A systematic approach for data validation using data driven visualisations and interactive reporting

Keywords: *Data reuse and sharing, data validation, automation, visualisation, design, outliers detection*

1 Introduction

Eurostat supplies European citizens, governments and institutions with reliable statistics and data on Europe. One of its most important missions is ensuring the finest statistical quality. As stated in the *European Statistics code of pracice* [1], "quality [i]s the basis of [Eurostat's] competitive advantage in a world experiencing a growing trend of instant information which often lacks the necessary proof of quality". Statistical excellence is essential because most of economic speculation, social analysis and political decisions are based on statistical foundations [2].
The purpose of data validation is to ensure a certain level of quality of the final data. Eurostat commits to a continuous improvement framework, by always challenging its processes and looking for possible weaknesses. Modernisation of data validation is thus at the center of our mission.
The core principles of quality as defined in [1] are: relevance, accuracy, timeliness and punctuality, accessibility and clarity, as well as comparability and coherence. **In this article we propose innovative methods aiming at controlling the relevance, the accuracy and the comparability of the data.** These methods apply a systematic approach based on data driven visualisations and interactive reporting. With the final goal of ensuring the finest data quality, this article aims at integrating innovative statistical techniques to control the plausibility of data, and use modern technologies to produce simple but straightforward errors.
Within the European Statistical System (ESS), Eurostat and the member states have already invested a lot in data validation. The objective of this article is to suggest further developments to the data validation in the context of the financial accounts domain, which can however be easily generalised and applied to any other domain.

2 Methods

2.1 The need for a systematic and reliable approach

In the financial accounts domain we deal with over 25,000 time series of annual data, since 1995. Even though many time series are not mandatory for transmission by Member States to Eurostat, observing and analysing by hand each of them is impossible, making it necessary to implement a systematic control of the data. Our approach includes already implemented consistency checks, and relies on state-of-the-art methods for outliers detection, such as [3] or [4]. The approach presented in this article includes the following innovations.

2.1.1 Select the most relevant revisions

This point is determinant for the quality of the report. Many revisions are usually observed in every transmission of data, but what we want rather than spotting all of them is to identify the abnormal ones. This is done by using dynamic threshold and sorting them by order of significance, and by verifying if revisions are consistent with previously detected outliers.

2.1.2 Outliers detection

The validation of the data related to its accuracy proceeds in two parts. The first one consists in controlling data from previous years that have already been validated in the past. This data is subject to revisions and might change. As explained in the previous section we deploy an automatic selection of the most relevant revisions, so that we know which data is valid or not.
The second part consists in checking the newly observed value (if included in the transmission). This could be done building a SARIMA model with the past values and a confidence interval for the next-step prediction of this model, and observing if the newly observed value lies in this interval (this method is inspired from [5]). The difficulty we encountered is that the financial accounts data is by nature very volatile, therefore an outlier detection method might be misleading.
Let us consider a time series (X_t). We use a seasonal ARIMA$(p,d,q)(P,D,Q)_s$ model for this series:

$$\phi_P(B^S)\Phi_P(B)(1-B)^d(1-B^S)^D y_{t^*} = \theta_Q(B^S)\Theta_P(B)\varepsilon_{t^*}$$

where $t^* = t - r$ and r is the number of last observations we removed to build the model. Then, SARIMA forecast interval are built using:

$$\hat{x}_{t^*}(h) = z_{\frac{\alpha}{2}}\Sigma_{e_{t^*}(h)}$$

where $\hat{x}_{t^*}(h)$ is the punctual forecast at time $t^* + h$, $z_{\frac{\alpha}{2}}$ is the percentile of a standardised normal distribution and $e_{t^*}(h)$ is the forecast error at time $t^* + h$, and $\Sigma_{e_{t^*}(h)}$ the standard error of the latter. If the observed value at time $t^* + h$ is not inside the forecast interval at time $t^* + h$, then the value is considered as an outlier. This generates an automatic visualisation.

2.1.3 Data-driven visualisations

As stated in [6], the major issues for selecting relevant visualisations are scale and utility. We need to be able to evaluate a large number of time series, and selecting the most appropriate metric to allow the easiest interpretation of the data.
The previous methodologies allow to spot the most relevant events, revisions and outliers. Based on these results, we offer smart and coherent visualisations that allow to retrieve the most useful insights from the data. This article is inspired from state-of-the-art methods for data driven visualisations, such as [6], [7] or [8].

2.2 Interactive reports

The data validation process is a back and forth exchange between the member states and Eurostat's teams. Making the process effective and reliable requires to minimise the burden for the respondents and to develop a good cooperation with them. To answer to these requirements, we put a lot of effort in designing very clear data reports that need to have the following characteristics:

1) The errors included in the report are selected by order of importance thanks to the approach presented previously. To avoid the explosion of the amount of information, the number of reported errors should be kept to the minimum.

2) The report displays clear (easy to read) errors or warning messages, and it includes the most relevant charts which help observing the issues.

3) The report should be an editable and interactive object, not a static one. This will ease the validation by allowing to add explanations, comments and corrections to the report, but most importantly by allowing the users of the report to investigate the data rather than dealing with a fixed output.

3 Results

In this section we provide an example of data visualisation used for outliers detection. Figure 1 presents the non consolidated liabilities of transferable deposits for a member state's total economy from 1995 to 2019. The model presented before allows to build a confidence interval for the data and identify a suspect value in 2017. This visualisation is a key for the user of the report to observe it and react to it in the appropriate way.

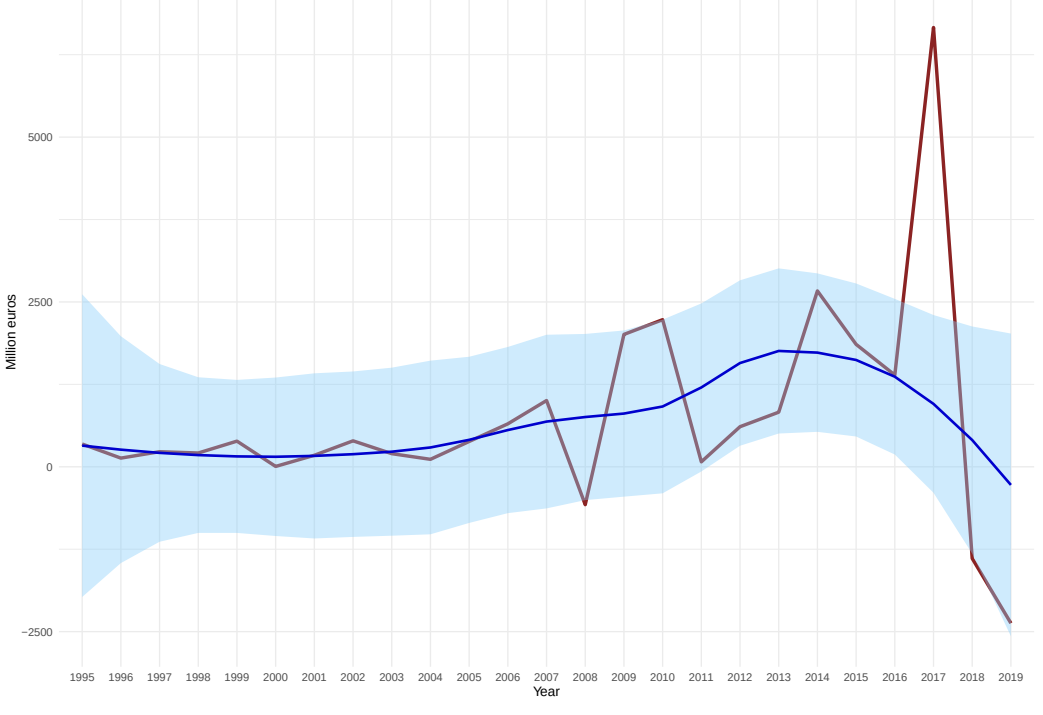

Figure 1: Example of a member state's financial accounts: Transferable deposits of the total economy, liabilities (non consolidated), data from 1995 to 2019

4 Conclusions

In this article we offered an innovative approach for data validation. We proposed a modern and systematic approach that allows for a precise evaluation of the data and most importantly leaves the possibility to the users to explore the data and understand the reported errors.
The added value of this paper lies in the fact that it offers a quick and effective

validation, which is helpful for Eurostat's staff but also for the member states: thanks to clearer and simpler reported errors, the latter can reply on shorter notice. In the end, this greatly contributes to ensure better data quality.
Even though this article was inspired from the needs of the financial accounts domain, it aims at being generalised, so that it could be used by other domains to further develop their data validation.

References

[1] *European Statistics Code of Practice*, 2018. URL http://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142. Revised edition 2017.

[2] J.B. Cohen. Misuse of statistics. *Journal of the American Statistical Association*, 33(204):657–674, 1938. DOI: 10.1080/01621459.1938.10502344.

[3] Stephen Bay, Krishna Kumaraswamy, Markus G Anderle, Rohit Kumar, and David M Steier. Large scale detection of irregularities in accounting data. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 75–86. IEEE, 2006.

[4] Rick J Lenderink. Unsupervised outlier detection in financial statement audits. Master's thesis, University of Twente, 2019.

[5] Dario Buono and Enrico Infante. New technique for predictability, uncertainty, implied volatility and statistical analysis of market risk using sarima forecasts intervals. 03 2013.

[6] Manasi Vartak, Sajjadur Rahman, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis. Seedb: Efficient data-driven visualization recommendations to support visual analytics. In *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, volume 8, page 2182. NIH Public Access, 2015.

[7] K. Stockinger, J. Shalf, K. Wu, and E. W. Bethel. Query-driven visualization of large data sets. In *VIS 05. IEEE Visualization, 2005.*, pages 167–174, 2005.

[8] M. C. Hao, Umeshwar Dayal, D. A. Keim, and T. Schreck. Importance-driven visualization layouts for large time series data. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 203–210, 2005.