Machine Learning tool for editing in the Italian Register of the Public Administration, a proposal

Organisation: Istat

Author(s): Fabiana Rocci, Roberta Varriale

Date: 14/06/2020

Version: draft

1. Background and why and how this study was initiated

In the new system of statistical registers of the Italian Institute of Statistics (Istat), the satellite register of the Public Administration (Frame PA) is under construction. This register aims at releasing economic variables of the Italian PA institutions, that are defined by the Register S13 (specific part of the business register related to the PA). The information contained in final Statistical Register Frame PA will be, for each unit, both structural information coming from the Register S13, and some economic variables respecting accountancy definitions. These variables should be obtained as the result of integration of data coming from administrative and survey sources.

Nowadays, Istat is working on the subpopulation of local authorities. The main administrative sources concerning these units are the Public Administration Database (BDAP) and the Information system on the operations of public bodies (SIOPE), BDAP collects information on a stock, while SIOPE on a flow. Except for specific controls of single values, these data sources are not automatically used in the existing Istat production processes.

Two main issues have to be faced during the design of the Frame PA architecture and process:

- ***To gain a proper knowledge of the "new" AD sources** to learn/to identify the relationship between the two sources and to achieve the complete map from the AD variables to the statistical ones.*
- ***How to use the "new" AD in the Frame PA production process**. Source BDAP is regarded as the primary source, because from a theoretical point of view it releases variables according to the statistical target definition of accountancy. Nevertheless, because Frame PA is with reference to two years before, the two administrative sources should be equal. However, some differences are registered between the two AD sources and we need to understand what are they due to. In particular, SIOPE is considered as the auxiliary variable to judge whether such differences are due to an error in BDAP or to any other reason.*

At the present stage, an efficient process of validation is needed to guide us through the data, in order to be able to classify each record among being: "correct", "potentially affected by errors" and "anomalous to be further analysed".

It is important to underline that so far the implementation of a complete automatic procedure to detect and treat errors/anomalous data is not the final aim. This long term result will be achieved only after some tests on the final releases of the overall validation process that would allow to assure the stability of the design.

At the present moment we are at the very exploring step of data. Following the GSDEM scheme (Unece, General Statistical Data Modeling, Unece, 2019 https://statswiki.unece.org/display/sde/GSDEM), errors in data can be classified according to their nature (deterministic or random), their effect (systematic or not) and their impact (low or high) on the final estimates. The definition of different typology of errors can overlap. Different methods can be proposed to detect errors, sometime they are very well focused for specific type of errors, depending on the criteria they are based on, some other time they detect different type of errors to be further evaluated.

Hence, form the Editing and Imputation (E&I) perspective, in the present project we pointed out the following issues:

- *There are some differences in the AD sources, it is necessary to analyse them to classify them in correct data, which difference is only due to the different timeliness, or errors to be treated.*
- *The need to identify specific patterns in data to define group of records according to given features (edit rules), to be further analysed to assess the presence of potential errors.*
- *Structured AD may allow to gain in efficiency and decrease time consuming controls.*

*We focused on **Random Errors**, i.e. type of errors that can happen across the data at random and can be distinguished according to the method of detection and to their impact on the final estimates.*

A. *Usually, methods to identify random error can typically results in identifying two types of errors:, Methods to identify **low impact errors:** edit rules are mostly defined to identify intra record incoherence, without any evaluation about the impact of the errors on the final estimates. In this application, we set up edit rules for which results are expected tobe about errors with **low impact** on the final estimates.*
B. *Methods to identify specifically **influential errors.** Influential errors are errors with an high impact on the finale estimates and are usually detected through methods explicitly using into the criteria the weight of the potential errors on the final estimates. Usually, they are based on some analyses of the distribution of the group of data.*

In Frame PA project, we settled the E&I scheme to identify both low impact and influential random errors. The present work describes how Machine Learning (ML) methods can help in assessing these methods to improve the setup of the entire E&I scheme to gain in efficiency and time resourcing.

In a synthetic way, E&I scheme has been at first settled in a traditional way to identify both type of errors. Hence, several application of ML have been used to find hidden patterns in data (behind potential errors) to extrapolate some edit rules to understand which incoherence in the data is a "true error", hence to be treated.

The obtained results will help in setting the process steps of the entire E&I scheme in a more efficient way.

2. Data

2.1 Input Data (short description)

Frame PA includes different subpopulation. Nowadays, Istat is working on the subpopulation of local authorities (municipalities, unions of municipalities, provinces, mountain communities, metropolitan cities).

The first step to build Frame PA is to select statistical units from Register S13, together with some structural information (address, number of employees, etc).

The main administrative sources concerning the economic variables of these units are the Public Administration Database (BDAP) and the Information system on the operations of public bodies (SIOPE). BDAP records the accounting variables of balance sheets according to the Financial Statement Management Schemes; SIOPE is a system of digital collection of profits and payments made by treasurers and cashiers of all public administrations. Therefore, BDAP collects information on a stock, while SIOPE on a flow.

The first variables we are treating relate to the revenues of the institutions. For simplicity, we will name them E1, E2, E3 and E4. BDAP collects all variables, while SIOPE collects only E4, that we will name E4_SIOPE. Since we are dealing with AD with reference to two years before, subject matter experts expect that the two sources provide the same information on E4.

Variable E4 represents the total amount of revenues that each institution efforts during the year. The other variables represents specific components of the total E4. The revenues E4 are declined across 148 "items". We will refer to the 148 items for the variable E4, for both AD sources, as the "theoretical scheme". The 148 voices are grouped in 9 Titles, representing different types of items for which expenses are done. As subject matters described, the items of the first two Titles are the most commonly fulfilled ones, therefore we classified the items in Title 1, 2, 3 (others).

Table 1 represents the theoretical scheme of the balance sheet including 148 items on the economic variable of interest for each statistical unit; N is the total number of local authorities, for each year. In 2017 data, N is equal to 8229.

Table 1. - Theoretical scheme of the balance sheet for the N statistical units (local authorities), for each year.

Statistical unit	***Structural information***	***Item (Theoretical scheme)***	***Title***	***E1***	***E2***	***E3***	***E4***	***E4_Siope***
1		*I1*	*1*					
1		*I2*	*1*					
.		*.*	*.*					
1		*.*	*1*					
1		*.*	*2*					
.		*.*	*.*					
1		*.*	*2*					
1		*.*	*3*					
.		*.*	*.*					
1		*I147*	*3*					
1		*I148*	*3*					
2		*I1*	*1*					
.		*.*	*.*					
2		*I148*	*3*					
.		*.*	*.*					
.		*.*	*.*					
N		*I1*	*1*					
.		*.*	*.*					
N		*I148*	*3*					

In the data:

- *information on E1, E2, E3, E4, E4_Siope is not necessarily present for each items;*
- *if E4 is present, E4_Siope should be present, and equal, and vice versa;*
- *if E4 is present, also E1, E2 and E3 should be present.*

Variable E4 is the only common information over the two AD sources; therefore, it represents the reference variable to make the comparison between the two sources. Comparison can be performed for each single item or at an aggregate level, i.e. at a Title or Total level.

2.2 Data Preparation

(e. g.: Data Cleaning, Normalisation… – or: none)

We applied different methods to identify statistical units affected by two type of potential[1] errors, Low impact random errors and Influential random errors:

A. *edit rules - low impact random errors: we compared the variables E4 and E4_SIOPE for each statistical unit (local authority), using information from the theoretical scheme (see Table 1), both at item and Title level. We also compared information from the two sources at a Total level, by including or not in the computation of the Total the residual item I149, outside the theoretical scheme).*

   Several group of edit rules have been defined:

   i. *a group of edit rules related to two kind of Totals (with and without I149)*
   ii. *a group of edit rules related to the totals computed at the Title level*
   iii. *three group of edits related to the number of missing items in BDAP in comparison to SIOPE.*

   These rules are run in order to identify internally inconsistent data: a statistical unit not respecting at least a group of edit rules is defined to be potentially erroneous. The total set of internally inconsistent data represent the potential set of data containing errors.

B. *Method to identify influential random errors: a probabilistic model was applied in order to identity statistical units (local authorities) affected by potential influential errors, to be treated interactively. A probabilistic model using mixture modelling has been applied (SeleMix package, R software). For each record, this method releases:*

   i. *a score variable that highlights the impact on the final estimates in terms of both its weight and its probability to be an error*
   ii. *a flag to indicate if it is suggested to be revised, according to a given acceptance threshold of the error accepted in the final estimates. Table 2 shows the result of some exploratory analyses.*

[1] It is worthwhile to note that we refer to *potential* error because we are still at a first proposal of edit rules and they need still to be assessed both by statistical tests and by a proper evaluation from subject matter experts evaluating the correctness of each information.

Table 2. – Incidence of error type A and B.

		B.		TOTAL
		0	1	
A.	0	4470 (53.9)	955 (12.0)	5425 (65.9)
	1	2207 (26.7)	597 (7.0)	2804 (34.1)
TOTAL		6636 (81.1)	1593 (18.9)	8229 (100)

The total incidence of error type A is equal to 34.1%, while the incidence of error type B is equal to 18.9%. The two errors may overlap: 597 units are potentially affected by influential errors and are internally inconsistent, representing21.3% of units affected by error type A and 37.5% of units affected by error type B. It is necessary to understand what is behind these signals: a proper assessment of methods for error identification could help in organizing the entire E&I process, i.e. to set up its steps and the investment on human resources in interactive editing to make it more efficient both in terms of results and costs (i.e. time and human resources).

2.3 Feature Selection

(yes/no, if yes: how, why)

2.4 Output data

(short description)

The output data of the Statistical Register Frame PA project is the list of local authorities with the full record of economic variables according to the theoretical scheme of the balance sheet (see Table 1). The final process will foresee, at every release, a phase of validation to identify potential errors and to correct them, if necessary. This project has been exploited to study how, together with a scheme of traditional E&I, the application of ML technique can release:

i. the list of statistical units together with two flags for item, identifying potential Type A and Type B errors;

ii. a list of the variables that mostly cause the errors in the data: relationships and thresholds would help in defining more precise edit rules and patterns behind data to lead to more efficient E&I process steps and their combination in the entire E&I process.

3. Machine Learning Solution

3.1 Models tried

(e. g.: Multi-Layer-Perceptron, Random Forest, SVM, ...)

Decision trees and Random forests

3.2 Model(s) finally selected and the criterion

(i.e.: which model was why seen being the best?)

ML models used are both DecisionTree and .Random forest.

Several ML models have been performed, testing different set of auxiliary information.

A binary classification problem is set up:

for every type of error J: (A, B):

$$Sel(J)=\begin{cases} 1 & \textit{recorded labeled as potential erros} \\ 0 & \textit{otherwise} \end{cases}$$

Auxiliary information:

For type error A.: structural variables (region, typology of institution, etc.) and Economic variables from both AD Sources

For type error B.: structural variables (region, typology of institution, etc.) and Economic variables from both AD Sources plus flags about type error A., to assess the presence of a statistical relationship between the two type of errors.

The choice between different models is based both on the output statistics (cfr. Results) and, overall, on content reasons. Since the results of these analyses should be used to design a more efficient E&I process, they need to be discussed with subject matter experts.

3.3 Hardware used

(e.g.: Intel Core i5-6300U, 2.4GHz)

The "amount" of available information does not constitute a problem in terms of hardware and runtime.

3.4 Runtime to train the model

(e.g.:2 hours for 500,000 training samples and 25 features)

The "amount" of available information does not constitute a problem in terms of hardware and runtime.

4. Results

(e. g. in terms of RMSE, MAE, distributional accuracy [*], F1 (micro or macro),recall, accuracy, (threshold,) ..., perhaps as a table for different situations (if available))

[*]: If used: How did you measure distributional accuracy? By proportions, moments, quantiles, correlations, …?

ML models used are both Decision Tree and Random forest. Model accuracy, defined as the fraction of correct predictions out of the total predictions, is evaluated to choose the best ML model. Model accuracy has been evaluated both on the training and test data.

Type error A.

- *Decision tree. We tried different composition of the auxiliary variables. No model (trying) showed a good capacity to identify a solid pattern behind the errors: this can be explained by the fact that the differences between the two sources not always are errors but are explainable as different accountancy rules. It is worthwhile to underlie that this result seems to confirm the first comments from the subject experts.*
- *Random forest. Because of the results obtained with decision tree, we did not deep the analysis with random forest tools.*

Type error B:

- *Decision tree. Show very good accuracy: it is possible to identify a group of variables that most commonly can explain the "peculiar" behavior of the given records(potential influential units).*
- *Random forest: show very good accuracy both on the training and the test set.*

From the analyses, we can identify the variables that have a high impact in explaining the investigated phenomenon. In particular, we found that several variables belonging to Title 3 and

- *some variables of belonging to Title 1 are "important".*
- *for these specific variables, we are able to define thresholds to distinguish correct/potentially not correct units and, therefore, to define a king of profiling of record to be regarded as influential errors.*

Summarizing the first results from a content point of view:

- *Among the 148 items, a small group of variables have the major impact on the final prediction, especially some variable of the Title 3. It is important to note that the items composing Title 3 are defined by subject experts as the items with a "residual" importance in terms of economic results*

- *Regions have different characteristics: different group of variables (Title 1 or Title 3) impact the predictions in different ways*
- *We tested several ML model according to different scenario of influential data, released by the probabilistic model implemented in the R package Selemix: the ML model has always very good accuracy fit and mostly lead to the same variable importance output release.*

These results need to be deeply discussed with subject experts. From that, we will be able to set better and more stable ML models to profile potential dangerous data and, therefore, to improve the setup of the entire E&I scheme to gain in efficiency and to reduce the cost of the design and of the implementation of the resulting E&I scheme.

Code/programming language

(e.g. the Python code is stored in GitHub)

Software R, packages: Rpart, randomForest

R code: Available on request.

5. Evolution of this study inside the organisation

(e. g.: Collaboration within the organisation? Has this study advanced ML within the organisation?)

ML tools are already used in Istat for classification, as for other national statistical institutes. The most common settled use is to classify enterprises according to their economic activity. With regards to other fields, web scraping to capture new information have been tested and some experiments on the use of ML tools have been implemented for imputing missing values. With the present project, we are exploring the opportunity to use ML tools specifically for editing procedures.

Since our final aim is to design an entire E&I process, results obtained through the application of ML tools for specific type of errors, should be carefully assessed with the collaboration of the subject matter experts.

6. Is it a proof of concept or is it already used in production?

(If it is a proof of concept: Was it successful? How will its results prospectively be used in the future?)

The described work is a proof of concept.

Frame PA is an innovative process that needs to be designed and implemented. Therefore, it is the first veer that the administrative sources BDAP and SIOPE have been used in a such structured way and machine learning tools are used in the editing process to identify edit rules and/or pattern of features in data to detect different types of errors to be properly treated.

7.1 What is now doable which was not doable before?

(**e. g.: Is something faster or cheaper or more exact? What is the added value using this machine learning?**)

This project is under construction under the supervision and collaboration of statistical methodologists and subject matter experts.

Even though experts have a deep knowledge of the balance sheet theoretical scheme, how the variables can change year by year and are usually able to identify and explain many types of errors, the E&I activity is very time consuming and cannot be extended to all situations.

The combined use of different AD sources represents a new way of treating data and it is considered a great opportunity to produce "better" statistical results, both in terms of time resources and quality of final data. Anyway, experts still need to "learn" how those AD sources together and how to evaluate the observed differences between the AD source in the data. Machine learning tools should help us to detect which are the most problematic situations and to understand if there are some "patterns" behind the inconsistencies observed between the different AD sources.

7.2 Is there already a roadmap/service journey available how to implement this?

(as many sentences as necessary, as few as possible)

No. Nowadays, Istat is still designing and implementing Frame PA, therefore the results of the present work will be used in this process in an iterative way.

7.3 Who are the stakeholders?

Stakeholders of this activity will be the users of Frame PA, and the people responsible for the processes that, to date, produce economic data on local authorities.

7.4 Fall Back

Is a fall back plan in place or planned to mitigate the risk of the ML solution failing in production? Will there be resource left in place to go back to e.g. manual imputation or the use of rule-based scripts?(as many sentences as necessary, as few as possible)

Frame PA is an innovative process that needs to be designed and implemented. Therefore, we are still evaluating the usability of ML tools.

7.5 Robustness

What fail checks are in place or planned to ensure that the ML solution is consistently meeting or exceeding the set gold standard?(as many sentences as necessary, as few as possible)

The results of ML solutions will be discussed with subject matters and people responsible for the processes that, to date, produce economic data on local authorities.

7. Conclusions and lessons learned

(e.g.: ML can be used for editing but one has to have the following points in mind ...)

The Statistical Register Frame PA project has to face the problem of design a new statistical process using different new AD sources. Subject matter experts are very well prepared about the accountancy variables and about any issues related to specific kind of statistical units (for example, PA local authorities). Nevertheless, the result of the AD sources integration can deliver some additional information worth to be explored, in order to better understand and treat any type of suspicious data. Before the availability of several AD sources, the phase of understanding any problem in data was very time consuming, because it was completely related to the expert knowledge of specific institutions.

This first application of ML methods in this context has shown the possibility to use ML to support the design of an E&I scheme to make it more efficient. Indeed, by exploring hidden patterns in the data with ML tools can drive us to help in understanding how to classify units in more efficient way in erroneous/not erroneous in terms of different error types and, therefore, how to combine the different E&I process steps.

8. Potential organisation risk if ML solution not implemented

(as many sentences as necessary, as few as possible)

There is not risk.

9. Has there been collaboration with other NSIs, universities, etc?

(yes/no, if yes: which ones?)

No.

10. Next Steps

(as many sentences as necessary, as few as possible)

The next steps of the work will be to discuss the produced results with subject experts, to improve the use of ML tools to identify error patterns and their treatment. Subsequently, analyses will be performed using data from the following year.

These results will be used to design an E&I process for the Statistical Register Frame PA.