Selective editing of multisource data based on latent class models

Marco Di Zio, Ugo Guarnera, Roberta Varriale[1]

Abstract

Statistical data editing and imputation are important phases of the statistical production process. Nowadays, the entire production of statistics is more and more based on multisource data and statistical data editing and imputation must be studied in this specific context. Among the issues of using multisource data, an important one concerns the redundancy of information when different sources overlap in terms of units and variables. This frequently happens when surveys and administrative data are integrated. In general, two different approaches can be adopted in this context: 1) survey data are considered as actual (possibly contaminated) data, and administrative sources are used as auxiliary variables to improve the quality of statistics based on survey data, 2) all data are considered as measures of the same target variables with possible different unknown level of reliability. The focus of the paper is on the use of latent class models for selective editing of multisource data in both approaches. In the first approach, the auxiliary variables are used to predict influential errors in the survey data. In the second one, influential errors are predicted by fitting a latent model on survey and administrative data simultaneously analysed. The results of an application to real data are presented.

Key Words: Statistical data editing, data integration, mixture models, influential errors

1. Introduction

In recent years, statistical analysis based on different data sources has been considered as an important alternative to the traditional approaches based on considering only survey data as primary source of information. Massive use of secondary data has become an active area of research in both theoretical and applied statistics, in all phases of the statistical production process, such as editing and imputation (E&I) phase. Among the issues of using multisource data, an important one concerns the redundancy of information when different sources, such as survey and administrative data, overlap in terms of units and variables. In general, two different approaches can be adopted in E&I when surveys and administrative data are integrated: 1) survey data are considered as actual (possibly contaminated) data, and administrative sources are used as auxiliary variables to improve the quality of statistics based on survey data, 2) all data are considered as measures of the same target variable with possible different unknown level of reliability.

The focus of the paper is on the use of latent class model for selective editing in multisource context in both approaches. The model aims at providing "predictions" of the true latent values of the target variable given the available information. Once predictions for the target variable are available, they are compared with the observed values in survey data and the largest discrepancies are selected for interactive editing.

The paper is organized as follows. Section 2 presents two different models to obtain predictions in a selective editing approach. An application to real data on investment coming from Istat Annual Survey on Economic and financial accounts of large enterprises is described in Section 3. Conclusions are reported in Section 4.

[1]Marco Di Zio, Istat, Via Cesare Balbo 16, Roma, email: dizio@istat.it. Ugo Guarnera, Istat, email: guarnera@istat.it. Roberta Varriale, Istat, email: varriale@istat.it.

2. Models

In this section we describe two modeling approaches for selective editing. In both approaches, the goal is to identify the units of the survey sample affected by the most influential errors in order to perform manual review. While in the first approach ($M1$) described in subsection (2.1), administrative data are modeled as merely auxiliary information, in the second approach ($M2$) described in Subsection 2.2, also external sources are treated as primary sources of information and are considered as alternative measures of the target variable. A common feature of the two models is that measurement errors are modeled through an intermittent mechanism, that is, it is assumed that the target variable is correctly measured with strictly positive probability. The last subsection briefly describes the use of predictions from the models $M1$ and $M2$ in the selective editing process. For further details on both model $M1$ and general selective editing procedure, see Di Zio and Guarnera (2013).

2.1 Model $M1$

Let us suppose that a variable Y_i^* is associated to each unit $i \quad (i = 1, \ldots, n)$ of a population of size n. We assume that in absence of measurement errors, data are independent realizations from Y_i^* and that Y_i^* is a Gaussian variable with mean μ_i and variance σ^2 $(i = 1, \ldots, n)$. We also allow for the possibility of a linear dependence of the means μ_i on some set of q covariates $x_i = (x_{i0}, x_{i1}, \ldots, x_{iq})'$ observed without error. True data are modeled via the ordinary linear regression model:

$$Y_i^* = \beta' x_i + U_i, \quad i = 1, \ldots, n \tag{1}$$

where U_i are *iid* Gaussian random variables with zero mean and variance σ^2; β_j $(j = 0, \ldots, q)$ are unknown coefficients to be estimated and, as usually, we set $x_{i0} \equiv 1$. In real applications on economic data, logarithms of data instead of data in their original scale are often assumed to be normally distributed. This does not imply substantial changes in the proposed methodology. We model the "error presence" in data through n independent Bernoulli random variables Z_i $(i = 1, \ldots, n)$ with parameter π, i.e., $Z_i = 1$ if an error occurs on unit i and $Z_i = 0$ otherwise. Furthermore, given that an error is present on the ith unit (i.e., given the event $\{Z_i = 1\}$), its action is described through an additive random noise represented by a Gaussian variable ϵ_i with zero mean and variance σ_ϵ^2. If Y_i denotes the random variable associated with the observed (possibly contaminated) value on the ith unit, and ϵ_i the corresponding error term, we can formally express the error mechanism as:

$$Y_i = Y_i^* + Z_i \epsilon_i, \quad f(\epsilon_i) = N(\epsilon_i; 0, \sigma_\epsilon^2), \quad \sigma_\epsilon^2 = \alpha \sigma^2, \tag{2}$$

where α is a numeric constant greater than 0. Equivalently, we can specify the error model through the conditional distribution:

$$f(y_i | y_i^*) = (1 - \pi)\delta_{y_i^*}(y_i) + \pi N(y_i; y_i^*, \sigma_\epsilon^2). \tag{3}$$

where π (mixing weight) represents the *a priori* probability of contamination and $\delta_t(\cdot)$ is the delta-function with mass at t.

Once the true data distribution and the error mechanism have been specified, the distribution of the observed data can also be easily derived multiplying the true data density by the error density (3), and integrating over Y^*. The resulting distribution is:

$$f(y_i) = (1 - \pi) N(y_i; \mu_i, \Sigma) + \pi N(y_i; \mu_i, (1 + \alpha)\Sigma). \tag{4}$$

Expression (4) represents a mixture of two regression models having the same coefficient matrix B but different (proportional) residual variance-covariance matrices. The last distribution refers to observed data and the parameters can be estimated by maximizing the likelihood based on n sample units via an Expectation Conditional Maximization algorithm (see Di Zio and Guarnera, 2013).

The separate specification of true data model and error model allows, contrarily to the direct specification of the observed data distribution, to derive, for $i = 1, \ldots, n$, the distribution $f(y_i^*|y_i)$ of the true data conditional on the observed data. Note that hereafter the reference to the X covariates is omitted in the notation for the sake of simplicity. A straightforward application of the Bayes formula provides:

$$f(y_i^*|y_i) = \tau_1(y_i)\delta_{y_i}(y^*) + \tau_2(y_i)N(y_i^*; \tilde{\mu}_i, \tilde{\Sigma}) \tag{5}$$

where

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

where $\tau_1(y_i)$, $\tau_2(y_i)$ are the posterior probabilities that a unit with observed values (y_i) belongs to correct and erroneous data group, respectively.

The expected values $\tilde{y}_i = E(y_i^*|y_i)$ from the distribution (5) can be used as predictions $\tilde{y}_i$ of the true values. From equation (5) it follows:

$$\tilde{y}_i = \tau_1(y_i)y_i + \tau_2(y_i)\tilde{\mu}_i, \quad i = 1, \ldots, n. \tag{6}$$

Correspondingly, we can define the expected error as

$$y_i - \tilde{y}_i = \tau_2(y_i)(y_i - \tilde{\mu}_i).$$

The previous methodology can be easily adapted to the lognormal case and can be extended to situations where observed data are incomplete and the nonresponse mechanism is assumed to be MAR (see Di Zio and Guarnera, 2013).

2.2 Model $M2$

Guarnera and Varriale (2016) introduced a latent variable model that could be used in a "multisource approach", that is when the informative context is represented by one target variable measured in G data sources, including both a survey and $G-1$ administrative sources. In other words, both survey and administrative data are considered as measures of the target variable with possible different unknown level of reliability, and influential errors are predicted by fitting a latent model on survey and administrative data simultaneously analysed. Furthermore, some other sources of information could be considered as auxiliary information. To this regard, we adopt a practical criterion to distinguish X as auxiliary variables (*covariates*) from Y as target variables measured with error. We say that a variable Y is a target variable measured with error if its conditional distribution (density) $f(y|y^*)$ given the true unobserved variable Y^*, can be expressed as:

$$f(y|y^*) = (1-\pi)\delta_{y^*}(y) + \pi h(y|y^*), \tag{7}$$

where $0 < \pi < 1$, and $h(y|y^*)$ is an arbitrary density function. As in the case of model $M1$, Equation (7) together with the strict inequality $\pi < 1$ express a measurement model based on the assumption of an intermittent error mechanism, implying that only a proportion of the data are affected by error. Differently from the Y variables, the X variables are such that, intuitively speaking, the events $X = Y^*$ have probability zero. Furthermore, the

probability distribution of the X variables is not of interest, and the X variables can be considered as genuine auxiliary information.

According to the above terminology, the model is specified through the conditional distributions of the Y^* variable given the covariates ("true data model"), and of the contaminated target variable Y given the Y^* variable (one error model for each source). Note that, due to the intermittent nature of the error mechanisms characterizing the measurement processes, there is a strictly positive probability that measures from the different data sources are error-free. Moreover, since the relevant probability distributions are supposed to be continuous, true values are *surely* observed in the special case of coincidence of corresponding observations from different data sources. We assume for the true data the same model as in $M1$ (Equation (1)), and denote with Y_i^g the variable corresponding to the value observed in the source S^g for the unit i ($i = 1, \ldots, n$). In order to complete the modeling, we have to specify the measurement error model for each source, that is, the conditional distribution of Y_i^g given the true value y_i^*. In analogy with the "single source" case, we model the intermittent nature of the error on the different data sources via independent Bernoullian variables Z_i^g with parameters π_g, i.e., $Z_i^g = 1$ if an error occurs for the unit i in the source S^g, and zero otherwise. Also, given the event $Z_i^g = 1$, we assume that $Y_i^g = Y_i^* + \epsilon_i^g$ where ϵ_i^g are mutually independent Gaussian variables with zero mean and variance $\alpha_g \sigma^2$, where α_g is a positive constant ($g = 1, \ldots, G$). In short, the measurement error model can be described through the equation:

$$Y_i^g = Y_i^* + Z_i^g \epsilon_i^g, \quad g = 1, \ldots, G;\ i = 1, \ldots, n. \tag{8}$$

Equations (1) and (8) completely specify the model.

In order to estimate the parameters of the model specified via Equations (1) and (8), we need to derive the observed data distribution. Since we treat the case of partially overlapping sources, where for some units less than G sources may be available, the observed data for each unit i are the measures $y_i^{j_1}, \ldots, y_i^{j_m}$ corresponding to the m available sources $S_i^{j_1}, \ldots, S_i^{j_m}$, where $(j_1, \ldots, j_m) \subset (1, \ldots, G)$. From the above model assumptions it follows that the distribution $f(y_i) = f(y_i^{i_1}, \ldots, y_i^{i_m})$ of the random vector $Y_i^{i_1}, \ldots, Y_i^{i_m}$ associated with the measures from $S_i^{i_1}, \ldots, S_i^{i_m}$ available for the ith unit is a mixture of probability distributions corresponding to the different error patterns across the sources. Formally:

$$f(y_i) = \sum_{k=1} w_k h_k(y_i; \beta, \sigma^2, \alpha),\ \alpha \equiv \alpha_{j_1}, \ldots, \alpha_{j_m};\ \beta \equiv \beta_0, \ldots, \beta_q, \tag{9}$$

where the sum is over the 2^m error patterns across $S_i^{j_1}, \ldots, S_i^{j_m}$, and for the kth pattern, the "mixing weight" w_k is the product of m factors of the form π_g or $1 - \pi_g$ depending on whether the pattern k corresponds to an erroneous or correct value in the source S^g. The densities h_k in (9) are suitable products of Gaussian distributions possibly degenerated in mass points. For instance, for three data-sources S^1, S^2, S^3 the mixture component associated with the pattern where y_i^1 is correct and y_i^2, y_i^3 are erroneous, is a 3-variate Gaussian density with mean vector $(\mu_i, \mu_i, \mu_i)'$ and a covariance matrix $\Sigma_{1.23}$ where all the elements are σ^2 except the Σ_{22} and Σ_{33} that are $(1 + \alpha_2)\sigma^2$ and $(1 + \alpha_3)\sigma^2$ respectively.

The log-likelihood function based on the observed data distribution is obtained by taking logarithm of (9) and summing over the units ($i = 1, \ldots, n$). We implemented an appropriate Expectation Maximization algorithm for the maximum likelihood estimation of the model parameter $\theta \equiv \beta_j, \sigma^2, \pi_g, \alpha_g$ ($j = 0, \ldots, q$; $g = 1, \ldots, G$).

Also in the multisource case, a straightforward application of the Bayes formula allows us to derive predictions using the conditional distribution $f(y_i^* | y_i^1, \ldots, y_i^g)$ of the true data

given the available information, i.e., the observations in the different sources (as for the previous section, the reference to the X covariates is omitted in the notation for the sake of simplicity). Note that this distribution is trivial whenever two (or more) values from different sources are equal. In fact, in this case the "true"value is known without uncertainty. In the other case, the conditional distribution is a mixture of suitable (possibly singular) Gaussian distributions where the mixing weights are the posterior probabilities corresponding to the different error patterns. It is worthwhile noting that a similar approach has been adopted by Sander *et al.*, (2015), where the error mechanism is assumed not intermittent.

2.3 Selective Editing

Predictions from models $M1$ and $M2$ are used to obtain the score functions to identify influential errors. Specifically, if the quantity of interest is the total t_y^* of the target variable Y, the score function is defined in terms of the ratio between the expected error and the reference estimate t_y^{ref} of t_y^*, that is:

$$s_i = \frac{|y_i - \hat{y}_i|}{t_y^{ref}}, \tag{10}$$

where the predictions $\hat{y}_i$ are obtained plugging-in the parameter estimates in the conditional expectation $E(y_i^*|y_i)$ in $M1$, or $E(y_i^*|y_i^1, \ldots, y_i^g)$ in $M2$. This definition of the score function allows to link the threshold for interactive editing to the residual expected error in the data. Details can be found in Di Zio and Guarnera (2013).

3. Application

3.1 Data

Selective editing based on predictions from model $M1$ is currently used in Istat to edit microdata on *gross investment* (see Di Zio *et al.*, 2015) from the 2012 SCI survey, that is the Istat Annual Survey on Economic and financial accounts of large enterprises. The survey is actually a census, covering all enterprises operating in Italy with at least 100 employees, and concerns all enterprises of industrial and services sectors excluding financial services. Statistical units units are enterprises drawn from the Italian Statistical Business Register, ASIA.

Survey SCI is carried out according to the normative guidelines of the European Community Structural Business Statistics (SBS), and collects data concerning profit-and-loss accounts and balance sheets, employment, investment and personnel costs. The periodicity of data collection and of the estimates is yearly. The analysis presented in this paper considers only responding units (5770 observations).

From administrative data available to Istat, three variables can be used as proxy of enterprise gross investments. Two of them are directly obtained from data sources:

- the information on expenditure for amortizable goods reported in *Value Added Tax declarations* (VAT variable hereafter),
- the derived variable that can be calculated from *financial statements* data source, based on the assets at the end of the year minus assets at the beginning of the year, plus depreciation and revaluation (Δ_S variable hereafter).

The third administrative variable is indirectly obtained by exploiting the information on investments in the *Explanatory Notes to the Financial Statements*, that are notes comprising

a summary of significant accounting policies, details of the reported values and explanations concerning the economic situation of the company. Istat may access the explanatory notes of corporations and limited companies in the form of both non-standardized text files (one for each company) and an experimental dataset reporting the investment value obtained using a software for automatic optical recognition from the non-standardized text files. It is worthwhile noting that the variable on total investment reported in the dataset is exactly the target variable of the selective editing procedure. However, it cannot be used to produce SBS data or to automatically correct the data because of the errors due to the automatic optical recognition, and because not rarely the automatic procedure is not able to classify and recognise data. For this reason, we use the value of total investment from the experimental database on the explanatory note (X_N variable hereafter) as a third variable in the selective editing procedure.

Beside the described role of the exploratory notes, in this work they have also been used to assess the validity of the editing procedure. Indeed, exploratory notes contain important information and explanation of the economic behavior of the enterprise, and they have been used by subject matter experts to estimate the real value of the target variable.

A selective editing procedure has been applied to microdata on gross investment in order to identify influential errors. In particular, the models described in Section 2 are applied to data in two settings:

1. the variable concerning investments directly observed in the survey (I) is used as Y in model (1) and the three variables VAT, Δ_S and X_N are used as covariates;

2. the investment observed in the survey I and the administrative variables Δ_S, X_N are considered as different measurements of the same latent variable, and VAT is assumed to be a covariate.

In the first and second setting, model $M1$ and $M2$ is applied to obtain predictions for the selective editing procedure, respectively. In both settings, an always observed stratification variable is also used, so that the selective editing procedure is applied separately within each stratum. The stratification variable is the *enterprise size* in terms of number of employees. More precisely, three size classes are used corresponding to the number of employees belonging to the intervals $(100-249, 250-499, +500)$, respectively.

The predictions obtained through $M1$ and $M2$ have been used to estimate influential errors on I and select the units to be clerically reviewed by subject matter experts with the help of exploratory notes. The estimation domains on which the impact of errors has been evaluated are 64 Industries, corresponding to the classification of economic activity A*64 that is used to disseminate National Accounts data (see Eurostat, 2013). We remind that the selective editing procedure allows to estimate the residual error in nit selected data. The threshold of the residual expected error in the data currently adopted in the official selective editing procedure is 4%; some additional analysis have been performed with different values of the threshold (1%, 2%, 6%).

3.2 Results

In this section we provide some results about the comparison of the selective editing approaches using $M1$ and $M2$. A summary of the results is reported in Table 1. The column *"Nobs"* reports the number of units observed in the sample, the columns *"Sel1"* and *'Sel2'* report the number of units identified by $M1$ and $M2$ as influential errors, respectively, with a threshold equal to 4%. The overall percentage of selected units in $M1$ is 1.8%, while in $M2$ is 2.4%.

Table 1: Number of selected with $M1$ and $M2$ by Industries

A*64	Industries	Nobs.	Sel1	Sel2
4	Mining and quarrying	20	0	0
5	Manufacture of food products	220	2	1
6	Manufacture of textiles	227	1	7
7	Manufacture of wood	36	0	1
8	Manufacture of paper and paper products	64	2	9
9	Printing and reproduction of recorded media	26	0	0
10	Manufacture of coke and refined petroleum products	17	1	0
11	Manufacture of chemicals and chemical products	125	3	1
12	Manufacture of basic pharmaceutical products	97	1	1
13	Manufacture of rubber and plastic products	154	0	6
14	Manufacture of other nonmetallic mineral products	121	0	0
15	Manufacture of basic metals	122	1	0
16	Manufacture of fabricated metal products	246	4	1
17	Manufacture of computer, electronic and optical products	90	3	7
18	Manufacture of electrical equipment	125	2	11
19	Manufacture of machinery and equipment n.e.c.	445	1	14
20	Manufacture of motor vehicles, trailers and semitrailers	128	3	1
21	Manufacture of other transport equipment	53	2	1
22	Manufacture of furniture; other manufacturing	117	0	0
23	Repair and installation of machinery and equipment	36	1	2
24	Electricity, gas, steam and air conditioning supply	68	2	2
25	Water collection, treatment and supply	40	5	2
26	Sewerage; waste collection	113	1	5
27	Construction	161	8	2
28	Wholesale and retail trade	64	2	2
29	Wholesale trade, except of motor vehicles and motorcycles	326	0	2
30	Retail trade, except of motor vehicles and motorcycles	324	5	7
31	Land transport and transport via pipelines	178	3	1
32	Water transport	30	3	2
33	Air transport	8	2	2
34	Warehousing and support activities for transportation	262	3	1
35	Postal and courier activities	8	0	0
36	Accommodation and food service activities	151	7	10
37	Publishing activities	30	2	1
38	Motion picture	16	0	0
39	Telecommunications	17	0	0
40	Computer programming	175	1	1
44	Real estate activities	9	1	1
45	Legal and accounting activities	88	4	4
46	Architectural and engineering activities	48	3	4
47	Scientific research and development	11	0	1
48	Advertising and market research	20	2	1
49	Other professional, scientific and technical activities	12	2	2
50	Rental and leasing activities	14	3	2
51	Employment activities	55	1	1
52	Travel agency, tour operator	15	0	0
53	Security and investigation activities	504	4	5
55	Education	17	1	2
56	Human health activities	181	1	3
57	Social work activities	275	6	4
58	Creative, arts and entertainment activities	18	2	2
59	Sports activities and amusement and recreation activities	15	5	5
61	Repair of computers and personal and household goods	2	0	0
62	Other personal service activities	46	0	0
Total		5770	106	140

The overall percentage for different values of threshold γ is reported in Table 2. We notice that as the level of accuracy increases, the ratio ($Sel2/Sel1$) of the number of units selected by using $M2$ and $M1$ increases.

The ranks representing the order of the units based on the score of the potential influential errors via the two procedures are compared by means of the Spearman's rank correlation coefficient. The concordance is high given that the Spearman's ρ is equal to 0.91. Nevertheless, there are some differences (see Table 3).

The selective editing models studied in this paper are based essentially on the prediction of the expected true values, hence it is worthwhile analysing the differences of the predicted values. The correlation between the predicted values is high and equal to 0.99. Table 4 shows some statistics computed over the prediction differences. In Figure 1, the scatter plot of the logarithm of predictions highlights that they are very close to each other.

An ideal assessment of a selective editing procedure would consist in recovering the true values for all units. In fact, if "true values" were available on all the observations, we could find the error left in each not selected unit (residual error) and thus the total estimation error resulting from using data before and after the selective editing procedure.

Table 2: Number of selected with $M1$ and $M2$ by threshold

	γ			
	1%	2%	4%	6%
$Sel1$	376	227	106	75
$Sel2$	700	347	140	78

Table 3: Statistics on rank's differences

Min	1st Quartile	Median	Mean	3rd Quartile	Max
-439	-13	2	0	18	216

Table 4: Statistics on prediction's differences

Min	1st Quartile	Median	Mean	3rd Quartile	Max
-274200.0	-66.2	0.0	74.1	0.0	368200.0

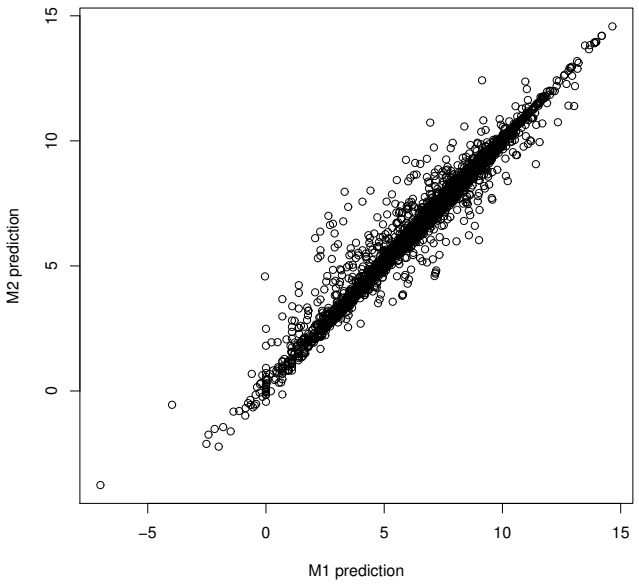

Figure 1: Scatter plot of the logarithm of predictions by $M1$ and $M2$

In this application, we use the textual information on the exploratory notes in order to recover the true values.

In order to evaluate the methodologies, we compute the residual error according to different thresholds for six estimation domains, namely the Industries: 10 (Manufacture of coke and refined petroleum products), 23 (Repair and installation of machinery and equipment), 32 (Water transport), 38 (Motion picture), 39 (Telecommunications), 52 (Employment activities).

Table 5 shows the impact of errors in raw data and the impact of residual errors after

Table 5: Relative percentage difference between estimated total of investments computed on raw or edited data and true data

		Industries					
	γ	10	23	32	38	39	52
Raw data		2.08	1.26	-35.29	-0.12	26.10	-1.06
$M1$	6%	2.08	1.66	-28.93	-0.12	26.10	-1.06
$M2$	6%	2.08	1.26	-28.93	-0.12	26.10	-1.06
$M1$	4%	2.08	1.66	-2.03	-0.12	26.10	-1.06
$M2$	4%	2.08	2.41	-20.63	-0.12	26.10	-1.06
$M1$	2%	-0.91	1.66	-0.01	-0.12	26.10	1.06
$M2$	2%	2.10	2.26	-20.63	-0.12	0.50	1.79
$M1$	1%	-0.91	0.73	-0.01	-0.12	26.10	1.06
$M2$	1%	-0.98	2.26	-1.50	-0.12	0.50	1.79

Table 6: Number of observations selceted by $M1$ and $M2$

		Industries					
	γ	10	23	32	38	39	52
$Sel1$	6%	1	1	1	0	0	0
$Sel2$	6%	0	0	1	0	0	0
$Sel1$	4%	1	1	3	0	0	0
$Sel2$	4%	0	2	2	0	0	0
$Sel1$	2%	2	2	6	1	0	1
$Sel2$	2%	1	5	2	0	1	1
$Sel1$	1%	2	3	6	2	0	1
$Sel2$	1%	3	7	5	0	1	1

selective editing. The (percentage) relative impact of errors is computed as $RE = (t - t^*)/t^* \times 100$, where t is the estimation of the total investments based on raw data or data after selective editing, and t^* is the estimate obtained by editing all data. We remark that, in the computation of t and t^*, missing values are imputed according to the official procedure.

The choice of a selective editing procedure should take into account both the indicators expressing the residual errors and the number of selected units (see Table 6). The results in Tables 5 and 6 do not provide strong evidence in favor of one of the methods. In fact, for γ at 6% and 2% the behavior of $M2$ and $M1$ is similar (also due to compensation), while $M1$ is preferable with a $\gamma = 4\%$ and $M2$ is better for a threshold at 1%. In terms of the number of selected units, on the contrary to the overall behavior (see Table 2), $M1$ and $M2$ identify a similar number of observations to be reviewed.

3.3 Parameters estimates and their interpretation

Tables 7 and 8 show the parameter estimates obtained through $M1$ and $M2$. We restrict to the case when information on the variable VAT is missing.

In Table 7, the estimates of parameters of $M1$ are reported. They can be intuitively interpreted as the coefficients of a robust linear regression model (β), the probability of error (π) and the variance inflation factor (α). Table 8 shows the estimates of parameters $M2$, π_g and α_g, that represent the error probability and the impact of error (variance inflation)

of the source S^g, respectively, and that can be used as quality indicators. Notice that in this application g varies in the set (I, X_N, Δ_S).

In both approaches, estimates are conditionally on the stratification variable $N.Empl$. We can notice that stratum $+500$ shows a different pattern of the parameters from those obtained in the other strata: from Table 7, it results that X_N is the reference source for strata $(100 - 249)$ and $(250 - 499)$, while in stratum $+500$ the source with the highest coefficient is Δ_S. This behavior is confirmed also by other results not reported in this paper.

In Table 8, X_N performs quite well in terms of error probability, while it shows a very high α value, corresponding to a big error size. Roughly speaking, the *Exploratory notes* are affected by a lower rate of errors, but the errors present in the source are quite big. On the contrary, for the source Δ_S the error probability is higher but the magnitude of error is low. It is worthwhile noting that from the comparison of Tables 7 and 8, it results that the error probabilities for the source I (survey data) are very similar, that is the two models provide similar assessment of the quality of the source I.

Table 7: Estimates of parameters with $M1$ approach

$N.Empl.$	β_0	β_{X_N}	β_{Δ_S}	π	α
+500	-0.09	0.13	0.88	0.39	37
250–499	-0.16	0.85	0.15	0.40	48
100–249	0.17	0.74	0.22	0.38	46

Table 8: Estimates of parameters with $M2$ approach

$N.Empl.$	π_I	π_{X_N}	π_{Δ_S}	α_I	α_{X_N}	α_{Δ_S}
+500	0.42	0.21	0.48	0.25	0.61	0.13
250–499	0.38	0.23	0.42	0.45	0.25	0.12
100–249	0.40	0.17	0.39	0.55	0.46	0.12

4. Conclusions

The results show that, in this application, there is no evidence in favor of the two methods $M1$ and $M2$. In fact, for the analysed *Industries*, the results in terms of both residual error and number of selected units are comparable for all the considered levels of the selective editing threshold.

The general idea that should guide the researcher in choosing one of the two approaches is that the administrative data sources are to be considered as covariates or as measurements affected by errors: i.e., the true value of the target variable is observed at least in a subset of units, the variable should be considered as a measurement of the target latent variable; on the contrary, if there exists a statistical relation between the administrative variable and the target variable, but the probability of observing the true value in administrative data is zero, then administrative variables should be considered as covariates. The fact that in some circumstances, in a subset of data, measures of the variables coincide in different data sources could lead to adopt the multisource approach ($M2$). On the other hand, this behavior should be carefully considered, since it may be due to dependence of errors among

the sources. For instance, the respondent could provide in the survey the same erroneous value already declared in one administrative source.

REFERENCES

Di Zio, M., and Guarnera, U. (2013), "A Contamination Model for Selective Editing," *Journal of Official Statistics*, 29 (4), 539–555.

Di Zio, M., Guarnera, U., Iommi, M., and Regano, A. (2015), "Selective editing of business investments by using administrative data as auxiliary information," *UNECE work session on Stat. Data Editing* Budapest, Hungary, 14–16 September 2015.

Eurostat (2013). European system of accounts (ESA 2010).

Guarnera, U., and Varriale, R. (in press), "Estimation from Contaminated Multi-Source Data Based on Latent Class Models," *Statistical Journal of the IAOS*.

Sander, S., Bakker Bart,F.M., Van Delden A., (2015) "Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables", Technical report, Report number: 2015-17, Centraal Bureau voor de Statistiek