

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Rome, Italy, 2-4 June 1999)

Topic (iii): New methodological and technological developments in statistical data editing

COMBINING MACROEDITING AND SELECTIVE EDITING TO DETECT INFLUENTIAL
OBSERVATIONS IN CROSS-SECTIONAL SURVEY DATA

Submitted by the National Statistical Institute, Italy ¹

Invited paper

I. INTRODUCTION

1. The presence of missing and outlying values in sample surveys may unduly affect inferences from sampled data to the parameters of interest in the population. The notion of outliers in survey sampling has a somewhat different meaning with respect to, for example, the corresponding notion in regression analysis, where the observations are assumed to be a random sample from an infinite population with a given parametric distribution and outliers are unique observations generated from some contaminated distribution. In the former, typically, samples are selected from the finite population and each sampled unit is assigned a weight, which usually equals the inverse of the sampling fraction. Therefore, a sampled unit which is "distant" from the bulk of the observed data may not affect the estimates of the population parameters if it has a small weight. In order to make inferences on samples from the finite population, it is then important to detect, among extreme values, the influential observations, i.e. those outliers which can greatly influence the parameter estimates, because of their large weights. These are what Chambers (1986) defines as *representative* outliers, keeping them distinct from *nonrepresentative* outliers, which are either sampled values incorrectly recorded or unique elements of the population (having unit weight).

2. In this classification, such as in some other papers dealing with outlier treatment in sample surveys, it is implicitly assumed that the data are clean and may contain only genuine outliers. Indeed, from the point of view of the data collection agency, which is responsible for releasing public-use data files that provide non-response weights and imputation, the notion of influential observations has to be extended to include possibly other non-sampling errors such as missing values or gross errors due to measurement errors, which will be distinguished from true outliers during the data editing steps.

3. The influence of a sampling unit may also depend on the estimator used. While, for example, an influential unit for the classical Horvitz-Thompson estimator may depend on the sample weight, this is not true for the ratio estimator. This is more evident when dealing with business surveys, where the variable of interest often has skewed distribution. In this case, estimates of population totals are often as badly affected by samples which do not contain any large values as by those that do.

4. When compared to the vast range of the publications on outliers in such diverse areas as linear models or time series analysis, the contributions to the study of outliers in sample surveys are rather sparse. Lee (1995) provides an excellent review on the outliers problem in sample surveys. For a review of the literature

¹ Prepared by Orietta Luzi and Alessandro Pallara.

on outliers in regression analysis the reader is referred to Chatterje and Hadi (1986). Moreover, while the research on the outlier problem for infinite populations is fairly balanced between methods for outlier detection and outlier treatment, most of the published papers on outliers in sample surveys concern estimation methods in the presence of outliers. These include methods based on either winsorization and trimming (Searls, 1966; Potter, 1990) or weight reduction techniques (Hidioglou and Srinath, 1981; Ghangurde, 1989) and methods based on robust estimation techniques. These may be further distinguished between robust M or GM-estimators using a model-based approach (Chambers, 1981; Lee, 1991) and robust versions of classical design-based estimators (Gwet and Rivest, 1992; Hulliger, 1995).

5. Methods for outlier detection in finite population samples have been mainly motivated from the need to rationalize the data editing phase in a survey process, in order to realize timeliness and cost effective editing and follow-up strategy for survey data containing item non-responses and outlying values. As an alternative to complete editing of all sampling units failing one or more edit rules, several techniques have been proposed to limit recontact and follow-up to suspicious units that may have a significant effect on survey estimates. These techniques include *aggregate* and *top-down* methods (cfr. Granquist, 1992), statistical editing (Hidioglou and Berthelot, 1992) and *selective* editing (Latouche and Berthelot, 1995). The latter two are variations of micro-editing, that is editing of the actual elementary data item, with the notable feature that validation actions for the outlying records is confined to sample unit showing a large change from the previous period. Application of these techniques requires, therefore, the availability of historical information at micro-level, which can be achieved only for periodic surveys.

6. In this paper, a method is proposed for extending the selective editing approach to cross-sectional surveys, where information on the sampled units is available only for one survey occasion. This is a situation typical of many households and business surveys, whose sample design does not contain any panel or rotated panel component. Indeed, information on previous sample occasion is used at aggregate level, in order to locate suspicious variations of survey variables for subsets of data deemed to hold a stable relationship over time for the variables of interest. Then, a modified version of the score function of selective editing is computed at micro data level, adapting to the analysis of survey data the idea of *case deletion* used in linear regression diagnostics for detection of influential observations (Smith, 1987).

7. The paper is organized as follows. Section II contains an outline of the proposed methodology. In section III the results of an application to data from the annual survey of small manufacturers are illustrated. Finally, section IV includes a discussion of some open questions and further developments of the technique. The selective method presented in this paper is part of an editing strategy aimed at releasing complete "clean" data file, which involves a selective recontact and follow-up activity together with automatic imputation of non-influential errors.

II. OUTLINE OF THE PROPOSED METHODOLOGY

8. Outliers detection is performed at a subpopulation level. This involves that a preliminary stratification of the data is carried out, to achieve a partition of the sampled data into subsets (*domains*), deemed to be homogeneous with respect to the survey variables to be analysed. In the application discussed in the next section, stratification is obtained using an *a priori* criterion, depending on size and classification of economic activity of each statistical unit.

9. The editing procedure is built up in three main steps:

- (i) the *domains with the largest variation* with respect to the previous period (e.g., last year, for yearly survey) for the variables analyzed are selected;
- (ii) a search is performed for observations suspected of containing either *outlying values* or anomalous variation for some of the variables, using a measure of extremeness of the differences between ordered observations (a variation of the *quartile method*);
- (iii) the *influential observations* are selected among the potential outliers identified in the previous steps, namely those observations having a notable impact on the population total estimates.

10. Both steps (i), for the selection of domains, and (iii), for the definition of influential observations, use modifications of the multivariate *score function* proposed by Latouche and Berthelot (1992) for selective

editing. However, selective editing approach has been proposed for repeated surveys where variation of each elementary unit can be traced through time for all the survey occasions. In this paper the score function is adapted to *cross-sectional survey* using the idea of *case deletion* used in linear regression diagnostics for detection of influential observations (Chatterje and Hadi, 1988, Smith, 1987). A detailed description of the proposed editing strategy now follows.

11. The selection of domains within which restrict the search for influential observations is accomplished using an aggregate score function depending on: *i*) the amount of variation for the subset of study variables from time $t-1$ to time t for domain D_v , ($v=1, \dots, V$); *ii*) the influence of the variation in each domain on the total variation of the variables from $t-1$ to t .

12. Suppose to be interested to an estimate of the population total of a variable X_j . With m variables and V domains, an estimate of the total of j -th variable in the v -th domain at time t is given by:

$$\hat{X}_{jt}^{D_v} = \sum_{i=1}^{n_{vt}} X_{ijt} / p_{it} \quad (j=1, \dots, m; v=1, \dots, V) \quad [1]$$

where n_{vt} are the sampled units belonging to domain D_v at time t , X_{ijt} denotes the value of the j -th variable for the sampled unit i at time t , while p_{it} is the sampling fraction of i at time t . Then

$$\hat{X}_{jt-1}^{D_v} = \sum_{i=1}^{n_{vt-1}} X_{ijt-1} / p_{it-1} \quad (j=1, \dots, m; v=1, \dots, V) \quad [2]$$

is the estimate of the total of X_j at time $t-1$.

13. A simple indicator of the influence of the variation of the variable j from $t-1$ to t in each domain is defined by:

$$D_j^{D_v} = \frac{|\hat{X}_{jt}^{D_v} - \hat{X}_{jt-1}^{D_v}|}{\sum_v |\hat{X}_{jt}^{D_v} - \hat{X}_{jt-1}^{D_v}|} \quad (j=1, \dots, m) \quad [3]$$

Note that the sum of [3] over all V domains does not equal the total variation of j -th variable from time $t-1$ to time t .

14. A global measure of the influence of each domain D_v is given by:

$$D^{D_v} = \sum_{j=1}^m D_j^{D_v} t_j \quad [4]$$

where t_j is a weight ($0 \leq t_j \leq 1$) related to the importance of the variable j .

15. Having selected a critical value D , all domains D_v with $D^{D_v} > D$ are labeled as suspicious, since they have anomalous variation for one or more variables from the previous to the current period. This critical value would be defined using, e.g., the empirical distribution of D^{D_v} , in a way to approximately achieve a predetermined recontact rate (the actual number of recontacts will be determined only at the end of the editing procedure). Further steps of the procedure are then restricted to the observations belonging to domains D_v labeled as suspicious.

16. Within each domain with anomalous variation, a subset of observations is then selected as containing suspicious data for one or more variables examined. The check for possible errors is performed through *ratio edits*, based on the relationships among the survey variables; namely, if for one respondent the values of one or more ratio lay outside predetermined bounds then the observed data item should contain an error. The lower and upper bounds for the ratios have been defined using the program D-MASO (*Distance Measurement Algorithm for Selection of Outliers*, Paletz, 1993).

17. The suspicious units localization process takes the following steps:

- (i) definition of the subset W of ratios R_{jk} ($j, k=1, \dots, m$) and for each R_{jk} definition of the bounds $[\text{Inf}_{jk}^{D_V}, \text{Sup}_{jk}^{D_V}]$ of the tolerance interval in each suspicious domain D_V ;
- (ii) for each suspicious domain D_V , selection of units i having at least one R_{jk}^i located outside the corresponding acceptance bounds.
- (iii) An error indicator $z_{R_{jk}^i}$ is then created assuming value 1 for unit i whose R_{jk}^i lies outside the acceptance bounds and zero otherwise.

18. An editing procedure is then carried out for the selection of influential observations among the subset E_{D_V} of sampling units containing at least one potential error, detected through the suspicious units localization process of steps (i) - (iii) described in para 17. The influence of each unit with respect to the variables of interest is measured trying to account for the following elements: *i*) size of the potential errors located through the ratio edits; *ii*) influence of each potential error on the estimates of the average value of the ratio in each domain; *iii*) influence of each unit on the estimates of the population total of the variables of interest; *iv*) sampling weight; *v*) relative importance of the ratio.

19. For each unit $i \in E_{D_V}$ a score function is computed which is defined as:

$$f_i = f(f_i^{jk} \cdot z_{R_{jk}^i}) \quad [5]$$

where:

$$f_i^{jk} = g[e(R_{jk}^i); I_{jk}(e), t_{jk}; W_{D_V}^i; w_{it}] \quad [6]$$

is a score computed for each couple j, k ($j, k=1, \dots, m$) of survey variables entering any of the ratios included in the editing procedure.

In expressions [5] and [6]

- $e(R_{jk}^i)$ is a measure of the potential error in the ratio R_{jk}^i ;
- $I_{jk}(e)$ is a weight ($0 < I_{jk}(e) \leq 1$) related to $e(R_{jk}^i)$ which measures the influence of each potential error on sample estimates defined on the ratio;
- $z_{R_{jk}^i}$ is the indicator variable for ratio R_{jk}^i ²;
- t_{jk} is a weight, determined subjectively, which measures the relative importance of the ratio R_{jk} ;
- $W_{D_V}^i$ is a weight ($0 < W_{D_V}^i \leq 1$) which measures the influence of each unit on the estimates of the population total of the variables of interest;
- $w_{it} = 1/\rho_{it}$ is the sampling weight of i at time t .

20. The measure of the potential error in any ratio R_{jk}^i is given by:

$$e(R_{jk}^i) = \frac{R_{jk}^i}{\bar{R}_{jk}^{D_V(-i)}} \quad [7]$$

where $\bar{R}_{jk}^{D_V(-i)}$ is the average value of the ratio R_{jk}^i computed in the domain D_V with the i -th unit excluded, namely:

² Clearly, only the ratios with $z_{R_{jk}^i} = 1$ will contribute to f_i .

$$\bar{R}_{jk}^{D_v(-i)} = \hat{m}(R_{jk}^l; l, i) = \frac{\bar{X}_j^{D_v(-i)}}{\bar{X}_k^{D_v(-i)}} \quad [8]$$

where

$$\bar{X}_j^{D_v(-i)} = \frac{\sum_{l \in D_v(-i)} X_j^l \times w_{lt}}{\sum_{l \neq i} w_{lt}}$$

and a similar expression holds for $\bar{X}_k^{D_v(-i)}$.

21. The weight $I^{ijk}(\epsilon)$ associated to the potential error is defined as:

$$I^{ijk}(\epsilon) = \frac{|\bar{R}_{jk}^{D_v} - \bar{R}_{jk}^{D_v(-i)}|}{\max(\bar{R}_{jk}^{D_v}, \bar{R}_{jk}^{D_v(-i)})} \quad [9]$$

where

$$\bar{R}_{jk}^{D_v} = \hat{m}(R_{jk}^i) = \frac{\bar{X}_j^{D_v}}{\bar{X}_k^{D_v}} = \frac{\sum_{i \in D_v} X_j^i \times w_{it}}{\sum_{i \in D_v} X_k^i \times w_{it}}$$

and $\bar{R}_{jk}^{D_v(-i)}$ is given by [8].

22. The influence of each unit on the estimates of the population totals of the variables of interest is given by:

$$W_{D_v}^j = \frac{\tilde{X}_1^{D_v} + \tilde{X}_2^{D_v} + \dots + \tilde{X}_m^{D_v}}{m} = \hat{m}(\tilde{X}_j^{D_v}) \quad [10]$$

where:

$$\tilde{X}_j^{D_v} = 1 - \frac{\sum_{l \in D_v(-i)} X_j^l \times w_{it}}{\sum_{l \in D_v} X_j^l \times w_{it}} \quad j=1, \dots, m$$

and $0 < \tilde{X}_j^{D_v} \leq 1$. $\tilde{X}_j^{D_v}$ attains its maximum when the value X_j^i of the i -th unit account for the variable X_j in the domain D_v . The parameter $W_{D_v}^j$ is therefore a global measure of the influence of i -th unit on the estimates of the population total of all the variables of interest.

23. Special cases of the score function [5] are obtained for sampling unit containing item nonresponse and for unit nonresponse. In the first case, the score f_j will be determined from the f_i^{jk} score functions of the ratios not involving the missing variables for the i -th unit, from the influence of the i -th unit on the population totals of the variables with no item nonresponse for the i -th unit, from the sampling weight w_{it} , and the factor t_{jk} . For sampling units containing missing responses for all the m variables of interest the score function will take a missing value: this kind of units can be considered unit nonresponse, and they could be excluded from the editing process. In our application we used only two of the survey critical survey, so in this case we decided to include unit non response in the interactive editing process.

24. The procedure for the localization of the influential observations within each domain then takes the following steps:

- (i) the score function f_j is computed for all $i \in E_{D_v}$;
- (ii) f_j 's are sorted in ascending order: a critical value F of the empirical distribution is determined and units with $f_j > F$ are selected as influential and the suspicious data item (variable) is identified;
- (iii) all sampling units considered as influential has to be recontacted. Therefore, incorrect data will be corrected, while genuine outliers will be accounted for in the estimation phase.

25. The selection of a suitable score function form may depend on operational aspects of the editing procedure, related to the characteristics of the survey variables entering the ratios which contribute to the score. Latouche and Berthelot (1992) use an additive form for their global score, whose modified version is represented from [5]. For the application an additive form for [5] and a multiplicative form for [6] were used. The explicit form for the score function used in application at enterprise level is then:

$$f_i = \sum_{j,k} [(e(R_{jk}^i) \cdot I_{jk}(e) \cdot t_{jk} \cdot y_i) \cdot z_{R_{jk}^i}] \quad [11]$$

III. AN APPLICATION TO THE SURVEY OF SMALL MANUFACTURERS

26. The editing procedure described in the previous section has been applied to the sample survey on Small Manufacturers (SM in the following) conducted every year by the Italian National Statistical Institute (ISTAT, 1997). The purpose of SM survey is to provide information about a very important part of Italian production system (firms having less than 20 employees in industrial and service sectors) and to supply estimates of some National Accounts aggregates.

27. Data collection is by mailout/mailback questionnaire. The sample, containing about 100,000 units, is stratified by size, industry group and geographical localization. The survey is characterized by a high non response rate (about 28% in 1993-1994 years) because of the data collection mode and business demography (high rates of births and deaths, changes of industry division or size class).

28. In the current data editing process each questionnaire is first submitted to a manual pre-editing, checking for the presence of a few basic questionnaire items. Then each questionnaire is submitted to an automatic micro-editing procedure, consisting of the application of a complex set of intra-record coherence rules. Whenever an edit failure occurs, according to the number and type of failed rules, records are either automatically filled using deterministic imputation or selected for questionnaire manual review and follow-up.

29. This interactive micro-editing process is obviously very costly and time consuming. The application of the proposed editing approach is restricted to a subset of divisions (two-digits) of the NACE REV. 1 classification, (NACE in the following), namely NACE 15 (*Foods and drinks industry*), 17 (*Textile industry*), 24 (*Chemical industry*) and 29 (*Mechanical industry*), for a total of slightly less than 5000 observations. The selection of variables and ratios used in the editing exercise have been guided from the requirements of the *COUNCIL REGULATION (EC) No. 58/97 concerning structural business statistics*, which asks each member state to provide preliminary estimates of a set of survey variables within a specified transmission period.

30. The survey variables selected for the application will be *Revenues* and *Value added*. The application uses data from the 1993 and 1994 surveys. "Clean" data for 1993 are used as *historical* information on enterprises while 1994 will be the reference year for *current* estimates. Since only clean data were available for 1994, an artificial raw data file has been generated, simulating among 1994 data a pre-defined amount of non sampling errors (item non responses and extreme values), reproducing the probability structure of non-sampling errors in the raw data file of the 1995 survey.

Raw data simulation

31. Raw data of 1995 survey were checked on the basis of a set of deterministic rules in order to verify internal consistency of main information on enterprises within or between questionnaire sections. The main purpose was the exclusion of out-of-scope units and gross data entry errors from subsequent steps. Then the probability of item non responses and extreme values among 1994 data was estimated through the frequency of these events in 1995 sample. This analysis was carried on separately in each NACE division considered in the application, for two size classes (1-9 employees, 10-20 employees).

32. Extreme values for each variable of interest in each domain were localised using a simplified version of the D-MASO algorithm (Paletz, 1993). The algorithm was applied to univariate distributions, in order to identify the acceptance ranges of each variable in each domain.

33. An error simulation process was then performed on data for 1994 using an algorithm (Generalised Error Simulation System, GESSY; Della Rocca, Luzi, 1997) for random generation of item non responses and extreme values for each variable of interest in each domain. The resulting artificial raw data file will be denoted as *raw data* (for the current year).

Selection of critical domains

34. In the first step of the editing strategy (*selection of suspicious domains*) the D_j^{Dv} score were computed on pre-defined domains of sampled units in order to determine a list of critical domains. To this end, the sample of current data has been stratified by three firm size (1 employee, 2-9 employees, 10-20 employees) and industry group (three digits of NACE classification), resulting in 85 sub-samples.

35. As mentioned previously *Revenues* and *Value added* were used to calculate the D_j^{Dv} score defined in the [3]. Preliminarily, both t_1 and t_2 in [4] were set to 1. Table 1 summarizes the results of domain selection process. The table shows the first 14 domains ordered by descending D^{Dv} values. The first 12 domains (15% of the total number of domains) have been selected as critical for further steps of the editing procedure. They sum up to about 60% of the absolute difference between historical and current totals of both *Revenues* and *Value Added*. The choice of the threshold for the selection of a domain as “critical” will of course depend on a predefined cut-off value of the maximum accepted difference between historical and current estimates in each domain as well as on the resources available for follow-up.

Table 1. Global D^{Dv} scores, partial D_1^{Dv} and D_2^{Dv} cumulate scores by domain ordered by descending D^{Dv} values, *Revenues* and *Value added* percent differences between historical and current raw data and percentage of simulated errors by domain.

Domain (Nace group- size class)	Δ^{Dv}	Δ_1^{Dv} <i>Cumulate</i>	Δ_2^{Dv} <i>Cumulate</i>	% Total <i>Revenues</i> difference	% Total <i>Value</i> <i>Added</i> difference	% of <i>Revenues</i> simulated errors	% of <i>Value</i> <i>Added</i> simulated errors
158-1	0.197	0.112	0.084	23.96	19.24	7.3	15.2
292-3	0.152	0.198	0.150	43.35	35.38	9.9	16.7
177-1	0.147	0.203	0.292	17.33	83.95	10.4	16.7
295-2	0.145	0.282	0.358	46.59	41.81	8.0	17.0
246-3	0.094	0.337	0.397	56.87	45.78	5.9	14.7
295-3	0.079	0.379	0.434	26.36	22.77	5.4	20.6
155-3	0.071	0.416	0.467	35.93	30.15	5.4	12.7
172-2	0.069	0.456	0.496	46.6	37.39	4.9	6.2
292-2	0.068	0.497	0.523	30.39	21.78	7.3	15.5
176-3	0.066	0.536	0.550	58.51	47.50	13.0	17.4
172-3	0.063	0.570	0.579	31.75	26.17	8.6	17.2
241-3	0.048	0.596	0.600	53.85	46.24	8.8	29.4
159-3	0.047	0.613	0.630	16.87	25.11	9.4	23.4
171-3	0.044	0.633	0.654	34.15	35.22	10.0	16.7

Localization of suspicious data and selection of influential units

36. The first step of this phase has been the calculation a set of ratios involving the objective variables (*Revenues* and *Value added*) and some other correlated variables for each enterprise within each suspicious domain. In particular, four ratios have been considered: *Value added/Revenues*, *Wages and salaries/Value added*, *Revenues/Employees*, *Value added/Employees*.

37. Within each suspicious domain D_v , the bounds $[\text{Inf}_{jk}^{D_v}, \text{Sup}_{jk}^{D_v}]$ of the tolerance interval for each R_{jk} have been determined: units having at least one R_{jk} located outside the corresponding acceptance bounds are considered suspicious and will go into next step of the editing procedure. This step consisted in the selection of influential units on the basis of the selective function [11]. For each critical domain units with $f_j > 0$ were

selected as influential and submitted to a simulated interactive editing and correction process. The procedure selected 380 observations (69 of which were classified as unit non responses), i.e. 7.7% of the total of 4932 observations considered in the application. Table 2 shows the number of suspicious units selected in each critical domain and the corresponding percentage of total number of units in the domain.

Table 2 - Number of suspicious units selected by domain

Domain	158-1	292-3	177-1	295-2	246-3	295-3	155-3	172-2	292-2	176-3	172-3	241-3
N	155	47	13	34	9	28	12	14	41	7	10	10
%	15.9	24.6	27.1	18.1	26.5	15.2	21.8	17.3	17.7	15.2	17.2	29.4

Preliminary estimates and results evaluation

38. For each outlying observations located in the previous step and corresponding to units with artificially generated errors, the “interactive” correction process has been carried out, substituting for the “raw” values of *Revenues* and *Value added* the “true” values, obtained from the 1994 clean data set. In this way, 221 (28.1%) of the 786 artificially generated errors were localised and corrected by the procedure. This will allow to measure the performance of the proposed editing strategy in terms of its *efficacy*, i.e. the capability of correctly identifying influential errors, artificially introduced among the 1994 clean data. The efficacy has been measured in terms of the *distance* between the preliminary estimates produced by the selective editing procedure and the corresponding “final” estimates obtained after the current micro editing procedure. To measure efficacy the following indicator was used:

$$\text{Eff}(D_1, D_2) = \frac{|Est_1 - Est_2|}{\text{Max}(Est_1, Est_2)} \times 100 \quad [12]$$

where Est_1 and Est_2 are the estimates of a given population parameter in the two different data sets D_1 and D_2 .

39. The efficacy index has been calculated for the sampling estimates of *Revenues* and *Value Added* total at three-digits level of the NACE REV. 1 classification, which is the detail of the preliminary estimates required from the *COUNCIL REGULATION (EC) No. 58/97 concerning structural business statistics*. Table 3 in next page shows the index [12] computed between 1994 final estimates (i.e. estimates obtained using the actual micro editing procedure), 1994 “raw” data and 1994 preliminary estimates obtained applying the proposed selective strategy. The table shows also the contribution of the total of each industry group to the overall sampling total of the two variables. The Nace group in the table with associated $\text{Eff}(\text{Clean}, \text{Orig}) > 25\%$ (shadowed in the table) have a low contribution to the overall totals: this is why the corresponding domains were not classified as critical in the first step of the editing strategy. It is immediate to verify that the higher is the strata contribution, the better the editing strategy works.

IV. DISCUSSION AND FURTHER DEVELOPMENTS

- Domain definition: the procedure herein proposed for detecting influential observations involves a preliminary stratification of the data into homogeneous domains with respect to the variables under analysis, among which in the first step of the procedure the aggregates with the largest period-to-period variation are selected as suspected to contain outlying observations. The application presented in the paper utilizes a stratification based on *a priori* criterion depending on size and classification of economic activity of each statistical unit. Different criteria for this stratification could be explored, e.g. adapting the solution proposed by Rosenbaum and Rubin (1983) and Little (1986) for survey nonresponse adjustments based on a model of the probability to respond or, alternatively, using a decision rule for homogeneous class definition based on recursive partitioning (Breiman *et al.*, 1984).
- Because the procedure is aimed at identifying influential observations, namely sampling units having a significant effect on the estimates, tentatively it seems reasonable to use a non robust estimate of location, such as the mean, in order to measure, e.g., in the [7] the error in the ratio R_{jk}^i . This is because an estimator which is sensitive to extreme values will change significantly when the i -th unit is excluded from the calculation, such as it is the case in the denominator of the [7], if i is an influential unit.

However, if the outliers are somewhat clustered it could be the case to consider a more robust alternative to the sample mean with the i -th observation deleted in the denominator of the [7], in order to prevent *masking effects*. To this end, a possible alternative could be building up an outlier detection procedure based on *cross-validation* (Breiman *et al.*, 1984). Case deletion diagnostics is a special case of v -fold *cross-validation*.

- Outlier test: improve score function proposed using formal statistical testing of outlyingness under specified distributional assumptions (Barnett, 1993)
- Integrate the score function proposed for detecting influential observations within a dynamic graphical environment in order to start building a generalized system for exploratory analysis and statistical editing of survey data, cfr. Weir *et al.* (1997).

Table 3 – Efficiency index values between original, raw and clean estimates and percentage contribution by publication domains

Stratum	Revenues			Value Added		
	<i>Eff(Orig,Raw)</i>	<i>Eff(Clean,Orig)</i>	% Contr	<i>Eff(Orig,Raw)</i>	<i>Eff(Clean,Orig)</i>	% Contr
151	11.71	11.71	7.08	10.21	10.21	7.49
152	0.00	0.00	0.63	0.13	0.13	0.65
153	12.72	12.72	0.70	6.31	6.31	0.86
154	25.35	25.35	2.42	14.26	14.26	2.38
155	16.11	1.06	4.86	11.21	0.38	5.21
156	14.26	14.26	4.87	15.78	15.78	5.41
157	4.68	4.68	3.02	3.22	3.22	3.63
158	20.40	4.84	15.78	18.05	5.97	14.94
159	15.09	15.09	5.77	11.11	11.11	6.29
171	4.74	4.74	2.37	5.50	5.50	2.47
172	26.64	1.58	4.55	16.30	0.44	4.58
173	11.18	11.18	1.62	6.76	6.76	1.34
174	14.79	14.79	2.40	13.25	13.25	2.29
175	4.57	4.57	2.72	5.59	5.59	2.67
176	39.94	17.00	2.54	27.77	10.56	2.37
177	9.86	9.59	7.03	42.10	9.05	6.83
241	48.18	26.09	1.63	38.98	20.46	1.50
242	33.04	33.04	0.41	18.72	18.72	0.41
243	2.57	2.57	1.53	14.95	14.95	1.80
244	19.01	19.01	0.99	11.82	11.82	0.90
245	33.35	33.35	1.96	21.63	21.63	1.70
246	41.58	3.87	2.27	27.70	2.02	2.33
247	0.00	0.00	0.08	3.56	3.56	0.08
291	14.14	14.14	2.42	8.98	8.98	2.22
292	29.13	6.42	8.47	22.24	3.53	7.83
293	2.42	2.42	2.16	9.60	9.60	2.45
294	7.53	7.53	1.54	1.81	1.81	1.60
295	28.28	0.14	7.60	24.40	0.44	7.19
296	0.00	0.00	0.03	1.80	1.80	0.03
297	0.00	0.00	0.53	0.08	0.08	0.53

References

- BARNETT, V. (1993): "The problem of outlier tests in sample surveys", *Communications in Statistics - Theory and Methods, series A*, 22: 2703-2721
- BREIMAN L., FRIEDMAN J.H., OLSHEN R.A., STONE C.J. (1984): *Classification and Regression Trees*, Wadsworth International, Belmont, CA.
- CHAMBERS R.L. (1986) "Outlier Robust Finite Population Estimator", *Journal of the American Statistical Association*, 81: 1063-1069.
- CHATTERJE, S., HADI, A.S. (1986): "Influential Observations, High Leverage Points, and Outliers in Linear Regression", *Statistical Science*, 1: 379-416.
- CHATTERJE, S., HADI, A.S. (1988): *Sensitivity Analysis in Linear Regression*, Wiley, New York.
- GHANGURDE, P.D. (1989): "Outliers in sample surveys", *American Statistical Association, Proceedings of the Survey Research Methods Section*, 736-739
- GRANQUIST L. (1992): "A Review of methods for rationalizing the editing of survey data", in: United Nations Statistical Commission and Economic Commission for Europe, *Statistical Data Editing Methods and Techniques*, Vol. I.
- GWET, J.P., RIVEST, L.P. (1992): "Outlier resistant alternatives to the ratio estimator", *Journal of the American Statistical Association*, 87: 1174-1182.
- HIDIROGLOU M.A., BERTHELOT J.M. (1986): "Statistical Editing and Imputation for Periodic Business Surveys", *Survey Methodology*, 12: 73-83.
- HIDIROGLOU, M.A., SRINATH, K.P. (1981): "Some estimators of the population total from simple random samples containing large units", *Journal of the American Statistical Association*, 76: 690-695.
- HULLIGER, (1995): "Outlier Robust Horvitz-Thompson Estimators", *Survey Methodology*, 21: 79-88.
- ISTAT (1997): *Conti economici delle imprese con addetti da 1 a 19 - anno 1994*, Collana Informazioni, n.43.
- LEE, H. (1991): "Model-Based Estimators That Are Robust to Outliers", *Proceedings of the Seventh Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, 178-202.
- LEE, H. (1995): "Outliers in Business Surveys", in: COX B.G., BINDER D.A. CHINNAPPA B.N., CHRISTIANSEN A. COLLEDGE M.J., KOTT P.S., *Business Survey Methods*, John Wiley & Sons, Inc., New York.
- LATOUCHE M., BERTHELOT J.M. (1992): "Use of Score Function to Prioritize and Limit Recontacts in Editing Business Surveys", *Journal of Official Statistics*, 8: 389-400.
- LITTLE R.J.A. (1986): "Survey Nonresponse Adjustments fo Estimates of Means", *International Statistical Review*, 54: 139-157.
- LUZI O., DELLA ROCCA G. (1998): "A Generalised Error Simulation System to Test the Performance of Editing Procedures", *Proceedings of the SEUGI 16*, Prague , 9-12 June 1998.
- PALETZ D.(1993): "Documentation of Distance Measurement Algorithm for Selection of Outliers (D-MASO)", *U.S. Bureau of the Census Technical Report*.
- POTTER, F.J. (1990): "A study of procedures to identify and trim extreme sampling weights", *American Statistical Association, Proceedings of the Survey Research Methods Section*, 225-230.
- ROSENBAUM P.R., RUBIN D.B. (1983): "The central role of the propensity score in observational studies for causal effects ", *Biometrika*, 70: 41-55.
- SEARLS, D.T. (1966): "The Estimator for a Population Mean Which Reduces the Effect of Large True Observations", *Journal of the American Statistical Association*, 61: 1200-1205.
- SMITH, T.M.F. (1987): "Influential observations in survey sampling", *Journal of Applied Statistics*, 14: 143-152.
- WEIR, P., EMERY, R. and WALKER, J. (1997): "The Graphical Editing Analysis Query System", in: United Nations Statistical Commission and Economic Commission for Europe: *Statistical Data Editing - Methods and Techniques*, vol.2, 51-55.