UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
virtual meeting, 2 September 2020

Topic T1_B: Methods

ML to identify patterns behind errors in STS statistics

Fabiana Rocci ***- Roberta Varriale – Salvatore Coppola***

Context

- UNECE HLG-MOS created a Machine Learning Project in 2019

- The project aims to demonstrate and assess the:
  - ✓ the added value of ML
  - ✓ the advanced capability of NSOs to use ML
  - ✓ to identify common issues encountered when incorporating ML in the organisation

- Work packages: Pilot Studies (C&C, E&I, Imagery), Quality, Integration

  Istat attended several WPs: Quality and Editing & Imputation

About Editing

- Editing defined as “identifying suspicious data”
- It has been underlined that across the NSOs this is the less investigated area in ML
  - at first only the PoC by ONS was presented *(Claus Sthamer - Office for National Statistics, UK -)*
- During the first meeting we had a lot of discussion on how ML could help in editing

About Editing

Expectations Editing: starting from former editing results:

- ML may discover rules that have only been “known” by intuition at first and trained in previous experience
- ML could learn which units (records or even cells) in a data set are problematic : to predict
- ML may offer a valid and efficient new instrument for the not rule based perspective on editing
- ML offers a new approach to outlier detection

Istat projects at that time

- **The new Statistical register on economic variables of the Public Administration (PA) units**

i. Aim of the project: to release the architecture of the statistical register, methodological methods both for validation and for integration are under cosntruction. The annual statistical output on some economic results of Public Administrations would be based on this new register instead of using the current processes
ii. Features: many variables, several administrative sources, integration of data, validation of AD sources
iii. Current Editing methods: rules are known by theme expertise, validation of data completely interactive until now, not standardized E&I scheme

Istat projects at that time

- **The new Statistical register on economic variables of the Public Administration units**

i. Aim of the project: the release the architecture of the statistical register, through methodological methods both for validation and for integration. The annual statistical output would be based on this new register instead of using the current process
ii. Features: many variables, several administrative sources, integration of data, validation of AD sources
iii. Current Editing methods: rules are known by theme expertise, validation of data completely interactive until now, not standardized E&I scheme

- **Short Term Survey on Turnover in Services**:

i. Aim of the project: to re-engineerize the validation process of the survey in view of the new EU regulation, that will transform the timeliness from quarterly to monthly
ii. Features: very few variable, short time for validation
iii. Current Editing methods: well established process, most of rules are based on the longitudinal profile of each statistical unit, standardized E&I scheme

Expectations from the application of ML to those projects:

- ✓ **Statistical register PA:** to understand to which degree is possible to reach the automation of controls through standardized rules and the identification of influential data

Expectations from the application of ML to those projects:

- ✓ Statistical register PA: to understand to which degree is possible to reach the automation of controls through standardized rules and the identification of influential data

  we submitted this project to the ML E&I group : we hope that sharing ideas could help in driving us in the design of a completely new E&I scheme design

Expectations from the application of ML to those projects:

- ✓ Statistical register: to understand to which degree is possible to reach the automation of controls through standardized rules and the identification of influential data

  we submitted this project to the ML E&I group : we hope that sharing ideas could help in driving us in the design of a completely new E&I scheme design

- ✓ STS survey on Turnover in Services: to understand to which degree is possible to achieve a more efficient validation process transforming the current process applying less number of methods and procedure to gain in efficiency and time saving

Expectations from the application of ML to those projects:

- ✓ **Statistical register**: to understand to which degree is possible to reach the automation of controls through standardized rules and the identification of influential data

  we submitted this project to the ML E&I group : we hope that sharing ideas could help in driving us in the design of a completely new E&I scheme design

- ✓ **STS survey on Turnover in Services**: to understand to which degree is possible to achieve a more efficient validation process transforming the current process applying less number of methods and procedure to gain in efficiency and time saving

  starting from former results, we were fascinated about the idea to be able to predict the 'problematic/sospicious units' for new data on the current process
  We started to work together with the field expertise team

What we present today:

- the first results for the re-engineering of the STS survey of Turnover in Services

What we present today:

- the first results for the re-engineering of the STS survey of Turnover in Services
- We try to explain how we started to move from

  To learn how “to predict”

  towards

  To learn how “to give support to decision making” in a validation process
  Identifying hidden pattern in data to save time and resource for the interactive editing

Features STS Turnover in services (FAS)

- For the economic activity represented by the NACE Rev.2 sector G,H , I , J, M, N. Statistical results on the outcome index measuring the evolution of sales by service sector enterprises at current prices are 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
- The current process is characterized by several sub-processes
- We focus on the the FAS NACE 46 (G 46 - 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 of about 4.500 surveyed enterprises
- The E&I process follows a well established scheme
- New regulation is going to be introduced that will ask to move from quarterly to monthly releases

- 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 suspicious data that are detected

- Our aim: the re-engineering the process to re-organize human intervention during specific phases of the process, thus reducing costs, while safeguarding the timeliness requirements, and ensuring higher levels of efficiency
- We have: long history in longitudinal data
- Several releases of the survey output and of the E&I run for each of them

i. to predict suspicious data

ii. to extract rules or patterns behind the suspicious data

Current situation AS-IS of the E&I process

- First step: Domain Obvious and Systematic errors are corrected (editing DOS)
- Second step: huge investment on the selective editing are currently used in FAS, NACE 46
- The final selection of influential data are determined as the intersection of two separate procedures, both based on a transformed value of the longitudinal ratio:

$$\text{for each unit } i : Z_{i,T} = \frac{Y_T}{Y_{T-4}}$$

procedure A. it follows a Hidiroglou and Berthelot approach: the larger the size of the unit, the smaller the percent change we allow from one period to the next

procedure B. a score function is computed for each record

$$S_{i,T} = Z_{i,T} \times w_{i,T-4}$$

each unit i with $S_{i,T}$ outside the interval (q1-inter$[S_{i,T}]$, q3 + 1.5 inter$[S_{i,T}]$) is flagged

Set of influential data: I

Method I ≡ Proc A ∩ Proc B.

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**	

Set of influential data: I

$$\text{Method I} \equiv \text{Proc A} \cap \text{Proc B}.$$

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**

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

- The method for selective editing (Method II) to be tested is implemented in the SeleMix R package
- This methodology is currently used in many processes at Istat as a suggested approach for the identification of potentially influential errors in continuous variables
- It is based on a latent class model, taking advantage of a probabilistic specification of the true data and of the error mechanism
- 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). All the units above a given threshold are selected to be interactively treated since they potentially represent the observations affected by *influential* errors.

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

The model used by SeleMix results:

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 model used by SeleMix results in comparison with the current Method I:

- 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
- Hypothesis of new design we were going to test: the units detected by both methods can be interactively treated, and the remaining 70% of units (related to the 15% of the target variable) can be automatically treated.

ML application: our idea

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).

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

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

Covariate variables: a set of core variables:

- Turnover at quarter T
- Turnover at quarter T-4
- Employment at quarter T and T-4
- Growth rate of Turnover from T-4 to T

- Used ML tools: Random Forest (RF) and Classification Trees (CT)
- Software R

- Several models have been set up and tested:

i. given the set of core variables - related to the survey

ii. for each other model predictor(s) representing the result from the several procedures of Selective editing is(are) added

- 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 Random Forest (RF) model

- Several models have been set up and tested:

i. given the set of core variables - related to the survey
ii. for each other model predictor(s) representing the result from the several procedures of Selective editing is(are) added

- 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

- Several models have been set up and tested:

i. given the set of core variables - related to the survey

ii. for each other model predictor(s) representing the result from the several procedures of Selective editing is(are) added

- 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

- Several models have been set up and tested:

i. given the set of core variables - related to the survey

ii. for each other model predictor(s) representing the result from the several procedures of Selective editing is(are) added

- 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 + Method II

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%.

First conclusions …

- 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

- 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 models 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.

- At this stage, deeper analysis and further experimental study are expected to be developed:
  - using a greater amount of historical data and to elaborate specific longitudinal indicators as added explanatory variables
  - to compare the different selective Methods in FAS Survey and to asses different E&I design

- Variable importance, analysis of nodes and classification tree next step to identify patterns behind different classification of data according different degree of being suspicious

- 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

- Any further ideas or hints are welcome!

Thank you for you attention