UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (i): Selective and macro editing

Selective editing of business investments by using administrative data as auxiliary information

Prepared by Di Zio M., 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 in data can seriously affect the accuracy of estimates. However, in the estimation phase NA has additional information on this phenomenon from other sources so that further verification of the data is possible.

2. Investment data for the production of SBS for Italy are collected via a survey because no administrative data is available. In fact, firms 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.

3. Istat may access the explanatory notes of corporations and limited companies in the form of non-standardized text files (one for each company) and in the form of an experimental dataset reporting the value of investment that is obtained using a software for automatic optical recognition from the non-standardized text files.

4. The dataset cannot be used to produce SBS data or to automatically correct the data because of the errors due to the automatic optical recognition, but it is a valuable source for selective editing. On the other hand, the explanatory notes in the form of text file are very useful at the review stage, because consulting them allows to recover with a high reliability the 'true' value of investment.

5. From administrative data, two other variables show a high degree of correlation with investment. One is the information on expenditure for amortizable goods reported in Value Added Tax declarations, that is available both for unlimited and limited companies and for sole traders. The other is a derived variable based on the assets at the end of the year minus assets at the beginning plus depreciation and revaluation that can be calculated from financial statements (that are available only for corporations and limited companies).

6. In this paper we describe a selective editing procedure for investment data from SBS that uses the administrative variables described above as auxiliary information and we apply it to SBS data for the year 2011.

II. Selective editing by using SeleMix

A. Selective editing through SeleMix

7. 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).

8. 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 $\theta = (\mathbf{B}, \mathbf{\Sigma}, w, \lambda)$.

8. 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

9. 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.

10. 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 relating 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. Application to Investment Data

A. Data description:

10. In this section we describe how SeleMix is applied to microdata on gross investment that NA uses (together with many other data sources) to produce estimates of gross fixed capital formation by

industry. Data on investment 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).

11. 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 (5770 observations).

12. The results of selective editing depends crucially on the quality of covariates. Di Zio et al (2014) describe a selective editing procedure for investment data from SBS that uses as covariate 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).

13. Although the results are overall satisfactory, they show that using historical values as covariate for investment is far from ideal. On the one hand, the hit rate (i.e., the percentage of erroneous observation in the selected units) was not very high (44.5%). The result is not a surprise, because investment is an extremely unstable variable (a year with very low -or even zero- investment can be followed by a year of high investment and vice-versa). On the other hand, they find that it is difficult to select an observation with an erroneous investment that it is not atypical with respect to the historical value. This is risky when historical data are still contaminated by errors, as can be the case when both the current and the historical variables are affected by the same error mechanism. In these cases, the usage of other variables as covariate may alleviate this annoying situation.

14. In this paper we test the explicative power of two variables that are available to Istat from administrative data: the information on expenditure for amortizable goods reported in Value Added Tax declarations (VAT variable hereafter) and a derived variable based on the assets at the end of the year minus assets at the beginning plus depreciation and revaluation that can be calculated from financial statements (DELTA_STOCK variable hereafter).

15. In Value Added Tax declarations firms are requested to report their expenditure for amortizable goods. The main differences between VAT variable and the target variable are the following ones: the VAT variable does not include expenditures for non-amortizable assets (like land); it includes only part of expenditure for fixed assets and major maintenance produced on own account; in case of an asset acquired through a financial lease, VAT variable includes only the price paid when the ownership of the asset is transferred to the lessee at the end of the lease term (even if the company produces their financial statements according to the International Accounting Standard and, then, registers the acquisition of the asset at the commencement of the lease term). To sum up, VAT variable it is likely to be a good proxy of the target variable when business investment refers mainly to purchases of land and/or assets produced on own account and/or assets acquired when acquired through a financial lease (in case the company adopts IAS).

16. In companies' financial statements, acquisition of fixed assets (i.e. investment) is one of the components that explain the difference between the value of net asset at the beginning of the accounting period and the value at the end of the accounting period:

$$NETSTOCK_{t,end} = NETSTOCK_{t,beg} + INVE_t + REVAL_t - SOLD_t - AMOR_t - WOFF_t + MA_t \quad (1)$$

where $NETSTOCK_{t,end}$ and $NETSTOCK_{t,beg}$ are, respectively, net assets at the end and at the beginning of year t, $INVE_t$ is the acquisition of assets (investment) in year t, $REVAL_t$ is the revaluation of existing assets in the year t, $SOLD_t$ is the net book value (i.e., net of cumulated depreciation) of existing assets sold in the year t, $AMOR_t$ is depreciation of existing assets in the year t, $WOFF_t$ is write-off of existing assets in the year t and MA_t is the effect of mergers and acquisitions.

17. Companies usually report the whole set of variables in equation 1 (including 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). Instead, only a subset of variables is reported in the profit and loss account and in the balance sheet (and investment is not among them). The FS database available to Istat reports the following variables: $NETSTOCK_{t,end}$, $AMOR_t$ and $WOFF_t$. $NETSTOCK_{t,beg}$ can be made observable assuming that $NETSTOCK_{t,end} = NETSTOCK_{t-1,end}$.
Then from FS database we can compute the following proxy of investment:

$$DELTA\_STOCK_t = NETSTOCK_{t,end} - NETSTOCK_{t,beg} + AMOR_t + WOFF_t \quad (2)$$

Note that $DELTA\_STOCK_t$ is equal to $INVE_t$ plus $REVAL_t$ minus $SOLD_t$ plus MA_t. Then $DELTA\_STOCK_t$ is a good proxy of investment when revaluations and selling of existing assets and mergers and acquisition are not very important.

18. Revaluations and mergers and acquisitions are quite a rare event, while selling some of the existing assets it is more common. In practice, we should bear in mind that the value of SOLD is not the price at which the asset is sold but only its book value (that quite often is much lower than the actual selling price). On the other hand, when the asset sold is high valued and only partially depreciated, the variable SOLD will have a very high value and DELTA_STOCK could be very far from the true value of investment. If SOLD is so high that DELTA_STOCK becomes negative we have an observable signal that DELTA_STOCK is not a good proxy of investment and we can assume that for the company the proxy variable is not available. The most annoying situation is when SOLD is high but DELTA_STOCK is positive. In this case, DELTA_STOCK and INVE will be poorly correlated and no signal to detect the problem is observable.

19. A third covariate that we use has a very different nature. Istat may access the explanatory notes of corporations and limited companies in the form of non-standardized text files (one for each company) and in the form of an experimental dataset reporting the value of investment that is obtained using a software for automatic optical recognition from the non-standardized text files. Note 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. For this reason, we use the value of total investment from the experimental database on the explanatory note (EXNOTE variable hereafter) as a third covariate in the selective editing procedure.

B. **Procedure description**:

20. The model described in Section II has been applied using as X variables the three variables described above, i.e., VAT, DELTA_STOCK and EXNOTE.

21. Although as explained in Section III B these variables cannot be considered as direct measures of the target variable, however their predictive power is very high, so that their use as covariates should provide good estimates of the measurement error.

22. Before applying the contamination model to the selective editing procedure, some pre-processing step has to be done. An important pre-processing task is the distinction between missing values and "genuine zeros" in the target variable, in fact, in the survey missing values are registered with a 'zero'. By considering as "zero investments" all cases where no positive values are reported, could result in an underestimation of the aggregates. The underestimation is given by counting as zeros the missing values having an investment. We decided to consider missing values all cases where at least one of the three variables VAT, DELTA_STOCK and EXNOTE are greater than zero.

23. In order to avoid singularities, the contamination model is applied to all the data characterised by an observed value of the investment greater than zero and different -up to a tolerance factor- from the values of all the covariates. Specifically, the survey data where the investment values agree with at least

one of the covariates are considered correct (not subject to the selective editing procedure) and excluded from the estimation process. In fact, applying the estimation algorithm to data including a high number of cases with values of the covariates (almost) coinciding with the target variable, may result in almost perfect fit for the not contaminated data, implying very low value of the residual variance and consequently very large value of the variance inflation parameter.

24. Another important aspect is the one concerning the different missing pattern for the covariates, in fact for different units different sets of covariates may be available. This requires the estimation of different models for different patterns, where, for each pattern, the units included in the estimation process are all those where (at least) the covariates corresponding to the current pattern are available.
For instance, if the current pattern corresponds to the availability of the first two covariates, the data used to estimate the parameters are all those having the first two covariates observed regardless of whether the third covariate is observed or not. Of course, the estimated parameters are used to predict the errors only in the data with the first two covariates observed.

25. In addition to the three covariates, an always observed stratification variable is also used, so that the above procedure is applied separately within each stratum. The stratification variable is the enterprise size in terms of number of employees. Precisely, three size classes are used corresponding to number of employees belonging to the intervals (100-249, 250-499, >499) respectively.

26. A final remark is about the treatment of missing data in the target variable. The models used for selective editing are also applied in order to impute missing data in the target variable. In fact, as selective editing aims at identifying influential errors, imputation of missing investments is needed to determine the impact of the errors on the target estimates. 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 4%.

IV. Results

27. In this section we provide some results about the selective editing procedure. A summary of this results is given by Table 1. In this table, the column '*N.of 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 they result to be error free 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 Value*" and "*Post-editing value*" show the value of the total of investments (in thousands of Euros) computed on data before and after the selective editing procedure.

28. Finally, to make it easier the comparison, in the last column the relative percentage difference between these estimates is reported. This quantity can be viewed as an indicator measuring the impact of the selective editing process. Note however, that relative differences refer only to respondent units, so that these differences cannot be interpreted as the real impact on the final estimates of the total investments (that in fact include also imputed values).

29. The percentage of selected units is 1.8%, and the hit-rate, i.e., the percentage of erroneous observation in the selected units, is 72.6%, and the average of the absolute relative changes is 21.3%. These aspects give an idea of the efficiency and the efficacy of the procedure. The procedure seems to be efficient and effective.

30. 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, that is to compare the results of the procedure with the true values also for the non-selected units.

31. In the remaining part of this section we illustrate the evaluation of the method based on the availability of highly reliable data reported in 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. In fact, after recovering "true values" 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 respectively.

32. Of course, the manual revision of all the not selected observations is not feasible in practice unless a big investment in resources is planned (in that case the data would not need selective editing at all). Thus, in order to evaluate the methodology, we perform the previous analysis only on three estimation domains, namely the industries 32 (Water transport), 38 (Motion picture) and 39 (Telecommunications).

33. Table 2 shows the impact of errors in raw data, after a step of preliminary editing for dealing with the problem of zero-missing error, and the impact of residual errors after selective editing. The comparison is made with respect to the 'true' values obtained after having reviewed manually all the records in the industries. 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 (raw, pre-edited, selective edited), while t^{*} is the estimate obtained by editing all data observed in the analyzed industry and by imputing the non-observed units by means of SeleMix. This is in fact the procedure used for producing the estimates of investments.

34. Table 2 shows that pre-selective editing step leaves the estimates almost unchanged with respect to the ones one would obtain based on the raw data. After revision of the units selected as influential, the error is strongly reduced in industry 32, moving form -35.2% to -2.0%. In the industry 38, the estimation error is practically not affected by the editing procedure, but its "true" value is far below the nominal threshold value (4%). Interestingly, the reduction of the error in industry 32 is achieved by correcting only 3 observation (all the ones that have been selected!) while in industry 38 (correctly) no observation has been selected as influential (see Table 1).

35. If results for industries 32 and 38 are decidedly satisfactory, the same is not true for industry 39. In this stratum, in fact, the selective editing procedure does not select any observation, but the actual error is 26,1%.

36. A careful analysis however has showed that the error is almost completely due to only one unit that did not report correctly in SBS questionnaire the value of a reclassification from assets in progress and advances to the corresponding assets categories. Why the selective editing procedure did not select this unit?

37. Out of the three covariates, EXNOTE is not observable for this firm, while DELTA_STOCK is very different from the true value of investment because of the effect an important Merger and Acquisition activity. Unfortunately, the value of DELTA_STOCK is not very far from the (wrong) value reported in the SBS questionnaire, hence the model cannot see any discrepancy between target variable and covariate that in fact does not exist. It is worthwhile to notice that, there is a discrepancy between the target variable and the other covariate observed for this firm (i.e. VAT). The model gave more weight to the similarity of DELTA_STOCK than to the discrepancy of VAT. The values of this critical unit are the following:

Observed value	VAT	DELTA_STOCK	True value
2316654	1303462	2134424	663000

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

	Industries	N.of Obs.	Selected	No-Edited	Edited	SBS Original Value* (A)	Post-editing Value* (B)	(B-A)/A %
4	Mining and quarrying	20	0	0	0	1,549,041	1,549,041	0%
5	Manufacture of food products	220	2	0	2	1,652,667	1,406,106	-15%
6	Manufacture of textiles	227	1	0	1	579,315	588,957	2%
7	Manufacture of wood	36	0	0	0	73,309	73,309	0%
8	Manufacture of paper and paper products	64	2	1	1	500,077	283,803	-43%
9	Printing and reproduction of recorded media	26	0	0	0	100,141	100,141	0%
10	Manufacture of coke and refined petroleum products	17	1	1	0	643,181	643,181	0%
11	Manufacture of chemicals and chemical products	125	3	0	3	828,393	744,427	-10%
12	Manufacture of basic pharmaceutical products	97	1	0	1	826,628	831,095	1%
13	Manufacture of rubber and plastic products	154	0	0	0	575,771	575,771	0%
14	Manufacture of other nonmetallic mineral products	121	0	0	0	642,383	642,383	0%
15	Manufacture of basic metals	122	1	0	1	889,292	876,648	-1%
16	Manufacture of fabricated metal products	246	4	0	4	756,273	672,888	-11%
17	Manufacture of computer, electronic and optical products	90	3	1	2	374,818	357,631	-5%
18	Manufacture of electrical equipment	125	2	2	0	512,878	512,878	0%
19	Manufacture of machinery and equipment n.e.c.	445	1	0	1	1,323,962	1,256,509	-5%
20	Manufacture of motor vehicles, trailers and semitrailers	128	3	0	3	1,041,013	1,332,121	28%
21	Manufacture of other transport equipment	53	2	2	0	804,111	804,111	0%
22	Manufacture of furniture; other manufacturing	117	0	0	0	301,780	301,780	0%
23	Repair and installation of machinery and equipment	36	1	0	1	32,856	33,121	1%
24	Electricity, gas, steam and air conditioning supply	68	2	0	2	3,786,197	5,137,831	36%
25	Water collection, treatment and supply	40	5	1	4	1,558,931	833,507	-47%
26	Sewerage; waste collection	113	1	1	0	430,582	430,582	0%
27	Construction	161	8	4	4	626,796	650,799	4%
28	Wholesale and retail trade	64	2	0	2	246,986	336,207	36%
29	Wholesale trade, except of motor vehicles and motorcycles	326	0	0	0	1,096,003	1,096,003	0%
30	Retail trade, except of motor vehicles and motorcycles	324	5	2	3	2,779,362	2,210,083	-20%
31	Land transport and transport via pipelines	178	3	0	3	941,378	865,711	-8%
32	Water transport	30	3	0	3	1,214,412	1,918,078	58%
33	Air transport	8	2	1	1	125,697	151,297	20%
34	Warehousing and support activities for transportation	262	3	0	3	5,791,325	5,669,765	-2%
35	Postal and courier activities	8	0	0	0	448,808	448,808	0%
36	Accommodation and food service activities	151	7	3	4	400,126	462,302	16%
37	Publishing activities	30	2	1	1	123,474	135,105	9%
38	Motion picture	16	0	0	0	1,122,431	1,122,431	0%
39	Telecommunications	17	0	0	0	4,921,729	4,921,729	0%
40	Computer programming	175	1	0	1	1,141,965	1,419,564	24%
44	Real estate activities	9	1	1	0	125,673	125,673	0%
45	Legal and accounting activities	88	4	0	4	146,761	140,312	-4%
46	Architectural and engineering activities	48	3	1	2	80,032	71,505	-11%
47	Scientific research and development	11	0	0	0	23,896	23,896	0%
48	Advertising and market research	20	2	0	2	7,476	12,770	71%
49	Other professional, scientific and technical activities	12	2	0	2	12,387	27,027	118%
50	Rental and leasing activities	14	3	1	2	1,650,330	2,376,851	44%
51	Employment activities	55	1	0	1	16,317	16,338	0%
52	Travel agency, tour operator	15	0	0	0	10,360	10,360	0%
53	Security and investigation activities	504	4	2	2	673,554	375,609	-44%
55	Education	17	1	0	1	13,683	14,656	7%
56	Human health activities	181	1	0	1	282,434	268,231	-5%
57	Social work activities	275	6	2	4	144,203	130,804	-9%
58	Creative, arts and entertainment activities	18	2	0	2	148,232	185,984	25%
59	Sports activities and amusement and recreation activities	15	5	2	3	223,889	230,392	3%
61	Repair of computers and personal and household goods	2	0	0	0	267	267	0%
62	Other personal service activities	46	0	0	0	149,636	149,636	0%
	Total	5,770	106	29	77	44,473,221	45,556,011	2%
	* The values are in K €							

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

	Industries		
	32	**38**	**39**
%RE on raw data	-35.3	-0.2	26.05
%RE on pre-selective editing	-35.3	-0.1	26.1
%RE after selective editing	-2.0	-0.1	26.1

IV. Final Conclusions

38. The selective editing procedure proved to be quite efficient: strong improvements in the results have been obtained selecting few units and the hit-rate was quite high. The use of covariates specific for investment that are available from administrative data improves the efficiency of the procedure with respect to the use of historical values. The analysis performed to validate the method is encouraging and should be extended to other estimation domains.

39. A result of the validation analysis is that the explanatory power of DELTA_STOCK is not satisfactory for firms that have been involved in mergers and acquisitions. Further improvements of the procedure could possibly be obtained using as additional explanatory variable also information on mergers and acquisitions that are available from business register.

References

Di Zio, M., Forestieri, P., Guarnera, U., Iommi M. and Regano A. (2014). Use of administrative data for selective editing: the case of business investments. Paper presented to the UNECE Work Session on Statistical Data Editing, Paris, France, 28-30 April 2014.

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.

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

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