Istat | Istituto Nazionale di Statistica

Rome. 6-7 December 2023

SECOND WORKSHOP ON METHODOLOGIES FOR OFFICIAL STATISTICS

State of play and perspectives on machine learning at Istat

Marco Di Zio

Istat | DIRECTORATE FOR METHODOLOGY AND STATISTICAL PROCESS DESIGN

Outline

- Context
- Machine learning at Istat
- Lessons learned and questions to deal with

Background in NSO

Some machine learning papers in NSO in ’90s:

- Nordbotten S. (1995). Editing statistical records by neural networks, *Journal of Official Statistics*
- Nordbotten, S. (1996). Neural network imputation applied to the Norwegian 1990 population census data. *Journal of Official Statistics*
- Roddick, L.H. (1996). Data editing using neural networks. *Data Editing Workshop and Exposition*.

Background in NSO and Istat

- **1999** AUTIMP-EU project

Evalutaion of Tree based methods for imputation.

Developed an algo WAID to impute.

- **2000** Euredit EU Project. 12 participants (Universities and NSO)

Multilayer perceptron, correlation matrix memories, self-organizing maps, support vector machines compared to traditional methods for editing and imputation.

Background in Istat

Than, research papers…

…

a boost in Istat was given by the *Scheveningen Memorandum "Big Data in Official Statistics*" (2013) and Bucharest Memorandum "Official Statistics in a datafied society - Trusted Smart Statistics" – (2018)

They made the TSS "more concrete" in Istat and the nature of the data and problems are naturally dealt with machine learning techniques

Some TSS and ML at Istat

Remote sensing data for urban green. Unsupervised classification: Kmeans, Kmedians, KDE, Canny edge, …

AIS – Automatic Identification System. Imputation of missing data in the vessels' route: Xgboost, deep learning

Sentiment analysis on twitter/X data. Social mood on economy index, gender based violence, hate speech: NLP models (lexicon methods, word embedding, BERT,…), deep learning for clustering algorithm

Some TSS and ML at Istat

Web scraping and enterprise automatic classification: Gradient Boosting, Random Forest, Neural network, SVM

Smart Surveys (ESSNet Smart Surveys Implementation, 2023). ML for structuring unstructured data, classify objects acquired from the images, or physical activities using accelerometer data, or leisure activities using GPS data matched with street maps.

… and others…

Studies on ML for imputation – Integrated Survey and admin data

2019 *UNECE HLG-MOS Machine Learning Project*

Multisource data. Imputation based on integration of admin and survey data

Application. MLP for mass imputation of level of education in Population Census.

MLP results compared with official ones.

Lessons learned (1)

TSS are a natural environment for machine learning:

- Unstructured data (signals, texts, images..)
- big volume
- prediction and classification problems

Lessons learned (2)

Imputation with integration of admin and survey data

- Use a random imputation
- Sampling weights should be used in models. Sometimes difficult to include directly in the ML model estimation. We explored their use in the loss function, and 'exploding' data by replicating units according to the weights
- *Application study results. Similar to official procedure. Due to few explicative variables? However, important because of the use of sampling weights*

We intend to explore

- Longitudinal data

  Relevant for registers because admin data provides data at person level for each time *t*

- Clustered data

  Multistage sampling: e.g., survey of households

Open problems we need to deal with

Quality evaluation

Measures mostly developed for prediction

We need accuracy measure for aggregates, e.g., confidence intervals for mean, quantiles,..

Open problems we need to deal with

Bootstrap.

- Computationally demanding (feasible?)
- Need to deal with complex sampling designs

Multiple imputation.

- Less computationally demanding
- How to perform ‘Proper imputation’?

A risk

Risk: Think of delegating everything to the machine

ML powerful tool but researcher is still important in modelling, for instance how to deal with and model different problems

selection bias
missing data mechanism
dealing with sampling design
…..

Thank you

MARCO DI ZIO | dizio@istat.it

Istat | Istituto Nazionale di Statistica