
INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly
financed by the European Union
CASINO ESTORIL
HOTEL PALÁCIO
EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Administrative Data Quality Challenges through the Lens of e-Invoice

António Portugal, Bruno Lima,
João Poças, Paula Cruz,
Salvador Gil, Sofia Rodrigues

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly
financed by the European Union

Topics

1. Context
2. Data Quality: Challenges and Methodology
3. Collaborative efforts
4. Application to other cases
5. Conclusions

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

The conference is partly financed by the European Union

1. Context

- The use of administrative data has been a constant over the years at Statistics Portugal (INE), aiming a significant impact on reducing the statistical burden
- 2020 brought three new features to Statistics Portugal's production:
  - The negative impact of the COVID-19 on the response rates to business surveys
  - Receiving a huge amount of data from the electronic invoicing system (tax authority)
  - The creation of a new unit dedicated to the collection, analysis and quality treatment of administrative data
  - Management support for training in new skills (data science) tools and techniques, for all organisational units

The conference is partly financed by the European Union

1. Context

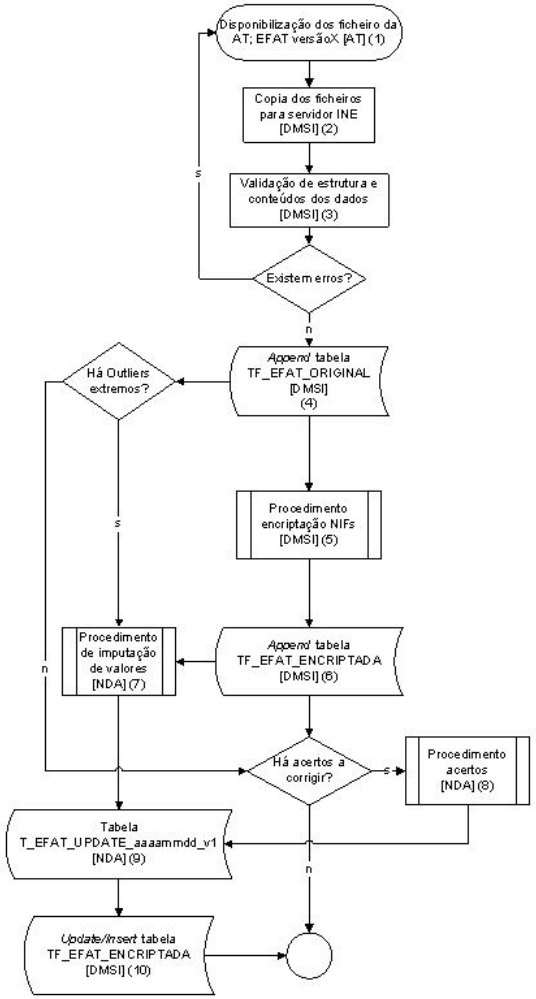
- INE receives monthly, from the Tax Authority, about 100 M records (x 3 months, since January 2024) regarding taxable amounts aggregated by issuer and buyers

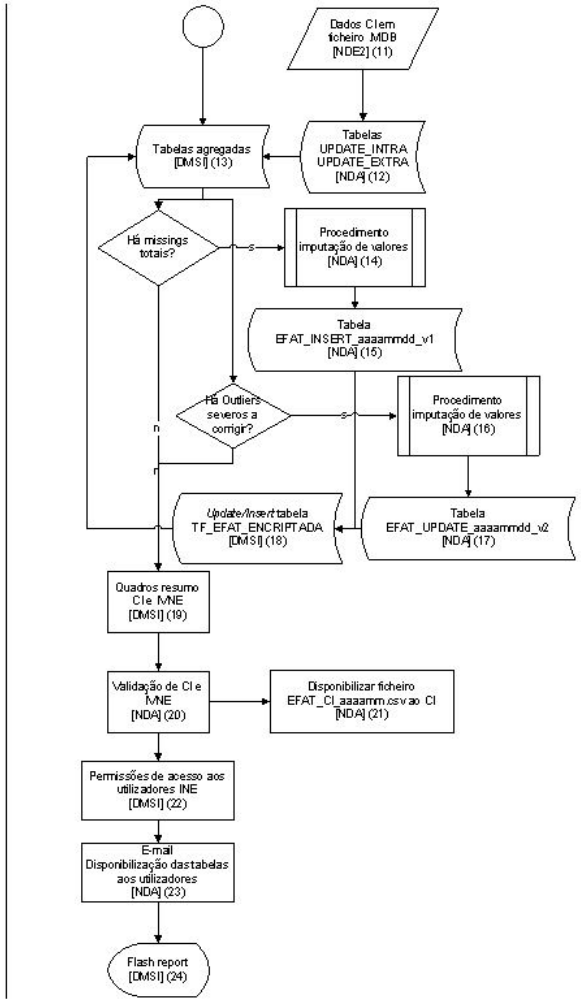
- These amounts result from invoices issued by legal entities (or natural) persons based in Portugal

The conference is partly financed by the European Union

2. Data Quality: Challenges and Methodology

- Validation of data structure
- Hashing of personal identifiers
- Normalisation of attributes (country codes)
- Adding attributes from other sources
- Imputation of outliers and missing values
- Correction of negatives values
- Comparison with other datasets
- 24h to treat each reference month





The conference is partly financed by the European Union

2. Data Quality: Challenges and Methodology

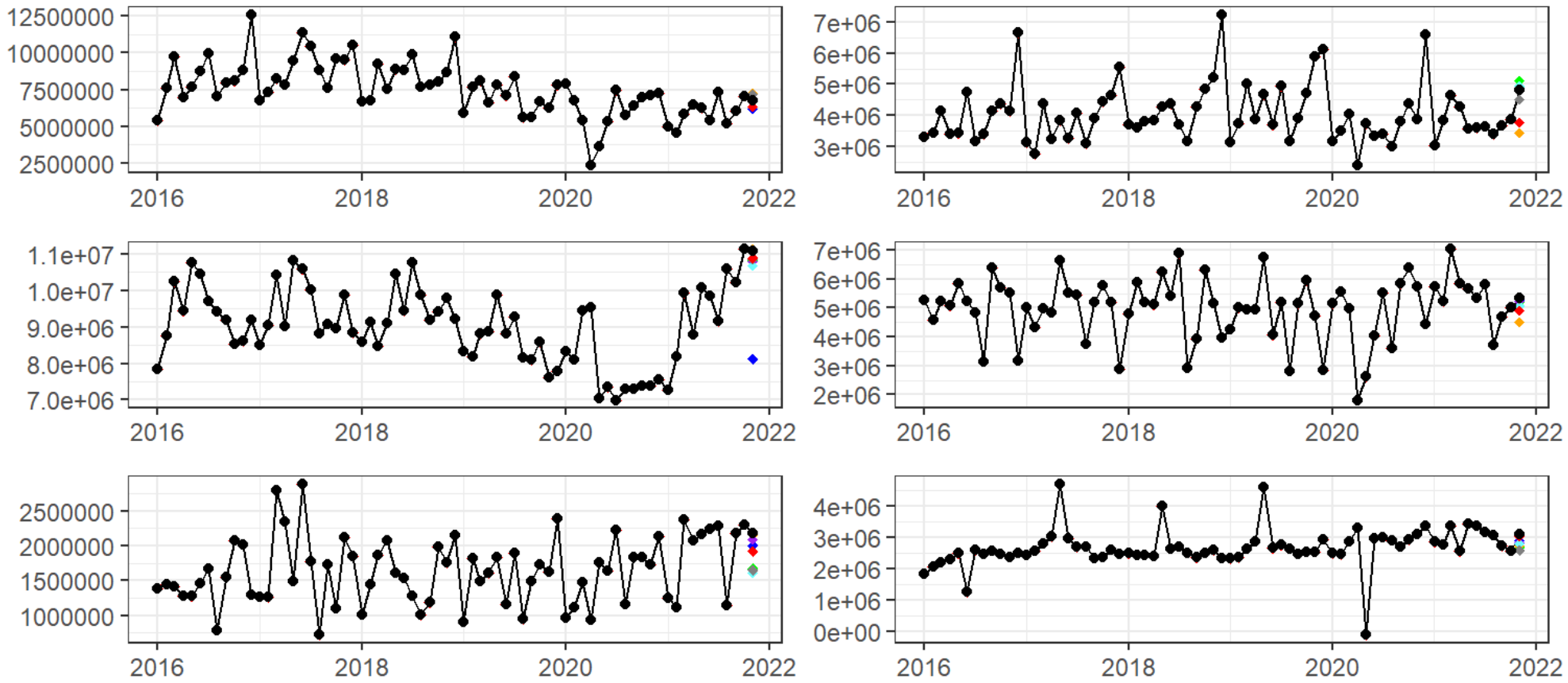
Missing values

- **Total missing**: in one period, an issuer does not have any value
- **Partial missing**: in one period, an issuer has, simultaneously, less invoiced value and less “buyers”
- When missing values are identified, for most relevant issuers, we need to **estimate**
- Using the time series of the taxable values (VT) of each issuer, we make forecasts and provide **imputed data** to the users

The conference is partly financed by the European Union

2. Data Quality: Challenges and Methodology

E-invoice time series are very heterogeneous (difficult to fit a unique model)



The conference is partly financed by the European Union

2. Data Quality: Challenges and Methodology

Evaluate, continuously, different nowcasting algorithms:

- Auto ARIMA, TRAMO/SEATS, X-13 ARIMA-SEATS, Kalman Smoothing, Prophet

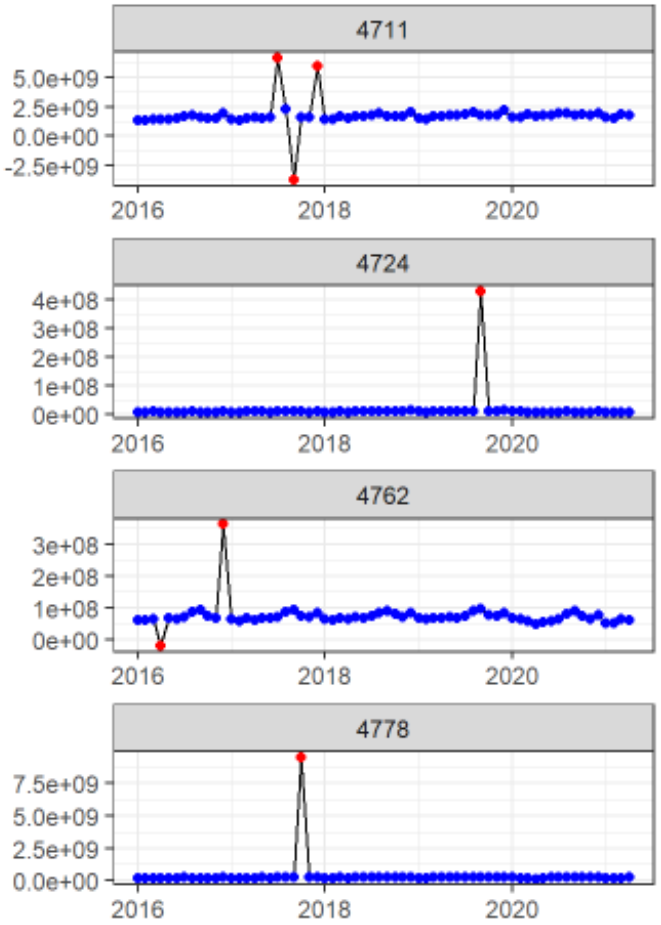
Apply those to each time series and comparing the estimation with the data received on a second version:

- Root Mean Square Deviation (RMSD)
- Mean Absolute Percentage Deviation (MAPD) (better metric heterogeneous T.S.)

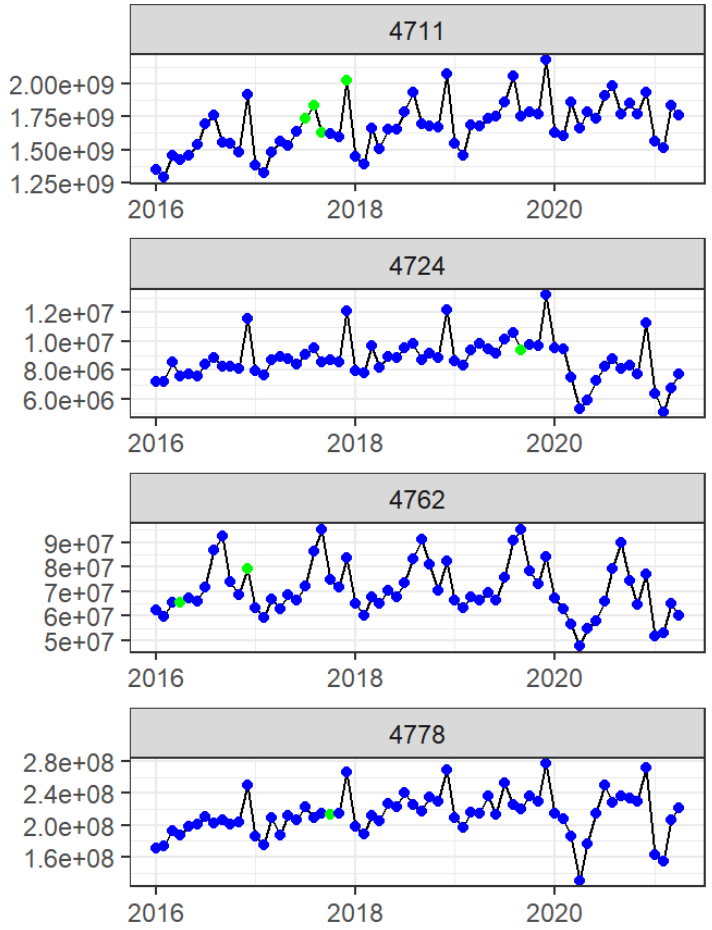
The conference is partly financed by the European Union

2. Data Quality: Challenges and Methodology

Original data



Imputed data



The conference is partly financed by the European Union

3. Collaborative efforts

- Using administrative data for Statistical purposes should **not be seen as a one-way** communication process
- In order to promote the use of data, it is important to know the **needs and expectations** of its users in the statistical production process
- To this end, **a close dialogue was/is promoted with data users** in order to consider and harmonize their needs in the adoption of a data treatment that is accepted by all
- To make it easier to analyse the data, a monthly 'Flash Report' is produced and delivered in 'R Flexdashboard' (soon using a 'shiny App')

The conference is partly financed by the European Union

4. Application to other cases

The methodologies and insights gained from managing e-invoice data can be effectively applied to other administrative data sources

- Adapting standardized procedures (data loading, normalisation, validation)
- Customizing anomaly detection (and imputation) algorithms
- Improving data integrity and consistency (cross-dataset comparisons and temporal consistency checks)
- Promotion of collaboration and engagement of other units

The conference is partly financed by the European Union

4. Application to other cases

Data sets being prepared for the application of these methodologies:

- Monthly pay statements (enhance the quality and reliability of employment and earnings data)
- International Trade, with identification of anomalies (speeding up correction work in data collection)
- Personal income tax data (IRS)
- Real estate transaction tax (IMT) and Property tax (IMI)
- …

Contributing to increase/improve the quality of the National Data Infrastructure

The conference is partly financed by the European Union

5. Conclusions

1. Strong collaboration between different Units of the statistical production process plays a very important role
2. Investment in acquiring new skills (Data Science), tools and techniques, in order to overcome the difficulties in processing a massive set of data

The conference is partly financed by the European Union

5. Conclusions

3. The e-invoice can be seen as a **a booster (PoC)** for the **treatment and use of administrative data** for statistical purposes:
   - Applies a set of procedures already in use in traditional sources (surveys)
   - Was recognized as the right way to apply for other data sources
   - Contributes to the construction and fulfilment of the objectives of the National Data Infrastructure
4. The work is not finished and is still evolving

Administrative Data Quality Challenges through the Lens of e-Invoice

António Portugal, Bruno Lima,
João Poças, Paula Cruz,
Salvador Gil, Sofia Rodrigues

Obrigado
Thank you

joao.pocas@ine.pt

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union