Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Geneva, Switzerland, 14-17 April 2020)

ML to identify patterns behind errors in STS statistics

Prepared by F. Rocci, R. Varriale, S. Coppola
Italian National Institute of Statistics, Italy

I. Introduction

1. New Eurostat regulation on enterprises are going to be launched, in order to achieve better results across several new statistics. One of the new regulations that are foreseen to be changed is that of the Short Term Survey on Turnover in Services (FAS), that is nowadays under the EU Regulation (EC) no. 1165/98 of the Council, and subsequent amendments, which define the level of detail, the standards and the frequency according which results has to be published as indices, showing the changes of the target variable in comparison with a fixed reference year (base year), for the economic activity represented by the NACE Rev.2 sector G,H , I , J, M, N). The Turnover in Services index measures the quarterly evolution of sales by service sector enterprises at current prices.
The change under evaluation is to move from a quarterly to a monthly release of the indicators.

2. Istat has started a new project that, from the assessment and analysis of the current STS FAS survey process, aims at identifying how the best features of the current methods and practices of the past experience can be exploited to design a new process, that would require to release data in a shorter time. This means that the survey managers have to translate the current efficiencies to the new process assuring the same quality results.
The project is expected to release methods that would allow important economies of scale, scope and knowledge to be applied in general to the STS productive context, usually working with a limited number of resources. In particular, a number of advantages could be achieved:

- simplifying and standardizing processes;
- reducing costs associated with operational aspects of surveys to increase efficiencies and improve timeliness;
- supporting the management of statistical production processes while eliminating redundancies and ensuring the overall coherence of the statistical processes in the business area.

3. In the current situation (*AS-IS model* hereafter), the FAS survey is characterized by several sub-processes, one for each NACE sector (G,H , I , J, M, N). Each of them applies similar flow models, some methodological solutions can be classified according to similar criteria, but they are implemented in different ways. It is worthwhile to note that each sub-process is also characterized by different E&I strategy.

4. The analysis of the AS-IS model revealed that the FAS survey incurs substantial E&I costs, especially due to intensive follow-up and interactive editing that is used for every type of errors detected. The results of this analysis can be used to direct human intervention during specific phases of the process, thus reducing costs, while safeguarding the timeliness requirements, and ensuring higher levels of efficiency.

In this view, the priority is considered to be *selective editing* phase aiming at identifying units potentially affected by influential errors. In FAS survey selective editing can help to identify, among the units potentially affected by errors, the most “dangerous” ones to be interactively treated, leaving the others to automatic treatments.

5. As a first part of the project, experimental analyses are planned on the FAS survey in order to both correctly design a selective editing strategy, and properly introduce some changes in the whole statistical production process.

6. As mentioned, FAS survey is organized into sub-processes, each designed accordingly a common E&I design, but characterized by different features. We selected for the first test the sub-process regarding the NACE activity 46 (G 46 - Wholesale trade, except of motor vehicles and motorcycles) since it already includes a selective editing procedure. Our aim is to evaluate the efficiency of the current strategy, eventually proposing some changes, and use the results to introduce a selective editing phase for the other sub-processes.

7. Historical data have been exploited. The current selective editing method have been analyzed and a new method have been tested. By identifying hidden patterns in data and “real” influential errors, it would be possible to identify the potential areas of improvement of the survey process. The results are promising, but there is still not a clear signal on how to use the selective editing methods (one of them or together) more efficiently.

8. In this view, we tried to exploit the lessons learned by participating to the High-Level Group for the Modernisation of Official Statistics (HLG-MOS, UNECE) about the Use of Machine Learning in Official Statistics. In this context, one of theme under study is the potential use of ML for editing procedures, since it has been underlined that across the NSO this is the less investigated area (Beck M., Dumpert F., Feuerhake J., 2018). With regard to this, in this work a first experiment to use Random Forest models, both to predict which units represent suspicious data both to assess it prediction potential use over new data and to explore data to identify hidden rules and patterns.

9. The paper is organized as follows. In Section II we describe the selective editing phase in FAS survey, NACE 46 and we propose and the an alternative method based on a probabilistic model. In Section III we use random forest modelling to compare the alternative methods in terms of prediction efficiency, Section IV concludes the work.

II. Selective editing in FAS survey, NACE 46

10. The FAS NACE 46. collects information on turnover in services of enterprises belonging to the NACE wholesale trade, except of motor vehicles and motorcycles. The survey sample is a panel of enterprises, selected at the base year on a quota sampling criteria, in order to reach the 70% of the total turnover as measured by the Italian Businesses Register (ASIA). This results in about 4.500 surveyed enterprises. Statistical results on the outcome index measuring the evolution of sales by service sector enterprises at current pricesare released every quarter, with 60 days of delay with respect to the end of the reference month. The survey process runs continuously during the period of every quarter, macro editing is run two weeks before the final release.

11. The AS-IS model for the E&I process has been represented in macro-phaseswith respect to the Statistical Data Editing flow model (GSDEM; UNECE, 2019), i.e. at first the domain and systematic errors are identified, through deterministic edit rules based on the change over the year of the target variable. After that, a Selective Editing Method is run.

12. Almost every criteria to detect every kind of errors are based on the longitudinal profile of the enterprises itself, based on the comparison between historical data of the same statistical units. For every kind of error(domain, systematic and influential), the treatment is run through interactive methods.

A. Current Selective editing method

13. Two different selective editing Methods are currently used in FAS, NACE 46., the main idea behind both methods is to have an acceptance boundary that varies according to the size of a unit in terms of the amount of the target variable. In particular, both methods elaborate a transformed value for each record that includes a factor for percent change, and a factor for size that is adjusted by the method parameters.

Procedure A. The first procedure follows a Hidiroglou and Berthelot approach,a common method used in periodic surveys (Belcher, 2003) to detect outliers. The larger the size of the unit, the smaller the percent change we allow from one period to the next.

Procedure B. The second procedure calculates score function for each record as the product between the *trend* variation of the company's turnover and the weight of the company in the stratum in terms of turnover in the previous year.
For each stratum the quartiles of the distribution of the scoring functions are calculated and then the limits of the acceptance region are identified as follows:

- Infer = q1-1.5 * interquartile difference;
- Super = Q3 + 1.5 interquartile difference *.

The units for which the control function is positioned outside of this region will be considered influential (Infl = 1).

14. The final classification of the potential influential errors is obtained as the intersection between Procedure A and Procedure B: Method I ≡ Procedure A ∩ Procedure B.
Table 1 shows the selective editing results over the year 2018.

Table 1. Distribution of number influential records (FAS NACE 46.)– year 2018

influential records	frequency	percentage
Yes	1507	9,4
No	14466	90,6
Tot	**15973**	

On average, the amount of the influential errors detected and treated every quarter is about the 9% of the total panel of the survey.

Table 2. Distribution of influential record of Selective Methods I (FAS NACE 46.) – year 2018

	Procedure B.		
Procedure A.	**Yes**	**No**	**Tot**
Yes	1507	4114	**5621**
No	695	9657	**10352**
Tot	**2202**	**13771**	**15973**

The number of influential errors detected is given by the intersection from the two procedures. Procedure A. (HB method) usually identifies much more units as affected by influential error than Procedures B.

B. Test of an alternative selective editing method based on the SeleMix package

15. The method for selective editing (Method II) to be tested is implemented in the SeleMix R package (Guarnera and Buglielli, 2013; Buglielli and Guarnera, 2016). This methodology is currently used in many processes at Istat as a suggested approach for the identification of potentially influential errors in continuous variables (https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processing-tools/selemix). It is based on a latent class model, taking advantage of a probabilistic specification of the true data and of the error mechanism. More specifically, a Gaussian model for true data and an

"intermittent" error mechanism are assumed, such that a proportion of data is contaminated by an additive Gaussian error (Di Zio and Guarnera, 2013).Observations are prioritized according to the values of a score function that expresses the impactof their potential error on the estimates of interest (Latouche and Berthelot, 1992). All the units above a given threshold are selected to be interactively treated since they potentially represent the observations affected by important errors.

16. The model used by SeleMix specifies:

- Target variable: Turnover at quarter T
- Covariate variable: Turnover at quarter T-4

The model is run over each quarter for each strata (given by NACE group activity 3 digit by size class of the enterprise).

17. Selective editing methods identify potential influential errors. This means that the selected units need to be interactively corrected. It is worthwhile to underline that in our experimental study, we could use only the corrections on the units that were selected by method I, i.e. it is not possible to have the corrected values for units selected by Method II and not selected by Method I.

For this reason, to evaluate of the efficiency of the two selective editing methods, we used the absolute overlapping number of selected units by both methods (potential influential errors)and the percentage of the total amount of the turnover covered by the set of overlapping units with regards the total amount of turnover related to the set of units selected by the Method I.

The first results (see Table 3) show how the model, which exploits the longitudinal behavior of enterprises, identifies subsets of flagged data.

Table 3. Distribution of influential data – year 2018

Quarter	**Influential errors**			
	Current Method (I)	**SeleMix (II)**	**I ∩ II**	**%(I ∩ II over I)**
1	370	195	109	29,5
2	403	228	138	34,2
3	343	190	105	30,6
4	391	209	127	32,5
Total	**1507**	**822**	**479**	**31,8**

The percentage of common data identified as being influential is around the 31%, but on average it explains the 85% of the total amount of the turnover. This means that probably the selective editing method based on SeleMix can be used as further instrument to detect the most dangerous errors among the ones identified with the current method: the units detected by both methods can be interactively treated, and the remaining70% of units (related to the 15% of the target variable) can be automatically treated.

III. Random Forest models, a proposal for the analysis of selective editing strategies

18. Machine Learning (ML) is the science of getting computers to automatically learn from experience instead of relying on explicitly programmed rules, and generalize the acquired knowledge to new settings. Given the information (explanatory variables) on a series of subjects, we want to "predict" the variable of interest. Generally speaking, predictive models must perform three essential tasks:

i. Predict new cases: build a model that relates the inputs (control variables/covariates) to a target variable (response variable);

ii. Select useful inputs: data mining problems are often characterized by important cardinality (both of variables and of units); the choice of the most relevant input variables is made in terms of redundancy and irrelevance;

iii. Optimize complexity: choosing between competing models; the selection of a model always involves a trade-off between bias (under-fitting) and variance (over-fitting).

ML methods include two phases: the learning phase and the application one. In the former phase, we have a training phase, where the procedure trains the proposed model, by pairing the input with the expected output, and a validation/test phase to estimate how well the model has been trained (depending on the size of data, the value to predict, input, etc.) and to estimate model properties (mean error for numeric predictors, classification errors for classifiers, recall and precision for IR-models etc.). In particular, the validation phase compares models and select the best performing and the test phase estimates the accuracy of the selected approach.

19. The ML tool Random forests (RF) are an ensemble learning method mainly used for classification and regression. RFs construct several decision trees, and the learning phase is carried out on different randomly selected training and validation sets: the decision chosen by the majority of trees (mode in classification problems or mean in regression) is used as final decision. We applied the R package RandomForest.

20. As first step to run a Random Forest model, the two set of Training/Validationdata and Test data have to be built. In our study, we used the 15783 observations from year 2018 as the Training/Validationset, and 3000 observations from the II quarter of the year 2019 as the Test set.

21. The design of the test consists of choosing the target variable and the set of auxiliary variables:

Target variable: a vector for all units with the flag of influential errors from the current Method I(given by the intersection of Procedure A. and B., see Table 1):

$$Y_{i,T} = \begin{cases} 1 \text{ if unit } i \text{ resulted influential by method I} \\ 0 \text{ otherwise} \end{cases}$$

Covariate variables: a set of core variables:

a. Turnover at quarter T
b. Turnover at quarter T-4
c. Employment at quarter T and T-4
d. Growth rate of Turnover from T-4 to T

22. Several models have been run adding to the set of core variable by time the variable of each specific selective editing elaborated for the survey (both the Procedure A. and Procedure B. and the Method II based on Selemix), in order to test hypothesis about how to predict influential data on new data. Several models against the case not to use any selective editing method during a possible estimation process of suspicious.

Model 1: only core variables – to test the hypothesis whether it is possible to predict influential error on new data only through RF model

Model 2: core variable + flag of influential errors by Method A – to test the hypothesis whether it is possible to predict influential error on new data through RF + Procedure A

Model 3: core variable + flag of influential errors by Method B – to test the hypothesis whether it is possible to predict influential error on new data through RF + Procedure B

Model 4: core variable + flag of influential errors by Selemix – to test the hypothesis whether it is possible to predict influential error on new data through RF + Procedure B

Results of these analyses are reported in Table 4.

Table 4. Confusion matrix of model on the training set:

	percentage of error	
	Training/Validation set	**Test set**
Model 1	6,9	8.1
Model 2	5,6	6.7
Model 3	1,2	2.0
Model 4	6,5	8.1

As expected from the estimation phase in the Training/Validation set, model 3 performs the best results also on the Test set. Therefore, we can suggest that it is possible to use only the procedure B to achieve similar results in terms of influential errors identification and reduce costs of human intervention. In model 1, without the application of any selective editing method, the Test phase indicates a potential expected error of 8.1%.

IV. Conclusions and next steps

23. The design and implementation of any improvement of the production process of Short Term business surveys implies a deep analysis of the current production processes.

24. In the paper we show the analysis on the Short Term Survey on Turnover and Orders (FAS) conducted in Istat, that will be renewed by the introduction of a new Eurostat regulation. In particular, the outcome index measuring the evolution of sales by service sector enterprises at current prices will be produced monthly instead of quarterly.

25. The FAS AS-IS model reveals that the survey incurs substantial E&I costs, especially due to intensive follow-up and interactive editing that is used for every type of errors detected. A better E&I strategy could split errors in influential errors, that need an interactive treatment, and less dangerous errors, that can be automatically treated. In NACE 46, the first mapping of the data process flow of the survey reveals someg ood features of the current Selective Editing phase, based on the use of two methods.

26. As a first step, a new Selective Editing based on SeleMix methodology has been evaluated . The first test of the four quarter data of 2018 revealed that a wise use of this method as an additional instrument in the current flow could reduce the human intervention: an interactive treatment could be reserved to the 30% of the units currently detected, and an automatic treatment could be delegated to the rest of the units.

27. Nevertheless, to use three methods could still represent a huge amount of work for the given constraint of human resources and timeliness. To try to reach a complete different design, a Machine Learning method has been used to test how the historical data about the E&I process can guide in predicting to identify suspicious errors. To this aim Random Forest have been tested, by considering different models using a set of core variables and adding time by time information from the selective editing methods under consideration. The result on the Test set are encouraging, even if this work represents only a first step of the analysis. There are some suggestions to use only Method B. to select influential errors.

28. Summarising, these first results lead to two major ideas to increase the selective editing strategy efficiency, that need to be analysed further:
a. It could be possible to gain in efficiency using a new procedure based on a combination of the current Procedure B, that has the highest efficiency in terms of prediction, and the Method based on Selemix, that seems to focus on the most dangerous units and provides an order of the units in terms of riskiness;
b. On the other side, it could be possible to predict suspicious data using only the current Method B. together with a proper Random Forest model.
At this stage, deeper analysis and further experimental study, using a greater amount of historical data, to compare the different selective Methods in FAS Survey and to asses different E&I design are foreseen.

What is expected is to achieve clearer ideas on which change in model and process could ensures a significant improvement of operational aspects of the whole statistical production process.

References

Beck M., Dumpert F., Feuerhake J. (2018). Machine Learning in Official Statistics.

Breiman L., Cutler A., (2001). Random Forests for Classification and Regression, available at: https://cran.r-project.org/web/packages/randomForest/index.html.

Buglielli, M.T. and Guarnera, U. (2016). SeleMix: Selective Editing via Mixture Models, Version 1.0.1,available at: https://CRAN.R-project.org/package=SeleMix.

Di Zio M., Guarnera U. (2013). Contamination Model for Selective Editing. *Journal of Official Statistics*, Vol. 26, n. 4, pp . 539-556.

Guarnera U., Buglielli M.T.(2013). SeleMix: an R Package for Selective Editing, available athttps://www.istat.it/it/files/2014/03/SeleMix-vignette.pdf.

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.

Lawrence, D., McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earnings. *Journal of Official Statistics*, 10, 437-447.

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

Luzi O., De Waal T., Hulliger B., Di Zio M., Pannekoek J., Kilchmann D., Guarnera U., Hoogland J., Manzari A., Tempelman C. (2008). Recommended practices for editing and imputation in cross-sectional business surveys.

UNECE (2015). Generic Statistical Data Editing Models - GSDEMs, Version 1.0, October 2015, available at: https://statswiki.unece.org/display/sde/GSDEMs.