CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Collection Session 4

27-30 September 2021, Online 16 September 2021

The use of electronic invoice data in COVID time

Almiro Moreira, António Portugal, Bruno Lima, João Poças, Jorge Magalhães, Paula Cruz, Salvador Gil, Sofia Rodrigues (Statistics Portugal)

almiro.moreira@ine.pt; antonio.portugal@ine.pt; bruno.lima@ine.pt; joao.pocas@ine.pt; jorge.magalhaes@ine.pt; paula.cruz@ine.pt; salvador.gil@ine.pt; sofia.rodrigues@ine.pt;

Abstract

One of the main impacts of the COVID-19 pandemic was the significant decrease in response rates to business surveys, particularly during the second quarter of 2020. The monthly surveys were the most affected, with response rates dropping close to 10% in the collection carried out during April and May.In this context, information from the e-Invoice system became even more relevant, particularly to fill in missing responses to the short term statistics, and contribute to the consistency of the results obtained in the production of statistical indicators. The e-Invoice can be defined as a mandatory system for reporting invoices or receipts implemented by the Tax Administration as part of the administrative simplification and anti-fraud measures. It is mandatory to transmit electronically to the Tax Authority data relating to invoices issued by individuals or legal entities that have their head office or permanent establishment in Portuguese territory. Under the cooperation established between the Tax Authority and Statistics Portugal through a protocol, the AT sends INE monthly anonymised information on the taxable amount aggregated by issuer and acquirer, month of invoicing and country of acquisition.

During the first week of April 2020, a very large set of data on company invoicing from January 2018 to February 2020 was received from the Tax Authority. The extensive volume of data (about 80 million records per month) and the urgency of the information in the shortest possible time, required a significant effort by the Information Systems and Data Collection and Analysis teams to make the data received (and treated) available to internal users. The data received was treated for completeness taking into account the expected structure and primary validation. Standard processes were defined and developed, at the level of loading, pseudo-encryption of identifiers (when necessary), processing and availability of data that not only ensure data integrity, but mainly the consistency of the information to be used in different statistics.

The main tasks at the level of data processing, coherence and consistency analysis are as follows: a) Validation of received data structure, changes to the loading processes, verifications of the number of records, validations of the fiscal identification number at the check-digit level and pseudo-encryption of personal identifiers; b) Normalization of attributes; c) Elimination of outliers of very significant expression and identification of suspicious cases; d) Consistency tests and comparison with other data sets.

The definition of these processes also made it possible to streamline the entire process of manipulating, exploring, and analyzing the data received. In order to facilitate data analysis and exploration, a monthly Flash report was also created, developed in R Flexdashboard where several indicators are presented, which accompanies each internal data delivery.

The use of electronic invoice data in COVID time

Authors: 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; Statistics Portugal

1. The Covid impact on response rates

We cannot say it was not unexpected, but the COVID-19 pandemic has significantly decreased the response rates to business surveys, particularly during the second quarter of 2020. The monthly surveys were the most affected, with response rates dropping close to 10% in the collection carried out during April and May.

Monthly Surveys	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%

Table 1 - Effect of Covid-19 on response rates on monthly surveys

Annual Surveys	Response rates		
	2019	2020	Diff
Services provided to enterprises	88,7%	82,8%	-5,9
Trade	88,3%	79,3%	-9,0
ICT Survey	92,5%	84,0%	-8,5

Table 2 - Effect of Covid-19 on response rates of annual surveys

In this context, information from the e-invoice system became even more relevant, particularly to fill in missing responses to the STS, and contribute to the consistency of the results obtained in the production of statistical indicators.

2. Data collection during Covid: from threat to opportunity

The focus on the use of administrative data has been a constant throughout the years at Statistics Portugal (INE), aiming a significant impact on the reduction of the statistical burden, as well as the possibility of returning new statistical indicators to Society, more urgent and adequate to the decision-making needs. An important example of the use of administrative data at Statistics Portugal was the implementation of the Simplified Business information System, in 2006, that allowed, for instance, to achieve a complete coverage of the business population, the reduction of information availability and the elimination of one of the costliest surveys (the Annual Business Survey).

In 2019, the development of the National Data Infrastructure (NDI) at INE began. This project has as its main goal the adoption of 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. This chain ensures the protection and integrity of data, from the development of platforms, applications and algorithms, data collection and validation, to the analysis of statistical information. In practical terms, one of its main objectives is to create a single point of access to the various types of data and to make them available in order to serve multiple purposes or projects, whether for the production of official statistics or for scientific research purposes.

The beginning of 2020 brought, among several others, three novelties to INE's statistical production: the impact of the COVID-19 pandemic on the response rates to business surveys; the receipt of a huge set of data from the invoicing of companies and; the creation of a new unit dedicated exclusively to the collection and analysis of administrative data.

In this pandemic situation, Statistics Portugal found itself with a decrease in the response rates to business surveys and in the need to quickly be able to use a vast set of data that could suppress the lack of response and thus continue to support statistical production. It was time to process, validate, integrate, analyse, and treat, the data received from the Portuguese Tax Authority: the e-invoice system.

This was a big challenge because there was no previous experience in analysing and processing this huge amount of data and to make it useful and readily available for statistical production.

3. Improving data quality and increasing its value

During the first week of April 2020, a very large set of data on company invoicing from January 2018 to February 2020 was received from the Tax Authority. The extensive volume of data (about 80 million records per month) and the urgency of the information in the shortest possible time, required a significant effort by the Information Systems and Data Collection and Analysis teams to make the data received available to internal users.

The data received were treated for completeness considering the expected structure and primary validation. Standard processes were defined and developed, at the level of loading, pseudo-encryption of identifiers (when necessary), processing and availability of data that not only ensured data integrity, but mainly the consistency of the information to be used in different statistics.

The main tasks at the level of data processing, coherence and consistency analysis are the following:

a. 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
b. Encryption of personal identifiers;
c. Normalization of attributes (country codes);
d. Elimination of outliers of very significant expression and identification of suspicious cases;
e. Consistency tests and comparison with other datasets.

The definition of these tasks also made it possible to streamline the entire process of manipulating, exploring and analyzing the incoming data, serving as a case-study for future datasets.

In addition, further progress was also made in terms of integration with other data sources, thus allowing the statistical value of the initial dataset to be increased. For example, data from the Statistical Business Register and from the Resident Population Register were integrated, thus making it possible to describe both the type of entities issuing invoices and the purchasers, especially those of national origin.

Thus, it was possible to incorporate additional information about the economic activity of the issuer, its region, which institutional sector it belongs to, among others. At the same time, it was possible to typify the purchasers in relation to their economic activity status or by integrating information about their geographic location.

The last, but no less important step in improving the quality of administrative data is related to identifying less obvious outliers and missing values.

To achieve this goal, we used machine learning methods, in particular, and in a first step, the "Isolation Forest" algorithm. This is an unsupervised machine learning algorithm that identifies anomalies by isolating outliers in the data. In this particular case, a top-down (from the most aggregated to the most disaggregated values) and iterative approach to identifying outliers was used, thus allowing for good robustness and consistency in the outliers identified. The process of identifying missing values followed a more traditional approach, also using other sources in order to find evidence of cessation of activity.

Data imputation was performed using the Kalman Smoothing Algorithm (R package imputeTS), based on the method proposed by the mathematician Rudolph Kalman in 1960, where a solution to the linear filtering problem for data generated over time is presented.

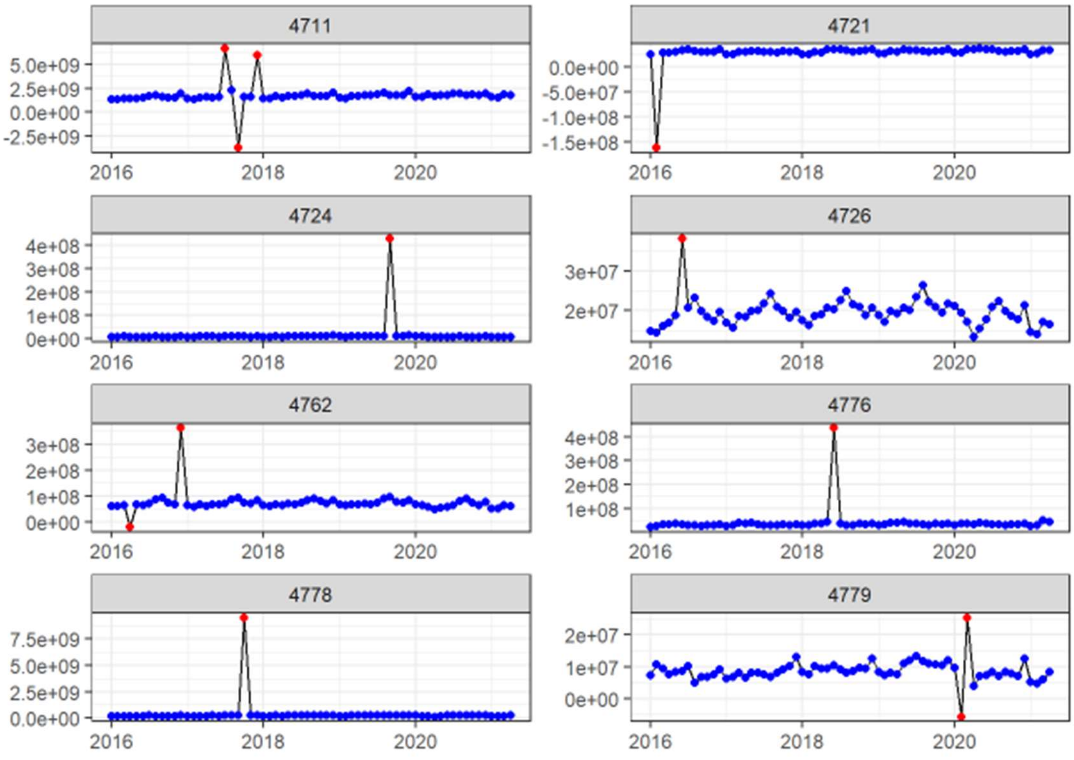


Figure 1 – Some activities with outliers detected under Trade activities (NACE 47)

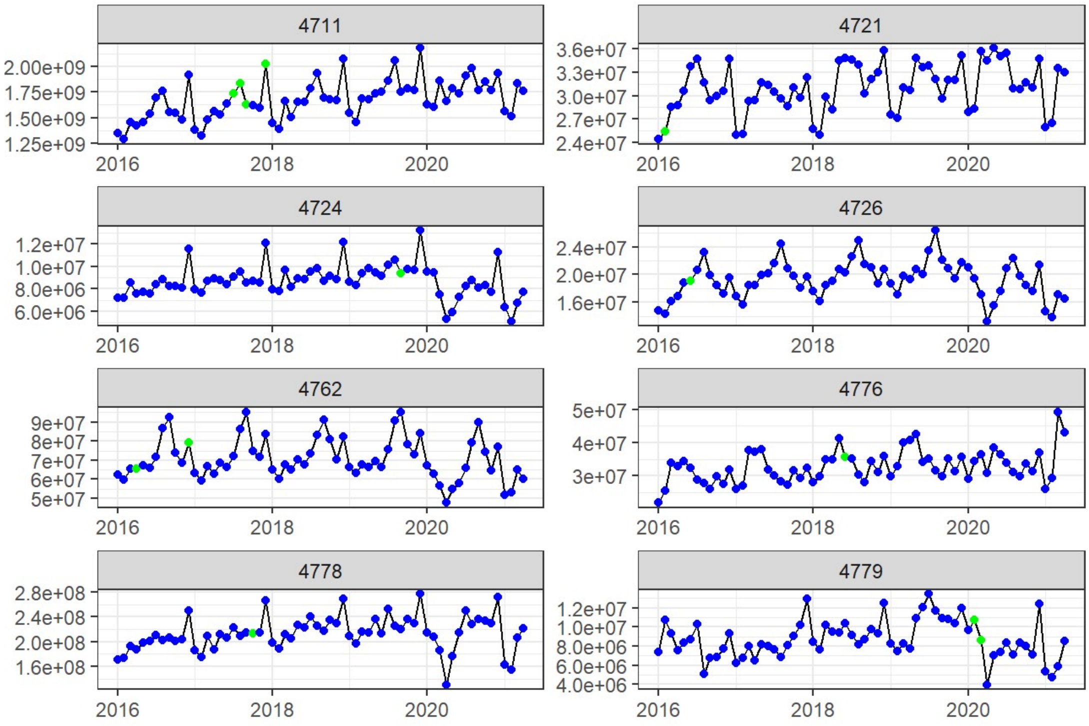


Figure 2 – Some activities with outliers imputed under Trade activities (NACE 47)

There is still a long way to go in this regard, but the work that has already been done allows us to face the future with a positive outlook.

4. Bi-directional communication: users' needs vs making data well-known

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. To this end, a close dialogue with data users was promoted in order to consider and harmonize their needs in the adoption of a data processing process that would be accepted by all.

In Statistics Portugal, for several years, flash reports are being developed for each data collection processes in all business surveys, monitoring the evolution of that process (namely, the total number of respondents), as well as the analysis of the main indicators collected, always bearing in mind the main objectives of each statistical operation – it is therefore a flexible tool, adapted to the specific needs of each project. In this manner, in a top-down approach, it is possible to identify the main trends right away and to direct the analysis of the data towards the variables identified as most relevant. It is also possible to breakdown information by region, by economic activity or by any other distribution that is considered relevant.

With a view to facilitate data analysis and exploration of administrative data, the production and use of flash reports (or similar reports such as dashboards) were extended to administrative sources. Following the same principles used in survey data collection, a key set of indicators are defined, both in terms of the analysis of information coverage (total number of respondents, main missing units, etc.) and in terms of the analysis of the indicators considered most relevant from the point of view of statistical production. In addition, information from other sources is included, namely information collected through surveys, which makes it possible to assess the degree of consistency. For example, in the Flash report created for this information – electronic invoicing -, indicators are included on the evolution of monthly and year-on-year invoicing; the distribution by activity groups and by consumer groups is presented and also the comparison of the evolution with the information collected in the monthly surveys in the scope of the STS.

To facilitate data analysis and exploration a monthly Flash report was developed in “R Flexdashboard”. This was the first time that Statistics Portugal used this tool to produce this kind of report and an important investment has been made in the learning process. These reports are made available in each internal publication of processed data and present several indicators that aim to provide a statistical picture about the quality of the received dataset in a more appealing and interactive way. The same template is now used for other administrative sources.

5. Conclusions

Despite the intensive use of administrative data that Statistics Portugal has made over recent years, the use of electronic invoice data has proved to be an opportunity to strengthen the procedures for processing and analyzing this type of data, in a more coherent, integrated, consistent, enriched, documented and widely used manner.

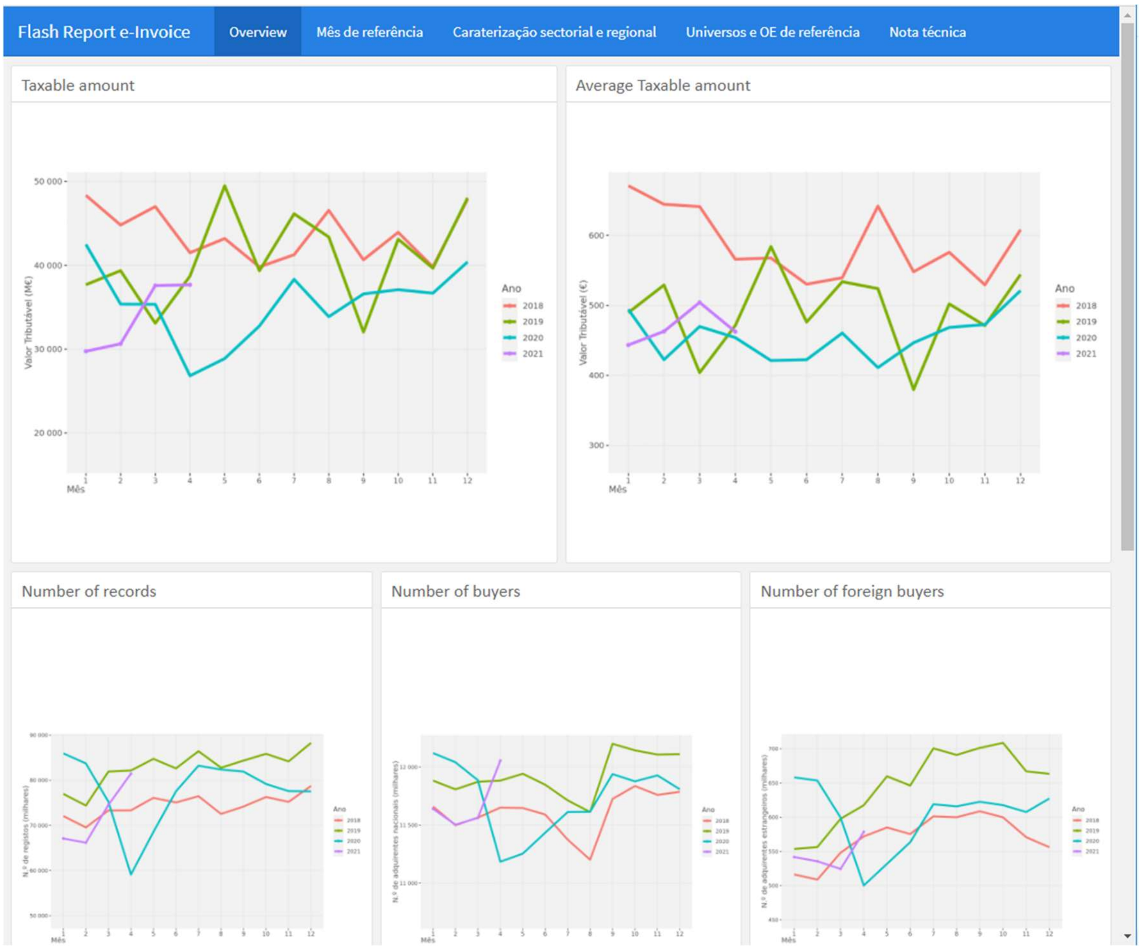
A significant investment was made in learning new skills, tools and techniques, in order to overcome the difficulties inherent in processing a massive set of data, to be made available in a very short space of time to internal users.

The strong collaboration between different areas of the traditional statistical production process also played a very important role.

It was recognized by the various types of users of this information that this way of working with the original data from this source would be the right way to go for other sources, thus avoiding duplication of effort and gaining advantages in the work of producing statistics. This experience, the availability of treated datasets, in a self-service model for users, proved to be an important contribution to the construction and fulfilment of the objectives of the National Data Infrastructure, a strategic project for Statistics Portugal

In the end, the worst period of the Covid-19 pandemic can be seen as a catalyst, as all these actions involving the processing, integration, availability and use of data in the statistical process took place in a shorter period of time than usual.

Annex – Example of the monthly Flash Report (“R FlexDashboard”) about e-invoice data


Flash Report e-Invoice
Overview
Mês de referência
Caraterização sectorial e regional
Universos e OE de referência
Nota técnica
Taxable amount
Valor Tributável (M€)
50 000
40 000
30 000
20 000
Mês
1
2
3
4
5
6
7
8
9
10
11
12
Ano
2018
2019
2020
2021
Average Taxable amount
Valor Tributável (€)
600
500
400
300
Number of records
N.º de registos (milhares)
90 000
80 000
70 000
60 000
50 000
Number of buyers
N.º de adquirentes nacionais (milhares)
12 000
11 500
11 000
Number of foreign buyers
N.º de adquirentes estrangeiros (milhares)
700
650
600
550
500
450