I can see clearly now outliers are gone - How to improve data quality of official statistics?

Sónia Mota[1], Susana Santos[2], Beatriz Amorim[3]

[1]Banco de Portugal, scmota@bportugal.pt, Portugal

[2]Banco de Portugal, smsantos@bportugal.pt, Portugal

[3]Banco de Portugal, bamorim@bportugal.pt, Portugal

Abstract

As statistical compilation relies further on highly granular data, and tighter deadlines, it is crucial to ensure that speed does not compromise data quality, and that decision-making rests on high-quality data.

In this context, we present indicators developed to assess the quality of our statistics. From information consistency to identifying anomalous values, each indicator contributes to enhance the value of the statistics we produce and can influence positively the trust that society has in them. This paper not only explores data management, but it represents an invitation to reflect on the standards that will define the next era of information.

We stress our commitment with data quality, covering the path for informed and reliable decisions. This is not just a use-case in data management, it is a commitment to excellence that will transform each statistic into more than a number, becoming a beacon of reliability in the vast ocean of information.

Keywords: quality control, quality indicators, web platform

1. Introduction

In a world where the velocity of data acquisition is indisputable and the management of microdata presents perpetual challenges, prioritizing information quality emerges as paramount. Granular data can enhance our understanding of the economy and facilitate a swift response to the current global context. However, it is imperative that this data meets high-quality standards.

At the core of this discussion lies the recognition that, in a data-saturated world, trust in information plays a fundamental role. The relentless pursuit of statistical quality is more than just a standard practice; it is an urgent requirement for informed decision-making.

The quality of information is not an abstract concept, but a tangible reality that requires constant awareness. To monitor and improve this quality, Banco de Portugal uses different analysis and indicators. The assessment of temporal, internal and external consistency, as well as the analysis of revisions and their impact on users, prove to be essential tools. They not only demonstrate the commitment to quality, but also provide reliable valuable insights for the continuous improvement of production processes.

This paper suggests a reflection on the inherent importance of excellence in statistical production, emphasizing the need to ensure that the immediate availability of data does not compromise the effective quality of this information. Additionally, it presents the fundamental concepts developed to assess statistical data quality, along with new initiatives aimed, by Banco de Portugal, at improving statistical data quality and implementing more effective data quality management in statistical systems.

2. How can we maintain statistical quality within the ongoing flux of information challenges?

The Statistics Department of Banco de Portugal has developed a model of quality indicators to systematically control the quality of its statistical systems and outputs. We will analyse this model with respect to the principles of Precision and Reliability, and Coherence and Comparability outlined in the European Statistics Code of Practice. This involves regularly measuring and monitoring quality indicators to document quality control procedures, facilitate comparisons across different statistical areas, identify system weaknesses, and set priorities for future statistical activities. The results from these assessments help to improve the quality of statistics, acquire a more comprehensive understanding of data integrity, anticipate potential revisions, and evaluate the impact of using different datasets in analyses.

2.1 Principle of Accuracy and Reliability – The impact of revisions

Why do we revise data? It is quite straightforward. Sometimes, new information arises after the initial data compilation, or we identify minor errors in the initial data. Due to the different availability schedules of our data sources, updates may need to wait until the subsequent production cycle to be published.

The revision process is like refining a diamond: we implement modifications to improve quality. This may involve incorporating new information, rectifying minor errors, or refining our statistical techniques and IT systems. In the end, our goal is to ensure that our data are as accurate and valuable as possible.

To ensure the transparency and consistency of statistical information, Banco de Portugal adopts two main strategies. On one hand, a recommended approach involves making the revision policy publicly accessible, promptly disclosing any relevant changes, and restricting revisions to instances where they substantially enhance quality. The other is to monitor these revisions, creating and analysing various indicators to improve data quality and reliability.

In the Statistics Department of Banco de Portugal, several revision indicators have been developed. These indicators are presented in the form of a dashboard, enabling a detailed analysis of the revision data. This dashboard identifies the most relevant revisions considering three versions of the data: the first-time data was published for that reference period and the version from the last and the second last published data. In the dashboard, we present metrics such as:

- Root Mean Square Relative Error (RMSRE), that quantifies the difference between initial and final estimates relative to the typical fluctuation of the data. A value of zero indicates precise initial estimates, while a value of 1 suggests an accuracy comparable to predicting the average. Values exceeding 1 imply lower accuracy compared to predicting the series average.
- Directional Reliability Indicator (Q), which examines the consistency of directional changes between initial and final estimates, ensuring alignment in directional shifts.
- Bias, that identifies systematic deviations between the averages of initial and final estimates. Significant deviations from zero indicate potential bias in our assessments.

To properly manage this kind of information, a robust data information solution, based on a data warehouse system, is essential to lead to higher quality standards and efficiency and respond to the challenges ahead. In reaction to this requirement, a database has been developed to store the historical values of all series published on the Banco de Portugal's statistics portal – BPstat[1]. This database covers all data versions since February 2020.

Given the vast number of series published, the database offers the capability to analyse either all series or selectively focus on a specific group. Additionally, a dashboard supported in PowerBI was developed for exploring and visualizing this information. It serves as a flexible tool, enabling analysis of the evolution of a specific observation for a given data series and reference date, either in table or graphic format. The dashboard's primary purpose is to evaluate the stability of data over time, providing a visual and accessible analysis of statistical information across different published data versions.

Through this dashboard, our aim is to analyse how a series has been revised at all its temporal points. It enables us to determine whether we have strayed from the initial value and, if revised, whether the revision was downward or upward. Furthermore, we can investigate if, upon multiple revisions, we reverted to a previously assumed value at any point in time. This analytical process helps us understand the trajectory of revisions over time and identify patterns or trends in the revisions of the data series.

[1] BPstat is the Banco de Portugal's statistics portal, through which statistical information about the Portuguese and euro area economies is provided. All information released in BPstat can be accessed free of charge, without the need for registration or subscription.

2.2 Principle of Coherence and Comparability – What is data consistency?

Consistency entails logical and numerical coherence, encompassing consistency over time, within datasets, across datasets, and in comparisons with other data. This concept is commonly recognized as internal, external, or temporal consistency.

The internal consistency indicators aim to ensure coherence within the same statistical domain or compliance with established reporting rules. Validation of internal consistency involves several procedures, such as calculating the difference between the total and the sum of the parts, checking whether the sum of monthly transaction values equals the quarterly values, or whether the end-year stocks are equal to the end-December stocks.

External consistency evaluates whether the data is comparable with other domains or sources of information. Comparative analysis procedures are performed using similar statistics from other sources, when available (cross-checking statistics for comparable phenomena) and between statistical information and administrative data (accounting data received for supervisory purposes) to ensure the overall quality of information disclosed by Banco de Portugal.

Another aspect we need to consider is temporal consistency, which aims to analyse the temporal evolution (month-on-month and year-on-year rates of change) and identify and manage outliers. Ensuring consistency over time helps to verify the absence of series breaks in the released data.

In the analysis of temporal data, both chain and year-on-year change rates emerge as distinctive methodologies, each offering unique benefits that enhance our understanding of temporal dynamics:

- The chain growth rates, when comparing current values to immediately preceding ones, stand out for their sensitivity to short-term changes.
- Year-on-year growth rates provide a direct annual perspective by comparing values to the same period of the previous year. This approach aims to offer a more stable and long-term view of trends. Year-on-year rates are relevant by attenuating the effects of seasonality and enabling the identification of potential patterns in the data.

Within the identification of outliers, to enrich Banco de Portugal statistics quality, an additional layer is added to the quality control procedure. This additional layer evaluates the plausibility of the last datapoint of a given series given the recent past behaviour of the corresponding series. By applying some checks, it is possible to flag outliers and identify series that deserve extra attention as their last observation presents a value that deviates from its recent past behaviour. Some examples of the outliers flagged are:

- Implausible zeroes, that identifies whether a series, which generally does not have zeros, suddenly has a zero value.
- Unexpected signal changes, that identifies signal changes in a series that normally maintains a constant direction.
- Unexpected changes in magnitude, that checks whether a significant change has occurred in the magnitude of the series, detecting values that are outside a selected average range.

These plausibility checks aim to highlight possible changes in statistical series, providing a valuable tool for early detection of unexpected variations or potential errors in the data to be released making quality control more efficient and reliable.

Furthermore, the combination of these two complementary approaches, not only enriches temporal analysis, but also provides a more balanced understanding of changes in the data over time.

2.3 How can we see clearly now – a fresh approach to quality control?

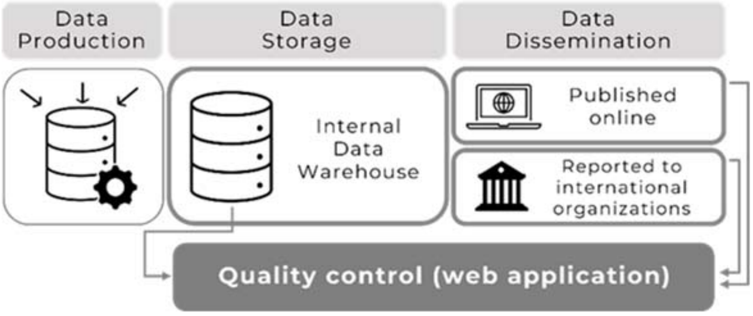
Following the implementation of the quality indicators delineated previously, the Statistics Department of Banco de Portugal has developed a new system that enables an integrated process of quality control throughout the production cycle of central bank statistics, leading up to their publication. This system has three fundamental components: a collection of statistical tests (which includes the checks previously outlined such as completeness, consistency, and plausibility checks), pre-defined operating rules to ensure data quality throughout every stage of the statistical production process, and a web platform that ensembles data from various storage system, performs the test previously defined and allows the analysis of the results.

It is important to highlight that the library of consistency and plausibility tests previously mentioned is quite generalizable and easy to apply to different statistical domains. For example, in terms of consistency checks we verify if the value of total assets equals the value of total liabilities. This check could be applied to investment funds, banks, and insurance corporation balance sheet. Having all the tests in a single platform brings efficiency gains to statistical quality control.

How is this integrated process of quality control applied in practice?

After receiving data from various data sources, statistics experts analyse it and compile the corresponding aggregates which are shared in an internal data warehouse. These aggregates are tested in the web platform which allows us not only to identify errors and correct them immediately, but also to quickly identify series whose last observation had presented an unexpected value. After applying the necessary corrections and analysing the flagged series data is ready to move to the next stage. Some data will be published online, and some data will be reported to international organizations. For these aims, data must be "translated" into specific "languages" and file formats. These transformations might impact data quality which imposes the need to run the tests again in the web application to verify that no errors arise, and all the flagged outliers have been previously analysed and justified (Figure 1).

Figure 1: Integration of quality control in the production cycle



Summing up, with a comprehensive library of consistency and plausibility tests, this integrated process of quality control, adheres to standardized protocols to ensure data quality across the statistical production cycle. Moreover, by centralizing quality control in a single platform, these protocols guarantee transparent, reliable, and efficient tool utilization.

3. Concluding remarks

In the contemporary data-centric environment, expeditiously acquiring information is imperative. However, we must remember that speed should never compromise quality. It is essential to prioritize both timely acquisition and data integrity.

Keeping in mind the crucial role of quality, let's not overlook its profound impact; it is what makes people trust our statistics and keep choosing them over others.

It is essential to monitor data quality by implementing robust quality indicators based on the principles of Accuracy and Reliability, and Coherence and Comparability. Moreover, considering the significant amount of data, it is crucial to ensure that the data quality control progresses towards enhanced integration, ensuring greater transparency and efficiency.

For this purpose, in the Statistics Department of Banco de Portugal there is an ongoing effort to develop and implement more flexible exploring tools enabling a more efficient, transparent, and reliable data quality control ensuring that our high-quality standard are met.

References

Amorim, B., & Silva, P. (2024). *Keeping data under control: a data management system for quality assurance*. JOCLAD 2024

European Central Bank. Public commitment on European Statistics by the ESCB. Retrieved March 2024 from https://www.ecb.europa.eu/stats/ecb_statistics/governance_and_quality_framework/html/escb_public_commitment_on_european_statistics.en.html

European Central Bank. (2024, January 25). *Why do statistics matter?* Retrieved March 2024 from https://www.ecb.europa.eu/ecb-and-you/explainers/tell-me-more/html/statistics.en.html

European Central Bank. (2017, June 27). *How does innovation lead to growth?* Retrieved March 2024 from https://www.ecb.europa.eu/ecb-and-you/explainers/tell-me-more/html/growth.en.html

European Central Bank. (2012). *Euro Area Balance of Payments and International Investment Position Statistics*. https://www.ecb.europa.eu/pub/pdf/other/euroareabalanceofpaymentsiipstatistics201203en.pdf

International Monetary Fund. (2003, June 25). *Data Quality Assessment Framework and Data Quality Program*. Retrieved from https://www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm

Silva, P., Pinto, M., & Agostinho, A. (2019). *How to Turn Quality into a Habit in the Statistical Production?* Statistika: Statistics & Economy Journal, 99(1), 97–103.