Zaccardi, Jules; Infante, Enrico

Article

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

Eurostat Review on National Accounts and Macroeconomic Indicators (EURONA)

Provided in Cooperation with:
Eurostat, Luxembourg

Suggested Citation: Zaccardi, Jules; Infante, Enrico (2021) : A systematic approach for data validation using data-driven visualisations and interactive reporting, Eurostat Review on National Accounts and Macroeconomic Indicators (EURONA), ISSN 1977-978X, Publications Office of the European Union, Luxembourg, pp. 67-89

This Version is available at:
https://hdl.handle.net/10419/309843

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

JULES ZACCARDI [1] AND ENRICO INFANTE [1]

Abstract: In this article we present the development of a tool that can serve most official statistics. This tool is fully interactive and based mostly on visualisations, which makes it a powerful ally to validate or analyse data. To make our tool a pragmatic solution in an era facing ever larger volumes of data, we integrated a data-driven analysis to identify the most unusual data and suggest them to the user. By defining the right metrics and smartly designing our report, we managed to develop an easily-accessible tool that can be used to validate or analyse data. We tested this tool for data validation in annual financial accounts, and it demonstrated its ability to identify the main issues in the dataset very quickly, therefore improving i) the productivity of the user of the tool ii) the long-term viability of the validation and iii) ultimately the quality of the data.

JEL codes: C02, C61, G20, P43

Keywords: data literacy, data validation, automation, visualisation, design, outlier detection

[1] Eurostat, Unit C2 — National accounts production.

1. Introduction

1.1. Research problem

Eurostat supplies citizens, governments and institutions with high quality statistics and data on the European Union (EU). One of its most important missions is ensuring the finest statistical quality. As stated in the *European statistics code of practice* (Eurostat (2018)), 'We see quality as the basis of our [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 economic speculation, social analysis and political decisions are based on statistical foundations' (Cohen (1938)).

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. The modernisation of data validation is thus at the centre of our mission. Besides performing data validation, official statisticians also use data for analysis, exploration and the monitoring of performance indicators. The scope of our tool therefore also includes the possibility to present and analyse data produced from official statistics.

The core principles of quality as defined in Eurostat (2018) 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 accuracy and the comparability of 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, the tool described in this article aims at integrating innovative statistical techniques to control the plausibility of data and using modern technologies to produce simple but straightforward error messages.

Within the European Statistical System (ESS), Eurostat and the EU Member States have already invested a lot in data validation. The objective of this article is to suggest further developments for data validation.

1.2. State of the art

We conducted a bibliographic review to establish what is the state-of-the-art regarding data-driven visualisations and outlier detection, with a special interest in financial accounts.

As outlined in Bay et al. (2006), there is a need for an automatic analysis of irregularities in large datasets (for example, fraud or errors). Thus, a variety of models currently exist to examine financial accounts. While Yang and Cogill (2013) considered a graph similarity algorithm using the string edit (or Levenshtein) distance to identify an unconventional financial statement, Buono and Infante (2013) considered using SARIMA forecast intervals for the predictability of financial time series. We decided to follow an approach based on the ideas presented in Lenderink (2019) in which Lenderink explains that a fraud or an error 'has to deviate from "normal" behaviour and from regular financial transaction patterns'. Following this idea, the most interesting approach seemed to us the one of Zhu (2006) which lies in estimating the deviation from a tested data point to a reference dataset.

Furthermore, given the increasing development of visualisation-based analytical tools and big data, a lot of work has been carried out to develop the design, the speed and the relevance of such tools. Mackinlay (1986) proposed a way of *Automating the design of graphical presentations of relational information* which we considered in the way we designed our tool overall. Hao et al. (2005) presented a framework for visualising large sets of time series. Although they faced the same problem as us, their approach was to find a way to show all the series while introducing a metric of importance that made it possible to prioritise the most useful ones. We considered this to be too heavy and therefore we preferred the approach of Vartak et al. (2015) and Kandel et al. (2012). Vartak et al. (2015) developed *SeeDB*, a visualisation recommendation engine which — given a dataset — intelligently explores the space of visualisations, evaluates promising ones and recommends those it deems more useful to the user. Similarly, Kandel et al. (2012) developed *Profiler*, a visual analysis tool for assessing quality in tabular data. The key to these approaches lies in the suggestion of interesting views and scalable visual summaries that support real-time interaction with the user across millions of data points.

2. About validation in official statistics

2.1. The generic procedure

The ESS has carried out a considerable amount of work to produce 'a generic framework for data validation in order to have a reference context, and to provide tools for setting an efficient and effective data validation procedure' (Di Zio et al. (2018)). According to the ESS definition, 'data validation is an activity verifying whether or not a combination of values is a member of a set of acceptable combinations' (Di Zio et al. (2018)), and its process includes:

- establishment of checking rules;
- detection of outliers or potential errors;
- communication of the detailed problems to the actors in best position to investigate them.

The purpose of data validation is to ensure a certain level of quality of the data. It focuses on specific dimensions of data quality, such as accuracy, comparability and coherence. Accuracy refers to the measurement of the difference between the 'target parameter' and the 'estimated parameter', while coherence and comparability refer to statistics being consistent internally (between related data points), over time and space (for example, comparable between regions and countries). Data validation can be divided in two categories.

- **Structural validation**, aimed at verifying the technical integrity of the file (in other words, consistency with the expected IT structural requirements).
- **Content validation**, aimed at verifying the logical and statistical consistency of the data. This includes:
  1. consistency within the dataset;
  2. consistency with other datasets within the same domain and within the same data source;
  3. consistency within the same domain between different data sources;
  4. consistency between separate domains for the same data provider;
  5. consistency with data from other data providers.

In this article we propose an improvement for points 1 and 4 of content validation. Following the definition of data validation (as provided at the beginning of this section), our approach has two main goals. The first is an improvement in detecting outliers or potential errors in a big set of time series by introducing advanced statistical methodologies. The second is to ease the communication process with statistical reporting agencies to facilitate the investigation and the correction of errors, by automating validation procedures as much as possible.

2.2. The case of financial accounts

Financial flows and stocks are the data described in Tables 6 and 7 of the *European system of accounts 2010 — ESA 2010 — Transmission programme of data* (Eurostat (2014)). The data consist of several thousand time series (up to 25 000) with annual observations since 1995. As presented in ESA 2010, financial flows and stocks are structured in the following ways.

- They can be stocks, transactions or other flows. Let us denote ***F*** = {stocks, transactions, other changes in volume, revaluation account} ([2]).
- They are split between institutional sectors (most of which are split into several subsectors). Let us denote ***S*** as the set of sectors, described by:
  - non-financial corporations (S.11),
  - financial corporations (S.12),
  - general government (S.13),
  - households (S.14),
  - non-profit institutions serving households (S.15).
- They are described by financial instruments, which constitute the total economy overall. Let us denote ***I*** as the set of financial instruments, described by:
  - monetary gold and special drawing rights (F.1),
  - currency and deposits (F.2),
  - debt security (F.3),
  - loans (F.4),
  - equity and investment fund shares or units (F.5),
  - insurance, pension and standardised guarantee schemes (F.6),
  - financial derivatives and employee stock options(F.7),
  - other accounts receivable/payable (F.8).
- They can be assets or liabilities. Let us denote ***A*** = {assets, liabilities}.
- Finally, they can be consolidated or not consolidated. Let us denote ***C*** = {consolidated, non-consolidated}.

([2]) As the other changes in volume and revaluation accounts are specific types of flows, in this article we will only mention either 'flows' or 'transactions', always referring simultaneously to transactions, other changes in volume and revaluation accounts.

Most of the institutional sectors and financial instruments mentioned above consist of different layers of subsectors (such as subsectors S.121, S.122 and so on) and different layers of instruments (such as F.51, F.511, F.512 and so on). Multiplied by the $|F| \times |A| \times |C| = 16$ possible categories, we obtain several thousand time series in the data. Financial accounts are by nature very volatile data, therefore it is hard to model trends in the data and predict future values.

The current validation system produces an automatic report based on implemented rules. Although this report includes all the inconsistencies, it is hard to read and to make use of. Weaknesses appear especially regarding the analysis of revisions and the detection of outliers (see next section for more details). The current system is a great basis to build on, but it demonstrates the need for a systematic approach using data-driven visualisations and interactive reporting, which is the focus of this article.

Our goal is to develop a reporting framework which is user-friendly, makes it possible for validators to spot the biggest inconsistencies easily, and reduces the burden for statistical reporting agencies to investigate and correct errors.

3. A framework for interactive and automatic validation reporting

3.1. General framework

Validation procedures usually include a reporting step during which the inconsistencies are listed in a document which is called the **validation report**. The latter generally includes the following sections.

- **Completeness**. This section shows the rate of missing values overall.
- **Confidentiality**. This section shows the rate of confidential data overall.
- **Zero and negative values**. This section shows counts of zeros and negative values. In many domains the latter should not occur, for example for statistics on the level of population. In the case of financial accounts, stocks should always be non-negative.
- **Internal consistency**. This section shows whether breakdowns add up to their main series. Most domains deal with series that are breakdowns and are faced with verifying that those add up to their relevant main series. In financial accounts, we check institutional sectors that are broken down into subsectors ([3]), to see if the sum of the values for the subsectors adds up to the value of the main sector.
- **Revisions**. This section is based on a comparison of the data to be validated with the data validated during the previous production round. It shows changes (called revisions) in the data.

In this article, we propose an updated framework for validation reporting that is more user-friendly, more straightforward and based on innovative statistical methodologies.

([3]) Formally, denoting by $n \in \mathbb{N}$ the maximal number of layers in a sector we can write that $S_n \subset S_{n-1} \subset \ldots \subset S$.

The main aspect that we will leverage to this purpose will be the visualisation of time series, which makes it easier for the validator to spot errors and understand how to act. However, it is impossible to visualise thousands of time series manually. Therefore, we propose in this article a method based on data-driven visualisation suggestions. To integrate this new approach, we propose the following new framework:

- **completeness**;
- **confidentiality**;
- **zero and negative values**;
- **internal consistency** ([4]);
- **macroeconomic indicators —** here we show key indicators used by institutions to have a first overview of the data ([5]);
- **generic visualisations** — here we propose a systematic visualisation of the most important mandatory series;
- **revisions visualisations** — here we propose visualisations of series, based on computations to spot the most unusual revisions;
- **detection of outliers** — here we propose visualisations of series based on computations to spot outliers.

Such a framework ensures that the report covers all the needs for official statistics. Indeed, not only does it improve data validation, but it also makes it possible to explore the data and to monitor macroeconomic indicators.

3.2. Metadata

3.2.1. DESCRIPTION OF THE PROCEDURE

As stated in Section 2, the process of data validation includes the communication of detailed problems to the actors in the best position to investigate and correct them. Once we have spotted errors (or inconsistencies) and contacted statistical reporting agencies, either the errors are corrected or the agencies provide us with a reason(s) for the inconsistency(ies). These explanations then become part of the **metadata**.

Given that the exchange of information described in the previous paragraph happens during every production round, we introduce in this article the use of a **metadata template** ([6]). This template is a formatted document that includes a list of known (by the statistical reporting agency) inconsistencies and detailed explanations for them. The inconsistencies are clearly identified in terms of series and year, and when possible categorised. Statistical reporting agencies would fill in this template and transmit it together with the data.

In the case of national accounts, the template is structured in such a way to clearly identify the series, the nature of the problem and the reason for it. It therefore contains fields to indicate the elements of *F*, *S*, *I*, *A* and *C* necessary to identify a series, one field to describe the nature

([4]) These elements are not refined further in the article because we consider that the current solution is sufficiently developed.
([5]) In the case of financial accounts we use the macroeconomic imbalance procedure indicators.
([6]) Such a template is not an invention of the authors. Indeed, this solution exists and is used in several domains across official statistics. Therefore, this proposal specifically concerns financial accounts, where such templates are not yet in use.

of the issue (it can be a negative value, a major event or an outlier, an important revision, or another reason) and an explanation for it.

3.2.2. ADDED VALUE OF THIS CONCEPT

The use of such a template has three main advantages.

The first one is to develop an easy communication between the different actors involved (in the case of financial accounts in the EU, these would mainly be the validator in Eurostat and the statistical reporting agencies in the EU Member States), therefore formalising, structuring and simplifying communication, making it more precise and unambiguous, ensuring overall a smooth validation. By using such a template, we give the Member States a chance to confirm and explain the known unusual values that are present in their data ([7]). We then know that any other unusual values that we identify will need to be investigated and possibly corrected.

The second one is that this template makes it possible for us to store the inconsistencies and reasons over time, therefore opening a world of possibilities, from improving the methodology of the data collection to ensuring a bilateral follow-up on data quality with the EU Member States. Importantly, it should also avoid data validators repeatedly asking statistical reporting agencies to confirm values that have already been confirmed, thereby reducing the burden on both actors.

Finally, this template makes it possible to use the metadata during the validation stage and include the information provided by countries in the validation report. This is a huge improvement in the sense that it takes the validation one step further. Whereas we used to only spot inconsistencies, we now have the possibility to investigate them immediately (at the time of the validation). Indeed, we will specify in the validation report whether an inconsistency has been mentioned in the metadata template or not.

The idea of introducing a metadata template is very generic and can easily be adapted to any domain of official statistics. In fact, several domains already make use of one, and in this article we propose the implementation of this concept for financial accounts. As metadata play a key role in the validation of official statistics, easing the way in which they are collected is a valuable innovation.

3.3. Cross-domain checks: the case of comparing annual and quarterly financial accounts

Cross-domain checks consist of 'checks between the data available in the data provider (institution) and the data available outside the data provider (institution). This implies no control over the methodology on the basis of which the external data are collected, and sometimes a limited knowledge of it.' In our case, annual financial accounts and quarterly financial accounts are based on different legal bases (regulations), and are often compiled by different institutions, although they are theoretically the same thing, just with different periodicity.

([7]) For instance, EU Member States may be aware of a big revision they made, or of an unusual value in their data, and explain it at the time of transmission.

3.3.1. DESCRIPTION OF THE PROCEDURE

Eurostat receives annual financial accounts (AFA) from each EU Member State, while the European Central Bank receives quarterly financial accounts (QFA) from them. To ensure the quality of the data, the consistency between AFA and QFA is checked. To do this, we implemented an automatic consistency check which generates individual reports for each Member State. Although this validation might be specific to financial accounts, it can be generalised to other domains considering any simple cross-domain check. Indeed, the added value lies not in the check itself, but in the automation of it.

The elements to be checked are the following:

- for transactions, $\sum_{i=1}^{4} Q_i = A$;
- for stocks, $Q_4 = A$;
- for all series, that the observation flags ([8]) and confidentiality flags ([9]) are identical for AFA and QFA.

In this case, the checks to be made are very simple, but there is a need for an automatic and elegant procedure. For this article we developed a tool that uses the raw data received through the SDMX procedure ([10]) and generates reports outlining the inconsistencies in a clear way. Both of the following rules need to be satisfied to consider a discrepancy as a proper inconsistency:

- $A - Q >$ 10 million of national currency;
- $\frac{A - Q}{GDP} >$ 0.3 %.

The code is structured into three main phases. Firstly, it reads and cleans the data, removing any information that would not be useful in this situation. Then it merges the AFA and QFA data into one single table and computes the gap between the values. Finally, it uses this information to generate individual reports for each EU Member State.

3.3.2. ADDED VALUE OF THIS CONCEPT

The implementation of an automated procedure is of high added value for three reasons.

Firstly, it is effortless for the validation team, as it only requires uploading the data and pressing a button. Everything from reading the data to computing the errors is done with R (R Core Team (2020)). It strengthens the **productivity** of the validation team.

([8]) Observation status flags (OBS STATUS): these provide information on the quality of a value or an unusual or missing value.
([9]) Confidential status flags (CONF STATUS): these provide information about the confidentiality status of the object to which this attribute is attached.
([10]) Standard for data and metadata exchange.

Secondly, it is an elegant document to send the statistical reporting agencies. We found a solution that can serve to easily read and interpret a long set of inconsistencies. By using a reactive document with different tabs, and listing separately the series presenting inconsistencies and the complete set of errors we made the report easy to read whatever the number of errors is ([11]). We started by an introductory paragraph which presents the situation, stating the purpose of the report, specifying when the AFA and QFA data were each received, and indicating the number of inconsistencies (all of this, as mentioned, is done automatically). We continued with a tab showing the number of errors grouped by institutional sector and by financial instrument, which helps identifying from where errors come. Then we added a tab listing the series for which an inconsistency has been observed. Finally, we listed all the inconsistencies, including the values and the gap (in absolute terms and relative to GDP). With such simple reports, this procedure encourages **reactivity** from the statistical reporting agencies and provides the basis for a simple communication between the various actors involved in data validation. In the long term, it should improve the quality of the data.

Finally and most importantly, it is a very reliable solution. Indeed, the procedure makes it possible to control automatically the consistency of each single time series, while a manual process might focus (only) on the most aggregated time series or on some indicators. Furthermore, this procedure also makes it possible to include an analysis of consistency, checking the observation and confidentiality flags. Therefore, it strengthens the **sustainability** of the validation.

4. An innovative way to visualise large datasets of time series

4.1. An approach based on interactivity

The main innovation of our approach is the interactivity of our report. Rather than generating a static list of problems, we give the opportunity for the user to explore the data. This is made possible by carefully considering the structure of the report, and giving the possibility to the user to experiment with it.

The user can click on, expand and extract graphics. By dragging the mouse over the graphics, the user can observe closely the values and additional information such as the percentage rate of change. The user can also zoom in or out and download graphics as images.

Furthermore, the sections of the report were carefully chosen to give the user the most relevant overview of the data. Starting from key indicators and the most aggregated data, it then offers a closer look at the revisions made on the data and possible outliers. The graphics of the outliers' section include a smoothed line and a confidence interval.

([11]) See Illustration 1 and Table 1 for more details.

Finally and most importantly, the user can choose what to see by way of a reactive drop-down menu, which gives the possibility to explore the data more fully. More specifically in the case of financial accounts, the user can choose by way of drop-down menus each element of *F*, *I*, *C* and *A* in order to isolate one specific series. The strength of this model is that it leaves the opportunity for the user to explore the data very closely, isolating each chosen series, but also making possible an exploration of the full dataset. Indeed, the report automatically recommends series to select, based on data-driven visualisations.

4.2. Data-driven visualisations recommendations

In this section we explain how we managed to identify the most relevant data in order to suggest its visualisation to the user of the report. The question we wanted to answer is how to find the most relevant visualisation among a large amount of data. Our idea to solve this problem is based on *SeeDB: efficient data-driven visualisation recommendations to support visual analytics* (Vartak et al. (2015)). Their approach to suggest relevant visualisations is to have one metric to **filter** out what is not relevant and one other metric to **order** series in terms of relevance. They adopted deviation-based metrics for a visualisation utility.

4.2.1. REVISIONS

The approach of this section is to examine by how much data have been revised. This can be done by comparing data that are currently disseminated and data that are to be validated during a production round. The main goal is to evaluate the changes made to a set of data X, that we will call the *reference*, to a set of data Y that we will call the *target*. Let us introduce the following notation:

1. n the number of observations in the reference ([12]);
2. m the number of observations in the target ([13]);
3. $X = \{x_1, ..., x_{n-1}\}$ the *reference* data;
4. $Y = \{y_1, ..., y_{n-1}, y_n\}$ the *target* data.

We will standardise both X and Y, in other words we will apply the transformation

$$f : x \to \frac{x - \min(x)}{\max(x) - \min(x)}.$$

We will refer to the standardised data as X_s and Y_s. We will then consider three different distances:

- The Euclidean or L_2 norm distance between the standardised distributions.
  $D_e(X_s, Y_s) = \|X_s - Y_s\|^2$
- The Chebyshev distance or L_∞ norm.
  $D_c(X, Y) = max|X_i - Y_i|$

([12]) In the case of financial accounts, time series usually start from 1995, so we have around 25 annual observations per series.
([13]) Most of the time, $m = n$ or $m = n + 1$.

- The Hamming distance which corresponds to the rate of revisions made.

$$H(X,Y,\delta)=\frac{1}{n}\sum_{i=0}^{n}\omega_i \text{ where } \omega_i=\begin{cases}0 \text{ if } |X_i-Y_i|<\delta\\ 1 \text{ otherwise}\end{cases}$$

The last of these is very relevant for our purpose given that it indicates the number of observations that have been revised, which is a key indicator to understand if the revision should be investigated or not. Often, data are revised following a 3-year cycle, meaning that revisions of a small amount for the last 3 years are not suspect. Thus, we want to focus on large revisions over the whole time span. Therefore, we computed the Euclidean distance to order the series based in terms of the magnitude of their revisions, and we weighted this by the Hamming distance to build the following score:

$$S(X, Y, \delta) = D_e(X_s, Y_s) \times H(X, Y, \delta)$$

With these metrics, we applied the technique mentioned above, in other words ordering and filtering:

1. **order** considering $S(X, Y, \delta)$;
2. **filter** considering $D_c(X, Y) > \delta$;
3. **filter** considering $\frac{D_c(X,Y)}{\text{Ref}} > \theta$.

Practically speaking for financial accounts, we considered δ = 100 million euro, and we used as reference value Ref the total financial assets or liabilities of the sector ([14]). Given that we receive the data not in euro but in national currency, we convert the 100 million euro into national currencies for each EU Member State using exchange rates from Eurostat's database.

The reference value is challenging to find. We first thought about using GDP of each EU Member State — which is what was done previously — but it implied necessarily cutting off some series that are very small compared with GDP. Finally, we decided to compute the total assets and liabilities for each institutional sector of the economy, and consider the value of the revisions in these sectors relative to the total economy. We used θ = 1 %.

4.2.2. DETECTION OF OUTLIERS

Let us consider similar notations to those in the previous section. However, this time our interest lies only in identifying yn as an outlier or not, compared with $\{y_1, ..., y_{n-1}\}$. We will therefore consider $\mathbb{E}(y_n|\{y_1, ..., y_{n-1}\})$. We considered a locally estimated scatterplot smoothing (LOESS) to model the time series. With a similar approach to before, let us introduce:

- $\bar{Y}$ the set $\{Y_1, ..., Y_{n-1}\}$
- the empirical average of $\bar{Y}$, $\hat{\mu}_Y = \frac{1}{n-1}\sum_{i=1}^{n-1} Y_i$
- the empirical standard deviation of $\bar{Y}$, $\hat{\sigma}_Y = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n-1}(Y_i - \hat{\mu}_Y)^2}$

([14]) The value of δ and θ have been chosen based on empirical knowledge of the domain. The two thresholds are complementary, as δ is the same for all EU Member States, while θ considers the size of the financial sector of each Member State.

With these notations, we applied the same technique as for the revisions coming from Vartak et al. (2015), in other words:

1. **order** considering the z-score of Y_n relatively to the distribution $\overline{Y}$,
$$z = \frac{y_n - \hat{\mu}_Y}{\hat{\sigma}_Y};$$
2. **filter** considering $\hat{\sigma}^2_Y > \varepsilon$;
3. **filter** considering $y_n > \delta$;
4. **filter** considering $\mathbb{E}(y_n | \{y_1, ..., y_{n-1}\})$.

The first rule ensures that we order based on the distance between the latest observation and the distribution of previous observations. Given that z is conversely proportional to $\hat{\sigma}_Y$, the risk is to select mostly constant series that have a standard deviation close to zero. To avoid this, we consider rule 2 to ensure that we select series with a variance larger than a certain ε. Furthermore, we filtered only the values larger than δ (the same as before), to avoid working with minor series. Finally, we filter out all the observations that belong to a confidence interval built using a regression on $\overline{Y}$, considering that those are not outliers. We are left with the most interesting series to observe, ordered by z.

5. Results

In this section, we present the output that we developed to answer our research question. As explained previously in Section 3, we developed a report structured in several parts. Here we will make a case study considering real data from the annual financial accounts sent by Belgium. We are comparing data sent in March 2021 for validation with data previously disseminated.

5.1. Overview of the report

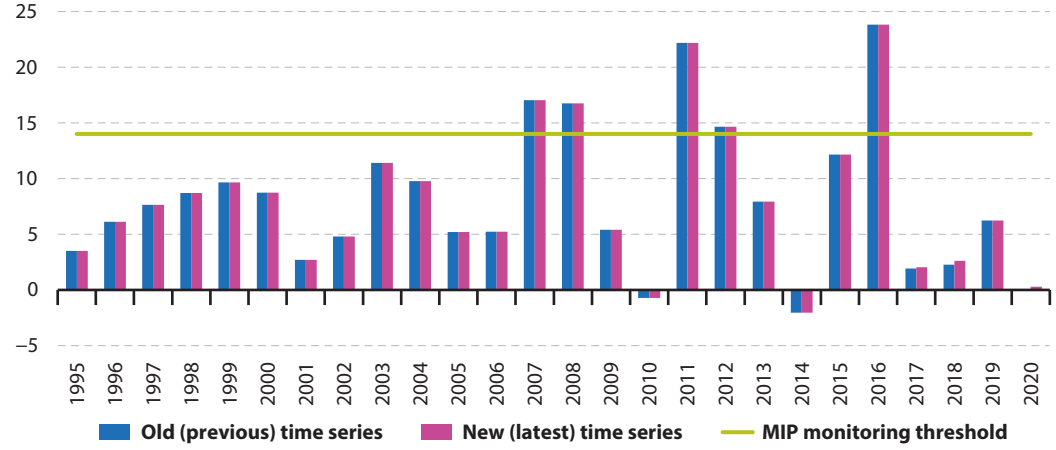
5.1.1. MACROECONOMIC INDICATORS

In this case study, we are dealing with three macroeconomic indicators used in annual financial accounts to monitor the data which are part of the *macroeconomic imbalance procedure* (MIP) indicators. Financial accounts MIP indicators are used for policy purposes. As they are the most used indicators in this domain, they have been selected to be in the report ([15]). Clearly, other indicators should be used for other domains.

Figures 1 and 2 show the visualisations we chose. We can easily compare the data transmitted with the data previously disseminated. Furthermore, we can simultaneously monitor the data, considering the green-line threshold on the figure. As for all the other figures in our report, these are interactive, meaning that we can zoom in and out, select a specific year or value, and export it as a *.png file if necessary.

([15]) See the following Eurostat webpage for details: https://ec.europa.eu/eurostat/web/macroeconomic-imbalances-procedure/indicators.

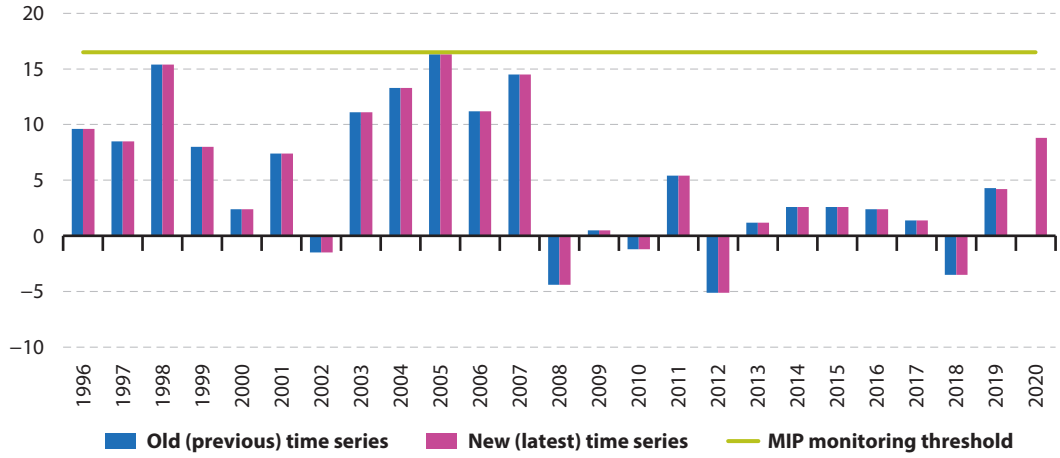
Figure 1: Macroeconomic indicators for annual financial accounts — private sector credit flow, Belgium, 1995-2020
(% of GDP)



Note: the private sector credit flow represents the net amount of liabilities that have been incurred by non-financial corporations (S.11), households (S.14) and non-profit institutions serving households (S.15) during the year. Old (previous) time series, not available for 2020.

Source: Eurostat (online data code: nama_10_gdp)

Figure 2: Macroeconomic indicators for annual financial accounts — annual change in total (non-consolidated) liabilities of the financial corporations sector, Belgium, 1996-2020
(%)



Note: old (previous) time series, not available for 2020.

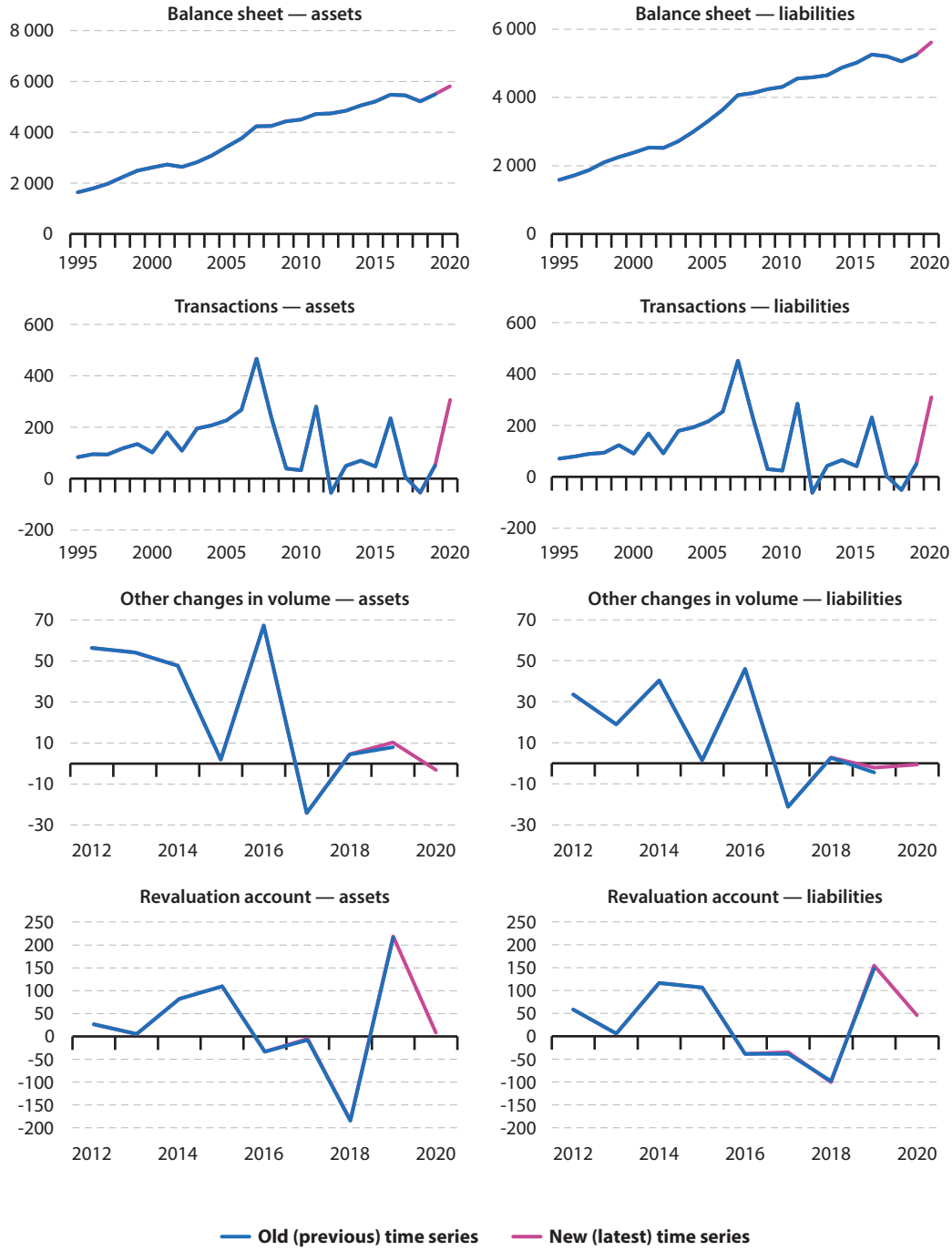
Source: Eurostat

5.1.2. GENERIC VISUALISATIONS

The second part of the report consists of generic visualisations that provide a quick overview of the data, considering only the most aggregated series. Figure 3 shows the eight elements of $F \times A$ in AFA, for the total economy. This section makes it possible to observe very quickly the tendencies of the data in the time series, and to identify if anything particularly unusual happened.

By having a quick overview of the most aggregated series, the user manages to have information on the trend for the most important series within the dataset (which are not excluded from the automatic selection). Trivial errors (for example scalar errors through the use of the wrong unit) are also easily spotted here.

Figure 3: Main series for annual financial accounts (non-consolidated), total financial assets/ liabilities, total economy, Belgium, 1995-2020 or 2012-2020
(billion national currency)



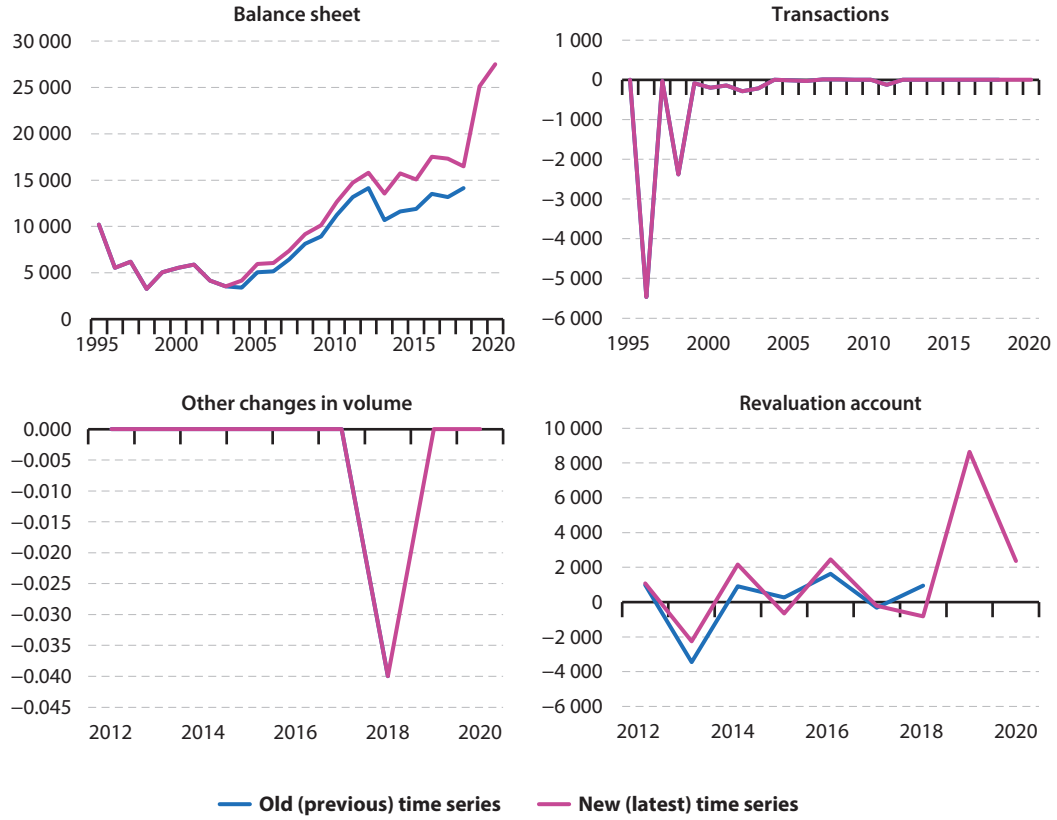
Note: old (previous) time series, not available for 2020.

Source: Eurostat

5.1.3. REVISIONS VISUALISATIONS

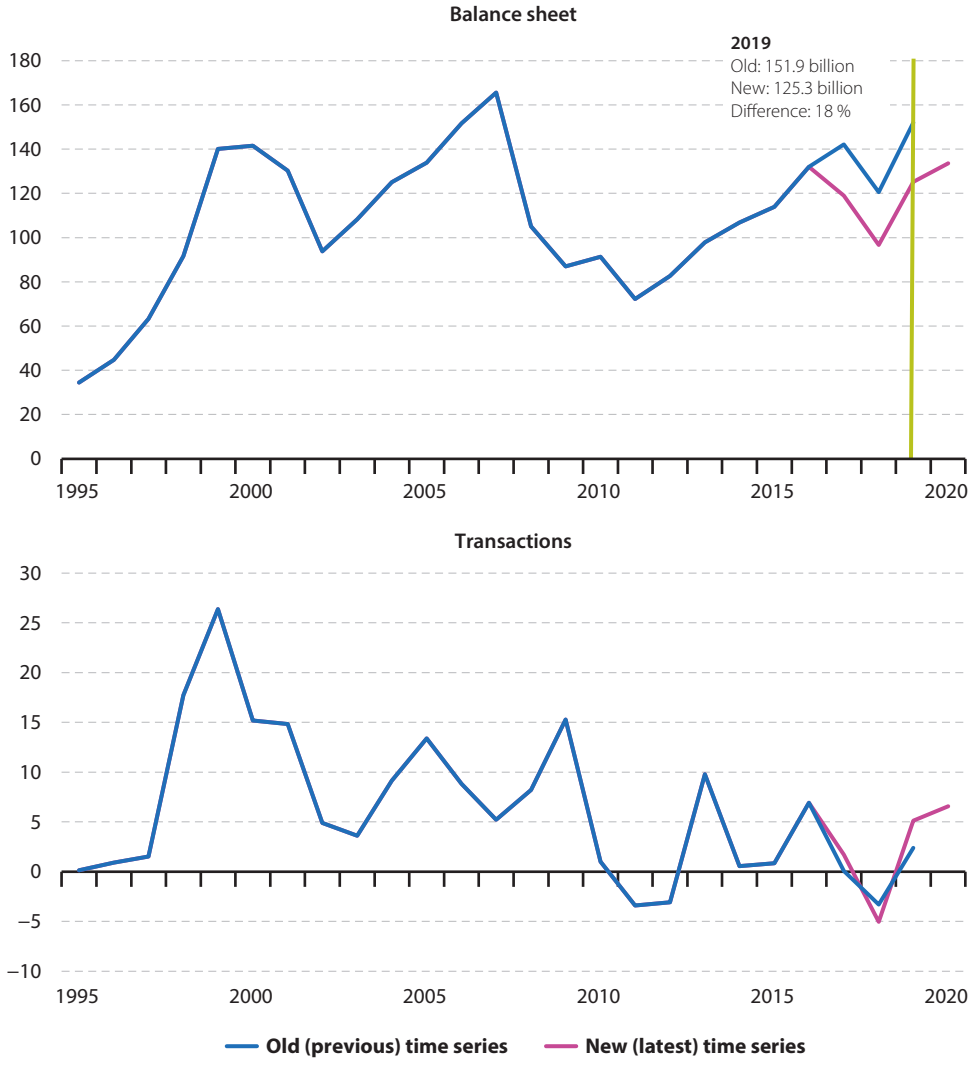
In this section, we show the most unusual revisions. In the example in Figure 4, it is possible to see revisions in the balance sheet of the consolidated other equity (F.519) liabilities of the Central Bank. The most recent revisions are triggered by the revaluation accounts. From the example in Figure 5, which shows the revisions of the consolidated listed shares (F.511) assets of the total economy, it is possible to note a feature of the tool which shows more details of a single data point when the cursor is moved over it.

Figure 4: Revisions for annual financial accounts (consolidated), other equity liabilities, central bank, Belgium, 1995-2020 or 2012-2020
(million national currency)



Note: old (previous) time series, not available for 2019 and 2020.
Source: Eurostat

Figure 5: Revisions for annual financial accounts (consolidated), listed shares assets, total economy, Belgium, 1995-2020
(billion national currency)



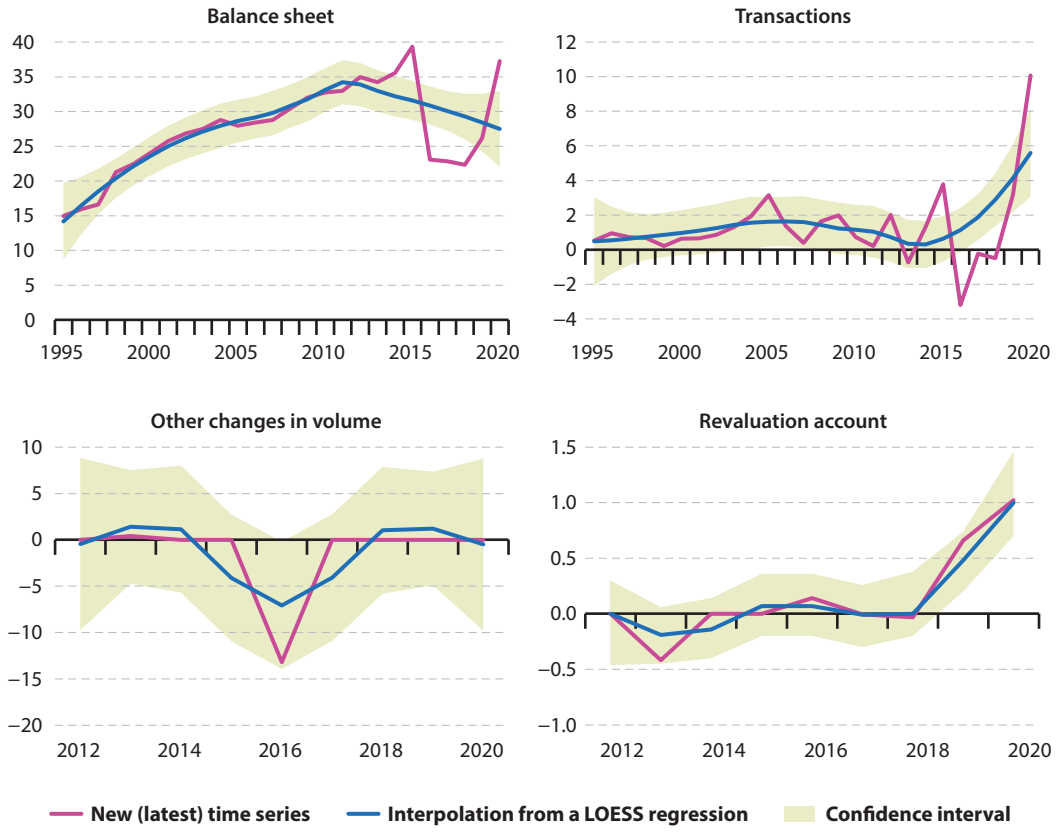
Note: old (previous) time series, not available for 2020.
Source: Eurostat

5.1.4. DETECTION OF OUTLIERS

This section shows the bigger outliers detected, such as those presented in Figures 6 and 7.

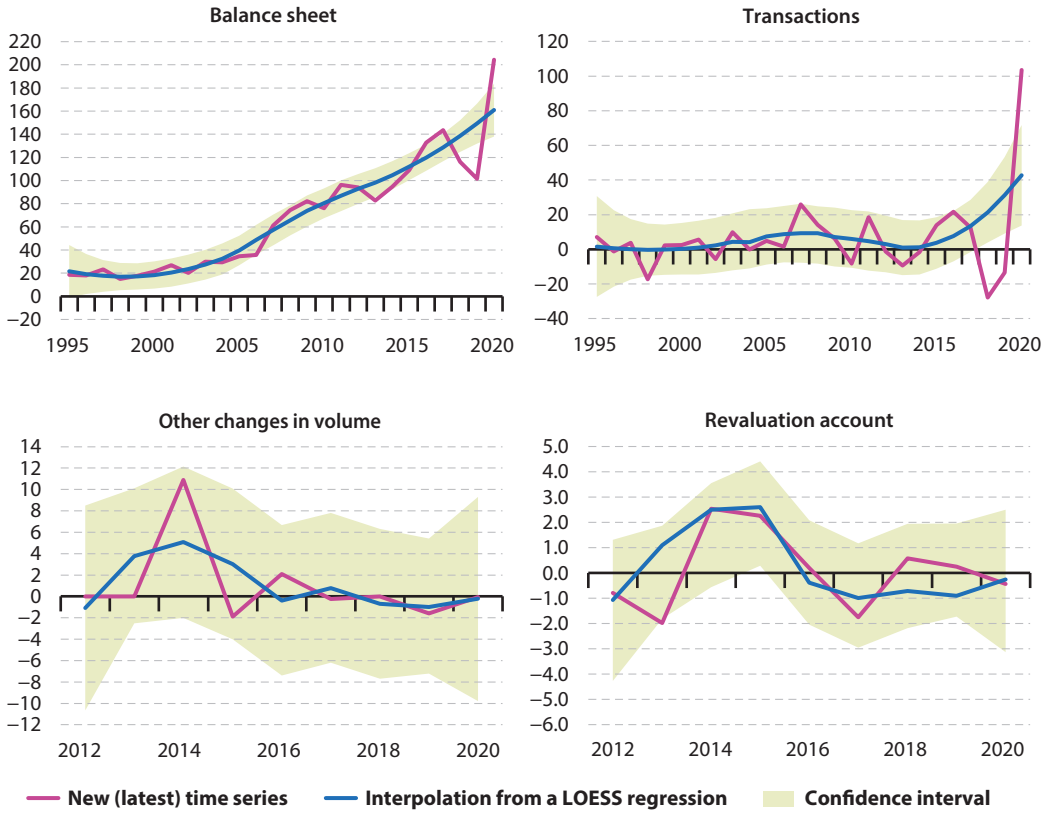
In the example of Figure 6, we can see that the latest value of the non-consolidated transaction non-life insurance technical reserves (F.61) assets of the Central Bank is clearly outside the interval and needs to be investigated. The balance sheet correspondent series is given alongside the sum of the transactions, the other changes in volume and the revaluation accounts; given that the other changes in volume and the revaluation accounts are in line with recent developments, it is (probably) enough to investigate the transactions figure. A similar conclusion can be made for the example shown in Figure 7, concerning the non-consolidated transactions of transferable deposits (F.22) as assets of financial corporations, which shows erratic values for the last three points of the series.

Figure 6: Detection of outliers for annual financial accounts (non-consolidated), non-life insurance technical reserves assets, total economy, Belgium, 1995-2020 or 2012-2020
(billion national currency)



Source: Eurostat

Figure 7: Detection of outliers for annual financial accounts (non-consolidated), transferable deposits assets, financial corporations, Belgium, 1995-2020 or 2012-2020
(billion national currency)



Source: Eurostat

5.2. Cross-domain checks: the case of annual and quarterly financial accounts

In this section we illustrate the cross-domain check presented in Section 3.3, using the case of AFA and QFA data from Germany which were transmitted in October 2020 and April 2021 respectively. As explained, we started by an introductory paragraph which presents the situation, stating the purpose of the report, specifying when the AFA and QFA data were received and indicating the number of errors/inconsistencies found, as shown in Illustration 1. We continued with a tab showing the number of errors grouped by institutional sector and by financial instrument, which helps to identify where errors come from, as shown in Table 1. This tab also included the values found in AFA and in QFA in order to be able to assess immediately the importance of the error and understand the nature of the problem.

Illustration 1: Outline of the report

AFA VS QFA REPORT

Dear colleagues from Germany,

We have received your annual financial accounts transmission on 2020-10-08 13:11:46 and the quarterly data was transmitted to us by the ECB on 2021-04-29. There are internal inconsistencies between the annual and quarterly financial accounts which will need your attention.

While reviewing both datasets we have identified **204** series which show inconsistencies and **0** series which show flag inconsistencies. Please find below details of methodology and the list of series and observations failing. We kindly ask you to review these inconsistencies and transmit the corrected annual dataset to us.

Please address any questions you might have to ESTAT-FINA@ec.europa.eu

Table 1: Cross domain checks — details concerning the inconsistencies for transactions

Value inconsistencies breakdown by sector | List of series with value inconsistencies | Value inconsistencies by Transactions | Value inconsistencies by Stocks | Inconsistencies in observation flags | Inconsistencies in confidentiality flags

Here, we checked that $\sum_{i=1}^{4} Q_i = A$. The rules we used to flag errors are the following ones:

1. $A - Q > 10$ Millions of National Currency.
2. $\frac{A-Q}{Q} > 1\%$.
3. $\frac{A-Q}{GDP} > 0.3\%$.

Series <chr>	Year <int>	A value <int>	Q value <int>	diff <int>	%diff (AFA-QFA)/QFA <dbl>	diff as % GDP <dbl>
A_S11_F	2013	178191	263973	85782	32.50	3.05
A_S11_F	2018	135578	204028	68450	33.55	2.04
A_S11_F	2016	173084	133176	39908	29.97	1.27
A_S11_F	2017	320864	284663	36201	12.72	1.11
A_S11_F	2012	188718	158695	30023	18.92	1.09
A_S11_F	2014	-11080	13586	24666	181.55	0.84
A_S11_F	2004	6430	18907	12477	65.99	0.55
A_S11_F	2015	187993	173135	14858	8.58	0.49
A_S11_F	2010	165119	174730	9611	5.50	0.37
A_S11_F	2001	133043	125612	7431	5.92	0.34
A_S11_F	2000	325867	332520	6653	2.00	0.32
A_S11_F	2011	115680	124382	8702	7.00	0.32
A_S11_F2	2017	46392	28417	17975	63.25	0.55
A_S11_F2	2013	13196	26382	13186	49.98	0.47
A_S11_F29	2017	21748	2625	19123	728.50	0.59
A_S11_F29	2013	-1859	11326	13185	116.41	0.47
A_S11_F4	2013	-19275	44739	64014	143.08	2.28
A_S11_F4	2013	-35407	11928	47335	396.84	1.68
A_S11_F4	2016	30196	-8952	39148	437.31	1.25
A_S11_F4	2016	35915	2120	33795	1594.10	1.08

Note: A = assets;
S11 = non-financial corporations;
F = total financial assets/liabilities;
F2 = currency and deposits;
F29 = other deposits;
F4 = loans.
AFA = annual financial accounts. QFA = quarterly financial accounts. GDP = gross domestic product.

Source: Eurostat

6. Conclusions

In this article we described the usual practices for data validation and analysis in official statistics, typically performed by statistical offices and international organisations. Building on a general procedure, we worked within a schematic framework where subsequent validation checks are performed on a very big set of time series, where exploration of single time series is impossible due to the limited time in which the validation should be done. We outlined the need for a systematic approach making it possible to process a large volume of time series.

Within such a framework, we developed an innovative data-driven solution for visualising the most relevant time series in terms of revisions and outliers, using a highly interactive application. Computationally speaking, this solution is based on three phases, namely defining a metric to order the data according to what we are looking for, establish rules to filter out inconsequential elements and finally rendering these results in an interactive dashboard. As validation and analysis might need domain specificities, this approach offers users the possibility to change some parameters according to their needs, such as a minimum threshold or the number of periods for which more revisions or outliers can be accepted. A reference time series should still be selected in order to quantify values relatively to this reference, according to the specific domain. While we have seen that GDP can be used as a reference value for many national accounts domains, this is not necessarily true for financial accounts. More generally, this work could be improved by quantifying the impact of the choice of the reference value, and testing the robustness of it.

Finally, we have shown a practical application of the methodology that we developed, based on annual financial accounts, showing that it made possible a very quick identification of the main issues in the dataset, ultimately leading to improvements in data quality. Furthermore, we suggest the introduction of a template for metadata — similar to practices in other domains — and we developed another reporting tool focused on cross-domain checks, which also has the potential to improve data quality.

This work could be taken one step further by fully integrating the metadata exchange during the generation of the dashboard, and most importantly by giving the opportunity to the user to extract content to be sent directly to the statistical reporting agency. Finally, the concepts developed in this article regarding data analysis could serve as the basis for new developments in official statistics with special interests in visualisations, such as Statistics coded.

Acknowledgements

We thank our colleagues Orestis Tsigkas (for his work in non-financial sector accounts from which our work is inspired), Luis Biedma (for his occasional support on development and coding matters) and Martin Turcan (for his valuable ideas). We thank John Verrinder for his advice. Finally, we thank Ani Todorova and Malgorzata Szczęsna for their guidance.

References

Bay S., K. Kumaraswamy, M. G. Anderle, R. Kumar and D. M. Steier (2006), 'Large scale detection of irregularities in accounting data' in *Sixth International Conference on Data Mining (ICDM'06)*, Institute of Electrical and Electronics Engineers, pp. 75-86.

Buono D. and E. Infante (2013), 'New technique for predictability, uncertainty, implied volatility and statistical analysis of market risk using SARIMA forecasts intervals' in *New Techniques and Technologies for Statistics 2013,* Eurostat.

Cohen J. B. (1938), 'The misuse of statistics', *Journal of the American Statistical Association*, Vol. 33, No. 204, pp. 657-674.

Di Zio M., N. Fursova, T. Gelsema, S. Gießing, U. Guarnera, J. Petrauskienė, L. Quensel-von Kalben, M. Scanu, K. O. ten Bosch, M. van der Loo and K. Walsdorfer (2018), *Methodology for data validation 1.1*, ESS Handbook.

Eurostat (2014), *European system of accounts — ESA 2010 — Transmission programme of data*, Publications Office of the European Union, Luxembourg.

Eurostat (2018), *European Statistics Code of Practice*, Publications Office of the European Union, Luxembourg.

Hao M. C., U. Dayal, D. A. Keim, and T. Schreck (2005), 'Importance-driven visualization layouts for large time series data' in *IEEE Symposium on Information Visualisation, 2005 (InfoVis 05)*, Institute of Electrical and Electronics Engineers, pp. 203-210.

Kandel S., R. Parikh, A. Paepcke, J. M. Hellerstein and J. Heer (2012), 'Profiler: integrated statistical analysis and visualization for data quality assessment' in *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI'12)*, Association for Computing Machinery, pp. 547-554.

Lenderink R.J. (2019), *Unsupervised outlier detection in financial statement audits*, Master's thesis, Faculty of Electrical Engineering, Mathematics & Computer Science, University of Twente.

Mackinlay J. (1986), 'Automating the design of graphical presentations of relational information', *ACM Transactions on Graphics,* Volume 5, Issue 2, pp. 110-141.

R Core Team (2020), R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna.

Vartak M., S. Rahman, S. Madden, A. Parameswaran and N. Polyzotis (2015), 'SeeDB: efficient data-driven visualisation recommendations to support visual analytics' in *Proceedings of the VLDB (Very Large Data Base) Endowment*, Volume 8, Number 13, pp. 2 182-2 193.

Yang S. and R. Cogill (2013), 'Balance sheet outlier detection using a graph similarity algorithm' in *2013 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, Institute of Electrical and Electronics Engineers, pp. 135-142.

Zhu T. (2006), 'An outlier detection model based on cross datasets comparison for financial surveillance' in *2006 IEEE Asia-Pacific Conference on Services Computing (APSCC'06)*, Institute of Electrical and Electronics Engineers, pp. 601-604.