# Comparisons of datasets to monitor data quality: Applications with banking data[1]

Martina Spaggiari, European Central Bank, Martina.Spaggiari@ecb.int
Angelos Vouldis, European Central Bank, Angelos.Vouldis@ecb.int
Stefano Borgioli, European Central Bank, Stefano.Borgioli@ecb.int

**Abstract**

*Cross-validation using different datasets can be a very effective way of monitoring data quality and may complement validation rules applied within a single dataset. In particular, cross-validation can identify data quality issues which would not be found with the validation rules applied at single dataset level, i.e. those related to the completeness or double-counting of information. This approach is especially relevant for banking data, given that various banking datasets have become available at central banks and to banking supervisors. These datasets differ with respect to their scope, method of consolidation and granularity. Two examples of the method are then provided.*

**Keywords:** Quality Assessment, Banking data, Macro-prudential policy

## 1. Introduction

High quality standards are a key factor in maintaining public trust in the ECB statistics upon which its policy decisions are based. In the course of the past years the ECB has developed its Statistics Quality Framework (SQF), which sets out the main quality principles and elements guiding the governance of ECB statistics. The SQF serves to ensure that ECB statistics remain fit for use for the ECB, its decision-making bodies and other ESCB users, as well as users outside the ESCB, such as financial market analysts, academia, journalists and the general public.

This paper focuses on a specific method for monitoring data quality, namely the cross validation across different datasets. The method consists in monitoring data quality of a dataset by comparing a subset of its values (mainly aggregate figures, but also more granular breakdowns) to the corresponding data reported in another

[1] The authors would like to thank Bogdan Chiriacescu, Francesco Donat, Patrick Sandars for helpful comments and suggestions. The views expressed in this paper are those of the authors and they do not necessarily reflect the views of the European Central Bank.

comparable dataset. This method is able to identify data quality issues which could not be captured when internal consistency checks are used. This aspect of data quality is especially important as it may lead to identify misreported aggregate figures even if internally the dataset seems to be consistent.

The paper also shows how this method of data quality monitoring is especially relevant in the field of financial and banking statistics. Banking data are especially sensitive to consolidation, sample and scope issues (e.g. whether data are consolidated or not, what is the perimeter of consolidation, which institutions are covered, the precise definition of the sample, which financial instruments are covered) and the proposed method is especially suitable to identify such issues.

The financial crisis which erupted in 2007-08 led to the identification of data gaps in the available statistics which needed to be filled to improve the surveillance of the financial system. The creation of the European system of financial supervision in 2010 gave further impetus to the collection of new datasets at European level. This proliferation of datasets which provide information on the financial and banking system from different angles renders the monitoring of their quality and the selection of the more suitable data to be used for each type of analysis a challenging but necessary task.

In this context, the cross datasets analysis gains in importance as a method of monitoring the data quality of datasets as it is able to identify issues of scope and coverage e.g. caused by differences in the perimeter of consolidation adopted or the impact arising from the the inclusion or exclusion of types of institutions in the reporting sample.

The paper is structured as follows. Section 2 presents the first test case comparing aggregate datasets. Section 3 compares data on financial interconnectedness and Section 4 concludes.

## 2. The ECB Consolidated Banking Data and the ECB Supervisory Banking Statistics

This Section presents a comparison of the data contained in two ECB datasets, namely the Consolidated Banking Data (CBD)[2] and the Supervisory Banking Statistics (SBS)[3]. Both datasets cover the European banking sector and should be comparable, permitting their use as alternatives by researchers who would like to analyse the European banking sector from a financial stability or a supervisory perspective. Both datasets report aggregated statistics and are based on the same methodological framework drawing on EBA Implementing Technical Standards (ITS), namely FINREP and COREP, as reported at the individual banking group/bank level.

The CBD are collected by the ECB since the early 2000s and contain aggregated cross-sectoral and cross-country consolidated data on banks' activities (balance sheets), solvency, profitability, and asset quality. The data are consolidated across all countries and financial sectors on a home-country basis. For a more detailed presentation of the CBD data see Barbic et al. (2017).

The SBS have been published by the ECB as of 2014Q4 and they are based on the firm-level supervisory data reported by the euro area 'significant institutions' (SIs[4]) i.e. banks that are supervised directly by the ECB. The information is then aggregated at national level. The comparison exercise presented below refers to the data contained in CBD and SBS with reference dates 2017Q1-Q2.

### *2.1. Methodology*

The aim is to formulate quality conditions which shall hold, when comparing the total amount of assets and the number of significant institutions contained in each dataset based on the underpinning methodologies.

Specifically, the SBS methodology requires that institutions are excluded from the SBS sample if they report at a higher consolidation level in another ECB Banking Supervision country. For example, data for Slovakia are not reported at all[5] given that all SIs in Slovakia are subsidiaries of other ECB-supervised significant institutions. In this way, SBS avoids double-counting when national data are added. On the other

[2] See: https://www.ecb.europa.eu/stats/supervisory_prudential_statistics/consolidated_banking_data/html/index.en.html.
[3] See: https://www.bankingsupervision.europa.eu/banking/statistics/html/index.en.html.
[4] Significant Institutions are determined on the basis of their size, economic importance, cross-border activities and, in case, direct public assistance (https://www.bankingsupervision.europa.eu/banking/list/criteria/html/index.en.html).
[5] We are referring here to data with a reference date 2017 Q1.

hand, CBD includes all institutions consolidated at the national level, therefore simply adding all national aggregates could lead to double counting at the European level.

As a result of this methodological difference, the two following conditions can be formulated at country level:

*Number of reporting SIs in CBD ≥ Number of reporting SIs in SBS (Condition 1)*

*AND*

*Total Assets of SIs in CBD ≥ Total Assets of SIs in SBS (Condition 2)*

In the analysis presented below, we refer to the total assets ratio and the number of institutions ratio referring to the SIs for both datasets. Consequently, the above conditions can be formulated as follows in terms of these ratios:

*Number of reporting institutions (SIs) ratio (CBD to SBS) ≥ 1 (Condition 1)*

*AND*

*Total Assets of SIs ratio (CBD to SBS) ≥ 1 (Condition 2)*

### *2.2. Results*

The comparison exercise aims to identify data that do not conform to conditions 1 and 2 to identify potential data quality issues in CBD. For many countries these two ratios are equal to 1, hence there are no differences between the two datasets. However, there are some cases for which discrepancies are observed.

The ratios are currently less than one only for one country. In this case, at least one large SI is not included in the CBD statistics because the entity is operating in a wind-down mode, in line with an orderly run-off restructuring plan that was agreed with the European Commission. This methodological choice has clearly an economic rationale, in the sense that such entities are no longer ‘active’ institutions and do not really compete with other credit institutions, and therefore its inclusion could distort the aggregate figures.

There are some other countries for which both ratios are higher than one. These cases do not violate conditions 1 and 2 above, however, especially when the ratio is much larger than one, we conducted further investigation to identify the causes. Investigations showed that these large discrepancies between CBD and SBS are explainable by the inclusion in the CBD of data of subsidiaries with a parent in another euro area country. Specifically, as was explained above, SBS includes only

data for SIs at the highest level of consolidation within the euro area. Therefore, subsidiaries of euro area banks hosted e.g. in Luxemburg are not included in SBS. Being Luxemburg a financial hub, there are many subsidiaries of banks which are consolidated in euro area countries outside Luxemburg. For example, Deutsche Bank Luxembourg S.A. has c. €80bn total assets, while the Luxemburgish subsidiary of the Belgian-based custodian bank State Street has c. €240bn. These two banks already explain to a large part the observed discrepancies in reported total assets between SBS and CBD.

Another particular case is the existence of large branches. In one country the number of institutions ratio is equal to one but the asset ratio is higher than one. Investigations showed that this phenomenon is caused by the inclusion in the CBD of a branch in the amount of total assets, however the branch is not counted in the number of institutions since it is not a separate legal entity.

Overall, the comparative analysis allowed to identify critical data quality issues which would not have been possible via internal consistency checks. In addition, the analysis enhanced substantially the understanding of the sources of differences between the two datasets and could inform further methodological work. Furthermore, understanding of the users regarding the optimal usage of the datasets has been also enhanced.

## 3. Comparison of two datasets on the interconnectedness of the European banks

This section presents a comparative analysis of the interconnectedness data contained in the CBD compared to the International Banking Statistics (more precisely the ‘Consolidated Banking Statistics’ subset) published by the Bank for International Settlements (BIS). Both datasets provide measures for interconnectedness of the national banking systems and are critical for macro-prudential analysis, especially for analyzing cross-border contagion or possible cross-country policies spillovers. Therefore, they represent alternative datasets that can be utilized by a researcher analyzing contagion risks and consequently their comparison is of great interest, as the use of one dataset over the other may have an impact on the policy conclusions thus reached.

This part of the work focuses mainly on the analysis of cross-country exposures, which are useful to assess the interconnectedness of the global banking system (e.g. see McCauley at al. 2012).

*3.1. Methodology*

As in the previous section, the one side of the cross-dataset-comparison involves the CBD data. The focus here, however, is on the subset of CBD data with information on cross border exposures. These data are available at quarterly frequency starting from 2014 Q4. As mentioned before, the items of interest in the context of this analysis are the cross-country exposures.

The BIS dataset focuses on the international financial interconnectedness dimension and contains the Consolidated Banking Statistics (BIS-CBS) which are compiled following the consolidation approach used by banking supervisors (i.e. positions of banks' foreign branches and subsidiaries are included, but intragroup positions are excluded). Therefore, BIS-CBS are conceptually close to the ECB-CBD as both datasets adopt a consolidation approach.

In assessing the comparability of the two datasets, the initial step was to identify the methodological differences underpinning the two datasets. First, the scope of the instruments included is different. BIS data include all items representing an “on-balance sheet financial” asset with an exception of the on-balance sheet derivatives with positive market value. CBD data include exposures which present credit risk, excluding some securitisation positions and equity exposures, including however off-balance sheet items with a percentage of its nominal value depending on its risk. These exclusions should tend to lead to BIS values being higher than CBD, however the inclusion of off-balance sheet items and on-balance sheet derivatives with positive market value would tend to lead to higher CBD values. Furthermore, the two datasets are closely aligned with respect to sample issues, however one could not exclude some discrepancies in the respective samples. In addition, methodological choices related to reporting thresholds differ between the two datasets. The COREP template which is the base for the CBD cross border data is reported with a 10% threshold (i.e. banks with international exposures less than 10% of their assets do not report this template).

The conceptual investigation which was presented above, shows that in contrast to the previous comparison between CBD and SBS logical relationships expected to hold to a large degree of certainty cannot be formulated. Table 1 summarises the sources of discrepancy and the expected impact on the relative values of CBD compared to BIS-CBS data.

*Table 1: Summary of sources of discrepancy between CBD cross border data and the BIS-CBS data.*

| Item | Approach in CBD | Approach in BIS-CBS | Impact on the CBD-to-BIS-CBS ratio |
|---|---|---|---|
| Definition of exposures | Exposures which present credit or counterparty risk | On-balance sheet financial assets | N/A |
| Off-balance sheet items | Included | Excluded | + |
| Derivatives with positive market value | In principle are included | Excluded | + |
| Securitisation positions | Excluded | In principle are included | - |
| Equity exposures | Excluded | In principle are included | - |
| Sample definition | 'Credit institutions' of the CRR | Financial institutions which receive deposits and grant credit | ? |
| Materiality thresholds | There is a 10% materiality threshold for foreign exposures | Accounting materiality thresholds if applicable | ? |
| Perimeter of consolidation | Prudential | Mainly accounting | ? |

Given the lack of unambiguous logical relationships, the analysis proceeded by investigating the properties of the datasets in relation to each other, aiming to understand how the methodological differences underlying the two datasets play out in practice.

### *3.2. Results and discussion*

The aim of the investigation empirically two questions. First, whether for each pair of countries $i,j$, the exposures from $i$ to $j$ are reported as being higher in the CBD dataset or the BIS-CBS dataset, in view of the various methodological discrepancies which were summarised above in Table 1 and which have an ambiguous overall effect on the relative values of the two datasets.

Second, we are interested also in the distribution and variability of the differences, both absolute and as a percentage, between the two datasets. This is important because the aim is to analyse the quantitative impact of the underlying methodological differences and which ones seem to be the drivers of the observed data.[6] A distinction is made between the differences which arise when the CBD value is larger and those that arise when the BIS-CBS is larger.

[6] We can only make inferences since we do not have access to the bank-level information, however these inferences could be quite useful when using the data.

In formal terms, the focus is on whether there is significant difference between the two following variables:

$$E\left(Exposure_{i,j}^{CBD} - Exposure_{i,j}^{BIS} | i,j: Exposure_{i,j}^{CBD} > Exposure_{i,j}^{BIS}\right)$$

and

$$E\left(Exposure_{i,j}^{BIS} - Exposure_{i,j}^{CBD} | i,j: Exposure_{i,j}^{CBD} < Exposure_{i,j}^{BIS}\right).$$

where the operator $E(\cdot)$ refers to the expected value (based on the sample of reported values), and the corresponding variables when percentages are used instead of the differences.

Therefore, the ratio of exposures reported in BIS to exposures reported in CBD for each pair of reporting-counterparty country is computed. The results are shown in Table 2. Given the existence of many pairwise links which are weak (e.g. less than 5% of the total international exposures and liabilities of the creditor or the borrower, respectively), the results for the 'important' links i.e. those which are higher than 5% are presented separately. As is shown in Table 2, these important links range between 181 and 182 of the total 630 pairs $i,j$ of countries considered, meaning that approximately 30% of all pairs of countries are characterised by links which are 'important' as defined above.

*Table 2: Comparison of corresponding CBD and BIS-CBS data.*

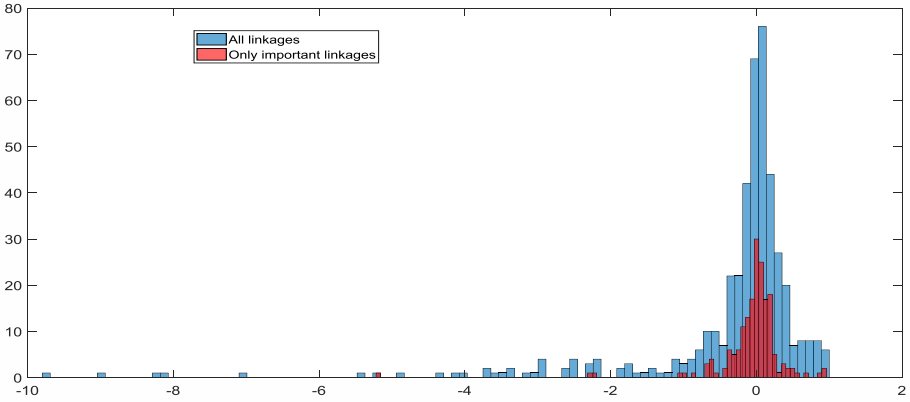
| | CBD higher | BIS-CBS higher | Total links existing | Empty pairs |
|---|---|---|---|---|
| **2016** | 230 | 265 | 495 | 135 |
| **2015** | 202 | 281 | 483 | 147 |
| **2014** | 196 | 292 | 488 | 142 |
| **Only important links considered** | | | | |
| **2016** | 98 | 84 | 182 | - |
| **2015** | 98 | 83 | 181 | - |
| **2014** | 86 | 95 | 181 | - |

When looking at the total sample, it is clear that the BIS-CBS values tend to be higher although in 2016 the difference between the pairwise links in which the BIS-CBS values are higher compared to those that the CBD values are higher is relatively small (265 versus 230 pairwise links). The situation is more balanced when only the important links are considered. In the latter case, the number of pairwise links in which one of the two datasets presents higher values is very close. The conclusion is that the sources of discrepancy listed in Table 1 tend to lead to higher values for the BIS-CBS values. However, we cannot say whether this effect is due to the securitisations or equity exposures which would unambiguously raise the BIS-CBS values, or if this pattern is caused by one of the other factors e.g. the perimeter of

consolidation, which a priori do not have a clear effect on the relative magnitudes of the two datasets.

In the second part of the analysis we focus on the difference between the two datasets, both in absolute and relative terms for each pair of reporting-counterparty country (see Figure 1).

**Figure 1**: The figure presents the percentage difference between CBD and BIS-CBS data (CBD minus BIS-CBS to the CBD value).




It is clear that the median is around 0 and that there is an asymmetry in the two tails. While the right hand tail is clearly bounded around 1 (i.e. a maximum difference of 100% when the CBD value is larger than the corresponding BIS-CBS value), the left hand tail (i.e. in the cases when the BIS-CBS value is larger than the corresponding CBD value) attains much larger values.

Tentatively, the difference in the consolidation perimeter have been identified as the main reason for the asymmetrical variability of the two tails as the difference in the consolidation perimeter seems to be the only source of discrepancy which is potentially unbounded (because a conglomerate can own many types of firms of different size and this affects the overall size of the balance sheet) while the other sources of discrepancy are all bounded by the size of the balance sheet.

The conceptual and quantitative analysis of the two datasets leads also to some conclusions about the optimal combined use of them for purposes of analysing interconnectedness of the financial system or potential cross-border contagion. Given that contagion and interconnectedness are multi-faceted phenomena which arise through exposures via a number of different instruments it seems optimal to include all these exposures in relevant analysis. Therefore, a suggestion to the users would be to use for every pair of countries, $i, j$, the exposure value from country $i$ to country $j$ as follows:

$$Exposure_{i,j} = max\left(Exposure_{i,j}^{CBD}, Exposure_{i,j}^{BIS}\right)$$

In this way, the data user is assured that all relevant exposures are captured. In addition, this way of aggregating information avoids double counting (which would arise e.g. if the values of the two datasets were added) and therefore it is not overly conservative.

## 4. Conclusions

The paper presented two examples of using the cross validation method to assess data quality in the context of banking data. It shows that the method provides significant value added to a data quality analysis by enabling the identification of data issues which would not have been captured by internal checks of consistency. The method is especially useful to identify issues of scope and coverage which are very relevant for banking data.

The two examples presented both use banking datasets and in both cases the CBD dataset represents one of the two datasets to be compared. The other datasets, the SBS data produced by the ECB and the CBS produced by the BIS are comparable to the CBD in a number of aspects but they are also characterized by differences in scope and underlying methodology.

The comparative studies led to a number of positive outcomes. First, we identified a number of data quality issues that could not have been identified otherwise. Second, the comparisons allowed us to gain a much better understanding of the respective datasets (which can be reflected e.g. in improving the methodological guidelines). Finally, the cross comparisons allowed us to formulate guidance to the users on how to best utilize the datasets examined. Therefore, as the two examples showed, the cross validation analysis provided significant value added with respect to the understanding how to utilize the information present in the examined datasets.

## 5. References

Barbic G., Borgioli S. and Klasco, J. (2017), The journey from micro supervisory data to aggregate macroprudential statistics, ECB Statistics Paper Series, 20, May 2017

McCauley R., McGuire P., von Peter G., (2012), After the global financial crisis: From international to multinational banking?, Journal of Economics and Business, 64, 7-23