scale, where $\mathbf{y}^* = \exp(\mathbf{x}^*)$ and $\mathbf{y} = \exp(\mathbf{x})$ represent error free data and contaminated data respectively, the contamination mechanism is multiplicative:

$$\mathbf{y} = \mathbf{y}^* \otimes e^{\boldsymbol{\epsilon}} \tag{6}$$

Thus, if the quantity to estimate is a linear function of the data in the original scale, the objective is to estimate the conditional bias $E(\mathbf{Y} - \mathbf{Y}^*|\mathbf{y})$ computed according to the pdf $f_{\mathbf{y}^*|\mathbf{y}}$:

$$0 \times \tau_1(\mathbf{y}) + \tau_2(\mathbf{y}) \int_0^\infty (\mathbf{y} - \mathbf{t}) f_{\mathbf{y}^*|\mathbf{y}}(\mathbf{t}) d\mathbf{t}.$$

From the model assumption it follows that, conditional on $\mathbf{y}^* \neq \mathbf{y}$, the distribution $f_{\mathbf{y}^*|\mathbf{y}}$ is log-normal with parameters $(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$. Thus,

$$E(\mathbf{Y} - \mathbf{Y}^*|\mathbf{y}) = \tau_2(\mathbf{y}) \left(\mathbf{y} - exp\left(\frac{1}{2}\tilde{\boldsymbol{\sigma}}^2 + \tilde{\boldsymbol{\mu}}\right)\right), \tag{7}$$

where $\tilde{\boldsymbol{\sigma}}^2$ is the p-dimensional vector composed of the diagonal elements of the matrix $\tilde{\boldsymbol{\Sigma}}$ The term $\tau_2(\mathbf{y})$ is the posterior probability that $\mathbf{y}$ is in error, i.e., that $\mathbf{y}$ is different from the true value $\mathbf{y}^*$ (given data). From the equalities

$$P(\mathbf{Y}^* \neq \mathbf{Y}) = P(e^{\mathbf{X}^*} \neq e^{\mathbf{X}}) = P(\mathbf{X}^* - \mathbf{X} \neq 0),$$

it follows that, in the original scale, the conditional bias for the jth variable in the ith unit Y_{ij} is:

$$E(Y_{ij} - Y_{ij}^*|\mathbf{y}_i) = \tau_2(\mathbf{x}_i)\left[\left(1 - \frac{1}{\alpha}\right)(\Sigma_{jj} + \mu_j) + \frac{log(y_{ij})}{\alpha}\right] \tag{8}$$

where $\tau_2(\mathbf{x}_i) = \tau_2(log(\mathbf{y}_i))$ is computed as in formula (5), and μ_j, Σ_{jj} are the jth component of the vector μ and jth element on the diagonal of $\boldsymbol{\Sigma}$ respectively. According to the previous observations, the score function is, for the ith observation jth variable:

$$Score(\mathbf{y}_{ij}) = \left|\hat{\tau}_2(\mathbf{x}_i)\left[\left(1 - \frac{1}{\hat{\alpha}}\right)(\hat{\Sigma}_{jj} + \hat{\mu}_j) + \frac{log(y_{ij})}{\hat{\alpha}}\right]\right|. \tag{9}$$

where $\hat{\tau}_2(\mathbf{x}_i)$, $\hat{\alpha}$, $\hat{\Sigma}_{jj}$, $\hat{\mu}_j$ are the maximum likelihood estimates of the corresponding parameters. In this case the two step algorithm is

(1) order the observation with respect to the $Score(y_{ij})$ (in a decreasing order);

(2) find the smallest $\bar{k}$ such that for any $k > \bar{k}$, $E|\sum_{i=k}^{n} \frac{(y_{ij} - Y_{ij}^*)}{\hat{T}_j^*}| < \epsilon$

where $\hat{T}_j^*$ is a robust estimate obtained as $\sum_{i=1}^{n} E(Y_{ij}^*|y_{ij})$

We remark that the model is the same of the one used by Ghosh-Dastidar and Schafer (2006), but in the latter the authors do not estimate the parameters p and α, in fact they estimate the parameters $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ conditionally on some values assigned to p and α. In this paper, the parameters p and α are estimated via the EM algorithm together with $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$.

IV. EXPERIMENT

In this section the proposed selective editing method is discussed by analyzing the results of an experimental application to real business survey data. The study is performed on a sub-sample of the ISTAT survey on *Economic Accounts of Agricultural Firms* (RICA-REA), year 2004.

The RICA-REA is an annual sampling survey which collects information on a set of economic variables which are necessary for microeconomic analyses and for fitting the National Accounts requirements. The annual sample consists of about 20.000 Italian agricultural firms, stratified by Unity of Economic Dimension (UDE). The survey collects information on costs, stocks, purchases and sale of fixed capital, re-investments, revenues, social contributions to firms, labor cost, and income of agricultural firms. The target parameters are the totals T of the surveyed economic variables. The variables considered in the application study relate to employment and labor cost for the non-household permanent workers. In particular, we considered the *number of worked days (gldti)*, and the *wages and salaries (wdti)*. The analysis is restricted to the subsample S of 999 firms which provided information on the variables of interest.

Concerning the current editing and imputation process of the RICA-REA, on 2006 a project aiming at re-designing the overall strategy has started. As for the detection of outliers and influential errors, the approach used until now consists of two steps: 1) graphical analysis of the ratio edits between correlated items, with the manual verification of all the values which appear to be far from the "centre" of the joint distributions, and 2) manual verification of the units having a *large* impact on anomalous domain totals, where a total is anomalous when it differs *too much* from the corresponding historical domain estimate. In the latter step, the thresholds at both micro and macro level are subjectively set by subject matter experts.

With our experimental study we want to assess the performance of the proposed selective editing approach in identifying non representative outliers and influential errors reducing the need of manual revisions.
The experimental application is based on the artificial contamination of the *final* RICA-REA subsample S (i.e. data "cleaned" through the old editing and imputation process, assumed as *true* data) and in the application on the artificial *raw* data set of the model illustrated in Section III. More in detail, the experimental application is performed as follows:

(1) given the sample S, the values of *gldti* and *wdti* are transformed in log scale to approach the multinormality assumption, and a percentage p of contaminated data is simulated by adding to log-transformed data an error component drawn from a normal distribution $N(0, (\alpha - 1)\mathbf{\Sigma})$, where α is the inflation parameter (see model 1);
(2) the contaminated model parameters are estimated on the artificial raw data obtained in step (1), and for each unit the score function $Score(\mathbf{y}_i)$ described in Section III (formula 9) is computed using the estimated error probabilities and the estimated predicted errors;
(3) the observations are ordered by descending values of the score function, and the threshold is set based on the pre-defined expected residual error ϵ on the target estimates T (totals of *gldti* and *wdti*).

In the experiment we set $\alpha = 2$ and $p = 0.05$. The contaminated data are reported in Figure 1 and the log-transformed contaminated data are shown in Figure 2.

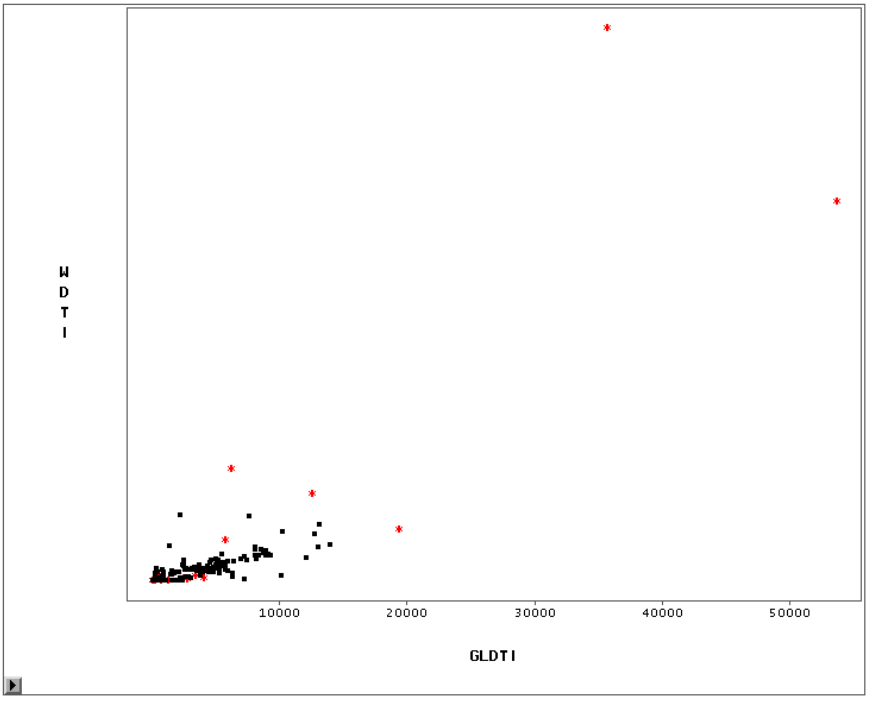
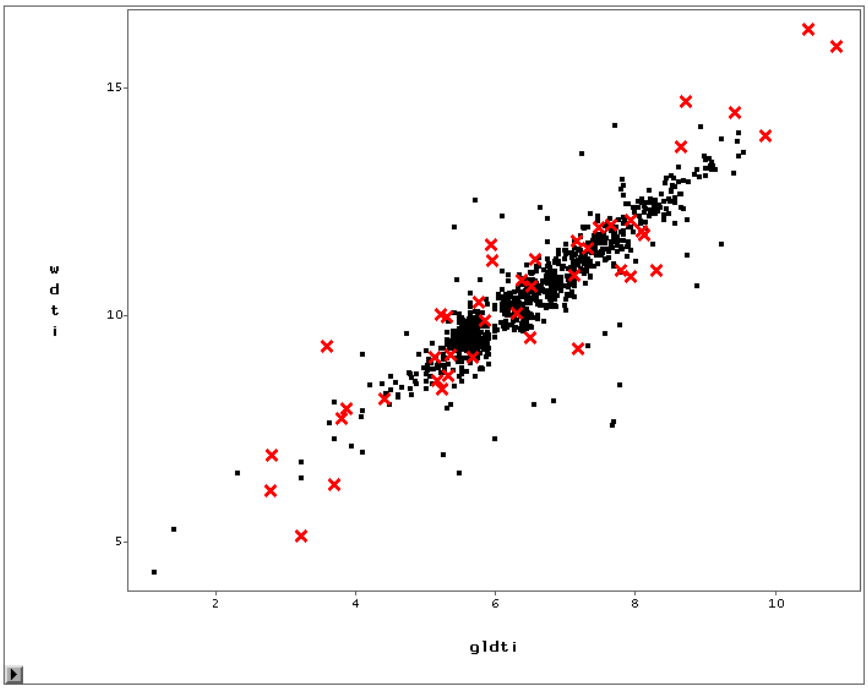


FIGURE 1. Contaminated data in original scale. The errors are depicted with a star.

Choosing $\epsilon = 5\%$, the proposed selective procedure identifies 18 critical units for *gldti*, and 28 for *wdti*. In Tables 1 and 2 the cross classification of data in errors and selected observations is reported. For



[h]

FIGURE 2. Contaminated data in log-scale. The errors are depicted with a star.

gldti the 50% of the selected units are erroneous, while for *wdti* the percentage of selected units which are erroneous is 36%.
An important consideration is about the fact that, once these critical units are corrected, also the true residual error is always below the expected residual error $\epsilon = 5\%$, in fact the true residual errors are 0.014 and 0.004 for *gldti* and *wdti* respectively. As mentioned in Section II, also the true absolute error for each single remaining unit is below 2ϵ. In Figures 3 and 4 the line plot of the true and estimated residual error for *wdti* and *gldti* are reported on the subset of the first 200 observations.

TABLE 1. Results for the experiment with $\epsilon = 0.05$ and *gldti*

	non sel	sel	
non err	947	9	956
err	34	9	43
	981	18	999

TABLE 2. Results for the experiment with $\epsilon = 0.05$ and *wdti*

	non sel	sel	
non err	938	18	956
err	33	10	43
	971	28	999

The difference in the functions in Figures 3 and 4 may be explained by the fact that the log-transformed original data are not exactly normal. In effect, if data fulfill the assumptions of model stated through formula (6), the estimated and the true residual error functions would be very close.
Figure 5, which is analogous to Figures 3 and 4, refers to log-data that are synthetically generated from a Gaussian distribution and, as in the previous case, with a Gaussian error. We note that the estimated and true residual error functions are almost the same, and this is an important evidence for the quality of the estimation methods.

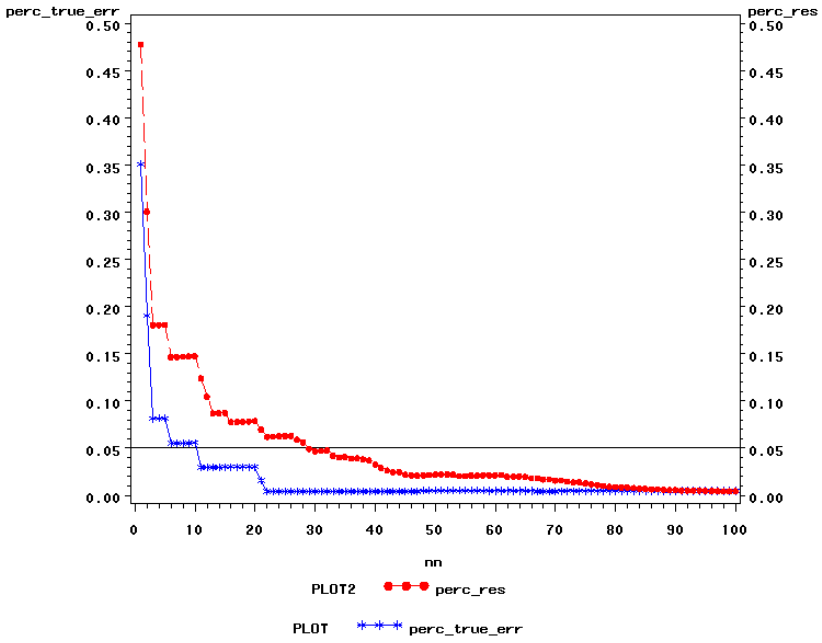


FIGURE 3. Estimated (perc_res) and true (perc_true_err) residual error functions for wdti with threshold 0.05.

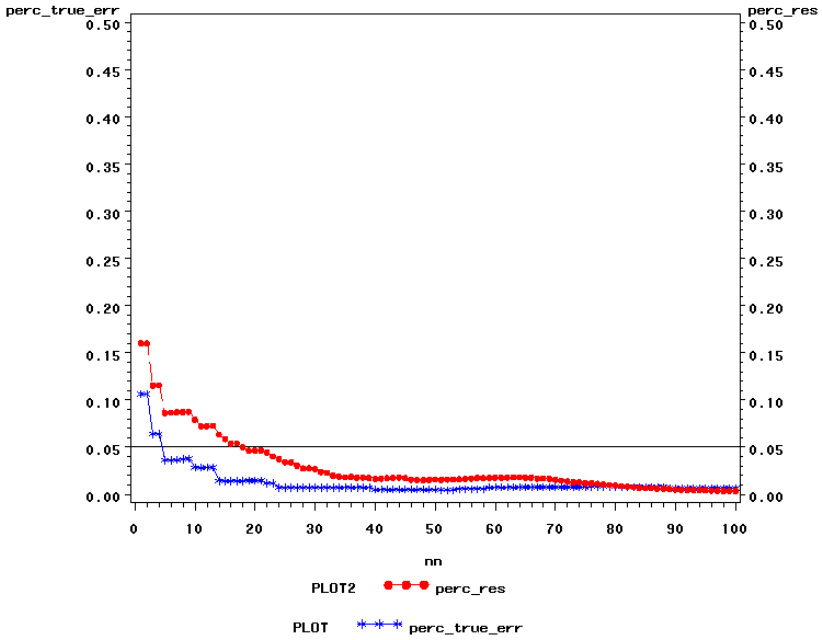


FIGURE 4. Estimated (perc_res) and true (perc_true_err) residual error functions for gldti with threshold 0.05.

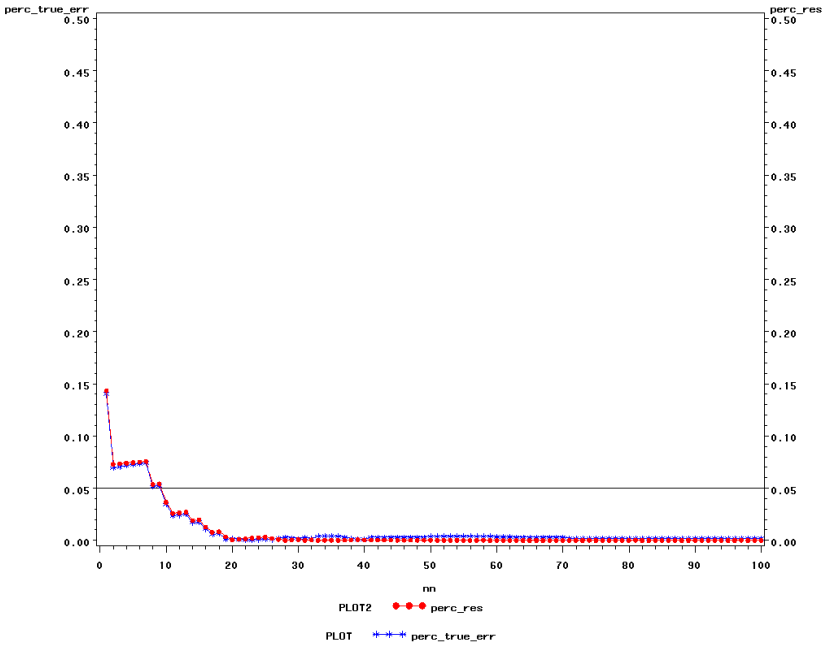


FIGURE 5. Estimated (perc_res) and true (perc_true_err) residual error functions for log(x) normally distributed and with threshold 0.05.

V. CONCLUSIONS AND FUTURE WORKS

In this paper we present a strategy for selective editing based on the assumption that log-transformed data are Gaussian and that a portion of log-transformed data are contaminated by an additive error that follows a Gaussian distribution with zero mean and with a variance proportional up to an inflation parameter to the variance of the data free of errors. It is worth to note that these assumptions imply that the original data are contaminated through a multiplicative error, and that the error does not have a zero mean. In this setting we propose a procedure dealing with all the issues of selective editing in the case of an estimate of a total (or a mean). In particular a score function and an algorithm for computing a threshold that ensure a certain level of reliability of the final estimates are proposed. It is also shown how to estimate the score function and when to stop the selection. The computation of the score function is based on maximum likelihood estimates of model parameters. It is also worth noting that, the score function requires the estimate of the true total, and the algorithm naturally gives a robust estimate for this quantity. Summarysing, once the researcher has set the level of reliability of estimates, and under the hypothesis that the model assumptions are fulfilled, the proposed procedure treats all the problems coherently.
At this stage, the observations are considered independent and identically distributed, this issue deserves further studies, even if a generalization to stratified sample designs seems not too difficult. Further comments concern the fact that this procedure is applied, in this paper, to a cross-sectional survey without external information. These surveys are critical because, when using selective editing without external information, errors are estimated comparing observed values to predicted values that must be estimated with the data at hand. However, when longitudinal surveys is analysed, historical data could be incorporated in the proposed procedure considering the outdated variables as covariates in the model. This aspect will be furtherly investigated in the future.
A final consideration is about the robustness of the procedure to departures from the assumption of the model. In this paper we have used data of a ISTAT business survey and we have contaminated them with the error mechanism of the model. Hence, we are in a mixed situation where the model is partially fulfilled. In the experiment reported, the proposed procedure gives satisfactorily results since it allows to select a number of units that effectively, once recovered the true values, let the estimates be below the chosen level of reliability. More experiments are needed for this issue.

References

Barnett V., Lewis T. (1994). *Outliers in Statistical Data*, New York: Wiley.

Chambers R. L. (1986). Outlier robust finite population estimation. *J. Am. Statist. Ass., 81, 1063-1069.*

Costa V., Cardno B. (2005). Use and editing of administrative data in the business indicators unit, *UN/ECE Work Session on Statistical Data Editing, Ottawa.*(http://www.unece.org/stats/documents/2005.05.sde.htm).

Ghosh-Dastidar B., Schafer J.L. (2006). Outlier Detection and Editing Procedures for Continuous Multivariate Data, *Journal of Official Statistics, Vol. 22, No. 3, pp. 487-506.*

Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986). *Robust Statistics*, John Wiley & Sons, Inc.

Hedlin D. (2003). Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics, *Journal of Official Statistics, Vol. 19, No. 2, 177-199.*

Jäder A., Norberg A. (2005). A Selective Editing Method considering both suspicion and potential impact, developed and applied to the Swedish Foreign Trade Statistics, *UN/ECE Work Session on Statistical Data Editing, Ottawa* (http://www.unece.org/stats/documents/2005.05.sde.htm).

Latouche M., Berthelot J.M. (1992). Use of a score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics, 8, n.3, 389- 400.*

Lawrence D., McKenzie R. (2000). The General Application of Significance Editing. *Journal of Official Statistics, 16, n. 3, 243-253.*

Lawrence D., McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earnings, *Journal of Official Statistics, Vol. 10, No. 4, pp. 437-447.*

Lee H. (1995). *Outliers in Business Surveys, in: Business Survey Methods*, Cox B.G., Binder D.A., Chinappa B.N., Christanson A., Colledge M.J. and Kott P.S. (Eds), John Wiley and Sons, Inc. 503-526.

Little, J.A. (1988). Robust estimation of the mean and covariance matrix from data with missing values, *J. R. Stat. Soc., Ser. C, Vol. 37, No. 1, 23-38.*