WP.3
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (i): Selective and macro editing

Multivariate selective editing via mixture models: first applications to Italian structural business surveys

Prepared by: Guarnera U., Luzi O., Silvestri F., Buglielli T., Nurra A., Siesto G. (ISTAT, Italy)

I. Introduction

1. The statistical use of administrative data is proven to produce well known benefits in terms of data quality and costs, both in case they are integrated in already existing statistical production processes in order to increase their efficiency, and in case they represent the main source of information for the target phenomena, possibly integrated with direct survey data for specific coverage/estimation problems.

2. The relevance of the additional problems deriving from the use of administrative data are also widely investigated in literature ([14]): the need of defining and measuring administrative data quality, data linking and data integration problems, data editing, estimation represent the field of study of a number of European cooperation projects, aiming at providing common guidelines and recommendations to Member States.

3. In addition, in the area of economic statistics, the new integrated EU Regulation (FRIBS) underlines the need of increasing output relevance, coherence and consistency across publication domains, of reducing respondents burden and survey costs due to limited resources, and of harmonising definitions and methodology.

4. In the economic area, the current general strategy at Istat is moving from statistical production models where administrative data are used as complementary sources of information mainly for data validation and estimation purposes, to more efficient integrated systems where administrative data represent the core of information on the target population and surveys are used as complementary sources of data ([4]).

5. In this context, some specific applications aiming at exploring the potential benefits deriving from the use of administrative data in the context of the detection of influential errors have been performed, starting from two key surveys: 1) the annual survey on *Information and Communication Technology usage and e-commerce in industry* (ICT) and 2) the annual survey on *Small and Medium Enterprises* (SME).

6. In both surveys, a model-based robust selective editing approach for continuous variables is considered [3], [1], [5]): in particular, the selective editing methodology implemented in the R package SeleMix (*Selective Editing via Mixture models*) [6] has been considered. The SeleMix multivariate editing approach is based on the use of contamination models. A score function strictly related to the expected error in data is defined: differently from most selective editing methods, the threshold identifying the subset of influential units can be statistically interpreted and associated to estimates accuracy. For each unit, data predictions are also provided.

7. As the efficiency of the approach also depends on the reliability of auxiliary information used in models, the aim of current applications is to evaluate the benefits (in terms of quality improvements and costs reduction) deriving from the use, in model estimation, of external information related to the surveys' target phenomena.

8. Actually, the idea is to improve the effectiveness of error detection by directly incorporating the auxiliary information available in external (both administrative and statistical) data sources in the selective editing strategy.

9. In the paper, the applications of SeleMix to ICT and PMI data using as auxiliary information both administrative and statistical data are illustrated: target variables are in both cases enterprises' *Turnover* and *Costs*. Problems relating to harmonization of variables definitions, accuracy of administrative sources and data linking are not considered in the paper, as they are assumed with low impact on the results for the considered variables based on previous data analyses.

10. The paper is structured as follows. In section II the multivariate model-based selective editing approach and the SeleMix tool are illustrated. Section III contains the description of the applications on the ICT and the SME surveys. Concluding remarks and future work are reported in section IV.

II. The selective editing approach implemented in SeleMix

11. The main objective of selective editing is to limit the most demanding activities of data editing (manual reviewing, re-contact, etc.) to those cases where the expected benefit is highest ([9],[10]). In practice to each unit in a dataset a score function is assigned so that observations with the highest score functions are supposed to contain the most potentially influential errors ([8]). A specific value for the score function (threshold) is generally specified in advance so that the units with score above the threshold are selected for manual editing. Since usually the estimates of interest (e.g. totals) involve several variables, a different score functions is computed for each variable (local score) and a unique global score is obtained by suitably combining these local scores.

12. Typically the score functions are based on comparing the observed values of the variables of interest with the corresponding predicted values obtained through some explicit or implicit model. An important problem with this approach is that errors tend to be identified with residuals with respect to the assumed model, so that it is difficult to separate the natural variability of the investigated phenomenon from the extra variability due to the presence of errors in data. Moreover it is not obvious how to define the threshold determining how many units have to be selected for interactive editing.

13. A recent approach ([3], [6]) tries to overcome this difficulties by explicitly modelling both true data and error mechanism. In particular, in order to capture the intermittent nature of the model, a two component mixture model is used where the components are naturally associated with error-free and contaminated data respectively. In other word, errors are assumed to affect only a subset of data in such a way that each unit in the dataset is corrupted by an error with an (unknown) a priori probability π . In the following, the model is described in some detail.

1) True data model

14. We assume that two sets of variables are observed: the variables of the first group, say X-variables, are supposed to be correctly measured while the second set of variables, say Z-variables, corresponds to items possibly affected by measurement errors. In this set-up, which can be useful when some variables are available from administrative sources or are measured with high accuracy, it is quite natural to treat the variables that are observed with error as response variables and the reliable variables as covariates. This framework includes as a special case, the situation where reliable covariates X are not available, so that what is to be modelled is the joint distribution of the Z variables. In the following we model true data through a log-normal probability distribution. This seems a reasonable assumption in many cases where economic data are to be analysed.

15. According to the previous assumptions, true data corresponding to possible contaminated items are represented as a $n \times p$ matrix Z^* of n independent realizations from a random p-vector assumed to follow a log-normal distribution whose parameters may depend on some set of q covariates not affected by error. Thus, if $Y^* = \ln Z^*$, we have the regression model:

$$Y^* = XB + U \qquad (1)$$

where **X** is a $n \times q$ matrix whose rows are the measures of the q covariates on the n units, B is the $q \times p$ matrix of the coefficients, and **U** is an $n \times p$ matrix whose rows u_i $(i=1,..,n)$ represent normal residuals:

$$u_i \sim N(0, \Sigma) \qquad (2).$$

2) Error model

16. In order to model the intermittent nature of the error mechanism we introduce a Bernoulli r.v. I with parameter π, where $I=1$ if an error occurs and $I=0$ otherwise. In the sequel, Z and Y will denote possible contaminated variable in original and logarithmic scale respectively. Thus, given that $I=0$, it must hold $Z=Z^*$ ($Y=Y^*$). Furthermore, given that $I=1$, errors affect data through an additive mechanism represented by a Gaussian r.v. with zero mean and covariance matrix Σ_ε proportional to Σ, i.e., given the event $\{I=1\}$:

$$Y = Y^* + \varepsilon, \qquad \varepsilon \sim N(0, \Sigma_\varepsilon), \quad \Sigma_\varepsilon = (\alpha - 1)\Sigma, \quad \alpha > 1.$$

17. The error model can be specified though the conditional distribution of the observed data given the true data as follows:

$$f_{Y|Y^*}(y \mid y^*) = (1-\pi)\delta(y - y^*) + \pi N(y; y^*, \Sigma_\varepsilon) \qquad (3)$$

where π (*mixing weight*) is the "a priori" probability of contamination and $\delta(t'-t)$ is the delta-function with mass at t.
In case that the set of X- variables is empty, the variables Y_i ($i = 1,\ldots,n$) are normally distributed with common mean vector μ. From the above specifications it easily follows that the distribution of the observed data is given by:

$$f_Y(y) = (1-\pi)N(y; B'x, \Sigma) + \pi N(y; B'x, \alpha\Sigma) \qquad (4).$$

This distribution can be easily estimated by maximizing the likelihood based on a n sample units via a suitable ECM algorithm.

3) Selective Editing

18. The relevant distribution for selective editing is the distribution of error-free data Y^* conditional on observed data (including covariates X).
Straightforward application of Bayes formula provides:

$$f_{Y^*|X,Y}(y^* \mid x, y) = \tau_1(x, y)\delta(y^* - y) + \tau_2(x, y)N(y^*; \tilde{\mu}_{x,y}, \tilde{\Sigma}) \qquad (5)$$

where τ_1 and τ_2 are the posterior probabilities of belonging to correct and erroneous data respectively:

$$\tau_1(x_i, y_i) = \Pr(y_i = y_i^* \mid x_i, y_i)$$

$$\tau_2(x_i, y_i) = \Pr(y_i \neq y_i^* \mid x_i, y_i) = 1 - \tau_1(x_i, y_i)$$

$i=1,\ldots,n$

and

$$\tilde{\mu}_{x,y} = \frac{(y + (\alpha - 1)B'x)}{\alpha}; \qquad \tilde{\Sigma} = (1 - \frac{1}{\alpha})\Sigma.$$

19. With an obvious shift of notation the corresponding distribution in the original scale is:

$$f_{Z^*|Z}(z^* \mid z) = \tau_1(\ln(z))\delta(z^* - z) + \tau_2(\ln(z))LN(z^*; \tilde{\mu}_{x,\ln z}, \tilde{\Sigma}) \qquad (6)$$

where $LN(\cdot, \mu, \Sigma)$ denotes the lognormal density with parameters (μ, Σ), and for the sake of simplicity, we have suppressed the X variables in the notation whenever they appear as conditioning variables. Estimation of the distribution (6) is obtained by replacing the corresponding parameters with the estimates of $(\mu, \Sigma, \pi, \alpha)$ resulting from the ECM algorithm.

20. Once the target distribution (6) has been estimated, "predictions" of "true" values z_i^*, conditional on observed values z_i can be obtained for all the observations $i=1,...,n$ as:

$$\hat{z}_i = E(z_i^* \mid z_i) = \int z_i^* f_{Z^*|Z}(z^* \mid z) dz_i^*$$

The model predictions can be used to define the *expected errors*

$$\varepsilon_i = (\hat{z}_i - z_i), \qquad i=1,..,n$$

and to compute "robust estimates" of some parameters of a finite population U.

21. For instance, let assume that the target estimate is given by the total T_z of the variable Z, i.e. $T_z = \sum_{i \in U} z_i$ and an estimator $\hat{T}_z = \sum_{i \in S} w_i z_i$ is used where w_i are sampling weights attached to each unit of a sample S of size n. A robust version of $\hat{T}_z$ is given by $\hat{T}_z^* = \sum_{i \in S} w_i \hat{z}_i$, where the last estimator is obtained by the previous one by replacing observed values z_i with the predictions $\hat{z}_i$.

22. A suitable local score function can be defined in terms of expected errors and reference robust estimates. This definition is particularly useful in that it makes it possible to estimate the residual error remaining in the data after the units with the highest expected error have been corrected. It follows that the number of units to be interactively reviewed can be chosen such that the residual error is below a prefixed threshold. Specifically, if for a certain variable Z, r_i denotes the relative individual error defined as the ratio between the (weighted) expected error and the reference estimate T_z^*

$$r_i = \frac{w_i(\hat{z}_i - z_i)}{\hat{T}_z^*},$$

then the score function is defined as: $SF_i = | r_i |$. Moreover, let R_M be the absolute value of the (approximated) expected residual percentage error remaining in data after removing errors in the units belonging to the set M:
$R_M = | \sum_{i \in \overline{M}} r_i |$, where $\overline{M}$ denotes the complement of M in S.

23. Once an "accuracy" threshold η is chosen, the selective editing procedure consists of:

1. sorting the observations in descending order according to the value of SF_i;
2. selecting the first $\bar{k}$ units for reviewing, where:
$\bar{k} = \min\{k \in (1,..,n) \mid R_{M_j} < \eta, \forall j > k\}$ and M_m is the set composed of the first m units.

A suitable extension to the multivariate case is easily obtained by taking the maximum of the different local scores as global score.

24. The described method is implemented in the R package SeleMix (Selective editing via Mixture models) available on the website http://www.R-project.org. This package includes functions for the estimation of the model parameters via EM algorithm, computation of prediction of true values conditional on observed values, prioritization of units for interactive editing according to a user-specified

threshold. Missing values in the response contaminated variables are allowed. In this case SeleMix can also be used as a tool for (robust) imputation of incomplete data. The covariates included in the model are supposed to be error-free and not affected by non response. Thus, the efficiency of the SeleMix approach also depends on the reliability of the available auxiliary information.

III. First applications to structural surveys on enterprises

25. In this section, the applications of SeleMix to the Istat economic surveys ICT and SME are illustrated. For both surveys, the frame is represented by the Italian *Business Register of Active Enterprises* (BR in the following), resulting from the combination of both statistical and administrative information (*Tax Register, Social Security Register, Register of the Electric Power Board*, etc.). BR contains key variables such as *Economic activity*, *Turnover* and *Number of persons employed*. It counts about 4.5 million enterprises which employ approximately 17.6 million persons employed.

26. For our purposes, the following administrative sources of auxiliary information have been considered:

- *Financial Statements* (*FS*) - years 2008 and 2009;
- *Sector Studies survey* (*SS*) - year 2008.

Out of the 4.5 million frame enterprises, corporate companies (about 15.000 records) are liable to fill in *FS*: this is the best harmonized source with the SBS Regulation definitions, and the data from balance sheet are sufficiently reliable to be considered as "true" data. Besides *FS*, Istat gets directly from the Tax Authority the *SS* source, which includes about 4 million enterprises with a turnover greater than 30 thousand Euros and less than 7.5 million Euros: as the common part of *SS* questionnaires is like a financial statement, it can be used in a more effective way than other types of tax return data.

III.1. The application of selective editing to the ICT survey

27. The ICT survey on enterprises with 10 and more persons employed working in industry and non financial services is carried out annually, following criteria and methods shared by all European Union countries. The phenomena observed are those defined annually by EC Regulations. The survey collects information on a sample of small and medium firms (those with less than 250 persons employed), and all the large enterprises. Target variables relate for example to the use of computers and computer networks, to the Internet connection type, to Internet usage in the enterprise, to the electronic exchange of data, to the electronic and automatic sharing of information inside of company, to sale/purchase of goods or services via computer networks (*e-Commerce*).

28. The sampling design is stratified with one-stage selection of units with equal probability: strata correspond to the combination of economic activity, size class and administrative region of the enterprise's administrative office. The multivariate and multi-domain allocation is based on the standard Istat methodology ([2]).

29. Data collection is based on self-compilation of electronic questionnaires embedding a number of edit rules aiming at gathering information with a pre-determined level of accuracy. Editing of categorical data is based on the Fellegi-Holt paradigm ([12]), while the treatment of continuous data is based on a deterministic approach exploiting as much as possible external information from related surveys and FS. Parameter estimation is based on standard weighting and calibration techniques, based on BR auxiliary

variables. For each variable *Y* and a domain *D,* target parameters are the *Y* totals by publication domain *D*: $T_Y^D = \Sigma_{k \in D} \omega_k y_k$, where ω_k is the sampling weight of unit *k*.

30. Besides *FS* and *SS*, the selective editing strategy applied to the ICT survey involves also two statistical sources, integrated with the administrative ones:

- *The Annual Survey on the Economic Accounts of Enterprises* (*SCI*), year 2008;
- *The Annual survey on Small and Medium-sized Enterprises* (*SME*), year 2008.

The *SCI* survey is a total survey on enterprises with 100 or more persons employed (about 12,000 enterprises): together with the sample survey *SME*, which is carried on enterprises with less than 100 persons employed, it collects information on profit-and-loss statements and balance sheets, as well as information regarding turnover, purchases of goods and services, changes in stocks, employment and investments.

31. Target variables are enterprises' *Turnover* and *Costs*: in the current E&I procedure, these variables are checked by analysing acceptance regions of ratio edits whose bounds are determined based on the distributions of the same ratios in the SCI e and SME surveys (referred to the previous year): enterprises failing such ratios are selected and interactively revised, all the other errors are automatically imputed.

32. Based on the premise that target estimates are the totals of *Turnover* and *Costs* for pre-specified publication domains, two types of experiments have been carried out, on either simulated or observed (raw) data.

1) Evaluation on simulated data

33. In order to assess the performances of the proposed selective editing method in terms of trade-off between costs associated with manual reviewing and accuracy of the target estimates, an experiment on simulate data has been conducted. We have evaluated SeleMix by analyzing the variables *costs* and *turnover* on the subset of the ICT sample corresponding to corporate companies. For these companies we have considered the values of the analyzed variables available from the balance sheet source (FS) as "true" values and we have artificially introduced errors in data. In detail, we have treated as response variables (Y-variables) costs and turnover for the current year (2009), the same variables referring to the previous year (2008) have been considered (error free) covariates (X-variables). Records where the variables were not available for both years have been excluded.

34. A 1,000 simulation based Monte Carlo experiment has been conducted where at each run errors have been introduced with probability π =0.05 on Y variables by inflating the model variance by a factor α =10. For model parameter estimation and selection of influential errors the SeleMix functions `ml.est` and `sel.edit` have been used respectively. The threshold for selective editing has been set to η=0.005. Correction of the selected units has been simulated by replacing the Y-values identified by `sel.edit` with the corresponding values available from FS.

35. In Table 1 the results of the experiment are reported. The target estimates are supposed to be totals of 2009 turnover (*turnv*) and costs (*cost*) for different aggregations of economic activity (column *Dom* in Table 2). For these two variables relative bias and root mean square error are reported for each estimator based on raw data, edited data and predictions provided by the model. The estimates take into account the ICT sample weights provided by the survey manager. For each domain of estimation the sample size (*N*) and the number of contaminated records (*n.cont*) are also reported. The column *n.out* contains the number of observations whose probability of being erroneous is greater than 0.5 according to the

estimated contamination model. Finally "n.sel" is the number of units which, according to the prefixed accuracy threshold are identified as "dangerous" and "corrected" in the simulation study.

Table 1. Experiment on simulated data. Relative bias and root mean square error (RRMSE) for the estimates based on raw data (RAW), edited data (EDITED) and SeleMix robust predictions (ROB.EST)

					Relative Bias (%)						RR MSE (%)					
					RAW		EDITED		ROB.EST		RAW		EDITED		ROB.EST	
Dom	N	n.cont	n.out	n.sel	turnv	cost	turnv	cost	turn	cost	turn	cost	turnv	cost	turnv	cost
G	3497	336.7	515.3	116.0	2.8	2.6	0.0	0.0	0.9	1.2	4.2	3.7	0.2	0.2	1.0	1.3
F	3260	317.6	565.2	238.5	15.4	18.1	-0.2	0.0	-7.6	-7.0	22.3	32.8	0.4	0.2	7.7	7.1
DE	876	85.0	143.1	16.2	4.4	13.6	0.1	0.1	-0.2	-1.0	10.4	39.3	0.3	0.5	2.0	1.8
C	3691	362.0	494.0	231.1	13.7	16.3	-0.1	-0.1	0.9	0.3	19.4	23.9	0.3	0.3	1.0	0.7
H	653	62.9	144.6	20.3	2.7	3.3	0.1	0.0	-0.6	-0.8	8.8	10.5	0.4	0.5	0.9	1.0
L	133	13.0	25.8	16.6	44.5	166.7	0.0	-0.1	3.9	10.2	95.4	686.4	1.0	0.7	7.9	11.5
J	565	54.6	76.8	15.1	16.2	19.4	0.0	-0.1	-1.8	-3.6	35.0	50.1	0.6	0.4	2.1	3.8
I	224	22.4	35.5	16.8	6.4	4.6	-0.2	-0.2	2.1	2.9	15.6	12.3	0.8	0.6	2.3	3.0
NS	1156	111.3	211.2	18.4	6.8	6.5	0.2	0.1	0.5	-0.4	11.0	12.1	0.5	0.7	0.9	0.8
M	450	43.4	78.6	38.3	39.2	30.5	-0.1	0.0	-6.1	7.4	79.5	64.3	0.4	0.4	6.1	7.5

36. The table shows that the discrepancy between estimators based on edited data and true data (measured in terms of RRMSE) is quite close to the specified threshold (0.5%). Moreover, analysis of column *n.sel* shows that high accuracy is achieved by correcting a relatively small number of influential errors. In particular this number in much lower than the number of outlying observations (*n.out*) defined as observations with probability of being erroneous higher than 0.5). This should make clearer the difference between outlier identification and selective editing. We also notice using SeleMix predictions (*rob.est*) in the estimators in place of the observed values allows in general to improve the estimate accuracy. However, in some cases (strata F, L and M), inadequacy of the working model causes bad performance of the estimator based on robust predictions, while the model is still using in determining the units to be edited.

2) Evaluation on raw data

37. In order to assess the performance of the selective editing method in terms of potential benefits that could derive by integrating it in the current E&I procedure, an application to the raw *Turnover* and *Costs* data for the ICT responding units has been performed: the evaluation has been carried out by comparing the parameters' estimates obtained after selective editing and the corresponding ones obtained by the current editing approach. Auxiliary variables are *Turnover* and *Costs* available in at least one external source (either administrative or statistical), with a given priority.

38. In the experiment, the 20,028 responding units to the ICT survey in the year 2009-2010 are considered: as only enterprises also present in the corresponding ICT final (*edited*) set of data and available in at least one external source are considered, the resulting experimental data consist of 19,293 enterprises. Let $turnv^{ICT}$ and $cost^{ICT}$ be the *raw* (observed) values of *Turnover* and *Costs*, respectively, and $turnv^{edited}$ and $cost^{edited}$ the corresponding *edited* (final) values. Target estimates are considered for 28 different sectors of economic activity. Sampling weights adjusted for total non responses are used in estimation.

39. As in the experiments different sources are integrated, and given that some enterprises may be observed in more than one source, priorities among sources have been determined based on previous data analyses [4]. The auxiliary variables on *Turnover* and *Costs* ($turnv^{AUX}$ and $cost^{AUX}$) used in selective

editing are finally obtained in the following way: for each responding unit i (i=1,...,19,293), the values of $turnv^{AUX}$ and $cost^{AUX}$ are taken from one of the auxiliary sources based on the priority $SCI \rightarrow FS \rightarrow PMI \rightarrow SS$.

40. The model-based selective editing is applied to Y-variables $turnv^{ICT}$ and $cost^{ICT}$ using as auxiliary variables (X-variables) the triplet $turnv^{AUX}$, $cost^{AUX}$, and number of employees observed in ICT ($nempl^{ICT}$): for each pre-defined value of α, the set of influential units I_α is then identified based on the algorithm illustrated in section 2. Different thresholds have been used (α=0.01 and α=0.02) in order to evaluate the method effectiveness due to the expected accuracy on target parameters estimates.

41. The interactive revision and treatment of each $i \in I_\alpha$ is simulated by replacing the values of $turnv_i^{ICT}$ and $cost_i^{ICT}$ with the corresponding values $turnv^{edited}$ and $cost^{edited}$. Missing values of $turnv^{ICT}$ and $cost^{ICT}$ are imputed too, using either the corresponding $turnv^{edited}$ and $cost^{edited}$, or the model-based predicted values, in order to evaluate the method performance resulting from different imputation strategies.

42. Evaluation is based on: 1) relative distances between the totals' estimates after selective editing ($\hat{T}_D^{y^{sel}}$) and the corresponding estimates obtained from the current E&I procedure ($\hat{T}_D^{y^{edited}}$): $Diff_D^Y(Sel.edited) = \left(\hat{T}_D^{Y^{edited}} - \hat{T}_D^{Y^{sel}}\right) / \hat{T}_D^{Y^{edited}}$, where $Y \in (Turnv^{ICT}, Cost^{ICT})$ and D=publication domain; 2) number of influential units N_i, $i \in I_\alpha$, compared to the corresponding number of manually revised units based on the current outlier detection procedure (2,430 units, ~12% of responding units).

43. In tables 2 and 3 we show the results corresponding to α=0.01 and α=0.02, respectively, for the strategy in which both influential errors and missing values of $turnv_i^{ICT}$ and $costs_i^{ICT}$ ($i \in I$) are replaced by the corresponding $turnv_i^{edited}$ and $cost_i^{edited}$. For each domain (*Dom)*, the following information are reported in tables: the economic section (*sec*), the domain size (N_D), the number of units which, identified as "suspect" and "artificially" corrected (*n.sel*) according to the threshold, the number of missing values (*n.miss*) of each variable, the distances $Diff_D^Y$ (*Edited.Sel*), $Diff_D^Y$ (*Raw.Edited*), $Diff_D^Y$ (*Raw.Sel*) (the prefix $Diff_D^Y$ is omitted in tables for shortness).

Table 2. Experiment on ICT raw data. Distances between between estimates based on SeleMix (Sel), raw data (Raw), and ICT edited data (ICT), by domain and variable – α=0.01.

				Turnover				Costs			
Dom	sec	N_D	n.sel	n.miss	Edited.Sel	Raw.Edited	Raw.Sel	n.miss	Edited.Sel	Raw. Edited	Raw.Sel
1	C	745	3	9	0,90	-0,09	-0,99	6	0,55	4,42	3,89
2	C	338	6	3	1,62	2,46	0,85	5	1,70	-1,24	-2,99
3	C	293	0	0	-0,22	0,00	0,22	1	0,39	-0,25	-0,64
4	C	546	8	2	0,16	1,49	1,33	7	0,21	0,76	0,55
5	C	255	1	1	-0,88	-1,81	-0,92	2	-0,35	-3,88	-3,51
6	C	1036	33	8	0,89	12,79	12,01	8	-0,12	5,70	5,82
7	C	146	6	3	-0,21	-2,27	-2,06	2	0,26	2,91	2,66
8	C	603	19	8	0,34	6,36	6,04	10	-0,24	2,22	2,46
9	C	169	4	1	-0,29	96,71	96,72	1	0,37	5,58	5,23
10	C	416	5	4	0,26	-0,36	-0,62	5	0,96	2,87	1,94
11	D	201	5	0	-0,48	8,95	9,38	1	-0,40	6,58	6,95
11	E	747	17	15	0,27	3,31	3,04	17	0,56	-5,62	-6,22
12	F	5155	16	74	-1,68	77,80	78,17	97	-1,91	11,25	12,91
13	G	620	3	7	0,13	59,40	59,35	9	-0,61	-0,27	0,33
14	G	2795	13	22	-0,94	9,57	10,41	29	0,26	8,43	8,19
15	G	1174	5	17	0,06	3,68	3,62	22	0,30	3,34	3,05
16	H	752	8	5	0,07	50,28	50,25	8	-0,01	53,56	53,57
17	H	29	0	0	0,00	0,00	0,00	1	0,20	-0,06	-0,27

18	I	205	6	4	0,23	39,13	38,99	4	1,97	38,06	36,81
19	I	131	1	4	0,04	-1,50	-1,55	5	3,69	-2,42	-6,35
20	J	120	0	2	0,35	-1,15	-1,50	2	-0,32	-1,32	-1,00
21	J	47	0	0	0,01	0,00	-0,01	0	-0,25	0,00	0,25
22	J	36	2	0	0,18	9,37	9,21	0	0,19	0,17	-0,02
23	J	406	2	1	-1,77	-7,89	-6,02	4	-2,28	-9,20	-6,76
24	L	149	22	2	0,00	0,54	0,54	2	0,42	-1,83	-2,26
25	M	613	16	10	1,17	29,16	28,32	12	1,33	38,13	37,29
26	N	1124	25	16	0,32	63,03	62,91	17	0,39	96,24	96,22
27	N	176	0	2	0,18	-0,49	-0,68	3	0,00	-0,94	-0,94
28	S	74	9	0	0,20	0,00	-0,20	0	0,01	0,00	-0,01
Total		**19,101**	**235**	**220**				**280**			

44. Based on results, it is possible to evaluate the effectiveness of selective editing in terms of both the impact of identified influential errors on final estimates, and of costs/burden reduction (regardless of the ICT imputation strategy currently adopted for predicting item non responses).

45. As it can be seen, when α=0.01 (Table 2), a set of 235 units are selected as influential (~1.2% of the experimental ICT sub-sample) and then imputed, providing totals' estimates having low distances from the corresponding edited estimates for the majority of domains, and for both *Turnover* and *Costs*. Note that for the most part of domains, *Raw.Sel* is less or equal to *Raw.Edited* for both variables: this can be seen as a positive element in the sense that useless re-contacts/changes (in terms of impact on final estimates) of information seem to be avoided.

46. When α=0.02 (Table 3), a set of 118 influential units is selected (~0.6% of the experimental sub-sample): however, based on the same artificial correction strategy, the results are found less stable than for α=0.01 in terms of estimates' distances with respect to the edited aggregates.

Table 3. Experiment on ICT raw data. Distances between between estimates based on SeleMix (Sel), raw data (Raw), and ICT edited data (ICT), by domain and variable – α=0.02.

				Turnover				Costs			
Dom	**sec**	**N_D**	**n.sel**	**n.miss**	**Edited.Sel**	**Raw. Edited**	**Raw.Sel**	**n.miss**	**Edited.Sel**	**Raw. Edited**	**Raw.Sel**
1	C	745	1	9	0,90	-0,09	-0,99	6	1,15	4,99	3,89
2	C	338	3	3	2,05	2,88	0,85	5	2,07	-0,86	-2,99
3	C	293	0	0	-0,22	0,00	0,22	1	0,39	-0,25	-0,64
4	C	546	3	2	-0,71	0,63	1,33	7	0,58	1,13	0,55
5	C	255	0	1	-0,88	-1,81	-0,92	2	1,43	-2,04	-3,51
6	C	1036	12	8	-0,09	11,93	12,01	8	0,18	5,98	5,82
7	C	146	1	3	-0,06	-2,12	-2,06	2	-0,75	1,92	2,66
8	C	603	12	8	0,54	6,55	6,04	10	1,01	3,44	2,46
9	C	169	2	1	-0,29	96,71	96,72	1	-0,56	4,69	5,23
10	C	416	1	4	-0,06	-0,68	-0,62	5	2,36	4,25	1,94
11	D	201	3	0	-1,04	8,44	9,38	1	0,90	7,80	6,95
11	E	747	5	15	1,13	4,14	3,04	17	1,09	-5,06	-6,22
12	F	5155	13	74	-2,33	77,66	78,17	97	-3,14	10,18	12,91
13	G	620	3	7	0,13	59,40	59,35	9	-0,61	-0,27	0,33
14	G	2795	10	22	-0,94	9,57	10,41	29	0,26	8,43	8,19
15	G	1174	3	17	0,06	3,68	3,62	22	0,30	3,34	3,05
16	H	752	2	5	-0,81	49,84	50,25	8	-0,20	53,47	53,57
17	H	29	0	0	0,00	0,00	0,00	1	0,20	-0,06	-0,27
18	I	205	2	4	-2,58	37,42	38,99	4	3,06	38,74	36,81
19	I	131	0	4	0,04	-1,50	-1,55	5	3,69	-2,42	-6,35
20	J	120	0	2	0,35	-1,15	-1,50	2	-0,32	-1,32	-1,00

21	J	47	0	0	0,01	0,00	-0,01	0	-0,25	0,00	0,25
22	J	36	2	0	0,18	9,37	9,21	0	0,19	0,17	-0,02
23	J	406	1	1	-1,77	-7,89	-6,02	4	-2,28	-9,20	-6,76
24	L	149	9	2	0,00	0,54	0,54	2	0,42	-1,83	-2,26
25	M	613	12	10	1,02	29,05	28,32	12	1,63	38,32	37,29
26	N	1124	16	16	-0,21	62,84	62,91	17	0,06	96,22	96,22
27	N	176	0	2	0,18	-0,49	-0,68	3	0,00	-0,94	-0,94
28	S	74	2	0	0,20	0,00	-0,20	0	0,01	0,00	-0,01
Total		**19,101**	**118**	**220**				**280**			

III.2 The application of selective editing to the SME survey

47. The SME sample survey is carried out annually to investigate profit-and-loss accounts of enterprises with less than 100 employees, as well as information on employment, investments, personnel costs and the regional breakdown of some variables in the industrial, construction, trade and services economic activities. As for the ICT survey, for each variable *Y* target parameters are the *Y* totals by publication domain (*D*). The domains are defined by economic activity, size class and administrative region.

48. For our evaluation purposes, we considered the 2008 SME survey data: about 105.000 enterprises were included in the sample, selected based on a one stage stratified random sampling without replacement and with equal probabilities, with the strata defined in terms of economic activity, size classes of persons employed and regions. The sample selection uses the JALES to achieve a negative coordination among samples drawn from the same selection register and thus reduce response burden procedure ([13]). Standard calibration estimation and post-stratification are adopted at the estimation stage.

49. The 2008 response rate was close to 40% in terms of reliable replies. In the current data processing strategy, a preliminary recovering of key variables for non responding enterprises is performed first (mostly from *FS* and social security archives). In the current E&I procedure, the identification of outliers is based on quantile algorithms, while the remaining errors are treated using a deterministic editing procedure. Item non responses are for the most part imputed based on within-cells *nearest-neighbor donor* algorithms.

50. In this study *FS* and *SS* (referred to years *t*=2008 and *t-1*=2007) are used as auxiliary sources with the following priority: for each responding unit *i*, auxiliary variables $turnv_i^{AUX}$ and $cost_i^{AUX}$ are taken from one of the sources based on the priority $FS \rightarrow SS$. In particular, combined *FS* and *SS* 2007 data are used at the model estimation stage, while the combined 2008 data, considered as *true* data, are used at the evaluation stage.

51. The experimental study consisted in the application of selective editing to the *t*=2008 raw values of *Turnover* and *Costs* (Y-variables), ($turnv^{SME}$ and $cost^{SME}$ in the following) on the subset of SME responding enterprises also available in at least one external source at <u>both</u> *t* and *t-1* (29,873 sample units). The estimation domains (*D)* correspond to 72 Divisions of economic activities. For the response variables (Y-variables) $turnv^{SME}$ and $cost^{SME}$, the selective editing model is estimated based on the auxiliary items (X-vector) $turnv_i^{AUX}$, $cost_i^{AUX}$, and number of employees observed in SME ($empl_i^{SME}$).

52. The two thresholds values α=0.01 and α=0.02 are considered again. Let I_α be the subset of influential units identified at the threshold level α. For each influential unit $i \in I_\alpha$ an interactive revision and treatment is simulated by replacing the values of $turnv_i^{SME}$ and $cost_i^{SME}$ with the corresponding 2008

values $turnv^{AUX}$ and $cost^{AUX}$, assumed as *true* data values. Item non responses on target variables are not imputed in this case, in order to isolate the effect of influential errors identification and treatment.

53. The indicators used to evaluate the performance of selective editing are (see section III.1, experiment on raw data): 1) the relative distances between the totals' estimates after selective editing ($\hat{T}_D^{y^{sel}}$) and the corresponding 2008 "true" estimates resulting from the combined administrative data ($\hat{T}_D^{y^{True}}$); 2) the number of influential units N_i, $i \in I_\alpha$, compared to the corresponding number of manually revised units based on the current outlier detection procedure (about 1,500 units each year, 3,6% of responding units).

54. In table 4 the results corresponding to α=0.01 and α=0.02 are shown for *Turnover*. For each domain (economic division *Div)*, the same information as those reported in tables 2 and 3 are shown: economic section (*sec*), domain size (N_D), number of units identified as influential and "artificially" corrected (*n.sel*), number of missing values (*n.miss*) of each target variable; the relative distances $Diff_D^Y$ (*True.Sel*), $Diff_D^Y$ (*Raw.True*), $Diff_D^Y$ (*Raw.Sel*) are also reported[1].

55. When α=0.01, 869 units are selected as influential (~2.9% of the experimental SME sub-sample) and then imputed, providing totals' estimates having distances from the corresponding "true" estimates which in the 89% of domains are less than 1.5 (the median of *Diff*(*True.Sel*) is 0.65). When α=0.02, 382 influential units are selected instead (~0.01% of the experimental sub-sample), with estimates having distances less than 1.5 from the corresponding "true" ones for the 75% of domains (the median of *Diff*(*True.Sel*) is 0.9).

56. The consistent reduction of expected revisions is then balanced, when α=0.02, by less accurate results for a higher number of domains. In order to gather more information on this aspect, the absolute difference between *Diff(True.Sel)* when α=0,01 and *Diff(True.Sel)* when α=0,02 is computed for all domains.

57. In Figure 1 the absolute differences between *Diff(True.Sel)* when α=0,01 and when α=0,02 for *Turnover* are plotted: as it can be seen, for the 17 domains having absolute differences ≥ 1.0, the effect of threshold on the accuracy of selective editing seems to be remarkable. However, for all the other 65 divisions (75% of domains) having absolute differences less than 1.0, estimates accuracy can be considered as comparable. Taking into account the second indicator (*number of identified influential errors*), at this stage of the analysis, it seems preferable a threshold α=0,02 which implies the treatment of 382 influential units (saving about 56% of interactive revisions w.r.t. α=0,01).

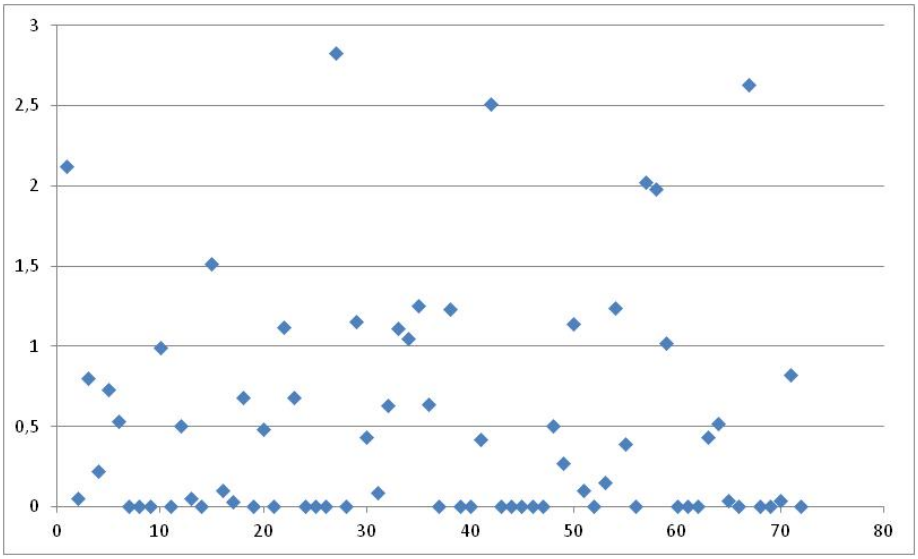
[1] also in this case the prefix $Diff_D^Y$ is omitted in tables for shortness.

Table 4. PMI data. Distances between *Turnover* estimates based on SeleMix (Sel), raw data (Raw), and administrative data (True), by domain.

			α=0.01				α=0.02			
Div	Sec	ND	n.infl.	True.Sel	Raw.True	Raw.Sel	n.infl.	True.Sel	Raw.True	Raw.Sel
10	C	812	3	-1,00	1,12	2,14	0	1,12	1,12	0,00
11	C	122	4	-0,60	0,68	1,29	3	-0,65	0,68	1,34
13	C	495	28	-0,73	6,49	7,28	14	-1,53	6,49	8,15
14	C	542	13	-1,18	-1,96	-0,79	6	-1,40	-1,96	-0,56
15	C	276	3	-0,04	-0,77	-0,73	0	-0,77	-0,77	0,00
16	C	413	6	0,23	-1,11	-1,34	3	-0,30	-1,11	-0,82
17	C	159	1	-0,52	-4,15	-3,64	1	-0,52	-4,15	-3,64
18	C	257	0	-1,31	-1,31	0,00	0	-1,31	-1,31	0,00
19	C	59	0	-0,26	-0,26	0,00	0	-0,26	-0,26	0,00
20	C	459	4	0,02	-0,97	-0,99	0	-0,97	-0,97	0,00
21	C	79	1	0,19	0,19	0,00	0	0,19	0,19	0,00
22	C	278	1	0,42	-0,08	-0,49	0	-0,08	-0,08	0,00
23	C	595	3	0,07	0,12	0,04	0	0,12	0,12	0,00
24	C	355	1	-0,59	3,52	4,14	1	-0,59	3,52	4,14
25	C	901	6	-1,46	4,14	5,68	1	0,05	4,14	4,09
26	C	400	7	-0,80	18,17	19,12	1	-0,70	18,17	19,00
27	C	391	3	-2,60	-2,63	-0,04	1	-2,63	-2,63	0,00
28	C	766	4	-1,92	-1,24	0,69	0	-1,24	-1,24	0,00
29	C	122	0	-0,13	-0,13	0,00	0	-0,13	-0,13	0,00
30	C	173	6	-2,15	-5,34	-3,26	5	-2,63	-5,34	-2,78
31	C	285	3	-1,15	-6,43	-5,34	3	-1,15	-6,43	-5,34
32	C	529	16	-1,02	14,07	15,24	3	-2,14	14,07	16,56
33	C	534	1	-2,84	-3,52	-0,69	0	-3,52	-3,52	0,00
35	D	232	3	0,53	-7,42	-7,90	2	0,53	-7,42	-7,90
36	E	60	3	-0,02	19,59	19,62	2	-0,02	19,59	19,62
37	E	54	1	0,39	0,39	0,00	1	0,39	0,39	0,00
38	E	264	6	0,44	-2,39	-2,82	0	-2,39	-2,39	0,00
39	E	24	2	-0,27	-0,27	0,00	1	-0,27	-0,27	0,00
41	F	753	94	-0,22	0,79	1,01	7	-1,37	0,79	2,19
42	F	317	10	-1,20	-1,63	-0,44	5	-1,63	-1,63	0,00
43	F	1009	9	-0,15	-0,24	-0,09	0	-0,24	-0,24	0,00
45	G	577	9	-0,29	-1,22	-0,93	4	-0,92	-1,22	-0,30
46	G	3237	124	-1,01	-5,64	-4,67	59	-2,12	-5,64	-3,59
47	G	2093	181	-0,88	-10,71	-9,92	109	-1,93	-10,71	-8,95
49	H	600	28	-0,19	-7,62	-7,44	11	-1,44	-7,62	-6,28
50	H	123	6	0,19	-0,45	-0,64	1	-0,45	-0,45	0,00
51	H	22	2	0,97	-13,68	-14,51	1	0,97	-13,68	-14,51
52	H	543	1	-0,70	-1,93	-1,24	0	-1,93	-1,93	0,00
53	H	77	0	-1,52	-1,52	0,00	0	-1,52	-1,52	0,00
55	I	344	1	-1,08	-1,08	0,00	0	-1,08	-1,08	0,00
56	I	316	6	-0,83	5,47	6,35	4	-1,25	5,47	6,80
58	J	251	2	-1,74	-12,81	-11,27	1	-4,25	-12,81	-8,95
59	J	220	1	-0,43	0,17	0,61	1	-0,43	0,17	0,61
60	J	101	3	0,32	0,32	0,00	2	0,32	0,32	0,00
61	J	113	6	-0,12	-2,70	-2,58	6	-0,12	-2,70	-2,58
62	J	373	2	-0,76	-0,75	0,00	2	-0,76	-0,75	0,00
63	J	390	0	-0,62	-0,62	0,00	0	-0,62	-0,62	0,00
66	K	366	6	-0,52	-1,02	-0,49	2	-1,02	-1,02	0,00
68	L	1442	147	0,64	-5,37	-5,97	77	0,91	-5,37	-6,22
69	M	382	1	0,32	-0,82	-1,14	0	-0,82	-0,82	0,00
70	M	412	8	-2,01	-9,21	-7,35	4	-2,11	-9,21	-7,25
71	M	627	0	0,95	0,95	0,00	0	0,95	0,95	0,00
72	M	102	12	-0,48	-3,93	-3,46	4	-0,63	-3,93	-3,32
73	M	485	2	-0,75	-1,99	-1,24	0	-1,99	-1,99	0,00
74	M	495	4	0,16	17,64	17,45	1	-0,23	17,64	17,91
75	M	185	1	-0,26	-0,26	0,00	0	-0,26	-0,26	0,00
77	N	356	16	-0,65	-2,04	-1,39	6	-2,67	-2,04	0,65
78	N	67	1	-1,05	0,93	2,00	0	0,93	0,93	0,00
79	N	265	8	-2,05	-5,08	-3,09	6	-3,07	-5,08	-2,08
80	N	72	3	-0,27	-2,68	-2,42	3	-0,27	-2,68	-2,42
81	N	231	1	-0,61	-1,80	-1,19	1	-0,61	-1,80	-1,19
82	N	646	2	-1,50	0,17	1,69	2	-1,50	0,17	1,69
85	P	315	7	0,63	1,06	0,43	2	1,06	1,06	0,00
86	Q	465	1	-0,71	-1,23	-0,52	0	-1,23	-1,23	0,00
87	Q	187	7	0,74	-0,68	-1,41	4	0,78	-0,68	-1,45
88	Q	158	0	-0,26	-0,26	0,00	0	-0,26	-0,26	0,00
90	R	205	5	0,66	-1,97	-2,61	0	-1,97	-1,97	0,00
91	R	77	0	-1,84	-1,84	0,00	0	-1,84	-1,84	0,00
92	R	144	4	-0,78	-0,98	-0,21	4	-0,78	-0,98	-0,21

93	R	323	7	-0,58	-0,62	-0,05	2	-0,62	-0,62	0,00
95	S	398	8	-0,79	0,03	0,83	0	0,03	0,03	0,00
96	S	373	4	-0,39	-0,07	0,32	3	-0,39	-0,07	0,32
Total		**29,873**	**869**				**382**			

Figure 1. *Turnover*: Plot of the absolute difference between *Diff(True.Sel)* when □=0,01 and when □=0,02.



IV. Conclusions

58. In the paper, the multivariate *Selective Editing via Mixture models* (SeleMix) approach to the identification of influential errors in continuous data is illustrated. The aim of the paper is to assess the possible benefits in terms if estimates' accuracy and burden/costs reduction deriving from the use as auxiliary information of external sources (both administrative and statistical archives) in the robustly estimated models.

59. The paper contains the description of the application of the approach to the Istat structural sample surveys on ICT and on SMEs. In both applications, the target variables are enterprises' *Turnover* and *Costs*. Auxiliary information is gathered from both fiscal archives (*Financial Statements* and *Sector Studies survey*) and other structural surveys.

60. For both applications, encouraging results for the considered variables are obtained. However, especially for SME, the complexity of surveys in terms of publication domains makes it necessary further analyses. For ICT, the integration of the method in the current E&I procedure is already in progress.

Referencers

[1] Bellisai D., Di Zio M., Guarnera U., Luzi O. (2009), A selective editing approach based on contamination models: a comparative application to an Istat business survey. *UN/ECE Work Session on Statistical Data Editing*, Neuchatel, 5-7 October.

[2] Bethel, J. (1989), Sample allocation in multivariate surveys. *Survey Methodology*, 15 (1989), pp. 47-57.

[3] Buglielli M.T., Di Zio M., Guarnera U., (2010). Use of Contamination Models for Selective Editing, *Q2010, European Conference on Quality in Survey Statistics*, 4-6 May 2010, Helsinki. (http://q2010.stat.fi/sessions/session-19/).

[4] De Giorgi V., Luzi O., Oropallo F., Seri G., Siesto G. (2011). Combining administrative and survey data: potential benefits and impact on editing and imputation for structural business surveys. *UN/ECE Work Session on Statistical Data Editing*, Lubjana, 9-11 May.

[5] Di Zio M., Guarnera U., Luzi O. (2008). Contamination Models for the Detection of Outliers and Influential Errors in Continuous Multivariate Data. *UN/ECE Work Session on Statistical Data Editing*, Vienna (http://www.unece.org/stats/documents/2008.04.sde.htm).

[6] Di Zio M., Guarnera U., (2011). SeleMix: an R Package for Selective Editing via Contamination Models, *Proceedings of Statistics Canada Symposium 2011*. Ottawa, 3-6 November 2011

[7] Hoogland J. (2010), Editing Strategies for VAT Data, paper presented at the *Seminar on Using Administrative Data in the Production of Business Statistics - Member States Experiences*, Rome, 16-18 March.

[8] 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.

[9] 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.

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

[11] Luzi O., Rinaldi M., Seri G., Guarnera U., De Giorgi V. (2011) Estimating structural business statistics based on administrative data: the case of the Italian small and medium enterprises, *Proceedings of the 2011 European Establishment Statistics Workshop (EESW 2011)*, Neuchâtel, 12 – 14 September.

[12] Nurra A., Luzi O., Salamone S., Silvestri F. (2012), Preventing and treating measurement errors and nonresponses in statistical surveys: the case of the Italian survey on ICT and e-commerce, *Conference of the Italian Statistical Society for Enhancing Public Statistics (SIS-VSP)*, Rome, European University, 19-20 April.

[13] Ohlsson, E. (1995). *Coordination of PPS Samples Over Time*, Stockholm University Mathematical Statistics, Stockholm University, S–106 91 Stockholm, Sweden.

[14] Wallgren A., Wallgren B. (2007), *Register-based Statistics: Administrative Data for Statistical Purposes*, John Wiley & Sons.