Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

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


by Fabiana Rocci, Roberta Varriale and Salvatore Coppola

Release date: October 15, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

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

Fabiana Rocci, Roberta Varriale, and Salvatore Coppola[1]

## Abstract

Istat has started a new project for the Short Term statistical processes, to satisfy the coming new EU Regulation to release estimates in a shorter time. The assessment and analysis of the current Short Term Survey on Turnover in Services (FAS) survey process, aims at identifying how the best features of the current methods and practices can be exploited to design a more "efficient" process. In particular, 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. The analysis of the AS-IS process 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 detected errors. 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 work, we present a first experiment using Random Forest models to: (i) predict which units represent "suspicious" data, (ii) to assess the prediction potential use over new data and (iii) to explore data to identify hidden rules and patterns. In particular, we focus on the use of Random Forest modelling to compare some alternative methods in terms of error prediction efficiency and to address the major aspects for the new design of the E&I scheme.

## 1. Introduction

Machine Learning (ML) methods are considered to hold a great potential also in the context of the National Official Statistical organisations (NSOs). To address this need, UNECE High-Level Group for the Modernisation of Official Statistics (HLG-MOS) launched a ML Project in 2019. The project aimed to demonstrate the added value of ML, i.e. whether it enables to produce more relevant, timely, accurate and trusted data in an efficient manner. One main issue for NSOs is to exploit how ML can help in automating certain processes or at least some phases or assisting humans to carry out the process.

The project started in April 2019 with 23 participants from 13 organisations and has grown to over 120 members from 23 countries, 31 national and 4 international organisations. It lasted for two years, many ideas were assessed and new ones took place. Reports and other documents were released on 19 pilot studies, early developments on the use of ML for data editing, and a generic pipeline for production of official statistics using satellite data and machine learning.
In this work, we would like to present the early ideas and the following developments to analyse to which extent ML algorithms can be used to efficiently improve Editing and Imputation. Specifically, the phase of editing, intended as all the methods in order to identify non sampling errors in data, is under study.

At first, the main aspects to be evaluated are related to the potential capacity of ML to learn from data, to predict potential errors automatically, and to deal with huge amount of data, for which to look for unknown underlying pattern can be very challenging. Along with some experiments, an unexpected further aspect came out: the potential gain of information about how to re-design statistical processes choosing among different strategies.

We will go through the re-engineering of the statistical processes for E&I in Short Term Statistics (STS) in the Italian National Institute of Statistic (Istat), to accomplish the new STS Eurostat Regulation requirements, that generally imply very strong constraints of time sources for the processes. Starting from the very first ideas of identifying and

[1]Fabiana Rocci, Istat, Methodology Department, Rome, Italy; Roberta Varriale, Istat, Methodology Department, Rome, Italy; Salvatore Coppola, Istat, Regional Production Department, Naples, Italy

predicting “suspicious erroneous data” using ML, we will present what we encountered on our way and some hints we are trying to grasp on how to proceed from an alternative use of ML methods for E&I.

The current Editing procedures of a STS have been analysed, that resulted in being strongly based on the identification and the treatment of Influential errors, through several Selective Editing methods. Subsequently, historical data have been exploited and a new Selective method have been tested, based on a latent class model (SeleMix). Applying, as further experiment, to use ML to predict Influential errors instead, we discover that to identify hidden patterns can be a major gain those methods can release. In the paper, we will show how we are trying to enlighten the potential areas of improvement of the survey process.

## 2. The project of re-engineering the E&I FAS scheme

New Eurostat Regulation of Economic Statistics on Enterprises are going to be launched, in order to achieve better results across several types of releases, such as the level of detail, the standards and the frequency according which results has to be published. In this context, the Short Term Survey on Turnover in Services (FAS) Regulation is foreseen to be changed, that is nowadays under the EU Regulation (EC) no. 1165/98 of the Council, and subsequent amendments. FAS releases quarterly indices of Turnover, representing the target variable, by service sector enterprises at current prices, showing the changes of the target variable, measure by the quantity of sales, 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 main change under evaluation in FAS is to move from a quarterly to a monthly release of the indicators, that would change the process in a significant way.

The 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. The new strategy should be more efficient, both in terms of human resources and time consuming.

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.

The AS-IS model for the E&I process has been represented in macro-phases with 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 to identify influential errors is run.

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.

The main foreseen change is to start to release data on a monthly base. To this aim, a new project started to design a new E&I scheme, that would maintain a certain quality level reducing human intervention.

Since the analysis of the current procedures has revealed that the FAS survey incurs substantial E&I costs, for which selective editing plays very important role, the results of this analysis are expected to direct human intervention during specific phases of the process, thus reducing costs, while safeguarding the timeliness requirements, and ensuring higher levels of efficiency. 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 their prediction potential use over new data and to explore data to identify hidden rules and patterns.

The strategy of the project started from the analysis of data of the current E&I strategy. Hence, the programmed steps are: i) to test an alternative method to asses the possibility to identify most "suspicious" units to be interactively treated maintaining a given degree of quality of the final estimates; ii) to apply the idea launched in the meanwhile by one of the groups working for the UNECE HLG-MOS project on ML in official statistics, to use ML to predict "suspicious" data to assess the actual potential capacity of ML as a predictive tool in the E&I phase.

The current Editing Method (Method I) in FAS-NACE 46 is based on the results of two editing procedures, Procedure A and Procedure B. The main idea behind both of them is to define 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, based on the Hidiroglou-Berthelot method usually identifies much more units as affected by influential error than Procedures B. The final classification of a unit in potential/not potential influential errors is obtained as the intersection between Procedure A and Procedure B: i.e. Method I $\equiv$ Procedure A $\cap$ Procedure B. 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 shows the results of Method I for the total number of records collected over the year 2018.

**Table 2.**
**Distribution of number influential records (FAS NACE 46.)– year 2018**

| Influential Records | Frequency | Percentage |
|---|---|---|
| Yes | 1507 | 9.4 |
| No | 14466 | 90.6 |
| Total | 15973 | |

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

The Method II for selective editing to be tested is implemented in the SeleMix R package (Guarnera and Buglielli, 2013; Buglielli and Guarnera, 2016). This method 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 impact of their potential error on the estimates of interest (Latouche and Berthelot, 1992), providing an order of the units in terms of their risk to be "true" influential errors. All the units above a given threshold are selected to be interactively treated since they potentially represent the observations affected by important errors. The model used by SeleMix for FAS - NACE 46 is designed as follows:

- 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).

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 the efficiency of the two selective editing methods, we used both the absolute number of units selected by both methods (overlapping units) as potential influential errors, and the percentage of the total amount of the turnover covered by the set of the overlapping units in comparison with the total amount of turnover covered by the set of units selected by the Method I.

The first results (see Table 3) show how the Selemix model, which exploits the longitudinal behavior of enterprises by using the Turnover at quarter T-4 as covariate, identifies a subset of data that are flagged by Method I. More specifically, the percentage of common data identified as being (potentially) influential errors is around the 32% of the total number of units selected by Method I. But, on average, the overlapping units explain the 85% of the amount of target variable checked/verified for errors. This means that probably the selective editing method based on SeleMix can be used as an instrument to detect the "most dangerous" (in terms of the target variable , i.e. turnover) errors among the ones identified with the current method.

**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 |

Thus, summarizing, the hypothesis of a new E&I strategy could be to interactively treat the units detected by both the selective editing methods I and II, and to automatically treat the remaining 70% of units (related to the 15% of the checked target variable). This strategy should be more efficient than the current one, even if it would involve to run three methods, that can be as much costly or even more. Nevertheless, the obtained results are encouraging toward the possibility to reduce the number of data to be interactively controlled, studying alternative automatic methods for the remaining errors. The possibility to check interactively less data maintaining the given quantity of the total target variable validated carefully is the main hint we grasp from the first results.

In the next section, we show additional analyses to design a more efficient E&I strategy. In particular, as introduced, we studied how to use ML methods both to predict which units represent "suspicious" data and to explore current data to identify hidden rules and patterns.

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

The ML method we used for FAS-NACE 46 is Random Forest models. As first step, to run a Random Forest model, the two sets of Training/Validation data and Test data have to be built. In our study, we used the almost 16000 observations from year 2018 as the Training/Validation set, and 3000 observations from the II quarter of the year 2019 as the Test set.

The design of this experiment of using ML in the context of E&I for STS consists of choosing the target variable and the set of auxiliary variables to reach the following aims:

- to learn a model that relates the inputs (control variables/covariates) to a target variable (response variable);
- to apply the same model upon new data to predict new cases;
- to select useful inputs, usually chosen in terms of redundancy and irrelevance;
- to extract edit rules and hidden patterns of potential errors in data;
- for data mining problems, which 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;
- optimize complexity: choosing between competing models, the selection of a model always involves a trade-off between bias (under-fitting) and variance (over-fitting).

To run our study, we construct a new variable of interest describing whether a unit is erroneous or not. In particular, the target variable is represented by the vector for all units *i*, observed at several quarter *T*, with the flag of influential errors from the current Method I (given by the intersection of Procedure A and B, see Table 2): $Y_{i,T} = 1$ if unit *i* resulted influential by method I, 0 otherwise. The covariates are represented by:

a. Turnover at both quarters *T* and *T-4*
b. Employment at both quarters *T* and *T-4*
c. Growth rate of Turnover from *T-4* to *T*

Different Random Forest models have been defined using different core variables, to test different hypothesis about how to predict influential data on new data:

- Model 1: RF with only core variables
- Model 2: RF with core variables + flag of influential errors by Procedure A
- Model 3: RF with core variables + flag of influential errors by Procedure B
- Model 4: RF with core variables + flag of influential errors by Selemix Procedure

Model 1 "learns" the classification of erroneous/not erroneous units from the results of Method I (Procedure A and B) on historical data only. It would assumes not to run other selective editing procedure on new data to predict the outcome variable. On the other hand, Model 2, 3 and 4 "learns" the classification of erroneous/not erroneous units from also from the results of applying an additional selective editing procedures (Procedure A or Procedure B or Selemix). Therefore, if we compare with the results from this first experiment of adding Selemix to the current method I, each of the Model 2,3,4 would assume to run only one selective editing method on new data.

## 5. Results and analysis

At first, the results from the confusion matrix for each of those models, reported in Table 5-1, show that for RF Model 1, without the application of any selective editing method for new data, the Test phase indicates a potential expected error of 8.1%. Furthermore, RF Model 3 performs the best results also on the Test, as expected from the estimation phase in the Training/Validation set. Therefore, we can suggest that a new E&I phase should include only Procedure B to identify influential errors in an efficient way.

**Table 5-1**
**Confusion matrix of model on the training set:**

| | Percentage of error | |
|---|---|---|
| | Training/Validation set | Test set |
| RF Model 1 | 6.9 | 8.1 |
| RF Model 2 | 5.6 | 6.7 |
| RF Model 3 | 1.2 | 2.0 |
| RF Model 4 | 6.5 | 8.1 |

Further analyses are drawn in order to better identify the "type" of units predicted as influential by each different models. Table 5-2 reports the number of units predicted as influential errors by RF Models 1, 2, 3 and 4 out of the units identified as influential errors by Method I, on the Training/Validation set. RF Model 3 shows a good performance by identifying 1423 units out of 1501. Nevertheless, it would imply to interactively treat more units (1526) than we would do with the current Method I.

**Table 5-2**
**Prediction of influential errors by Method I and by RF models. Number of units**

| | Prediction | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | RF Model 1 | | RF Model 2 | | RF Model 3 | | RF Model 4 | | Total |
| Method I | Not infl. | Infl. | Not infl. | Infl. | Not infl. | Infl. | Not infl. | Infl. | Influential |
| Not infl. | 13977 | 468 | 14023 | 422 | 14342 | 103 | 13996 | 449 | 14445 |
| Influential | 638 | 863 | 468 | 1033 | 78 | 1423 | 664 | 837 | 1501 |
| Total | 14615 | 1331 | 14491 | 1455 | 14420 | 1526 | 14660 | 1286 | 15946 |

As a further analysis, we study the quantity of the target variable each model would guarantee to validate: Table 5-3 shows the percentage of units and Turnover that would be corrected from each RF Model out of the total percentage of units and Turnover that is corrected by Method I. As an example, by using RF Model 1, we should interactively check 88% of units checked by Method I, and we would correct 95% of the total amount of Turnover corrected by using the results from Method I.

**Table 5-3**
**Prediction of influential errors by Method I and by RF models. Percentage of number of units and target variable (Turnover).**

| | Method I | Prediction | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | RF Model 1 | | RF Model 2 | | RF Model 3 | | RF Model 4 | |
| | | N.units | Turnover | N.units | Turnover | N.units | Turnover | N.units | Turnover |
| N.units | 1501 | 1331 | | 1455 | | 1526 | | 1286 | |
| % | | 88 | 95 | 96 | 101 | 101 | 103 | 89 | 99 |

From Table 5-3, we see that RF Model 1 and RF Model 4 would imply to analyze much less units, but with a high percentage of control of the total Turnover. Furthermore, it results that 825 units are identified by all methods, 322 units are identified both by RF Model 1 and RF Model 4, but not by RF Model 3. Therefore, to run RF Model 1 and RF Model 4, i.e. using Random Forest with core variables without any selective editing procedure or in addition to the results from Selemix procedure, is probably more efficient in predicting influential units in terms of number of units, with a similar amount of target variable checked and corrected. Hence, the hint is to investigate which kind of units the RF Model 1 and RF Model 4 identify, to assess whether identifying their features could deliver useful information to design a much more efficient E&I.

These results suggest that the current method could be substituted by another one, that would guarantee to identify and treat only the most dangerous data

## 6. Conclusions and next steps

As lessons learned, when dealing with non-sampling errors the comparison of different E&I strategies is not straightforward.

In the Short Term Survey FAS – 46 NACE, to use three selective editing methods could still represent a huge amount of work for the given constraint of human resources and timeliness. To design a new E&I strategy, , a Machine Learning method using Random Forest models has been studied to test how the historical data from the E&I process can guide in the identification of influential errors.

Summarising, the results described in the paper 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 Procedure B, that has the highest efficiency in terms of prediction of “true” errors, and the Method II based on Selemix, that seems to focus on the most dangerous units and provides an order of the units in terms of their risk to be “true” influential errors;
b. On the other side, it could be possible to predict suspicious data using only the Procedure B, together with a proper Random Forest model.

At this stage, the analysis should proceed in order to better define how to use Machine Learning with a different aim, i.e. to learn why different RF models behave differently, that is to study whether there are hidden patterns in the available information identifying influential data. Deeper analysis and further experimental studies using a greater amount of historical data are foreseen, to compare different selective editing methods in FAS Survey and to assess different E&I design.

What is expected, is to achieve clearer ideas on which change in selective editing models and, more generally in the E&I process, could ensures a significant improvement of operational aspects of the whole statistical production process.

## References


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

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

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

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

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

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

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