Administrative Data Quality Challenges through the Lens of e-Invoice

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

Statistics Portugal

Abstract

In Portugal, the Tax Administration has mandated electronic invoicing for reporting commercial transactions. Statistics Portugal receives this data monthly, aggregating taxable amounts per fiscal number for issuers and acquirers. However, challenges such as managing the huge amount of data, identifying outliers and missing values within a limited timeframe, and ensuring its integration into the statistical production process persist. This paper explores the strategies employed by Statistics Portugal to address these challenges, emphasizing collaboration and innovative approaches to enhance the quality of e-invoice data and its usability in statistical analysis. Furthermore, it discusses the potential for applying the experience and knowledge gained from e-invoice data processing to improve the quality of other administrative data sources.

Keywords: administrative data, e-invoice, large volume data, anomalies, quality improvement.

Introduction

In the modern digital era, the use and efficient management of administrative data has become crucial for statistical data production. Administrative data, encompassing various types of records collected by government agencies, plays a pivotal role in understanding and addressing socio-economic issues. Among these data sources, e-invoice systems have emerged as a significant innovation, providing detailed insights into commercial transactions and economic activities.

In Portugal, the Tax Administration has mandated the use of electronic invoicing (e-invoicing) for reporting all commercial transactions. This initiative requires businesses to submit their invoices electronically, enabling almost real-time monitoring and enhancing the accuracy of fiscal data. Each month, Statistics Portugal receives this huge amount of data, comprising aggregated taxable amounts per fiscal number for both issuers and acquirers. While the e-invoicing system offers substantial benefits in terms of data availability and granularity, it also presents significant challenges related to data quality.

The primary challenges associated with e-invoice data stem from its large volume and the need for timely processing. With around 100 million records received monthly, ensuring the quality, reliability, consistency, and completeness of this data within a short timeframe is a complex task. Incomplete or erroneous data, necessitates rigorous validation and imputation processes to maintain the data integrity and usability for statistical purposes.

This paper explores the methodologies and procedures developed to enhance the quality of e-invoice data and discusses how these approaches can be applied to other administrative data sources. By leveraging automated procedures and standardized checks, we aim to ensure the reproducibility and efficiency of data processing. This centralized and comprehensive approach not only improves the statistical quality of administrative data but also contributes to a more coherent and valuable National Data Infrastructure.

Through an examination of the e-invoice data processing workflow, this paper highlights the collaborative efforts required among various units within Statistics Portugal to manage and improve data quality. We also discuss the broader implications of these efforts, emphasizing how the lessons learned from the e-invoice system can be generalized to enhance other administrative datasets, ultimately contributing to a more coherent and valuable National Data Infrastructure. By improving data quality, these efforts not only reduce the statistical burden on data producers but also enhance the reliability of statistical outputs for informed policy-making and public administration.

Methodology

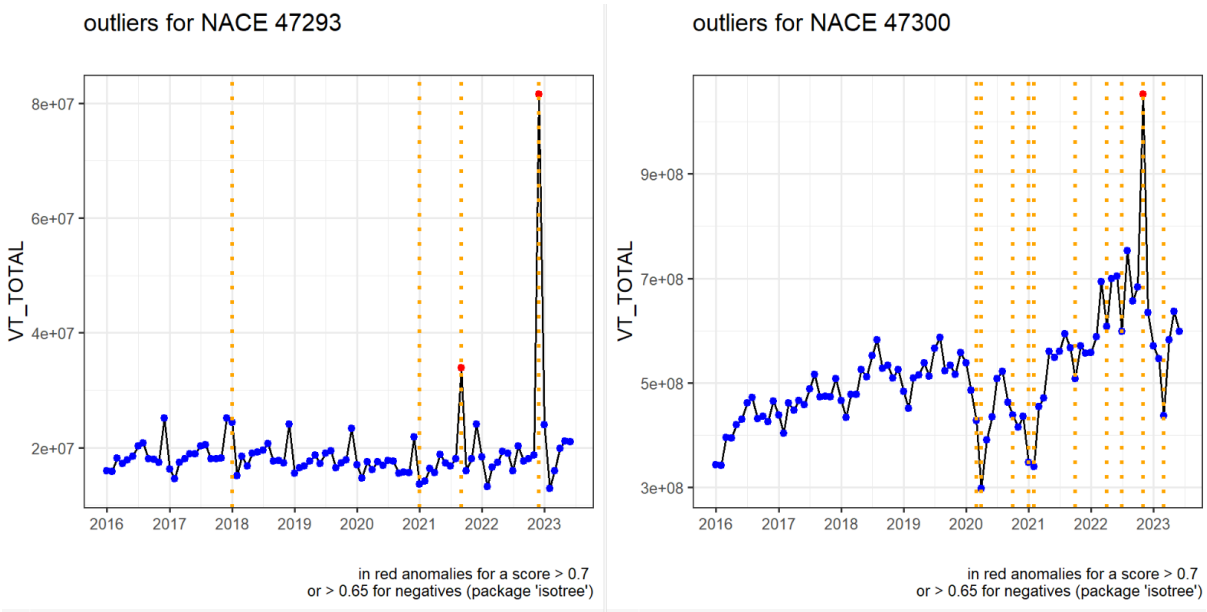
The methodology employed to improve the quality of e-invoice data in Portugal involves a systematic data processing workflow designed to handle large volume of records efficiently and accurately. The workflow begins at the data loading stage and includes several standardized procedures to verify and validate the data structure.

Initially, during data loading, a series of checks are performed to ensure the integrity of the dataset. This includes verifying the total number of records to ensure it aligns with expected volumes, validating fiscal identification numbers and country codification. Any significant discrepancies detected during these checks are flagged for further investigation.

Following the initial verification, key attributes such as transaction dates and amounts are normalized to a standard format, while identifiers such as fiscal numbers are encrypted to protect sensitive information while allowing for subsequent analysis without exposing sensitive information.

Anomaly detection algorithms are then employed to identify irregularities within the dataset, such as outliers, missing values, and inconsistencies. These algorithms utilize statistical methods and machine learning techniques to detect patterns deviating from the norm. Subsequently, imputation procedures

are applied to correct or fill in missing values detected during anomaly detection. This may involve techniques such as mean or median substitution, imputation based on historical data, or multiple imputation to ensure accuracy.



Integrity and consistency checks further ensure the reliability of the e-invoice data. The data is compared with other datasets, such as administrative records and survey data, to identify discrepancies and reinforce its reliability. Temporal consistency checks are also performed to ensure that the data remains consistent across different reporting periods, involving comparisons between current data and historical records to detect any unusual trends or anomalies.

Throughout the data processing workflow, collaborative coordination among different teams within Statistics Portugal is essential. A predefined workflow integrates all tasks, ensuring seamless collaboration and efficient resolution of issues.

By tailoring these methodologies to fit the specific requirements of different datasets, Statistics Portugal aims to enhance the quality of all administrative data sources, contributing to a more reliable and coherent National Data Infrastructure. This, in turn, supports better decision-making and policy formulation based on high-quality, consistent, and timely statistical data.

Application to Other Administrative Data Sources

The methodologies and insights gained from managing e-invoice data can be effectively applied to other administrative data sources, enhancing their quality and integration into the statistical production

process. The scalability and adaptability of the procedures developed for e-invoice data make them suitable for a wide range of administrative datasets, each with its unique characteristics and challenges.

- Adapting Standardized Procedures

The initial steps of data loading, normalization, and validation developed for e-invoice data can be tailored to fit the specific requirements of other administrative datasets. By adjusting the normalization and validation rules, the methodology ensures that each dataset is processed accurately according to its specific context.

- Customizing Anomaly Detection Algorithms

Anomaly detection algorithms, which are crucial for identifying outliers, missing values, and inconsistencies in e-invoice data, are being customized for other administrative datasets. These algorithms can be trained on dataset-specific characteristics to enhance their accuracy and effectiveness. By tailoring the detection algorithms to these patterns, the process of identifying and rectifying errors becomes more precise and efficient.

- Enhancing Data Integrity and Consistency

The procedures for ensuring data integrity and consistency, such as cross-dataset comparisons and temporal consistency checks, are universally applicable to various administrative data sources. For instance, the integrity of a dataset on employment records can be ensured by comparing it with data from monthly pay statements and social security databases. Similarly, temporal consistency checks can be applied to ensure that employment data remains consistent across different reporting periods, just as it is done for e-invoice data.

- Promoting Collaborative Coordination

The collaborative framework established for e-invoice data processing, which involves seamless coordination among various units within Statistics Portugal, can be extended to other administrative data domains. This approach promotes collaboration between different Units and ensures that data quality improvement processes are conducted in a unified and more efficient manner. By integrating tasks into a predefined workflow, the methodology ensures efficient and coordinated data processing across different administrative sources.

We are starting to apply these methodologies to monthly pay statements, which can significantly enhance the quality and reliability of employment and earnings data. This, in turn, improves the accuracy of Short-Term Statistics, providing valuable insights for economic policy and labor market analysis.

Conclusions

The application of the methodologies developed for e-invoice data to monthly pay statements and other administrative data sources is taking its first steps. However, the experience gained with the e-invoice case demonstrates the potential of standardised procedures and automated processes to improve the quality and usability of data in different administrative datasets.

Even at this preliminary stage, it is evident that the techniques used for e-invoice data processing are scalable and adaptable. The principles of data loading, normalization, validation, anomaly detection, and imputation have been effectively tailored to fit the specific characteristics of other datasets like, for instance, monthly pay statements. Early results show enhanced accuracy and completeness, with initial anomaly detection and imputation processes successfully addressing outliers and missing values. Cross-dataset comparisons and temporal consistency checks are starting to ensure data reliability, although further work is needed to fully realize these benefits.

Inter-departmental collaboration ensures that tasks are executed efficiently and that data quality improvements are made systematically. Early experiences highlight the importance of seamless collaboration for managing large datasets and addressing data quality issues promptly.

The application of these methodologies is laying the foundations for a more reliable and coherent national data infrastructure. As these techniques are refined and extended, they promise to support better decision-making, policy-making and public administration by providing higher quality statistical data.

The current focus is on refining these methodologies and expanding their application to other administrative data sources. Continued research and development will be essential to enhance anomaly detection and imputation processes, potentially incorporating advanced machine learning techniques and artificial intelligence. Ongoing collaboration and continuous monitoring of data processing workflows will be crucial to maintain and improve data quality over time.