

Detection of Potentially Influential Errors in Statistical Survey Data

Individuazione di Dati Errati Potenzialmente Influenti nelle Indagini Statistiche

Marco Di Zio

ISTAT, Via Cesare Balbo, 16, dizio@istat.it

Ugo Guarnera

ISTAT, Via Cesare Balbo, 16, guarnera@istat.it

Orietta Luzi

ISTAT, Via Cesare Balbo, 16, luzi@istat.it

Irene Tommasi

ISTAT, Via Ravà, 150, irtommas@istat.it

Riassunto. In this paper we discuss methodological problems and critical aspects relating to the identification of errors potentially influential on the estimates of target parameters in business survey. Potentially influential errors are generally identified through techniques that include outlier detection methods as well as approaches aiming at sorting units based on the impact of their potential errors on target parameter estimates. In general, these techniques allow to prioritize units to be manually revised/followed up, in order to optimize the trade-off between costs of data treatment and data quality. The assumption is that the remaining units can be treated by less expensive approaches without affecting substantially the quality of final data.

Keywords: Influential errors, multivariate outliers, robust estimation, selective editing.

1. Influential errors and outliers in Official Statistics

In order to produce statistical information of high quality, it is common practice at National Statistical Institutes (NSIs) to adopt specific procedures to identify non sampling errors in statistical data to reduce their effects on target estimates and analyses. One of these procedures is commonly referred to as *editing and imputation* (E&I). Due to the relevant costs of E&I, there is a need of balancing between resources spent and data accuracy. In effect, in order to provide statistical information of high quality, it is in general not necessary to identify all the errors affecting data (Granquist and Kovar, 1997): efforts are to be concentrated on errors having the highest impact on figures to be published (*influential measurement errors*). An influential measurement error (*influential error* in the following) is defined as a deviation of the observed value

of a survey variable from its *true* value having a significant influence on publication statistics computed on this variable. Influential errors represent a classical problem in Official Statistics, particularly in business statistics, due to their potential effects on estimation and inference.

Strictly related to the concept of influential errors is that of influential observations. 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. A variable value may be influential because of its magnitude, or since it is affected by an influential measurement error. In sample surveys, since the sample is drawn from a finite population, the detection of influential observations is generally performed taking into account sampling weights. In this case, influential observations are the ones for which the combination of the reported value and the sampling weight have a large influence on the estimate.

Strictly related to the concept of influential observations and errors is the concept of outliers. Outliers are generally defined as observations that deviate from a specified data model. Outliers have different sources: Chambers (1986) defines *non representative* and *representative* outliers: the former are either considered to be units corresponding to gross measurement errors, i.e., to possible influential errors, or unique in the population; the latter are true values which are originated by either the inherent variability of the data, or they may be generated by a different model. In sample surveys, outliers corresponding to extreme values may be influential on target statistics depending on their sampling weights. Outliers may be univariate or multivariate: multivariate outliers are observations appearing to be inconsistent with the correlation structure of the data, while univariate outliers are inconsistent with respect only to the marginal distribution of each variable.

In Official Statistics, the risk associated with outliers corresponding to gross measurement errors and other influential errors derives from the different ways in which they may exert undue influence on estimation and inferences: firstly, due to their potential effects on estimates, these errors need to be treated before performing data analyses; furthermore, in large scale surveys conducted by NSIs, usually some models are used to compensate for (partial) non response: erroneous values which are influential on model parameters should be properly treated in order to obtain reliable predictions; finally, outliers and more in general influential observation need to be accounted for at the estimation stage, e.g., through robust estimation.

At NSIs, it is common practice to deal with the first two aspects during the E&I phase. Identification of outliers actually due to errors is often complicated by the fact that in observed data this type of outliers is mixed with extreme but correct data values.

In complex surveys, the choice of the methods for the detection of outliers and influential errors depends, among others, on target estimates and the model possibly assumed for the population.

The aim of this paper is to discuss through an application to real survey data the issues related to the choice of methods for the detection of potentially influential errors. To this aim, we consider: 1) two classical univariate nonparametric approaches, one based on the distance of the observations from the center of the univariate distribution, and the other based on the *selective editing* technique; 2) two classical parametric approaches for multivariate normal datasets based on *robust regression* and on *robust Mahalanobis distance*. In robust regression a specific linear model is assumed between a response variable and a set of covariates, and robust parameter estimates algorithms are used to

identify genuine influential outlying observations. Methods based on robust Mahalanobis distance use robust estimates of multivariate location and scatter to identify values that are far away from the multivariate cloud of the data.

The application is carried out on employment variables collected in the ISTAT survey on *Economic Accounts of Agricultural Firms* (RICA-REA).

The paper is structured as follows. In Section 2 the applied methods are described. Section 3 contains the description of the results of the application of these methods on the RICA-REA data. Section 4 contains concluding remarks.

2. The experimented approaches

Let $X=(x_1, \dots, x_p)$ be a data matrix containing the values of p variables observed on a sample S of n units. Univariate methods for the identification of outliers and influential errors are based on the analysis of the (weighted) marginal distributions of each variable, or their proper transformations. In multivariate methods the covariance structure of the data is taken into account as the joint data distributions are analysed to identify (influential) anomalies in data. More specifically, in the field of outlier detection, the identification of multivariate outliers is more difficult than the detection of univariate outliers, since for $p > 2$ one can no longer rely on visual inspection. In particular, units that result to be outlier in one dimensional data may belong to the bulk of the data when several dimensions are considered. Vice-versa, in p -variate data some anomalies can result which cannot be detected in one-dimensional space. Furthermore, especially in case of multivariate outliers, the efficient detection of anomalies may result compromised by the so-called *masking* and *swamping* effects: masking occurs when a group of outlying points skews the mean and covariance estimates towards it, and the resulting distance of the outlying point from the mean is small, making them do not look like outliers; on the contrary swamping occurs when a group of outlying points skews the mean and covariance estimates towards it, and away from other non-outlying observations, and the resulting distance from these units to the mean is large, making them look like outliers (see Barnett and Lewis, 1994). For this reason, in the class of multivariate methods for the detection of influential outlier, those based on *robust estimation* of model parameters are preferred. Robust estimators can deal with data containing a certain percentage of outliers. The robustness of an estimator T is measured by its *breakdown point*, i.e., the smallest fraction of contamination that can have an arbitrarily large effect on T (see Lee, 1995). Other properties to take into account to evaluate robust estimators are bias and precision, and if the estimator is affected by location or scale transformations (equivariance).

Additional elements that are to be taken into account in the detection of outliers and influential errors are sampling weights and scale.

As already mentioned, in sample surveys extreme values may or may not be influential due to their sampling weights. For this reason, when looking for influential errors (like in *selective editing*) weights are to be taken into account.

In general, the effect due to weights should be reduced in outlier detection if they are calculated separately by domain, and if these domains are very similar to the sampling strata, so that units belonging to the same domain have similar design weights (as in simple stratified sampling design).

As for the scale, since outliers are defined as observations which deviate for a given model assumed for the (multivariate) data, transformations may be needed to make the data better fit the assumed model (generally, multinormal assumption is made). In the E&I phase, however, the main objective is detect those observations which are suspicious and potentially influential in the original variable scale, hence in these situations it is generally preferable to analyze the original data distributions, provided that the possible explicit model assumptions are still valid.

In Section 2.1 univariate methods for the detection of outliers are illustrated. In Section 2.2 the *robust multivariate regression* and the *robust Mahalanobis algorithm* considered in this paper are described. In Section 2.3, a *selective editing* approach for the detection of potentially influential errors based on robust regression is described.

2.1. Univariate methods for outlier detection

The underlying idea of outlier detection is to measure the relative distance of each observation from the center of the data. The distance is generally expressed by $d_i = |x_i - m| / s$, where m and s are location and scale parameters respectively. In a univariate context, where observations can be naturally ordered, and thus quantiles can be easily computed, the distance can be measured in terms of quantiles. This allows to obtain robust estimates of m and s , e.g., estimating s with the interquartile range and m with $(q_1 + q_3) / 2$, where q_1 and q_3 are the 1st and 3rd quartile respectively. The last choice is generally used in the box-plot graphical representation. Another frequent alternative is the sample median for the location parameter m and the median of the absolute deviations from the sample median (MAD), $MAD = \text{median}(|x - \text{median}(x)|)$. For the use of quantiles, this family of methods will be referred to as *quantile* methods. An extension of the *quantile* methods to the case of trend data, i.e., when units are observed in surveys carried out in different occasions, is proposed in Hidiroglou and Berthelot (1986). This class of methods is not only robust but simple and nonparametric.

2.2. Multivariate methods for outlier detection

2.1.1 Multivariate robust regression

In multiple regression, one of the target variables, say Y , is considered as response variable, and it is related to a set of p explanatory variables $X_1, \dots, X_p$ in the model $y_i = \theta_0 + x_{i1}\theta_1 + \dots + x_{ip}\theta_p + \varepsilon_i$, ($i=1, \dots, n$), where ε_i is an error term assumed normally distributed with mean zero and unknown standard deviation σ . The n observations (x_i, y_i) belong to the linear space of row vectors of dimension $p+1$. The unknown parameter θ is a p -dimensional vector $(\theta_1, \dots, \theta_p)^t$.

In last decades, the most used methods for outlier detection in multivariate regression are robust regression approaches and regression diagnostics. In robust regression, parameter estimators are used that reduce the impact of outliers that would be highly influential otherwise. A robust estimation procedure tries to accommodate the majority of the data, so that “bad points” lying far away from the pattern formed by good ones, will consequently possess large residuals from the robust fit, and can be easily identified. Hence, robust regression estimators are non only insensitive to outliers, but

have the advantage of allowing an easy detection of outliers which are influential on the model parameter estimates themselves, i.e., corresponding to leverage points (having large positive and large negative residuals).

In the class of robust regression estimators, Rousseeuw (1984) proposed the *Least Median Squares (LMS)* and the *Last Trimmed Squares (LTS)* methods, which should be preferred to other robust methods due to their high breakdown points and other statistical properties. The *LMS* estimator is given by

$$\text{Minimize}_{\theta} \text{med}_i r_i^2$$

where med_i is the median of the residuals r_i ($i=1, \dots, n$). This estimator is very robust with respect to outliers in y as well as in x : its breakdown point is 50%, the highest possible value. The *LMS* is equivariant with respect to linear transformations on the explanatory variables (Rousseeuw and Leroy, 1987, p. 117). Since *LMS* performs poorly from the point of view of asymptotic efficiency, Rousseeuw (1984, 1985) proposed the *Least Trimmed Squares (LTS)* estimator, given by:

$$\text{Minimize}_{\theta} \sum_{i=1}^h (r^2)_{i:n}$$

where $(r^2)_{1:n} \leq \dots \leq (r^2)_{n:n}$ are the ordered squared residuals. Since the largest residuals are excluded from the summation, the fit stay away from the outliers. The best robustness properties are achieved when h is approximately $n/2$, in which case the breakdown point is 50%. The *LTS* is equivariant for linear transformations on the x_i . It is also regression, scale and affine equivariant (Rousseeuw and Leroy, 1987, p.132). The resistance of both *LMS* and *LTS* to outliers is independent on p .

Leverage points, i.e., units corresponding influential outliers, are characterized by large positive or negative residuals. The identification of large residuals implies their standardization with respect to an estimate $\hat{\sigma}$ of the error scale. For *LTS* regression, Rousseeuw (1987) suggests different estimators, such as the following one:

$$\hat{\sigma} = C_2 \sqrt{\frac{1}{n} \sum_{i=1}^h (r^2)_{i:n}}$$

where r_i is the residual of case i with respect to the *LTS* fit, and C_2 is a correction factor used to achieve consistency at Gaussian error distributions. More refined estimators can be used for small samples (Rousseeuw and Leroy, 1987). In any case, the i th case is identified as an outlier if and only if $|r_i^2 / \hat{\sigma}|$ exceeds a predefined cutoff. In this paper we use the FAST-LTS algorithm of Rousseeuw and Van Driessen (2000).

2.1.2 Robust Mahalanobis distance

Robust estimation of multivariate location and scatter is the key tool to robustify multivariate techniques, such as those based on the use of the squared Mahalanobis distance. In these techniques, in which the centre of the data cloud is estimated, all variables are treated in the same way (unlike regression analysis). The focus is on the

dispersion of the data about this “centre”. To this aim, the values of the Mahalanobis distance can be compared with quantiles of the chi-squared distribution with $p-1$ degrees of freedom. However, in order to avoid the possible masking effect, high-breakdown estimators of multivariate location and scatter used to compute the Mahalanobis distances have been proposed. Among others, the *Minimum Volume Ellipsoid (MVE)* (Rousseeuw, 1984; 1985) looks for the ellipsoid with smallest volume that covers h data points, where $n/2 \leq h < n$. In other words, the *MVE estimator* is the centre and the covariance of a sub-sample of size h that minimizes the covariance matrix associated to the sub-sample. Formally:

$$MVE = (\bar{x}_J^*, S_J^*)$$

where $J = \{\text{set of } h \text{ instances: } Vol(S_J^*) \leq Vol(S_K^*) \text{ for all } K \text{ so that } \#(K)=h\}$, where $Vol(S)$ stands for the volume of S .

The *MVE* breakdown point is essentially $(n-h)/n$. The *MVE* belongs to the class of affine equivariant estimators, which are invariant with respect to linear transformations (Rousseeuw and Leroy, 1987, pp 249-250).

The *Minimum Covariance Determinant (MCD)* estimator is another method proposed by Rousseeuw (1984, 1985), whose objective is to find h observations (out of n) whose classical covariance matrix has the lowest determinant. The *MCD* estimate of location is then the average of these h points, and the *MCD* estimate of scatter is their covariance matrix. Formally:

$$MCD = (\bar{x}_J^*, S_J^*)$$

where $J = \{\text{set of } h \text{ instances: } |S_J^*| \leq |S_K^*| \text{ for all } K \text{ so that } \#(K)=h\}$.

The *MCD* breakdown point equals that of *MVE* estimator. However, *MCD* should be preferred to the *MVE* due to its higher statistical efficiency and higher convergence rate. Furthermore, robust distances based on *MCD* are more precise than those based on *MVE* and hence better suited to identify multivariate outliers.

The algorithms to compute the *MVE* and *MCD* estimators are based on combinatorial arguments. In this paper we use the *MCD* algorithm based on the FAST-MCD algorithm given by Rousseeuw and Van Driessen (1999).

2.3. Selective editing

In the recent past, the problem of targeting only those responses worth pursuing has been dealt with. A significant unit to edit is one which, if treated by amendment, would lead to an important change to an estimate. Since obtaining a correct value through clerical action (especially considering re-contact and follow-up) is particularly expensive, an approach considering both the quality aspects given by data accuracy and costs is needed, in other words an approach addressing the significant units to edit. This approach is generally called *selective editing* (Lawrence and McKenzie, 1994; Latouche and Berthelot, 1992). This approach is essentially based on a “score function”. This is a function that allows to rank observations according to their potential error impact on figures to be published.

For an estimate of a total like $X = \sum w_i x_i$, the expected unit contribution to change can be approximated by $w_i q_{x,i} Dx_i$, where Dx_i is the estimated error. For instance it could be $Dx_i = \text{abs}(x_i^* - x_i)$, where x_i^* is our expected or imputed value for the response (x_i) and $q_{x,i}$ is the probability that the response is erroneous for item x .

Local scores based on contributions can be expressed in terms of estimates that they target. This has the advantage of indicating the likely impact on an estimate. Local relative score for an estimate of the total (at a certain level of stratification) can be

$S_{L,i} = (w_i q_{x,i} Dx_i) / X$. This analyzes one variable, and it can be computed for different variables in the units. Then, it can be combined to produce an overall score (global score) for instance by taking the maximum of the local scores.

There are several problems concerning the elements involved in the score function. Since the *selective editing* is generally performed in an early stage of the data production process, the final sampling weights are not available, thus they are generally approximated by the initial sampling weights. Another problem is about the target estimate X . It is generally accepted to use a first guess (for instance obtained through robust methods or historical value). The q_i values are generally estimated through either the use of logical/mathematical constraints, and/or through the degree of outlyingness of the observation, or sometimes, in practice, it is assumed $q_i = 1$ leaving all the quantification of degree of errors to the quantity Dx_i .

Finally, it remains the problem of estimating the error, and thus x_i^* . Concerning this point, it is worth to remark that, as stated in Lawrence and McKenzie (1994), this prediction has not to be particularly accurate, but it has to give a clue about the magnitude of error. Of course, a suggested technique is that of predict x_i^* through the technique that will be used for imputing missing data, since in a certain sense, the final estimates will be based on those hypothetical values.

Apart these problems, the most difficult issue to solve is the choice of a cut-off for the score function giving the threshold that classifies observations in units to be recontacted (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.

3. The application study on RICA-REA data

Influential error detection can be dealt by using different methods and different perspectives. Some of the most important issues to be considered are:

- 1) Is it better to use a univariate or a multivariate method?
- 2) Is it better to apply techniques directly to the data at hand, or to an appropriate transformation of them?
- 3) Is it better to use a technique based on modeling the joint distribution or based on some regression model (conditional distribution)?
- 4) Should we use outlier detection methods, *selective editing* or both of them?

In Official Statistics, answers to these questions have to take into account not only theoretical aspects, but also operational concerns. In this paper, we do not have the pretension to say a final word on these problems, we want to explore these issues and discuss them by analyzing the results of an application to data from a real business survey. The study is performed on a sub-sample of data of the Istat survey on *Economic*

Accounts of Agricultural Firms (RICA-REA) in the year 2004. 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 sample consists of about 20.000 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, labour cost, and income of agricultural households. Among others, parameters of interest are the totals of the surveyed economic variables. The variables considered in the application study relate to employment and labour cost for the non-household permanent workers: *number of worked days* (*gldti*), *wages and salaries* (*wdti*), *social contributions* (*csdti*). We restrict our analysis to the subsample of 988 firms which filled in the corresponding section of the questionnaire.

In order to compare different approaches, a common threshold for the selection of observations is required. In the following, the number of selected observations is 50, i.e., about the 5% of the population.

As far as the use of univariate or multivariate procedure is concerned, there are some issues that should be stressed. Of course, when the variables are highly related, the use of a multivariate method is useful. However, it is not always simple to model joint distributions, and in practical applications, this is an important issue to consider. Univariate methods are less powerful than multivariate ones in presence of strong relationships among variables, but they generally have the advantage of being fast, easy to use and provide results which are directly interpretable. In order to discuss this issue, we compared the results obtained by applying a multivariate and a univariate technique. The multivariate technique is the *MCD* described in Section 2.1.2, computed on the variables $\log(wdti)$ and $\log(gldti)$. The univariate outliers refer to the distribution of $\log(wdti/gldti)$ and are found based on the distance defined in Section 2.1. where the center m is estimated by the median and the scale is the *MAD*.

In Figure 1, the log-transformed values of the variables *wdti* and *gldti* are plotted. In this picture, outliers resulting from the univariate method and *MCD* are depicted. In the figure, grey squares correspond to the 923 units not selected as outliers neither by the univariate nor by the *MCD* method. Dots are the 35 units selected by both methods. Crosses and triangles represent the 15 observations selected only by *MCD* and only by the univariate method, respectively. The absolute frequencies of the four types of units are reported in Table 1.

Figure 1: *MCD vs Univariate*

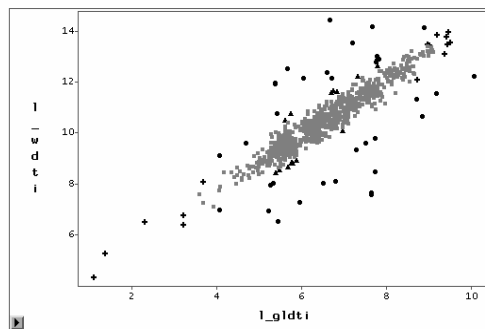


Table 1: *Number of selected units by method*

		Univariate		
		Selected	Not selected	Total
<i>MCD</i>	Selected	35	15	50
	Not Selected	15	923	938
	Total	50	938	988

As expected, the *MCD* method classifies as outliers, units having a remarkable distance from the bulk of the data even if they have an acceptable level of daily wage.

On the other hand, the use of *MCD* implies that data are (almost) multinormally distributed. To meet this assumption, it is common practice to transform data. In economic surveys, where data are typically positively skewed, the log-transformation is generally adopted. However, this is not a neutral task, since outlier detection performed in the original scale would provide different results, as shown in Figures 2 and 3. In these figures, grey squares correspond again to the 912 units not classified as outliers neither in the log-scale nor in the original scale. Dots are the 24 units selected in both scales. Crosses and triangles represent the 26 observations selected only in log-scale and only in the original scale, respectively. The absolute frequencies of the four types of units are reported in table 2.

Figure 2: *MCD on transformed data vs MCD on original data (log scale)*

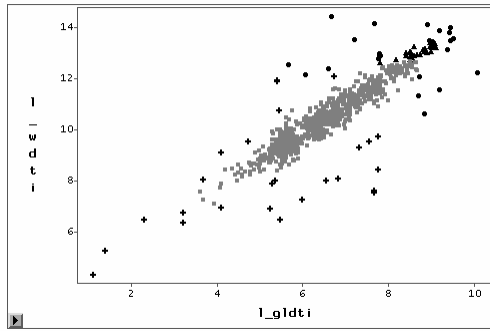


Figure 3: *MCD on transformed data vs MCD on original data (original scale)*

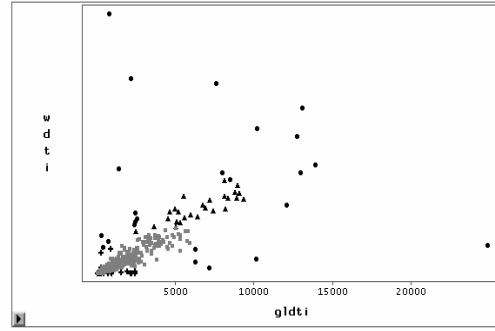
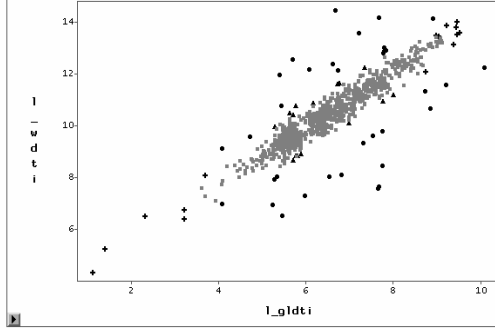
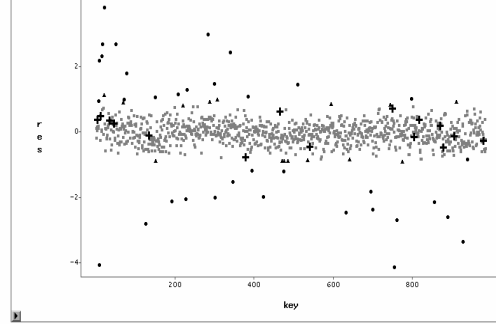


Table 2: *Number of selected units by MCD in log and original scale*

		MCD no log		
		<i>Selected</i>	<i>Not selected</i>	<i>Total</i>
<i>MCD log</i>	<i>Selected</i>	24	26	50
	<i>Not selected</i>	26	912	938
	<i>Total</i>	50	938	988

It can be seen that with the log-scale transformation the method selects outliers also in the left tail, while in the original scale outliers are mainly identified in the right tail of the joint distribution. Note that the units with extreme values, which are most likely the most influential on target estimates (totals), are identified in both the scales.

The third problem explored in the application concerns the choice between techniques based on modeling the joint distribution or methods based on some regression model. In the presence of linear relations, like in our application context, a natural comparison is between outliers found by means of the *MCD* and those found through a (robust) linear regression. Figure 4a represents the scatter plot of $\log(wdti)$ vs. $\log(gldti)$, while Figure 4b contains the residuals of the model $\log(wdti) = \theta_0 + \theta_1 \log(gldti)$, where (θ_0, θ_1) are estimated via the *LTS* method described in Section 2.1.1. In both figures, grey squares correspond to the 923 units not selected as outliers neither by *MCD* nor by *LTS*. Dots are the 35 units selected by both methods. Crosses and triangles represent the 15 observations selected only by *MCD* and only by *LTS*, respectively. The absolute frequencies of the four types of units are reported in table 3.

Figure 4a: MCD vs LTS**Figure 4b: Residuals of LTS regression****Table 3: Number of selected units by method**

		LTS		
		<i>Selected</i>	<i>Not selected</i>	<i>Total</i>
MCD	<i>Selected</i>	35	15	50
	<i>Not selected</i>	15	923	938
	<i>Total</i>	50	938	988

In the regression, the outliers are those observations which are far from the regression line (*regression outliers*). They correspond to “leverage points”, i.e., observations which greatly influence the slope of the regression line. On the contrary, in *MCD*, where the role of variables is symmetric, also observations that do not affect the slope of the regression can be classified as outliers when they lie far away from the center of the data. This evidence also results from the analysis of the *LTS* residuals in Figure 4b.

A final consideration is about the use of outlier detection methods and/or selective editing. As already mentioned, in the editing and imputation phase of business surveys, the objective is not only to identify errors which originate anomalies with respect to some data model, but also those errors which may influence the target estimates. The outlier detection methods so far analyzed are effective for the former objective, but are less suitable for the latter one. *Selective editing*, described in Section 2.3, is an approach which allows to identify potentially influential errors taking into account the sampling weights easily.

In *selective editing*, the selection of influential errors is based on a prediction of error and the impact that this error has on the target estimates. In this example, we assume to

be interested in the estimate of the total of the variable $wdti$: $T = \sum_{k=1}^n w_k wdti_k$, where w_k

is the sample weight of the k th observation and n is the sample size. For each unit k , the prediction ($wdti_k^*$) is determined by using *LTS* with $gldti_k$ and $csdti_k$ as covariates. In particular, the parameters of the regression model are estimated on log-data, thus obtaining preliminary predictions ($l_wdti_k^*$). Since this implies the assumption of log-normality, the actual prediction $wdti_k^*$ is obtained by the formula $wdti_k^* = \exp(l_wdti_k^* + 0.5 s^2)$, where s^2 is the estimated residual variance of the regression model. The score function used to rank units is: $S_{L,k} = (w_k | wdti_k - wdti_k^*|) / T$, ($k=1, \dots, n$).

In Figure 5, the first 5% of observations classified as outliers through the *LTS* method and those pointed out through *selective editing* are illustrated. In the figure, grey squares

correspond to the 901 units not selected neither by *LTS* nor by *selective editing*. Dots are the 13 units selected by both methods. Crosses and triangles represent the 37 observations selected only by *LTS* and only by *selective editing*, respectively. The absolute frequencies of the four types of units are reported in table 4.

Figure 5: *LTS vs Selective Editing*

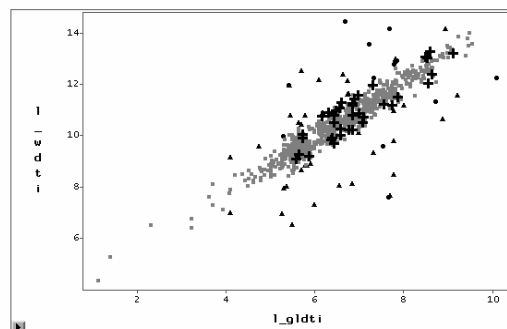


Table 4: *Number of selected units by method*

		<i>Selective Editing</i>		
		<i>Selected</i>	<i>Not selected</i>	<i>Total</i>
<i>LTS</i>	<i>Selected</i>	13	37	50
	<i>Not selected</i>	37	901	938
	<i>Total</i>	50	938	988

It is important to remark that there is a low overlapping of observations found by both the methods. *Selective editing* identifies as potentially influential errors not only observations which are far away from the regression line (used for obtaining the predictions), but also observations that are not regression outliers but which assume a relevant importance with respect to the target estimates due to their sampling weights. On the other hand, some regression outliers are not selected in that they are associated to small sampling weights. Nevertheless, outlier detection technique points out observations in which relations between variables are anomalous with respect to the data model assumed, and its use should be addressed for searching errors in data. This is particularly important since Official Statistics data are usually analyzed by end-users which may focus on target estimates different than those taken into account in the *selective editing* score function.

4. Conclusion and future work

In this paper we have discussed several issues concerning the identification of outliers and potentially influential errors through outlier detection methods and *selective editing*. In the practice of official statistics, univariate methods are generally preferred to multivariate since they are generally simpler. However they cannot detect units violating the correlation structure of the data. Despite of their higher theoretical complexity, the multivariate methods used in this paper are simple to use since they are available in commercial software. Based on the results obtained in this paper, we cannot recommend a unique method to detect errors because some methods are efficient for detecting certain types of errors but fail to detect others. As a general recommendation outlier detection and *selective editing* should be both carried out, especially when microdata are to be released, since they highlight different types of potentially influential errors. Since outliers due to errors in the data are mixed in with correct, but extreme, data values, in order to better identify non representative outliers, variability can be reduced by appropriate stratification.

The paper is a first attempt to analyze the potential of outlier detection methods in terms of capability of identifying suspicious data corresponding to non representative outliers. Further studies, for instance based on simulations, are needed to actually evaluate the performance of different approaches in correctly identifying errors.

The study performed in this paper does not analyze the actual impact of extreme values in the models used during the imputation phase. This is another important aspect to deal with when studying outliers/influential errors. Future works will be devoted to the inspection of this issue. Finally, future work should be devoted to the use of different multivariate approaches to outlier detection, in particular projection pursuit techniques that are particularly useful in high dimensional problems (Rousseeuw and Leroy, 1987).

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.
- Granquist L., Kovar J.G. (1997) Editing of Survey Data: How Much is Enough?, in: *Survey Data Editing: methods and Techniques*, Volume 1, United Nations, New York and Geneva, 127-137.
- Hedlin D. (2003) Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics, *Journal of Official Statistics*, 19, n.2, 177-199.
- Hidioglou M. A., and Berthelot J.M. (1986). Statistical edit and imputation for periodic surveys. *Survey Methodology*, 12, 73-83.
- 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.
- 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.
- Rousseeuw P. J. (1984) Least Median Squared Regression. *J. Am. Stat. Assoc.*, 79, 871-880.
- Rousseeuw P. J. (1985) Multivariate Estimation With High Breakdown Point, in: *Mathematical Statistics and Applications*, Vol. B, Grossmann W., Pflug G., Vincze I. and Wertz W. (Eds.), Dordrecht: Reidel, 283-297.
- Rousseeuw P. J., Leroy A. M. (1987) *Robust Regression and Outlier Detection*, Wiley, New York.
- Rousseeuw and Van Driessen (1999) A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41, 3, 212-223.
- Rousseeuw, P. J. and Van Driessen, K. (2000), An Algorithm for Positive-Breakdown Regression Based on Concentration Steps, in: *Data Analysis: Scientific Modeling and Practical Application*, Gaul W., Opitz O. and Schader M. (Eds), Berlin: Springer-Verlag, 335-346.