HLG-MOS UNECE project about the use of ML in Official statistics

WP1 Editing & Imputation

Istat Pilot Study

The potential use ML to design the Editing process of The Statsitical Register of economic variable of PA

Fabiana Rocci
Roberta Varriale

Virtual meeting, 1 April 2020

HLG-MOS UNECE project about the use of ML in Official statistics

- Istat in 2019 started to participate to the **E&I WP** group

- Aim: to learn how ML are used for E&I: features, methods, possible benefits and challenge to implement it in the actual process

- During the sprint meeting in London from an overview of literature and experiences in NSI offices of ML, has resulted
  - very few application have been done so far in the editing part regarded as the part **to detect** data containing *non sampling errors*
  - most applications are related to the imputation for missing data

HLG-MOS UNECE project about the use of ML in Official statistics

- Istat proposed a PoC about the imputation of the Attained Level of Education for the Statistical Register of the Population (Filippini R., De Fausti F.)

- ***About Editing*** *we were left with an homework:* to analyse the potential use of ML for the editing part

- Some ideas during the sprint meeting were suggested by the E&I group

- In this view, a cooperation among ONS (PoC on Editing), Destatis and Istat started to study the potential use of ML for editing, meant as only the functions to detect the suspicious/errors data

HLG-MOS UNECE project about the use of ML in Official statistics

- Notes were released: *Machine Learning for Data Editing Cleaning in NSI (Editing & Imputation): Some ideas and hints*

- The reasoning has been based on the *General Statistical Data Editing Model* (**GSDEM**, Unece) scheme:

  - a guide to classify **type of errors**
  - definition of methodology to detect them, basically based on the concepts:
    a. **edit rules**: express relationship among variables the data are expected to respect
    b. **functions** to analyse the data distribution
  - methods to be properly combined and adopted for each type of error during different phases of a statistical process

HLG-MOS UNECE project about the use of ML in Official statistics

- The idea behind the proposal is:

  the result of the editing phase can be regarded as a problem of classification of data in:

  i. correct

  ii. suspicious to be erroneous

  iii. surely erroneous

- In this sense, some ***hints*** have been given according to the two situations:
  - there exists a previous release of the process for which the Editing documentation is available, it can be modelled a test by a supervised approach:
    a. to predict the suspicious/erroneous data
    b. to extract the edit rules
  - there is not a previous release of the process, unsupervised approach to investigate the underline structure of data and to delineate some edit rules

In the meanwhile in Istat….

- The E&I group is always involved in the re-organization of statistical processes
- I am personally involved in the project to design a new Satellite Statistical Register
- I launched the challenge to test how to use ML on a new statistical process
- I was personally passionate about the idea to drive us in understanding **edit rules**…!
- Roberta Varriale, who is in charge of the design of this Statistical Register process accepted!!!!
- She also belongs to the E&I and she studied and applied ML in some other experiment, but this is the first time for the specific Editing part

HLG-MOS UNECE project about the use of ML in Official statistics

- The Project: The satellite Statistical Register of the Public Administration (Frame PA)

- Target variables: some economic variables respecting accountancy definitions

- The Target variables should be obtained as the result of integration of data coming from the new AD sources

- New Administrative Data (AD) sources are available that release variable useful for several statistical processes

- A new working team has been created: methodologists, IT, subject matter experts from National Accounts and Structural Economic Surveys

- Two main AD sources
  - BDAP that collects information on a stock and it is considered the primary source
  - SIOPE that collects data on a flow

  Because Frame PA is due to be released two years after the reference period, the two administrative sources should/are expected to be equal

- The new AD sources are much more structured (especially since 2017, with a new information system) than the previous ones and they deliver new option of operational processing

Main issues to be faced during the design of the register final architecture and process:

- **New AD sources:** subject experts know very well the target variables, nevertheless they still need to learn/to identify the complete map from the new AD variables to the statistical ones

- **Use of AD in existing processes**: information from BDAP is considered to respect the theoretical scheme of target variables. SIOPE is mainly used as auxiliary information to integrate the BDAP source <u>if</u> necessary (in the presence of errors or missing data)

HLG-MOS UNECE project about the use of ML in Official statistics

Main issues to be faced during the design of the register final architecture and process

- **From the E&I perspective**:
  - There some differences in the AD sources, it is necessary to understand what are they due to, to classify them in *correct* or *errors to be treated*
  - To find patterns in data in order to design efficient editing rules procedures to identify group of records to be treated interactively (complete automatic procedure is not the final aim until the process could be considered definitely stable)
  - Specific needs to exploit the much more structured data to gain in efficiency and decrease time consuming controls

General description of data

Pilot Study on ML for editing for the Statistical Register Frame PA

- Structural variables are delivered by other Statistical Registers: typology of institutions, geographical features, employment, etc.

- Data about the economic variables:
  Years : 2017 and 2018

  census of all the units belonging to the Base Frame of the Target Population

  (nevertheless there is a small percentage of missing data: solved through integration with SIOPE auxiliary AD source)

- Economic variables:
  variables on the expenditures of the institution
  E4 : total amount, present in both AD sources
  E1, E2, E3: variables components of E4, present only in BDAP

- BDAP is distributed across 148 Items, grouped in 3 major Titles (following the so called Theoretical scheme)

- SIOPE is properly re-conducted to this scheme (it releases more variables, but only the ones to re-produce the theoretical scheme are selected)

- Information of the 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 viceversa;
  if E4 is present, also E1, E2 and E3 should be present.

- Analyses are made on the variable E4

- Theoretical scheme may change every year (depending on regulamentation)

HLG_MOS UNECE project about the use of ML in Official statistics

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

What we did: we focused on **Random Errors,** that type of errors that can happen across the data at random (there is not ant deterministic feature behind them) and are distinguished according to the impact on the final estimates

- a first step was to find random errors with low impact on the finale estimates: usually they detected through edit rules

- as a second step, we focused specifically on the **influential errors**
this is an error that has an important impact on the estimates.
Units potentially affected by influential error usually need to be accurately analysed (e.g., interactively, recontacting,etc.)
Selective editing is a general approach for the detection of influential errors: mostly based on some analysis of the data distribution

- For both type of random errors, we applied ML techniques according to the ***Hints***

- Software R, packages:
  randomForest & Rpart (Decision Tree)

- R code: Available on request

- measure used to assess the model:
  - model accuracy on the overall **training and validation data** (confusion matrix)
  - classification accuracy, on the **test data**, external to that one used to train the model (confusion matrix)
  - distribution of the data (national and regional level)
  - variable importance (decision tree)

A. Random errors with low impact :

A1. records with differences in the two AD sources are identified: suspicious group of data

A2. general analysis of those records: only part of them are considered to be errors

A.3 Hence the classification between:

- Correct units
- Wrong units: units affected by errors

A.4 **Application of ML tools**:

i. **Random Forest**: to predict the erroneous data

At this stage of the experiment no model has shown the capacity of prediction

ii. **Decision Three**: the application of these models allowed to highlight some group of items having more impact than others on the error identification

B. **Potential influential errors :**

B1. A probabilistic model using mixture modelling has been applied (SeleMix package, R)

B2. A dataset of all the record labelled 0/1 as *correct/potential influential error* was released

B3. Starting from this result, our aim was to evaluate the possibility to predict which data are potential influential errors or not, by using ML tools

B4. **Application of ML tools**

i. **Random Forest** models to predict the potential influential units
At this stage there are some results, but they should be better modelled

ii. **Decision Tree**: to analyse the underline patterns to extract edit rules
this is the main aim now!

First Results:

- Among the 148 items, a small group of variables have the major impact on the final prediction, especially some variable of the Title 3 (representing the group of items with a residual importance in terms of economic results)

- Going region by region: different group of variables (Title 1 or Title 3) impact the predictions in different ways

Next steps

- Deep analysis on available data
- Discuss results with subject matter experts to focus to propose *automatic edit rules*
- To better identify, among the suspicious records representing potential errors, the actual errors (subject matter experts) and use ML tools to predict this target variable

…any advice, comment or constructive criticism is very welcome!!!

Thank you for your attention