INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Process automation for treatment of missing values in e-invoice data

Bruno Lima, João Poças, Sofia Rodrigues, Almiro Moreira, Paulo Saraiva

2023-06-14

Introduction

e-invoice:

- monthly data on mandatory e-invoice declaration from Tax Authority;
- 80M+ registries on issuer/acquirer transaction values;
- incomplete administrative data that needs to be treated to be usefull for statistical production;

Aim

Instituto Nacional de Estatística
Statistics Portugal

Implement an automated process to detect and correct missing values.

- Review data from 2000+ most economically relevant issuers;
- A reproducible pipeline allowing for robust data-driven decision making;

Methodology

Assumptions:

- The identification of the most relevant issuers was done with feedback from users;
- Missing values:
  - *total missing: no value is reported by the issuer;*
  - *parcial missing: both the reported value and the number of acquirers by issuer are much lower than expected.*

Methodology (cont.):

An isolation forest algorithm for anomalies' detection is applied to the identification of partial missing values both for transaction values and reported number of acquirers:

- the algorithm of isolation forest consists in splitting sub-samples of the data according to the feature;
- the rarer the observation, the more likely is that the outlier (anomaly) is put alone in one branch and fewer splits will take to isolate it;
- it allows us to have a conservative approach, only treating the more obvious (higher probability) anomalies.

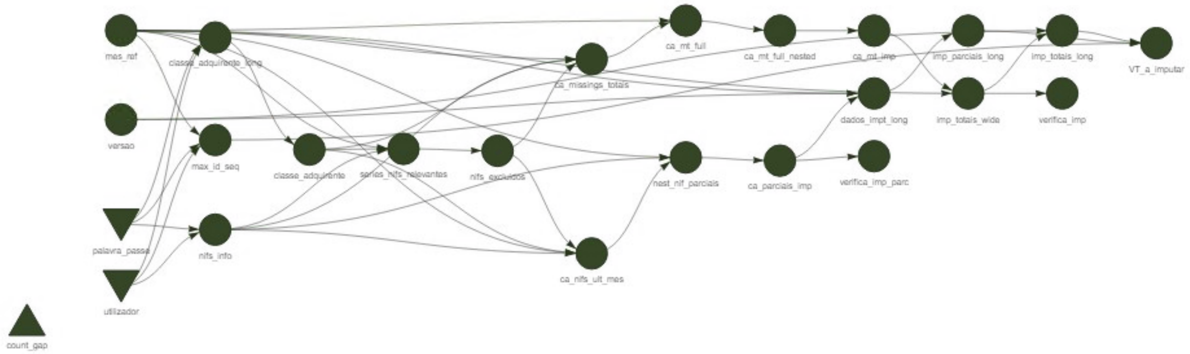
Methodology (cont.):

Kalman Smoothing on structural time series models is used for imputation:

- imputation is done on taxable amounts univariate time series at level issuer / acquirer's class;
- imputations are done both for total an partial missings:
  - *for partial missings, the reported partial value is subtracted from the Kalman Smoothing imputation.*

Results

Process network



Conclusions

- This procedure allows to identify and correct missing data and to ensure an increase in the quality of administrative data;
- When possible, the goal is to replace data collected through traditional surveys with administrative sources, in the field of short-term statistics.

Cheers from Portugal

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Thank you for your time.