2023 UNECE Expert Meeting on Statistical Data Collection
'Rethinking Data Collection' online (12 - 14 June 2023)

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

Bruno Lima, Statistics Portugal, bruno.lima@ine.pt
João Poças, Statistics Portugal, joao.poças@ine.pt
Sofia Rodrigues, Statistics Portugal, sofia.rodrigues@ine.pt
Almiro Moreira, Statistics Portugal, almiro.moreira@ine.pt
Paulo Saraiva, Statistics Portugal, paulo.saraiva@ine.pt

Abstract

Monthly data from a mandatory e-invoice declaration system must be analysed and treated before making it available for Statistics Portugal users. As any data obtained from administrative sources this data is sometimes incomplete and must be corrected. Here we describe an implement automated procedure for the detection of missing values (total and partial) and the consequent imputation of new values.

Introduction

Every month, Statistics Portugal receives data from the Tax Authority on a mandatory e-invoice declaration system. However, this data is sometimes incomplete, so to improve its usefulness for statistical production, we analyse more than 2000 of the most economically relevant issuers, in terms of turnover and/or persons employed, to detect any missing data. This analysis requires a great deal of effort in its processing, especially as it requires the analysis of the corresponding time series. To help with this, we have set up an automated procedure to identify and fill in any potential missing data.

In the context of the building a National Data Infrastructure, Statistics Portugal is developing a centralized treatment processes that aim to improve the consistency and quality of data for various purposes. In this particular case, one of the main goals is to replace data collected through traditional surveys with this administrative source, in the field of short-term statistics.

Methods

The implemented procedure is applied to a set of the most relevant issuers, capable of ensuring a remarkable quality improvement in the processed data. To classify an issuer as relevant, we compile their historical reported values and take into account the contributions from Statistics Portugal's subject matter units (National Account Department and Business Statistics Department, the most "interested users" in this data).

Firstly, we identify those relevant issuers that remain in business and had a reported value in the previous month. For these issuers, we compute the monthly taxable amount and the number of registers aggregated by both issuer and acquirers' class.

There are two types of possible missing data:

- total missing, when no value is reported by the issuer;
- partial missing, when the reported value and the number of acquirers is much lower than expected.

For the latter, an isolation forest algorithm is applied to detect partial missing (anomaly detection) for both the taxable amount and the number of records (acquirers) grouped by issuer.

Kalman Smoothing in structural time series models is utilized to fill in missing taxable amounts at the level of issuer/acquirers' class

Results

The 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.

Imputation of values for detected missing is done using Kalman Smoothing on structural time series models:

- imputation is done on taxable amounts univariate time series at level issuer / acquirer's class;
- imputations are done both for total an partial missing:

Particularly, for partial missing, the reported partial value is subtracted from the Kalman Smoothing imputation.

A reproducible pipeline is defined (Fig. 1), ensuring that we always have the same result for each monthly run, allowing for robust data-driven decision making.

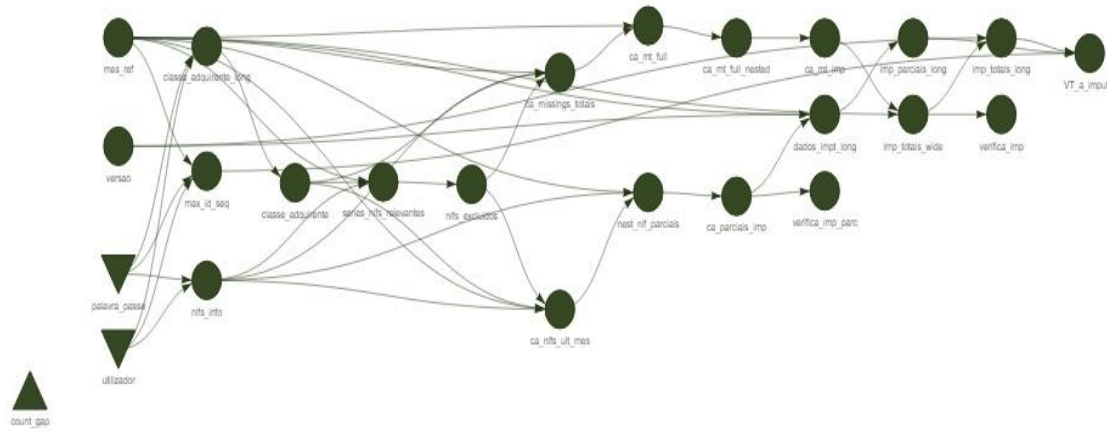

Figure 1. Procedure workflow for the detection an imputation of missing values

Conclusions

The described procedure allows to identify and correct obvious reporting errors and ensure an increase in the quality of administrative data, to be used by statistical procedures. The treatment of administrative data is a continuous process, and we keep working for improving these methods, to improve quality also with the “not so obvious anomalies”.

By applying these types of procedures in a centralized and comprehensive way (in terms of application to different data sources) we are contributing to an effective improvement in the quality of statistics, to a reduction in the statistical burden but also to a more consistent and useful National Data Infrastructure.