EUROPEAN CENTRAL BANK

EUROSYSTEM

Comparisons of datasets to monitor data quality: Applications with banking data

Martina Spaggiari

Angelos Vouldis

Stefano Borgioli

European Central Bank

27 June 2018

Cross validation as a method of quality monitoring for banking data

- **The banking crisis has led to a proliferation of datasets which cover the financial sector**

- **In this context, the Consolidated Banking Data (CBD), produced by the ECB, have been expanding considerably during the last years**.

- **Data quality monitoring for banking datasets can be substantially enhanced by cross dataset comparisons.** This method is especially suitable to address issues related to sampling, coverage and scope. **Two test cases are provided in which the CBD data are compared with two external datasets.**

- **The aforementioned data quality issues are especially relevant in banking datasets.**

Dimensions of data quality for CBD data

1) Completeness

Completeness at three levels:
i) data point;
ii) coverage of national banking systems;
iii) users' perspective (usage of published data)

2) Consistency – cross sectional dimension

Extent to which data conform to the validation rules and the adequacy of provided explanations for violations (comments database).

3) Plausibility – time dimension

Plausibility of the changes of variables across time (structural breaks and ad-hoc events).
Special focus on the CBD – CBD2 transition effects.

Dimensions covering aspects internal to the dataset

4) Cross validation (provides added value to data quality monitoring using external comparisons e.g. with respect to coverage issues)

Comparison with comparable external datasets.

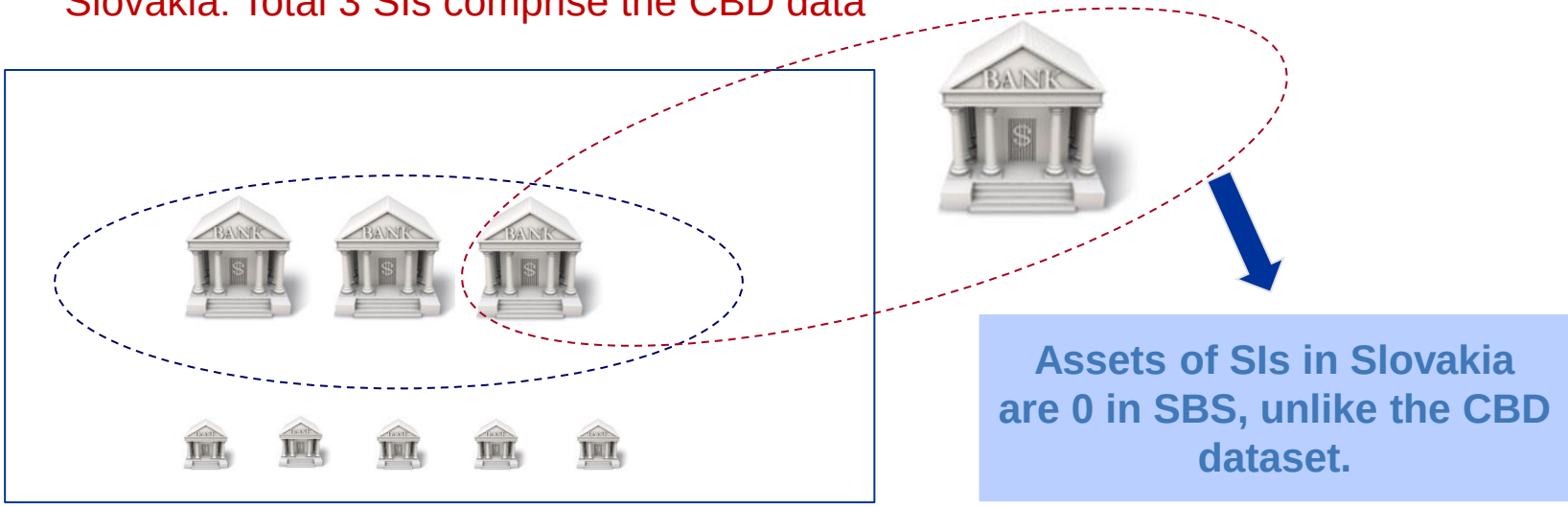
1st test case: CBD vs. Supervisory Banking Statistics (SBS)

Both datasets present consolidated data on national banking sectors, also separately for Significant Institutions

Main difference: SBS includes for each bank only information referring to its highest level of consolidation within the eurozone. In this case we can formulate **logical relationships** with respect e.g. to the total assets.

However, other SIs in AT, IT are the parents of the 3 Slovakian SIs => only the consolidated parents' data in AT, IT are reported in SBS!

Significant Institutions in the banking system of Slovakia: Total 3 SIs comprise the CBD data



Assets of SIs in Slovakia are 0 in SBS, unlike the CBD dataset.

1st test case: CBD vs. Supervisory Banking Statistics (SBS)

Similarly, there are differences between the two datasets for countries which function as financial hubs, such as Luxemburg.

Due to the existence of subsidiaries with eurozone parents, countries that act as financial hubs, like Luxemburg, present higher assets in CBD compared to SBS.

For example, Deutsche Bank Luxembourg S.A. has c. €80bn total assets.

The Luxemburgish subsidiary of the Belgian-based custodian bank State Street has c. €240bn

The comparative analysis allowed the identification of data quality issues and enhanced understanding of differences.

In addition, exceptional cases where an SI was not included in CBD but only in the SBS data were identified.

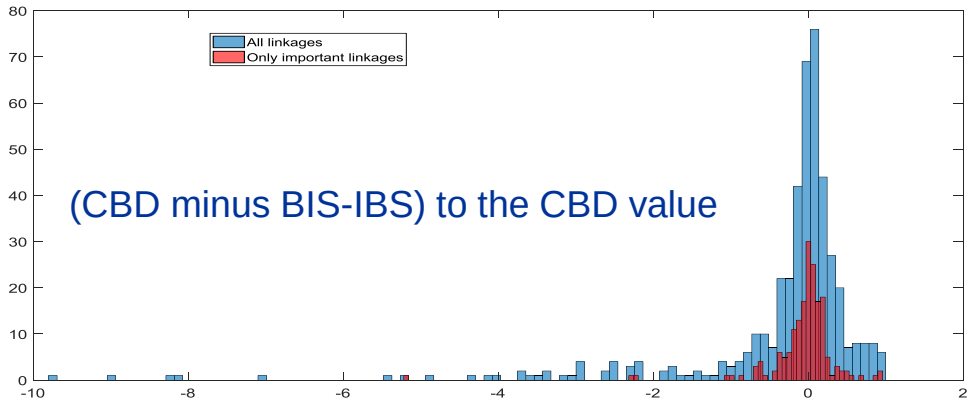
2nd test case: CBD vs. BIS International Banking Statistics (BIS-IBS)

Both datasets present data on interconnectedness of national banking sectors – pairwise exposures across countries. **Methodological differences:** Differences in **consolidation perimeter**, **scope of instruments**, and **others**. No unambiguous logical relationships.

Summary of differences and relative effects

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	?

Conceptual and quantitative investigation showed that the tail of differences in the part where BIS data are larger is much fatter due to consolidation effects. Optimal combination of data is proposed.



Cross validation as a method of quality monitoring for banking data

- **The paper presented two examples of using the cross validation method to assess data quality in the context of banking data.**

- 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 comparative studies led to a number of positive outcomes:**
  - First, we identified a number of data quality issues.
  - Second, the comparisons allowed us to gain a much better understanding of the respective datasets.
  - Finally, the cross comparisons allowed us to formulate guidance to the users on how to best utilize the datasets examined.