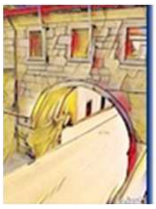


JOCLAD 2021

UNIVERSIDADE BEIRA INTERIOR

Data analysis during Covid time
the e-invoice case

Almiro Moreira, almiro.moreira@ine.pt;
António Portugal, antonio.portugal@ine.pt;
Bruno Lima, bruno.lima@ine.pt;
João Poças, joao.pocas@ine.pt;
Jorge Magalhães, jorge.magalhaes@ine.pt;
Paula Cruz, paula.cruz@ine.pt;
Salvador Gil, salvador.gil@ine.pt;
Sofia Rodrigues, sofia.rodrigues@ine.pt;

CLAD Associação Portuguesa de Classificação e Análise de Dados

Instituto Nacional de Estatística
Statistics Portugal

Instituto Nacional de Estatística
Statistics Portugal

Agenda

1. The Covid impact on response rates
2. Data collection during Covid: threat or opportunity?
3. Analysis and data treatment: Improving quality
4. Conclusions (learnings)

1. The Covid impact on response rates

- Response rates on **monthly** surveys
  (similar behaviour in the annual surveys)

Monthly Surveys	Response rates (%)				
	2019	2020			
		March	April	May	June
INTRASTAT	**80,5**	73,1	75,4	75,5	77,9
Qualitative - Trade	**93,5**	89,7	85,2	80,4	87,4
Qualitative - Industry	**92,5**	88,2	81,1	75,3	83,4
Qualitative - Services	**92,5**	89,5	83,5	79,3	86,0
Short-Term business Statistics - Trade	**79,0**	77,0	72,0	73,0	77,0
Short-Term business Statistics - Industry	**84,0**	80,0	80,0	81,0	82,0
Short-Term business Statistics - Services	**85,0**	83,0	82,0	82,0	83,0
Index Prices on products	**88,0**	77,0	81,0	78,0	81,0

1. The Covid impact on response rates

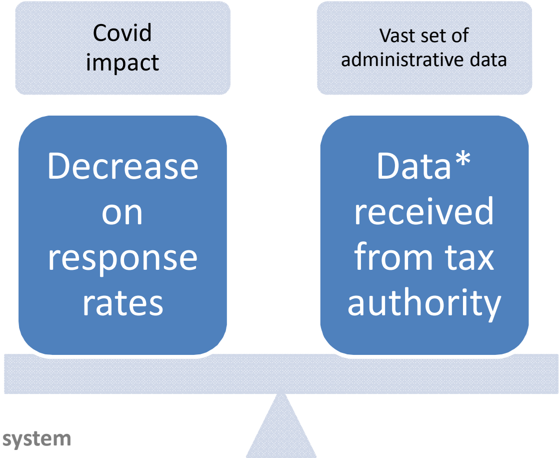
- It was not a surprise but simply **a fact we have to deal with**:

  The COVID-19 pandemic decreased the response rates to business surveys, particularly during the second quarter of 2020.

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

2. Data collection during Covid: threat or opportunity?

Two situations on the beginning of 2020:



Covid impact

Vast set of administrative data

Decrease on response rates

Data* received from tax authority

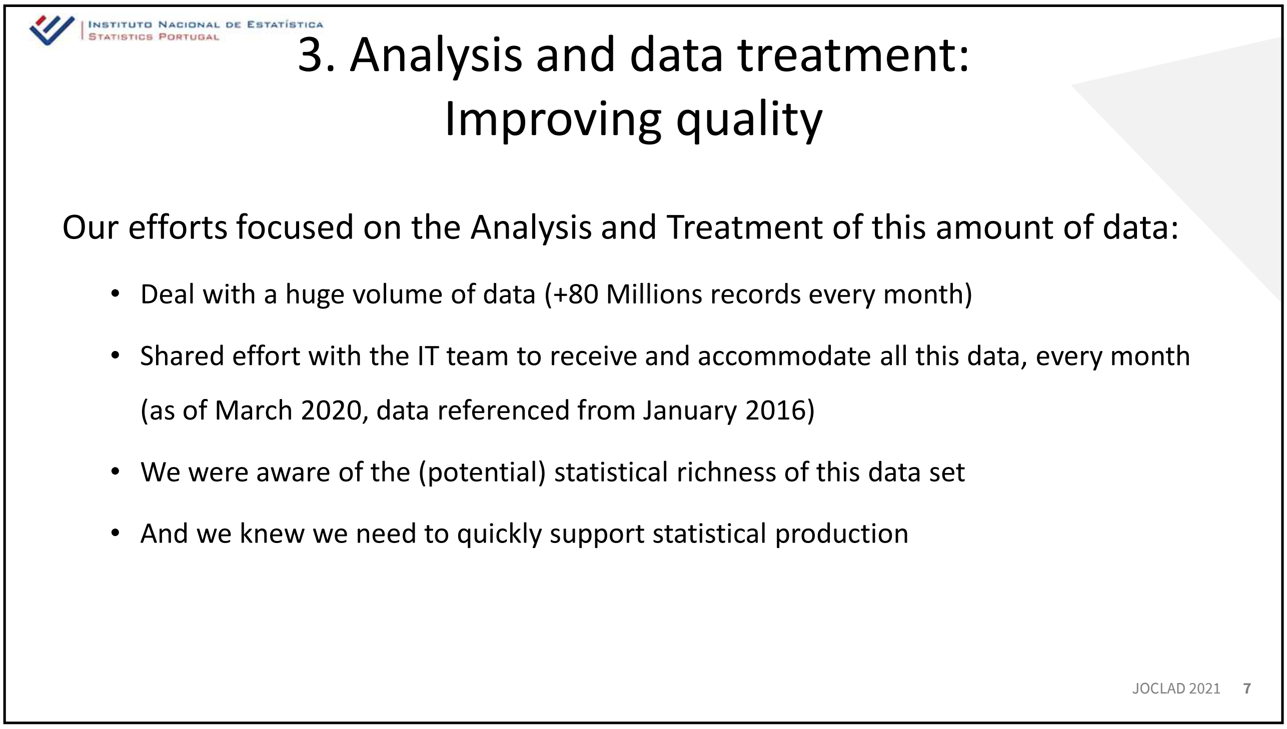
* Data from the e-invoice system

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

2. Data collection during Covid: threat or opportunity?

- We had two options:
  (1) cross our fingers and hope everything was going to be back to normal or…
  (2) **use administrative data** more intensively!

- The decision was obvious…
  – In this context, the information from the e-invoice system became more relevant, filling in the missing answers to the STS and contributing to the consistency of the results obtained in the production of statistical indicators



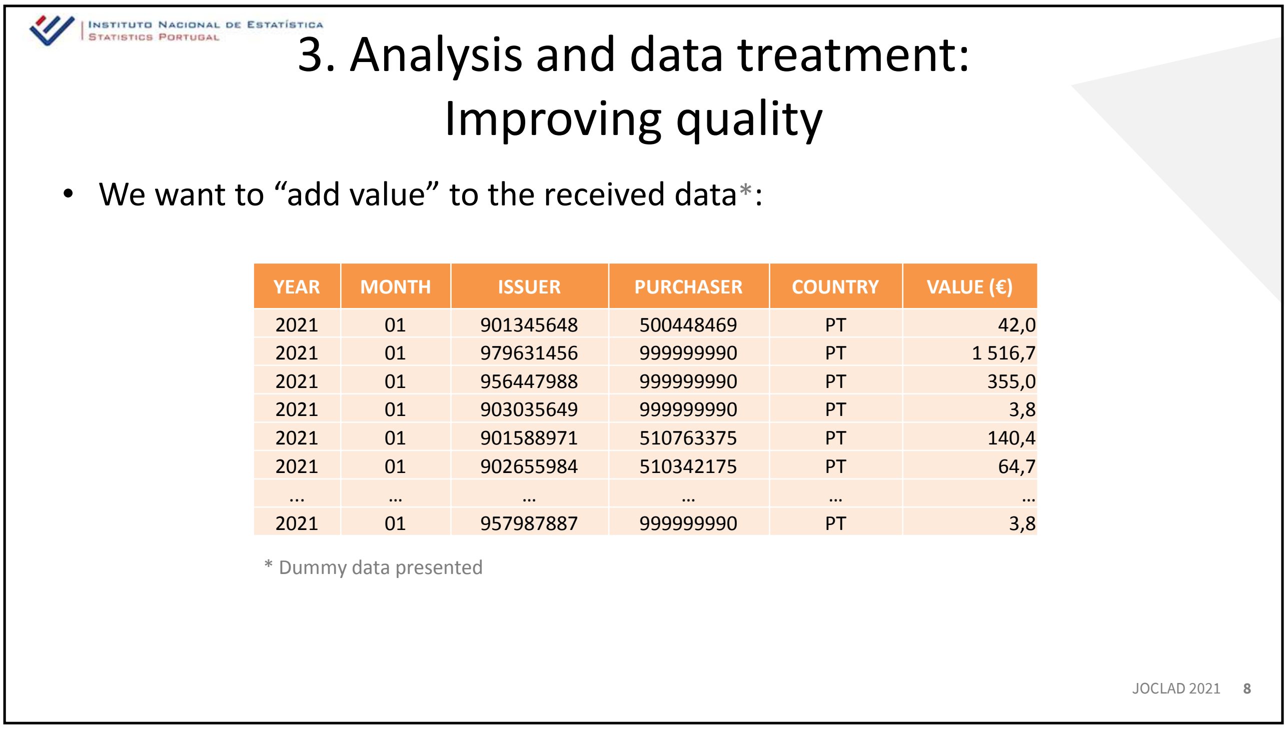
Instituto Nacional de Estatística
Statistics Portugal

3. Analysis and data treatment: Improving quality

Our efforts focused on the Analysis and Treatment of this amount of data:

- Deal with a huge volume of data (+80 Millions records every month)
- Shared effort with the IT team to receive and accommodate all this data, every month (as of March 2020, data referenced from January 2016)
- We were aware of the (potential) statistical richness of this data set
- And we knew we need to quickly support statistical production

JOCLAD 2021 7



Instituto Nacional de Estatística
Statistics Portugal

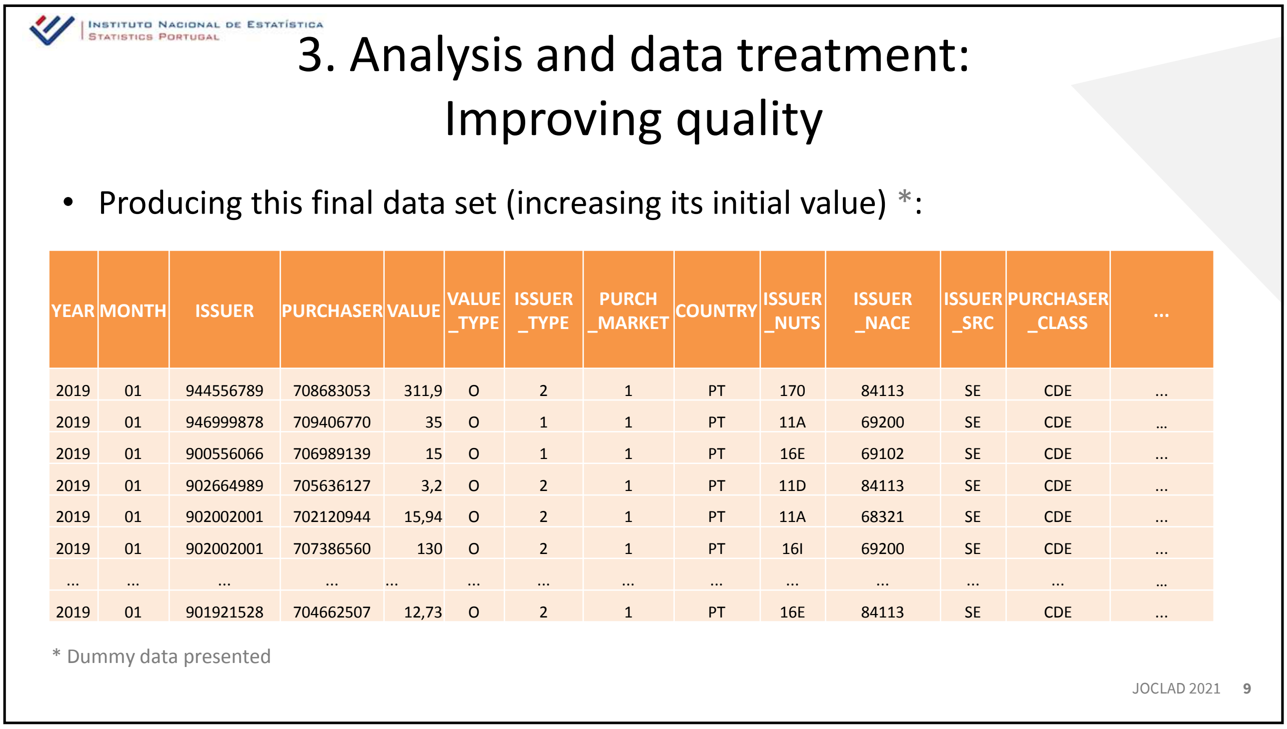
3. Analysis and data treatment: Improving quality

- We want to "add value" to the received data*:

YEAR	MONTH	ISSUER	PURCHASER	COUNTRY	VALUE (€)
2021	01	901345648	500448469	PT	42,0
2021	01	979631456	999999990	PT	1 516,7
2021	01	956447988	999999990	PT	355,0
2021	01	903035649	999999990	PT	3,8
2021	01	901588971	510763375	PT	140,4
2021	01	902655984	510342175	PT	64,7
…	…	…	…	…	…
2021	01	957987887	999999990	PT	3,8

* Dummy data presented

JOCLAD 2021 8

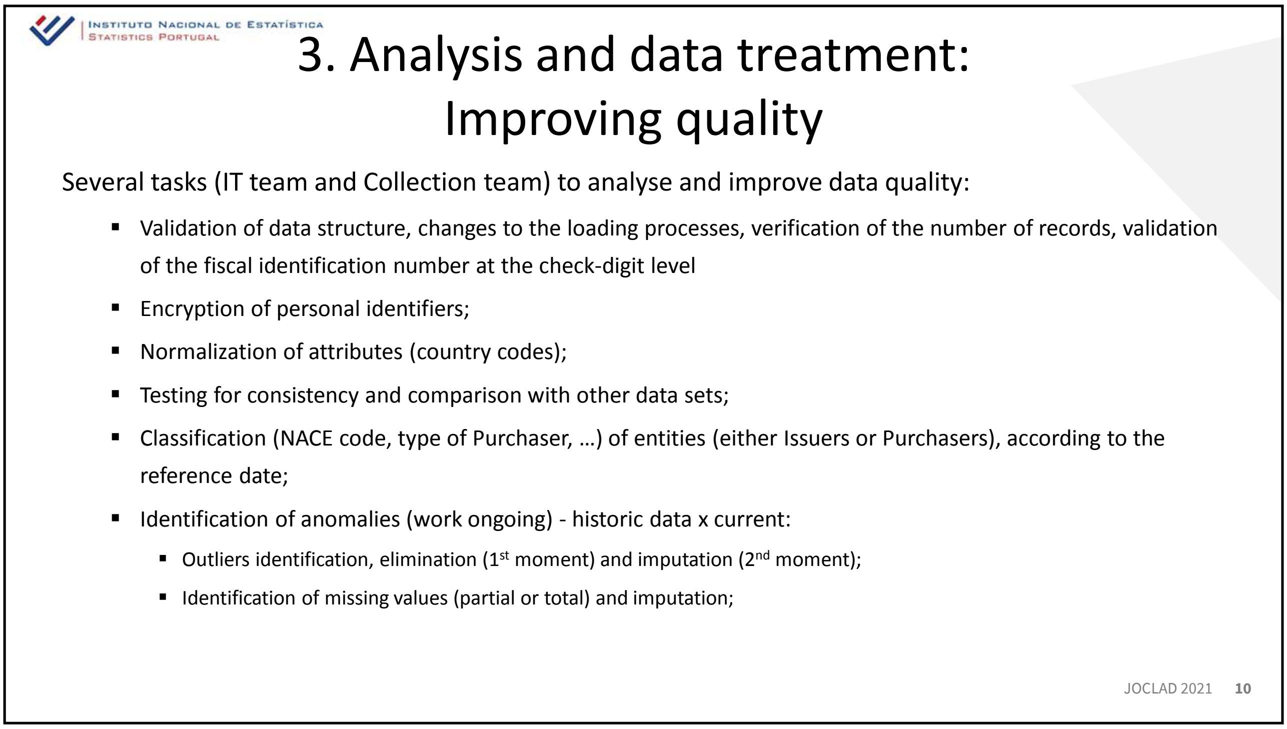


3. Analysis and data treatment: Improving quality

- Producing this final data set (increasing its initial value) *:

YEAR	MONTH	ISSUER	PURCHASER	VALUE	VALUE_TYPE	ISSUER_TYPE	PURCH_MARKET	COUNTRY	ISSUER_NUTS	ISSUER_NACE	ISSUER_SRC	PURCHASER_CLASS	…
2019	01	944556789	708683053	311,9	O	2	1	PT	170	84113	SE	CDE	…
2019	01	946999878	709406770	35	O	1	1	PT	11A	69200	SE	CDE	…
2019	01	900556066	706989139	15	O	1	1	PT	16E	69102	SE	CDE	…
2019	01	902664989	705636127	3,2	O	2	1	PT	11D	84113	SE	CDE	…
2019	01	902002001	702120944	15,94	O	2	1	PT	11A	68321	SE	CDE	…
2019	01	902002001	707386560	130	O	2	1	PT	16I	69200	SE	CDE	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…
2019	01	901921528	704662507	12,73	O	2	1	PT	16E	84113	SE	CDE	…

* Dummy data presented



3. Analysis and data treatment: Improving quality

Several tasks (IT team and Collection team) to analyse and improve data quality:

- Validation of data structure, changes to the loading processes, verification of the number of records, validation of the fiscal identification number at the check-digit level
- Encryption of personal identifiers;
- Normalization of attributes (country codes);
- Testing for consistency and comparison with other data sets;
- Classification (NACE code, type of Purchaser, …) of entities (either Issuers or Purchasers), according to the reference date;
- Identification of anomalies (work ongoing) - historic data x current:
    - Outliers identification, elimination (1st moment) and imputation (2nd moment);
    - Identification of missing values (partial or total) and imputation;

3. Analysis and data treatment: Improving quality

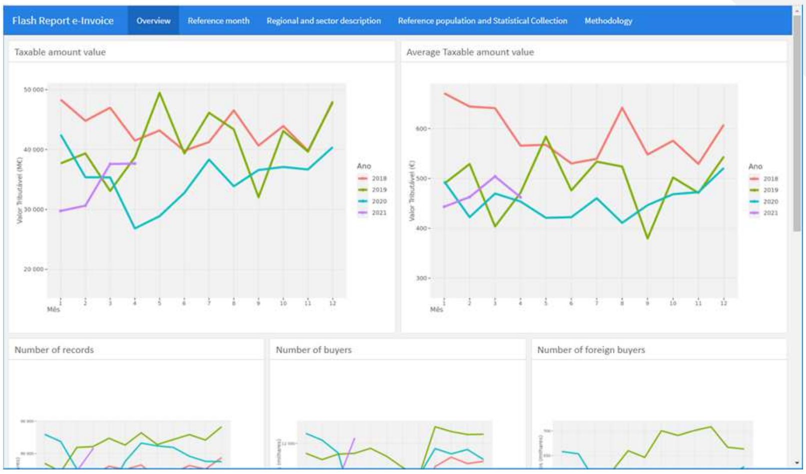
Importance of communicating with statistical users

- Handling 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 recipients in the statistical production process
- **A close dialogue with data users** was promoted in order to consider and harmonize their needs in the adoption of a data treatment that would be accepted by all
- To facilitate data analysis and exploration a monthly Flash report was developed in "R Flexdashboard"

3. Analysis and data treatment: Improving quality

Importance of communicating with statistical users

- Example of a Flash-report about e-invoice data produced and shared every month



3. Analysis and data treatment: Improving quality

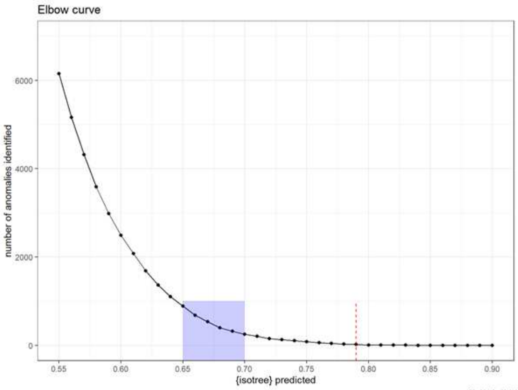
Outliers analysis (work in progress)

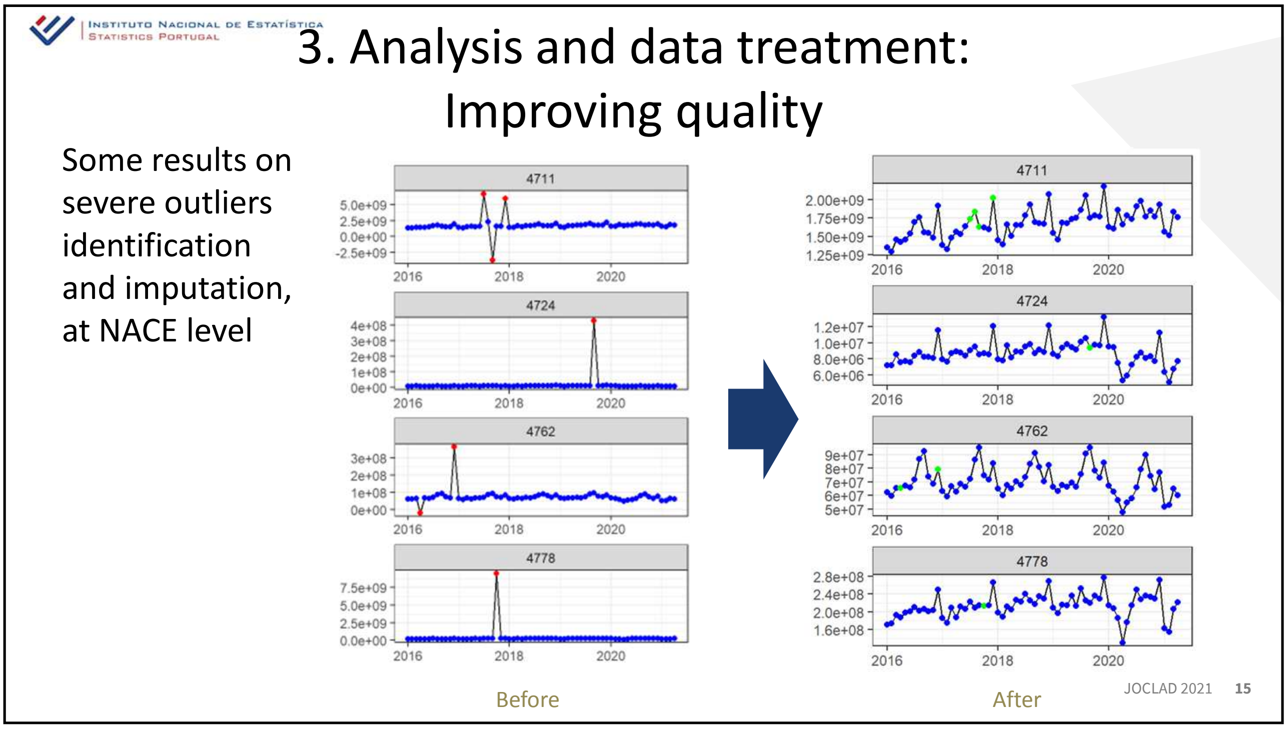
- Built 845 time-series (NACE level) for each issuer, grouping data by NACE and summing monthly total values.
- To each of these time-series applied the isolation forest algorithm (univariate analysis), in order to compute the probability of an observation being anomalous.
- Then, iteratively and using also Isolation Forest algorithm, search for anomalies at the issuer/buyer pair (pair level)
- Impute values when anomalies at pair level agree (on the same observation period) with those from NACE level (Kalman – Smoothing algorithm)

3. Analysis and data treatment: Improving quality

Outliers analysis (work in progress)

- To define a cut-off value for the probability of anomaly (isolation forest score), we use an elbow curve with the number of anomalies detected for different scores within each NACE.
- for the most severe anomalies we chose observations with a score > 0.79 at level NACE and score > 0.7 at level pair issuer / acquirer (buyer)





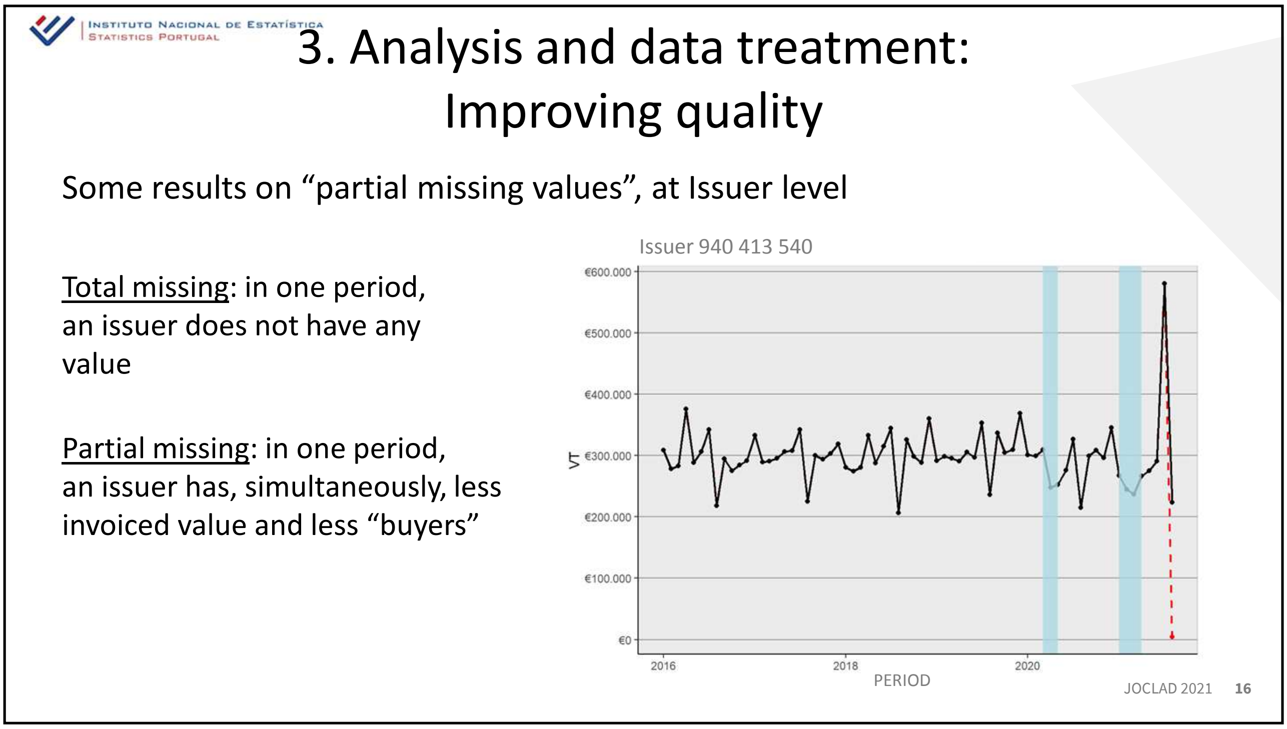
Instituto Nacional de Estatística
Statistics Portugal

3. Analysis and data treatment: Improving quality

Some results on severe outliers identification and imputation, at NACE level

Before

After



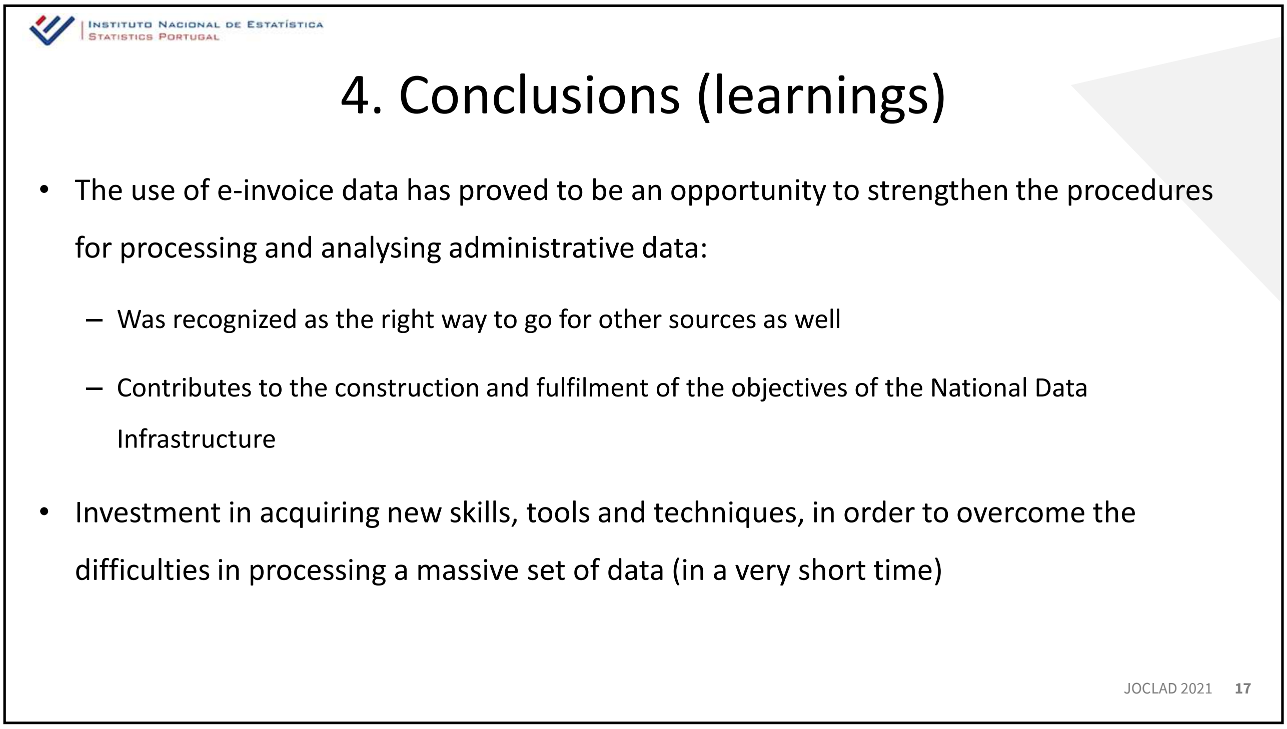
Instituto Nacional de Estatística
Statistics Portugal

3. Analysis and data treatment: Improving quality

Some results on “partial missing values”, at Issuer level

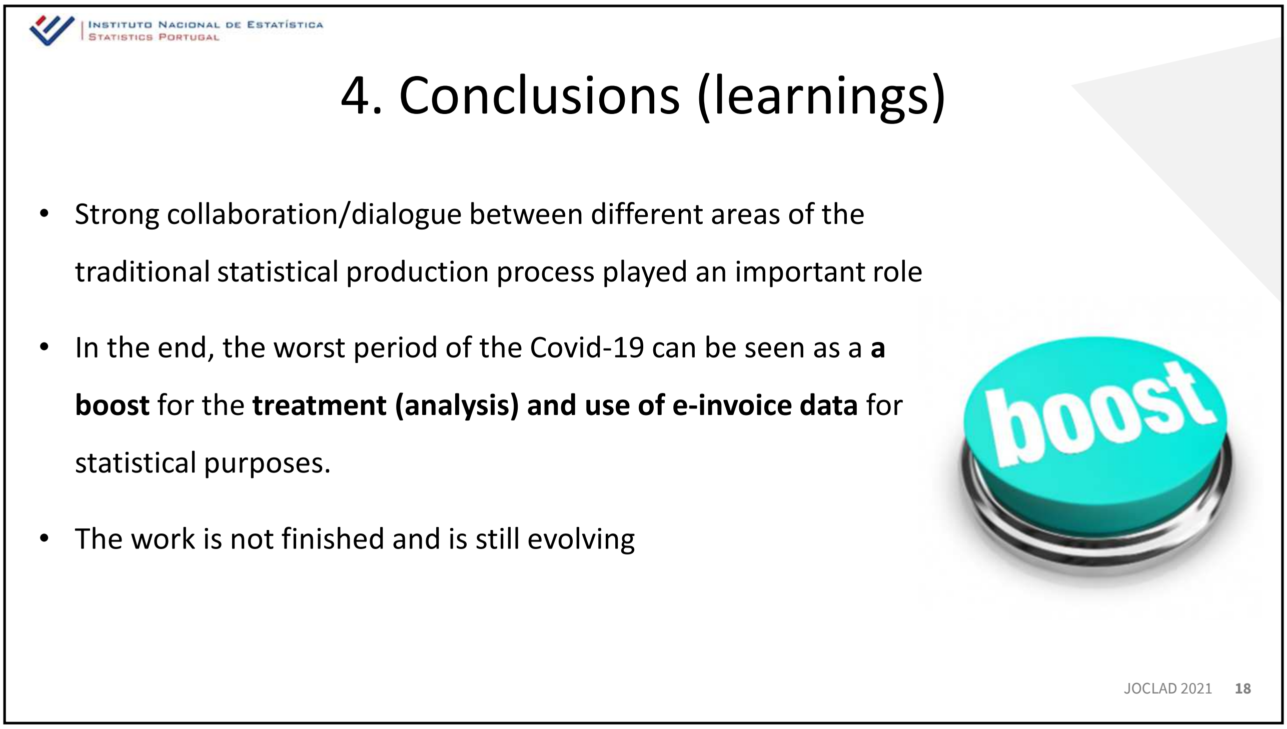
Total missing: in one period, an issuer does not have any value

Partial missing: in one period, an issuer has, simultaneously, less invoiced value and less “buyers”



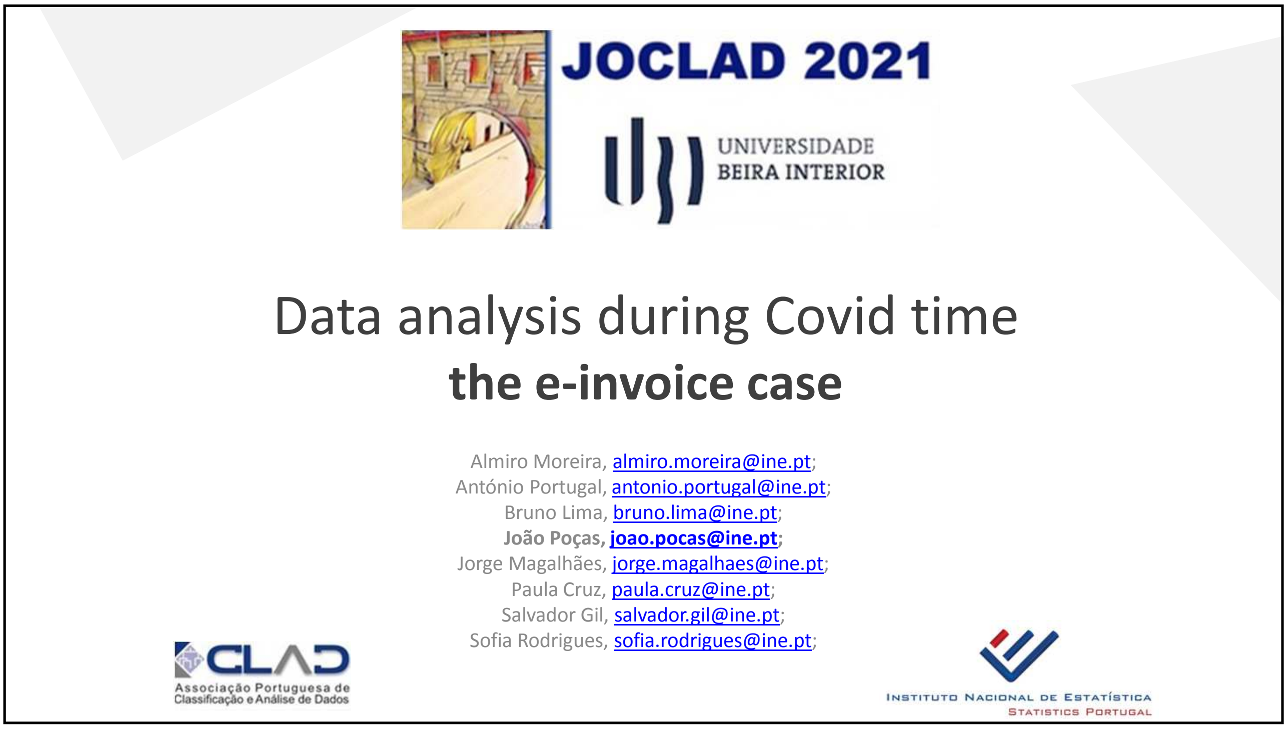
4. Conclusions (learnings)

- The use of e-invoice data has proved to be an opportunity to strengthen the procedures for processing and analysing administrative data:
  - Was recognized as the right way to go for other sources as well
  - Contributes to the construction and fulfilment of the objectives of the National Data Infrastructure
- Investment in acquiring new skills, tools and techniques, in order to overcome the difficulties in processing a massive set of data (in a very short time)



4. Conclusions (learnings)

- Strong collaboration/dialogue between different areas of the traditional statistical production process played an important role
- In the end, the worst period of the Covid-19 can be seen as a **a boost** for the **treatment (analysis) and use of e-invoice data** for statistical purposes.
- The work is not finished and is still evolving



JOCLAD 2021

UNIVERSIDADE BEIRA INTERIOR

Data analysis during Covid time
the e-invoice case

Almiro Moreira, almiro.moreira@ine.pt;
António Portugal, antonio.portugal@ine.pt;
Bruno Lima, bruno.lima@ine.pt;
João Poças, joao.pocas@ine.pt;
Jorge Magalhães, jorge.magalhaes@ine.pt;
Paula Cruz, paula.cruz@ine.pt;
Salvador Gil, salvador.gil@ine.pt;
Sofia Rodrigues, sofia.rodrigues@ine.pt;

CLAD Associação Portuguesa de Classificação e Análise de Dados

Instituto Nacional de Estatística
Statistics Portugal