Working Paper.
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (i): Selective editing / macro editing

Use of administrative data for selective editing: the case of business investments

Prepared by Di Zio M., Forestieri P, Guarnera U., Iommi M., Regano A.
Istat, Italy

I. Introduction

1. Microdata available to National Accounts (NA) for the estimates of enterprise gross investment in tangible goods (hereafter *investment*) come from Structural Business Surveys (SBS). Since NA has to produce estimates at a higher level of detail in terms of domains and estimation of variables, errors still present in data can seriously affect the accuracy estimates. Furthermore, in the estimation phase, NA has additional information on this phenomenon from other sources so that further verification of the data is possible.

2. Recently administrative data are entering the process of SBS in Istat, but unfortunately investment values are not observed in the available administrative sources, hence they are gathered only through the SBS survey. It is however worthwhile to remark that, in one of the most important administrative source that is the Financial Statements (FS) of enterprises, there is a very rich source of information; in fact firms report explanatory notes comprising a summary of significant accounting policies and details of the reported values in FS and explanations concerning the economic situation of the company. In these notes, data on investments are reported. Istat has access to the explanatory notes of corporations and limited companies only in the form of non-standardized text files (one for each company) and not in the form of statistical database that could be used to produce SBS data on investment or to automatically correct data.

3. Recovering investment data from the explanatory notes is a time consuming process. For this reason, an editing procedure based on the explanatory notes is grafted naturally in the selective editing framework, whose goal is just to minimize the number of checks by focusing on the units where editing has the highest expected benefit. Typically, selective editing is useful if at the review stage it is possible to recover with a high reliability the 'true' value, for example, when it is possible to quickly get back to the respondent units. In the case of investments, the revision of the values of the units is done by consulting the Explanatory Notes. The probability of recovering the true value from this note is high, in fact the presence of a textual explanation of the economic situation of the business is important to validate data that are reported in the note itself, and is indeed avoided the problem of memory effect that for instance may affect the quality of the data obtained through a follow-up, since the explanatory notes are filled out at the same time of FS.

4. In this paper, we describe the selective editing procedure applied to the SBS data for the year 2010 focusing on the *investment* variable. An assessment of the procedure is performed by analyzing the impact of remaining errors in data not selected by the procedure.

5. The paper is structured as follows. In Section II the contamination model implemented in the R package SeleMix and used for selective editing is introduced. The application and some results are described in Section III. The description of the assessment of the method is given in Section IV. Conclusions and future plans are in Section V.

II. Selective editing by using SeleMix

A. Selective editing through SeleMix

6. The selective editing method used for those data is based on explicitly modelling both true (error-free) data and error mechanism. Details on model specification and parameter estimation can be found in Di Zio and Guarnera (2013). The model assumptions can be summarized as follows. True data (possibly in log-scale) are thought of as n realizations from a random p-vector $\mathbf{Y}$ that, conditional on a set of q covariates $\mathbf{X}$, is normally distributed with mean vector $\mathbf{BX}$ and covariance matrix $\mathbf{\Sigma}$. The intermittent nature of the error, which is crucial to the present approach, is modelled through a Bernoullian r.v. $\mathbf{I}$, with parameter w, assuming value 1 or 0 depending on whether an error occurs in data or not respectively. The parameter w can be interpreted as the marginal probability of an observation of being affected by an error in at least one variable $\mathbf{Y}$. Conditional on $\mathbf{I}$=1 (presence of error), we assume a Gaussian additive error with zero mean and covariance matrix proportional to $\mathbf{\Sigma}$, the proportionality constant being some positive number λ. Thus the model parameters are $\theta = (\mathbf{B}, \mathbf{\Sigma}, w, \lambda)$.

7. The previous assumptions allow us to explicitly derive, via Bayes formula, the distribution of the true data conditional on the observed data. It is a mixture of a mass density corresponding to absence of error and a Gaussian distribution corresponding to presence of error. This mixture is the central object for the proposed selective editing method and is completely identified by the set of parameters θ. In order to estimate parameters θ we note that they also identify the (unconditional) distribution of the observed data which is another mixture whose components are non-degenerate Gaussians. The model parameters can be estimated by maximizing the likelihood function based on the observed data using an EM-type algorithm

8. Once the model parameters have been estimated, they can be plugged into the functional form of the conditional distribution of true data given observed data. The selective editing strategy consists in using this estimated distribution to build up a score function. Specifically, for each unit we compute an "anticipated" value as expected "true value" conditional on the observed value. The anticipated value is obtained by means of a weighted average of the observed value and a synthetic value. The weights are given by the probability of being in error. The synthetic value is in turn the weighted average of the observed value and a robust estimate of the regressed value. The weights are the inverse of the estimated covariance matrices of the true and erroneous data respectively. Hence, a score function can be defined in terms of difference between observed and anticipated value (expected error), and the units to be interactively reviewed can be selected as those having higher score function. Once all the observations have been ordered according to this score function, we are able to estimate the residual error remaining in data after the correction of the first k units (k=1,..,n). The number of most critical units to be edited can be chosen so that the estimate of the residual error is below a prefixed threshold (see Section III). This feature is an important point in the proposed method. In fact, differently from most selective editing procedures, our approach allows to explicitly relate the efforts in editing activities (number of units to be manually checked) to the accuracy of the target estimates.

9. The method is implemented in the R package SeleMix (Selective editing via Mixture models) available on the website http://www.R-project.org.

III. Description of the application of the selective editing to Istat data

A. Data

10. In this section we describe how SeleMix is applied to microdata on gross investment in tangible goods that NA uses (together with many other data sources) to produce estimates of gross fixed capital formation by industry. Data come from 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 persons employed and concerns all enterprises of industrial and services sectors excluding financial services. Reporting and analysis units are Enterprises (drawn from the Italian Statistical Business Register, ASIA). The survey is carried out according to the normative guidelines of the EC Structural Business Statistics. The periodicity of data collection and of the estimates is yearly. The survey collects data concerning profit-and-loss accounts and balance sheets, employment, investment and personnel costs. The analysis considered only responding units (5120 observations).

11. Data on investment for the production of SBS are collected via a survey because no administrative data is available. In fact, Istat currently has access to a rich set of administrative data regarding variables reported in profit and loss accounts and (only for corporations and limited companies) in balance sheets. However, this set of information does not provide data on investment or on the whole set of the accounting variables from which investment could be calculated (tangible assets at the beginning and at the end of the year, revaluation of the assets, depreciation, disposal, sells of investment goods and others).
Companies report data on investment in the Explanatory Notes to the Financial Statements (notes comprising a summary of significant accounting policies and details of the reported values and explanations concerning the economic situation of the company). In other words, explanatory notes report the "true" value of investment. However, Istat has access to the explanatory notes of corporations and limited companies only in the form of non-standardized text files (one for each company) and not in the form of statistical database that could be used to produce SBS data on investment or to automatically correct data. On the other hand, integrative notes can be very useful to check data that may contain errors potentially influent on the target estimates.

B. The selective editing procedure

12. The model in SeleMix and described in Section II is estimated separately for each economic activity at the level of NACE Rev2 sections (see Eurostat, 2008) while 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). Taking into account the reasonable number of units that can be checked with the available resources, the threshold used for the selection of critical units is chosen such that the estimated relative residual error for each estimation domain is 5%. The covariate used in the model is the *investment* of the same firm in the previous year, or in the case it is not available the variable *depreciation for the current year*. These variables are considered not affected by errors since they are treated and officially released.

13. It is worthwhile to remark that data are characterized by a high frequency of zeroes (zero-inflated), and in order to manage this characteristic, that is not explicitly modeled in SeleMix, the model is estimated on the subset of data with positive values of the analyzed variables.. However, predictions are computed also for units whose observed value is zero.
Finally, we notice that, although data should cover the whole population of businesses with more than 100 employees (it is actually a census), there are missing data; consequently a weight for taking into account the non-response is computed and used in the score function. A summary of the results of selective procedure is given by Table 1. In this table, the column '*n. obs*' reports the number of units observed in the sample, the column '*Selected*' shows the number of units selected by SeleMix, the column "*No-Edited*" lists the number of selected units that are not edited because free of errors on the basis of the information available on the explanatory notes, and on the contrary the column "edited" reports the number of units with errors on the variable *investment*. The "*SBS original Values*" and "*Post-editing value*" show the value of the total of investments (in thousands of euro) computed on data before

and after the selective editing procedure. Finally, to make it easier the comparison, in the last column the relative percent difference between these estimates is reported. This quantity can be viewed as a measure of the impact of the selective editing process.

Table 1. Number of selected and/or edited observation, and differences between estimates computed on raw and edited data

	Industries	n. obs	Selected	No-Edited	Edited	SBS Original Value* (A)	Post-editing Value* (B)	(B-A)/A %
4	Mining and quarrying	20	-	-	-	1.151.191	1.151.191	0%
5	Manufacture of food products, beverages and tobacco products	202	7	4	3	1.318.344	1.190.884	-10%
6	Manufacture of textiles, wearing apparel and leather products	207	-	-	-	337.909	337.909	0%
7	Manufacture of wood and of products of wood and cork	35	-	-	-	67.556	67.556	0%
8	Manufacture of paper and paper products	56	-	-	-	244.175	244.175	0%
9	Printing and reproduction of recorded media	30	1	0	1	137.446	132.293	-4%
10	Manufacture of coke and refined petroleum products	17	2	0	2	552.322	595.961	8%
11	Manufacture of chemicals and chemical products	115	-	-	-	852.556	852.556	0%
12	Manufacture of basic pharmaceutical products	80	-	-	-	567.440	567.440	0%
13	Manufacture of rubber and plastic products	117	1	0	1	415.946	405.519	-3%
14	Manufacture of other nonmetallic mineral products	114	3	2	1	758.972	749.010	-1%
15	Manufacture of basic metals	116	2	0	2	1.382.920	1.194.160	-14%
16	Manufacture of fabricated metal products, except machinery and equipment	199	2	1	1	468.006	468.625	0%
17	Manufacture of computer, electronic and optical products	89	4	2	2	175.821	315.799	80%
18	Manufacture of electrical equipment	124	1	0	1	435.102	397.545	-9%
19	Manufacture of machinery and equipment n.e.c.	370	-	-	-	792.479	792.479	0%
20	Manufacture of motor vehicles, trailers and semitrailers	116	-	-	-	1.066.322	1.066.322	0%
21	Manufacture of other transport equipment	52	7	3	4	348.113	349.585	0%
22	Manufacture of furniture; other manufacturing	110	2	1	1	175.608	155.702	-11%
23	Repair and installation of machinery and equipment	37	2	1	1	18.319	32.183	76%
24	Electricity, gas, steam and air conditioning supply	58	-	-	-	4.244.515	4.244.515	0%
25	Water collection, treatment and supply	52	3	1	2	746.065	596.687	-20%
26	Sewerage;	111	1	1	0	529.045	529.045	0%
27	Construction	202	5	4	1	928.630	913.935	-2%
28	Wholesale and retail trade and repair of motor vehicles and motorcycles	67	9	3	6	173.824	226.139	30%
29	Wholesale trade, except of motor vehicles and motorcycles	305	1	0	1	866.673	835.714	-4%
30	Retail trade, except of motor vehicles and motorcycles	304	1	0	1	4.893.648	1.888.379	-61%
31	Land transport and transport via pipelines	185	-	-	-	2.539.556	2.539.556	0%
32	Water transport	28	3	0	3	2.554.618	2.704.878	6%
33	Air transport	10	1	1	0	140.989	140.989	0%
34	Warehousing and support activities for transportation	241	1	1	0	4.286.565	4.286.565	0%
35	Postal and courier activities	8	-	-	-	242.070	242.070	0%
36	Accommodation and food service activities	143	3	1	2	246.890	241.850	-2%
37	Publishing activities	35	1	1	0	48.369	48.369	0%
38	Motion picture, video and tv programme production	24	2	2	0	387.524	387.524	0%
39	Telecommunications	18	-	-	-	3.462.770	3.462.770	0%
40	service activities	146	4	1	3	376.928	368.973	-2%
44	Real estate activities	11	2	2	0	20.230	22.308	10%
45	consultancy activities	101	9	4	5	90.545	103.515	14%
46	Architectural and engineering activities; technical testing and analysis	51	1	0	1	98.989	95.247	-4%
47	Scientific research and development	13	2	2	0	39.323	39.323	0%
48	Advertising and market research	19	3	3	0	15.001	15.001	0%
49	Other professional, scientific and technical activities; veterinary activities	9	5	3	2	1.485	3.669	147%
50	Rental and leasing activities	9	1	0	1	2.179.073	2.153.034	-1%
51	Employment activities	42	2	0	2	28.490	5.727	-80%
52	Travel agency, tour operator reservation service and related activities	9	-	-	-	8.342	8.342	0%
53	activities; office administrative	403	39	28	11	120.034	128.541	7%
55	Education	9	3	3	0	2.736	2.736	0%
56	Human health activities	144	2	1	1	332.562	306.211	-8%
57	Social work activities	258	2	1	1	94.471	85.331	-10%
58	Creative, arts and entertainment activities	16	1	1	0	66.850	66.850	0%
59	Sports activities and amusement and recreation activities	12	3	1	2	33.891	30.675	-9%
61	Repair of computers and personal and household goods	3	1	1	0	668	668	0%
62	Other personal service activities	28	1	1	0	137.882	137.882	0%
	Total	**5.280**	**146**	**81**	**65**	**41.205.798**	**37.929.914**	-8%

14. The percentage of selected units is 2.7%, and the hit-rate, i.e., the percentage of erroneous observation in the selected units, is 44.5%. These two pieces of information should be seen together, in fact the hit-rate is not high, but it is important to note that also the number of selected units is low. These two aspects jointly analysed give an idea of the efficiency of the procedure. However, the assessment of the quality of the procedure cannot be based only on them. In fact, the aim of selective editing is to

remove the influential errors, thus it is necessary to quantify the impact of the residual errors on the estimates.

IV. Validation of the procedure

15. In this application, we may compute the impact of residual errors on estimates by analysing the explanatory notes. The ideal validation of the procedure would consist in revising all the not selected observations of the sample by looking at their reported values in the notes. By means of this comparison, we could find the error left in each not selected unit (residual error) and compute the impact of residual errors on the released estimate by comparing the estimate of investment based on data after the selective editing procedure with the estimate obtained after having revised all the observations. The latter is taken as reference value and represents the estimate obtained in case data are not anymore affected by errors.

16. The manual revision of all the 5134 not selected observations is not feasible in practice unless a big investment in resources is planned. Instead, we decided to revise all the not selected observations in some selected industries. The analysis of the results in each industry is quite informative, because the threshold used (set equal to 5%) is referred to each industry. At this stage only three industries have been selected (industries 12, 17, 32).

17. Table 2 shows the impact of errors in raw data after editing the observations selected by SeleMix. 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 current data, i.e., raw data and data edited after the selective procedure, while t^* is the estimate obtained by editing all data observed in the anaysed *industry*.

Table 2. Relative percentage difference between estimated total of investments computed on raw and edited data compared with estimate computed on data that are all edited

	Industries		
	12	**17**	**32**
%RE on raw data	-12.6	43.6	15.5
%RE after selective editing	-12.6	7.2	-5.5

18. Before interpreting the results, we need to take into account that due to zero-inflation, data are considerably far from the model assumptions, so that the estimation of the expected impact of residual errors should be taken as an approximation of the truth.

19 In this context, results for the industries 17 and 32 can be considered satisfactory, since they are not so far from the 5% estimated with SeleMix. It is worthwhile to look at the strong improvement in industry 17 where the relative difference between the estimate computed on raw data and the values without errors is 43.6%. The results are obtained by selecting few units: in industry 17 only 4 units are selected out of the 89 observed in the sample, while in industry 32, 3 units out of 38.
On the contrary, the result for industry 12 cannot be considered satisfactory. The selective procedure does not select any observation in this industry, but the error is 12.6%. It is indeed interesting to see what happened.

20. Firstly we notice that the high difference is only due to one company. The cause is that expenditures for tangible fixed assets incurred in a given year are the sum of two components: expenditures for tangible assets paid and delivered during the accounting year (classified in categories as land, buildings, plants and machinery, etc.) and payments to suppliers for goods not yet delivered at year end (classified in the category tangible assets in progress and advances). When the good is delivered, the company registers in its balance sheet a reclassification from assets in progress and advances to the corresponding assets categories. Actually only the first two items are to be registered as investment of the year, while reclassifications were already included in investment of the previous years (as assets in progress and advances). One company instead registered as investment of the year also the

reclassifications made during the year. The company followed the same criteria (then made the same error) when compiling the SBS questionnaire both in 2009 and 2010 and the error in 2009 were not detected.

21. From the point of view of the statistical model, the interpretation of the failure is the following. The values of investments for this business in the two years are coherent each other, hence the model cannot see any discrepancy that in fact does not exist. In addition, since 2009 data were treated and officially released, they were used as error-free covariates in the model. This makes the selection of this error almost impossible. However, the problem is not in the model, but in data: The current value of investment contains an error but it is not atypical conditional on the historical value. Actually, it is difficult to imagine a selection procedure based on past officially released data that selects as affected by an important error an observation that is consistent with the past. We remark that SeleMix may deal with the case where there is not any auxiliary variable. In this application, we could have considered also the investments of the previous year a variable with errors, but the observation would have been selected only if the values of the investments were jointly considered outliers, i.e., both the values observed in the two years are far from the rest of the observations.

V. Final conclusions

22. The selective editing procedure is considered by the survey managers feasible in terms of costs. The analysis performed to validate the method is encouraging and should be extended to other estimation domains. A result of the validation analysis is that it is difficult to select an observation with an erroneous investment which it is not atypical with respect to the historical value. This is risky when historical data are still contaminated by errors. Since we have noticed that both the current and the historical variables are affected by the same error mechanism, that is in fact a deterministic misclassification, the usage of other variables as covariate may alleviate this annoying situation. To this aim, analysis about the explicative power of other variables such as expenditures for amortizable goods reported in Value Added Tax declarations, and a derived variable based on the assets at the end of the year minus assets at the beginning plus depreciation and revaluation will be performed.

23. In order to have a complete view of the quality of the selective editing procedure and to improve the accuracy of the procedure itself, the ideal situation would be that in which all the units are revised. In this framework the problems of errors in previously released data would be avoided and future application of this procedure would not be affected by those cases. In practice, such a validation procedure is rarely used because of its high cost. Hence, future works will be devoted to refine the validation analysis. A first step will be experimenting with an alternative validation procedure, based on the use of samples of not selected observations. With this evaluation scheme, we expect to have a complete picture concerning the quality of the selective editing procedure, but not to remove the errors from the historical values. A strategy for dealing with this last issue could be that of revising all the observations belonging to the domains where the residual error estimated with the previous scheme is considerably far from the prefixed threshold. In this way, we may also understand whether there are other important error mechanisms affecting the quality of the procedure.
A possible strategy for drawing samples is to resort to a PPS sampling proportional to the scores estimated with SeleMix. This makes the estimation procedure more efficient as suggested by Ilves and Laitila (2009) in a general context, and studied by Di Zio and Guarnera (2013a) in the specific case when SeleMix is used.

References

Di Zio, M., and Guarnera, U. (2013). A Contamination Model for Selective editing. *Journal of Official Statistics*, Vol. 29, No. 4, pp. 539-555.

Di Zio, M., and Guarnera, U. (2013a). A two-step selective editing procedure based on contamination models. *Rivista di Statistica Ufficiale*, No. 2-3, 2013.

Ilves, M., and Laitila, T. (2009). Probability-Sampling approach to Editing. Austrian Journal of Statistics, Vol. 38, No. 3, pp. 171-182

Eurostat (2008). Nace Rev. 2 Statistical classification of economic activities in the European Community

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