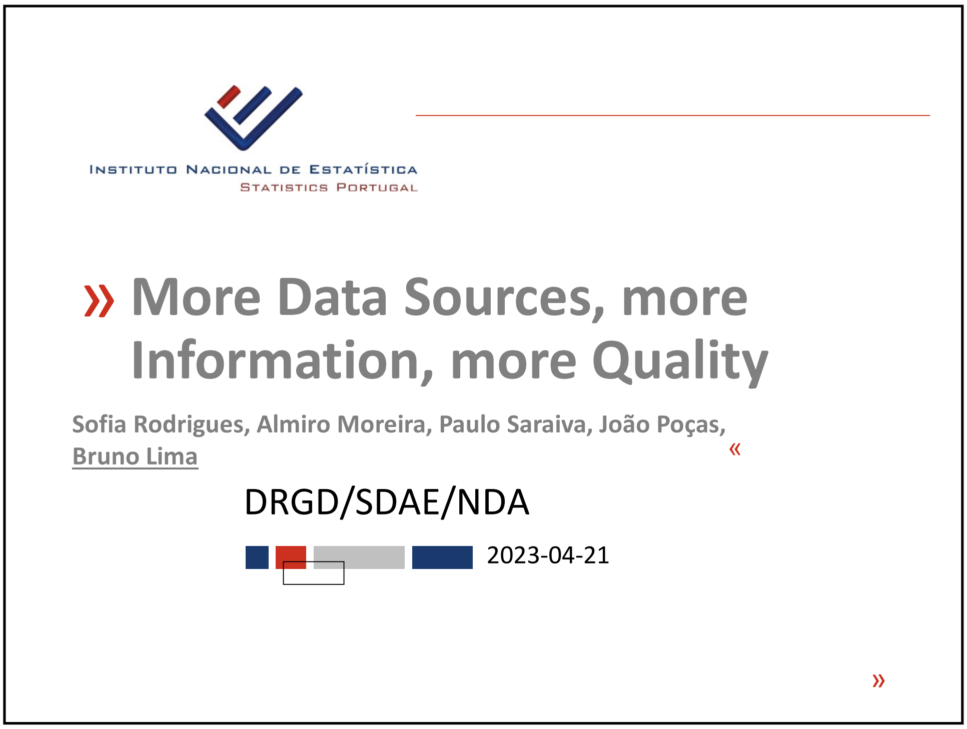


Instituto Nacional de Estatística
Statistics Portugal

» More Data Sources, more Information, more Quality

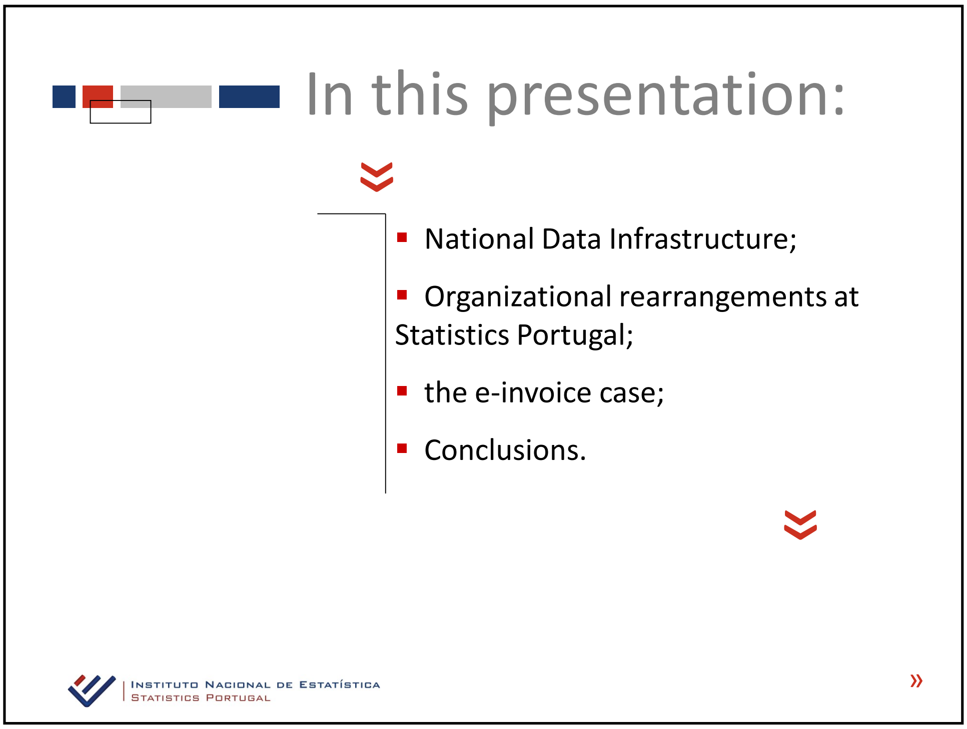
Sofia Rodrigues, Almiro Moreira, Paulo Saraiva, João Poças, Bruno Lima «

DRGD/SDAE/NDA

2023-04-21

»

1



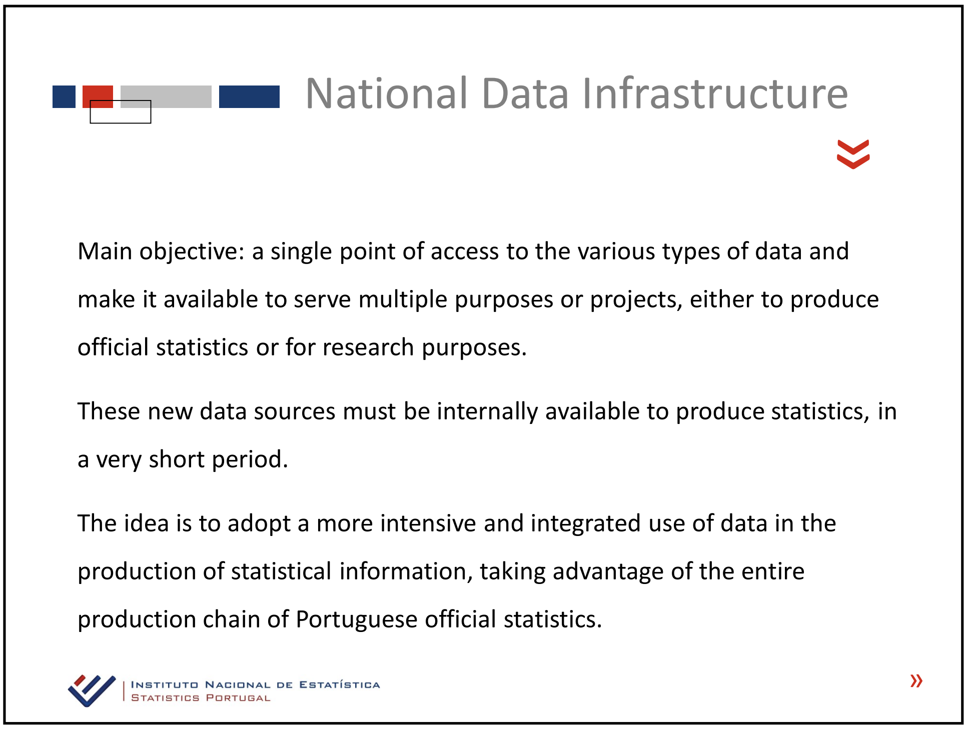
In this presentation:

- National Data Infrastructure;
- Organizational rearrangements at Statistics Portugal;
- the e-invoice case;
- Conclusions.

Instituto Nacional de Estatística
Statistics Portugal

»

2



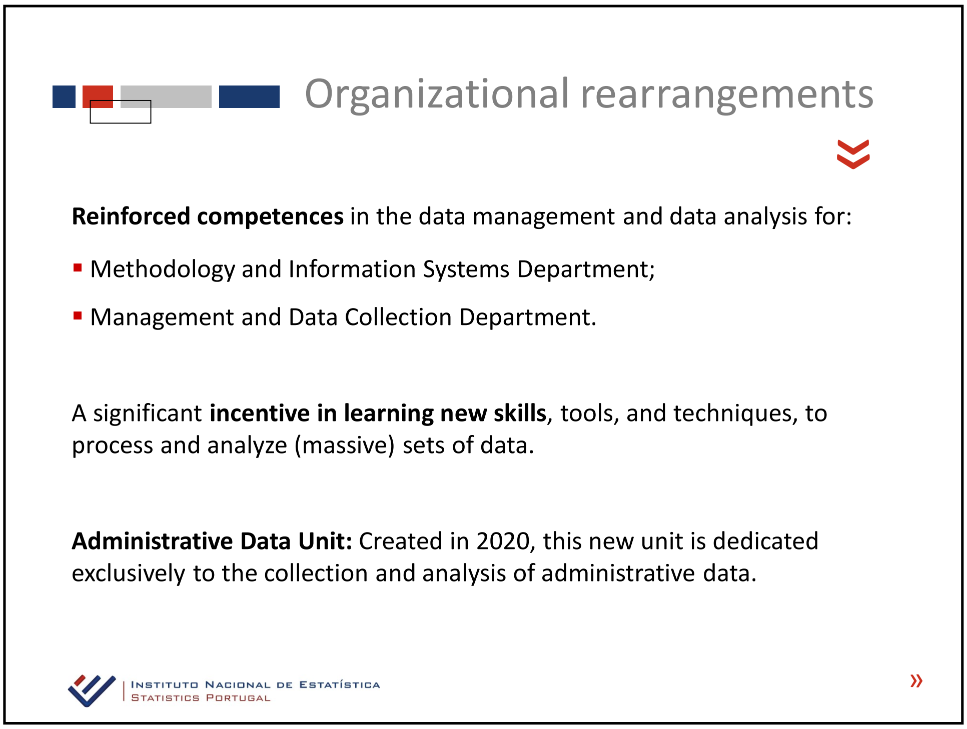
National Data Infrastructure

Main objective: a single point of access to the various types of data and make it available to serve multiple purposes or projects, either to produce official statistics or for research purposes.

These new data sources must be internally available to produce statistics, in a very short period.

The idea is to adopt a more intensive and integrated use of data in the production of statistical information, taking advantage of the entire production chain of Portuguese official statistics.

3



Organizational rearrangements

Reinforced competences in the data management and data analysis for:

- Methodology and Information Systems Department;
- Management and Data Collection Department.

A significant **incentive in learning new skills**, tools, and techniques, to process and analyze (massive) sets of data.

Administrative Data Unit: Created in 2020, this new unit is dedicated exclusively to the collection and analysis of administrative data.

4

The e-invoice case

Every month, Statistics Portugal receives data from the Tax Authority on a mandatory e-invoice declaration system.

More than 80 million records regarding taxable amounts aggregated by issuer and acquirer’s VAT number.

We ‘just’ have to process, validate, integrate, analyze, treat and make these data available for internal users.

The e-invoice case (cont.)

Data completeness:

Validation of data structure;

Encryption of personal identifiers;

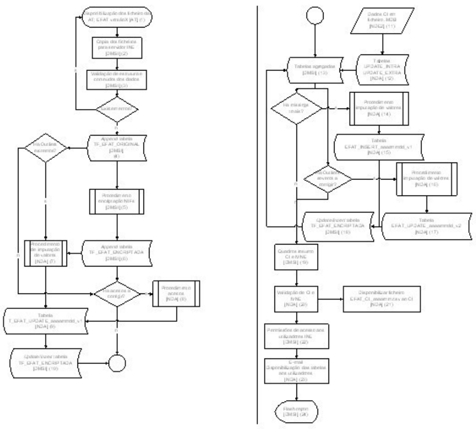
Normalization of attributes (country codes);

New attributes from other sources;

Elimination of outliers;

Imputation of missing values;

Consistency tests and comparison with other datasets.



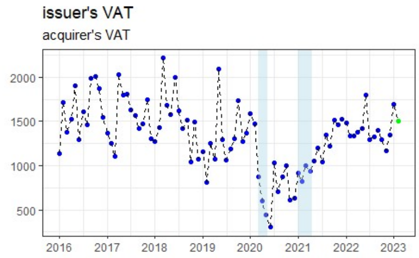
The e-invoice case (cont.)

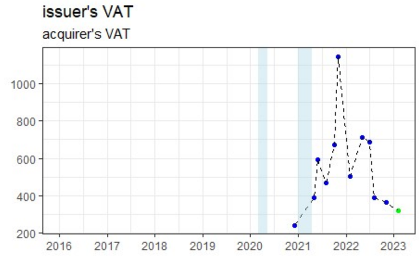
Data completeness:

Top 10 (positive and negative) values are used to identify issuer/acquirer time series;

Reported values with more then 3 SD from last year issuer mean are also identified;

Imputation of taxable amount is done with Kalman Smoothing in structural time series models.





»

The e-invoice case (cont.)

Most relevant issuers

More than 2000 of the most economically relevant issuers are monthly analyzed looking for any missing data.

To classify an issuer as relevant, we compile their historical reported values and contributions from Statistics Portugal's subject matter units.

This analysis requires a great deal of effort in its processing, especially as it requires the analysis of each time series. We have set up an automated procedure to identify and fill in any potential missing data.

»

The e-invoice case (cont.)

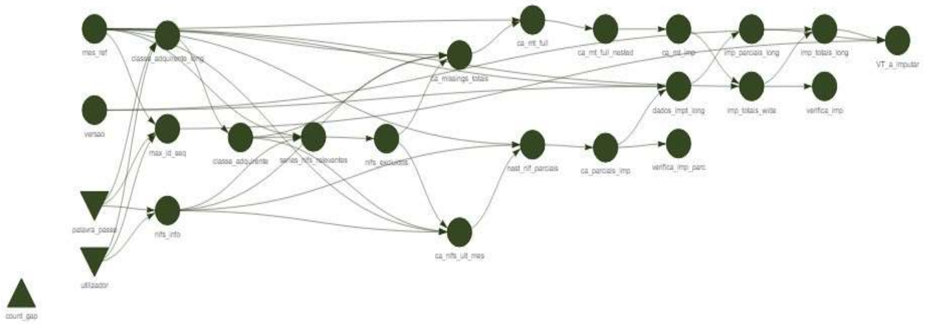
Missing data

» **total missing**, when no value is reported by the issuer;

» **partial missing**, when the reported value and the number of buyers are much lower than expected. In this case, an isolation forest algorithm is applied to detect partial missing (anomaly detection) for the taxable amount and for the number of records (buyers), aggregated by issuer.

The e-invoice case (cont.)

The workflow



The e-invoice case (cont.)

nowcasting methods used:

Kalman-Smoothing

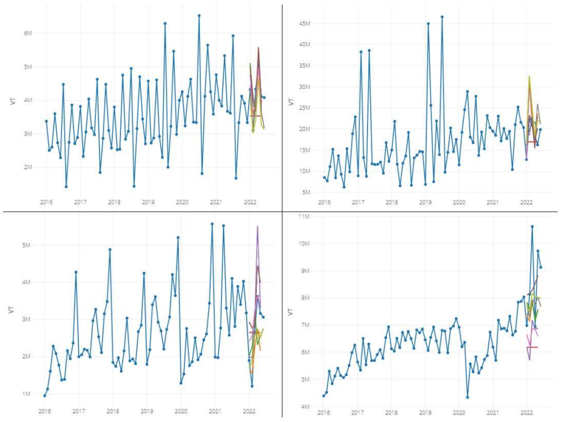
auto Arima

Exponential smoothing (ETS)

JDEMETRA

prophet

ensemble



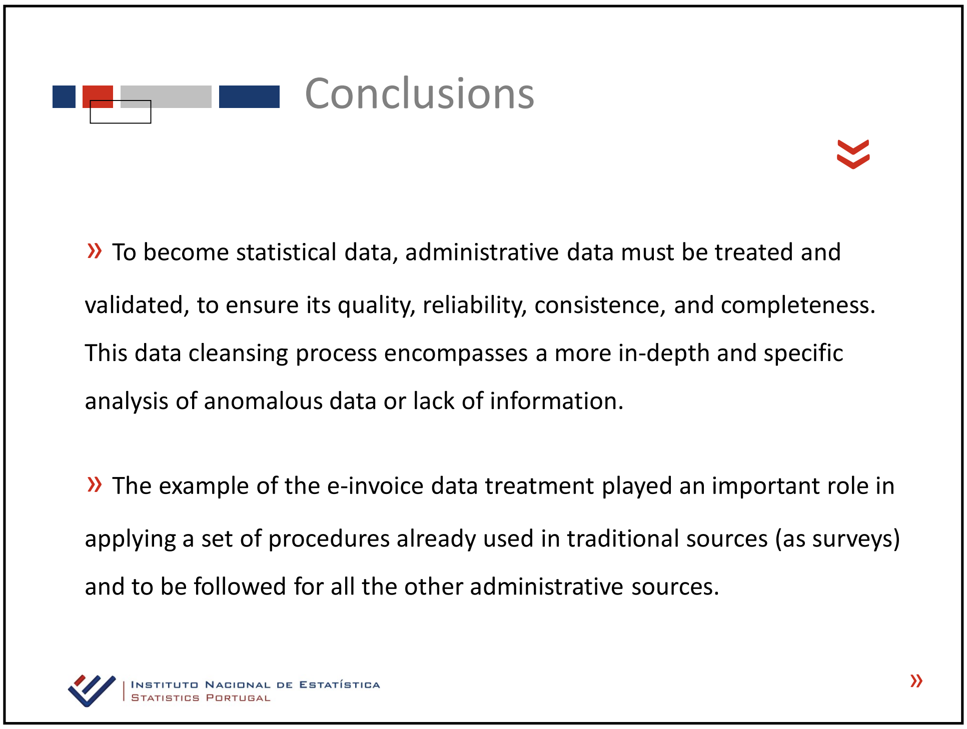
Next steps…

» Deepen the identification of anomalies and their treatment in the context of the e-invoice;

» Replicate the procedures for improving the quality of e-invoice data to other data sources;

» Continue to expand the use of quality administrative data in the support and / or replacement of traditional data collection;

» Increase the sets of processed data available for researchers.



Conclusions

» To become statistical data, administrative data must be treated and validated, to ensure its quality, reliability, consistence, and completeness. This data cleansing process encompasses a more in-depth and specific analysis of anomalous data or lack of information.

» The example of the e-invoice data treatment played an important role in applying a set of procedures already used in traditional sources (as surveys) and to be followed for all the other administrative sources.